**UNIVERSIDADE DE BRASÍLIA**
**Faculdade de Ciência da Informação**
**Programa de Pós-Graduação em Ciência da Informação**

# MINERAÇÃO E MODELAGEM DE CONCEITOS COMO *PRAXIS* DE GESTÃO DO CONHECIMENTO PARA INTELIGÊNCIA COMPETITIVA

**ETHEL AIRTON CAPUANO**

Tese apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de Doutor em Ciência da Informação.

**Professor Orientador: Dr. Rogério Henrique de Araújo Jr.**
**Professor Co-orientador: Dr. Cláudio Chauke Nehme**

**Brasília, DF**
**2010**

**UNIVERSIDADE DE BRASÍLIA**
**Faculdade de Ciência da Informação**
**Programa de Pós-Graduação em Ciência da Informação**

# MINERAÇÃO E MODELAGEM DE CONCEITOS COMO *PRAXIS* DE GESTÃO DO CONHECIMENTO PARA INTELIGÊNCIA COMPETITIVA

**ETHEL AIRTON CAPUANO**

Tese apresentada à Faculdade de Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de Doutor em Ciência da Informação.

**Professor Orientador: Dr. Rogério Henrique de Araújo Jr.**
**Professor Co-orientador: Dr. Cláudio Chauke Nehme**

**Brasília, DF**
**2010**

# FOLHA DE APROVAÇÃO

*Peirce entendeu, tanto como qualquer dos seus ancestrais filosóficos, que nunca é fácil ser um metafísico realista, particularmente se o mesmo abranger, como Peirce o fez, a necessidade de experimentação e as modalidades do possível, do provável, e suas chances de serem constituintes da realidade.* (LUKOSE *et al*., 1997)

*A essência do conhecimento científico são os conceitos. (...) Conhecimento científico é informação conceitual obtida de textos ensinados em aulas de ciências na escola. Conhecimento científico é, portanto, algo que é compreendido e é, por isso, em primeiro lugar um objeto semântico.* (FRAWLEY, 1993)

# AGRADECIMENTOS

*BEN DELL'INTELLETTO*

*Quando o "Bem do Intelecto" entra,*
*Então a paz está conosco, e um suave controle*
*De todos os pensamentos rudes; mas um desejo*
*Preenche cada peito, - para esquecer o ruído*
*Das coisas do lado de fora, e submeter a alma*
*Ao banquete da amizade em um clarão noturno.*
*Então é a estação neste mundo de pecado*
*Que traz nova força, e nos manteve confiantes*
*A despeito das mudanças que causam ansiedade e cansam;*
*E quando da sabedoria temos sido andarilhos,*
*Então aquele estupor roubado do espírito*
*Das coisas desconhecidas, com visões escuras e terríveis,*
*Em sua alta presença nos restabelecemos*
*Mais que por todos os volumes nas prateleiras.*

T. W. Parsons
(*apud* Roberts (1973))

Agradeço e dedico este esforço, em primeiro lugar, a Deus, grande arquiteto do universo que criou toda a matéria e energia existentes e nos concedeu a capacidade de perceber a organização dessas coisas – a informação. Esta tese também é dedicada a todos os entusiastas da Inteligência Artificial, Gestão da Informação e do Conhecimento e Inteligência Competitiva e, em especial, às pessoas e outros seres vivos a seguir relacionados, aos quais agradeço pelas suas especiais participações neste meu esforço intelectivo.

Ao Sr. Carmine Antonio Capuano e Sra. Elvira Bini Capuano, Sr. Cosmo Spinelli e Sra. Maria Zancan Spinelli, meus saudosos avós, e ao Sr. Waldomiro Capuano *(in memorian)* e Sra. Josefina A. Spinelli Capuano, meus pais, que me contaram as primeiras histórias interessantes sobre o mundo e me ensinaram os primeiros passos da convivência harmoniosa em sociedade, apoiando-me no início de minha jornada de aperfeiçoamento como pessoa e como profissional.

A Luzia, Bruno, Guilherme e Beatriz, minha querida família de humanos, por aceitarem minhas ausências e respeitarem a necessidade de separação temporária para concentração nas atividades acadêmicas. E a Astrus *(in memorian)*, Teka, Uly, Princesa e Boris, pelo prazer de sua companhia a qualquer tempo e inspiração cognitiva *(insights)* que me proporcionaram em momentos decisivos de minhas reflexões epistemológicas.

Aos Profs. Drs. Rogério Henrique de Araújo Jr., Cláudio Chauke Nehme, Kira Maria Antonia Tarapanoff, Ulf Gregor Baranow e Lillian Maria Araújo de Rezende Alvares, todos *Ben dell'Intelletto*[1], que me acompanharam nesta viagem pelo mundo da Inteligência Competitiva.

Agradeço, imensamente, ao Prof. Dr. Rogério Henrique de Araújo Jr., notável especialista na disciplina Recuperação da Informação e meu orientador, pela sua atenção, interesse, compreensão, confiança e valiosos conselhos ao longo do trabalho de orientação acadêmica que

[1] Parsons, *apud* Roberts (1973, p. 8), escreveu: *'Ben dell'Intelletto', no Inferno de Dante, significa 'Bem do Intelecto'*. Que poderia ser interpretado, numa versão contemporânea, como *Capital Intelectual*.

tornou possível a elaboração deste trabalho. E, como não poderia esquecer nesta oportunidade, agradeço-lhe também pela sua amizade, pelas interessantes conversas abertas sobre temas de conhecimento geral e pelo precioso bom humor que o fazem uma pessoa ímpar. Com sua presença de espírito marcante, nosso convívio tornou-se muito agradável e produtivo e um trabalho que já me era prazeroso, por natureza, tornou-se quase lúdico.

Agradeço, novamente, ao Prof. Dr. Cláudio Chauke Nehme, meu orientador de empreitada acadêmica anterior e sempre grande incentivador, por me acompanhar em mais este desafio como co-orientador, prestimoso conselheiro e cúmplice do meu entusiasmo pelos temas da tese. O seu amplo conhecimento sobre temas acadêmicos de meu especial interesse, como Inteligência Artificial e Gestão da Informação e do Conhecimento, tornaram o desenvolvimento do trabalho bastante estimulante.

Agradeço à Prof[a]. Dra. Kira Maria Antonia Tarapanoff pelas preciosas críticas construtivas à apresentação dos resultados da tese, agregando inestimável contribuição aos aspectos cognitivos da criação de significado e construção do conhecimento. Com uma notável bagagem acadêmica no mundo da Inteligência Competitiva no Brasil, verdadeira pioneira nessa área do conhecimento ainda em construção, sua participação em mais esta jornada acadêmica jamais será esquecida.

Ao Prof. Dr. Ulf Gregor Baranow, com seu notável conhecimento lingüístico, agradeço pelo seu interesse e disposição em participar da banca, acrescentando valorosa crítica à tese apresentada e aos seus desdobramentos futuros. A sua presença realça a dimensão lingüística do experimento desenvolvido e suas implicações epistemológicas no contexto da mineração de conceitos como *praxis* de gestão do conhecimento para Inteligência Competitiva.

E à Prof[a]. Dra. Lillian Maria Araújo de Rezende Alvares agradeço pelo seu interesse, disposição e atenção nos últimos meses desta jornada, quando a leitura e avaliação das últimas versões do texto pelos fisicamente mais próximos se tornaram de grande valor na exata medida da maior eficiência da interação pessoal no aperfeiçoamento dos aspectos de comunicação das ideias centrais da tese. As suas impressões como recém-doutorada em Ciência da Informação na Faculdade de Ciência da Informação da Universidade de Brasília (FCI/UnB) são de grande valia como referência qualificada para a avaliação de nosso esforço.

Agradeço, ainda, aos colegas discentes e demais professores e funcionários da FCI/UnB pelo apoio emocional, cognitivo e situacional que eventualmente me dispensaram. Cito a sempre prestimosa atenção dispensada por Jucilene Gomes e Martha Araújo, diligentes servidoras públicas que mantêm a Secretaria do Programa de Pós-Graduação em Ciência da Informação, da Faculdade de Ciência da Informação, da Universidade de Brasília (PPGCINF/FCI/UnB) sempre ativa e operante em qualquer tempo, a despeito das inúmeras condições institucionais restritivas que, infelizmente, conspiram contra a qualidade da gestão no setor público no Brasil.

E, finalmente, agradeço aos doutores com perfis multidisciplinares e transdisciplinares que estão contribuindo, decisivamente, para o desenvolvimento evolutivo da Ciência da Informação na

FCI/UnB pela sua coragem, determinação e entusiasmo que muito me contagiaram nos últimos anos e me estimularam para elaboração desta tese. O seu esforço não será em vão, pois com faculdades multidisciplinares e transdisciplinares teremos, certamente, seres humanos dotados de intelectos mais holistas no futuro, com uma mais larga capacidade de compreensão dos temas científicos e tecnológicos atuais e dos enormes desafios que se apresentam à humanidade nesta era de mudanças tão velozes e impactantes.

# RESUMO

CAPUANO, Ethel Airton. *Mineração e Modelagem de Conceitos como Praxis de Gestão do Conhecimento para Inteligência Competitiva*. Brasília: Universidade de Brasília (Tese de Doutorado em Ciência da Informação).

Com o crescimento assintótico do volume de informação textual digital disponível sobre as organizações e acirramento da competição empresarial, métodos e técnicas de *Engenharia do Conhecimento* para recuperação e modelagem da informação relevante em repositórios abertos se tornam cada vez mais importantes para o desenvolvimento de processos de negócios. Com essa motivação, apresenta-se uma metodologia inovadora para mineração e modelagem conceitual de informação textual relevante para suporte a uma *praxis* de Gestão do Conhecimento apoiando o desenvolvimento de Inteligência Competitiva nas organizações. A metodologia, testada experimentalmente com simulação computacional, emprega recursos de Inteligência Artificial integrados em *softwares* de Processamento da Linguagem Natural, combinando, num construto epistemológico e tecnológico multidisciplinar de Ciência da Informação, conteúdos de várias áreas do conhecimento científico como Linguística, Filosofia, Matemática, Psicologia, Ciência da Computação e Engenharia. Os conceitos de *auto-informação,* de Shannon (1948), e da *diferença que faz diferença* (relativo à informação mais relevante), de Bateson (2002), Weick (1995) e Choo (2003), são centrais no processo de criação de significado, e o princípio da *inteligência emergente*, da Inteligência Artificial baseada na natureza, norteia o processo de apoio à construção do conhecimento proposto na tese. Os produtos resultantes da aplicação da metodologia são modelos ontológicos pictóricos com classes, objetos e relações povoados com sintagmas complexos extraídos dos textos digitais, com inspiração nas idéias de Gottschalg-Duque (2005). Esses modelos apresentam os conceitos e relações mais relevantes em contextos de negócio com objetivo de estimular, cognitivamente, os engenheiros do conhecimento nos seus processos mentais de criação de significados e construção do conhecimento útil para Inteligência Competitiva. O experimento mostra, estatisticamente, que a metodologia apresenta um desempenho bastante satisfatório, com revocação de no mínimo 90% de um conjunto dos substantivos mais relevantes presentes em repositórios de textos digitais sobre as organizações.

INTELIGÊNCIA COMPETITIVA
MINERAÇÃO DE CONCEITOS
MODELAGEM DE CONCEITOS
ENGENHARIA DO CONHECIMENTO
INTELIGÊNCIA ARTIFICIAL

# ABSTRACT

CAPUANO, Ethel Airton. *Mineração e Modelagem de Conceitos como Praxis de Gestão do Conhecimento para Inteligência Competitiva*. Brasília: Universidade de Brasília (Tese de Doutorado em Ciência da Informação).

With the asymptotic growth of the digital volume of textual information available about organizations and the business competition, methods and techniques of *Knowledge Engineering* for modeling and retrieval of relevant information in open repositories become increasingly important for the development of business processes. With this motivation, we present an innovative methodology for conceptual mining and modeling of textual information relevant to a *praxis* of knowledge management to support the development of Competitive Intelligence in organizations. The methodology is experimentally tested with computer simulation, employing features integrated Artificial Intelligence software in Natural Language Processing and combining contents of several areas of scientific knowledge as Linguistics, Philosophy, Mathematics, Psychology, Computer Science and Engineering, compounding a technological and epistemological multidisciplinary construct of Information Science. The Shannon's (1948) concept of *self-information* and the Bateson's (2002), Weick's (1995) and Choo's (2003) concept of *difference that makes a difference* (on the most relevant information) are central to the process of creating meaning, and the nature-based Artificial Intelligence's principle of *emergent intelligence* guides the process of supporting knowledge construction proposed in the thesis. The products resulting from the application of the methodology are pictorial ontological models with classes, objects and relations populated with complex phrases extracted from the digital texts, inspired by the ideas of Gottschalg Duke (2005). These models feature the most relevant concepts and relationships in business contexts in order to stimulate, cognitively, the knowledge engineers in their mental processes of creating meaning and constructing knowledge useful for Competitive Intelligence. The experiment shows statistically that the methodology gives a very satisfactory performance, with recall of at least 90% of a number of nouns present in most relevant repositories of digital texts on organizations.

COMPETITIVE INTELLIGENCE
CONCEPT MINING
CONCEPT MODELLING
KNOWLEDGE ENGINEERING
ARTIFICIAL INTELLIGENCE

# SUMÁRIO

# RELAÇÃO DE ABREVIATURAS E SIGLAS

**AAAI – Association for Advancement of the Artificial Intelligence**

**ACF – Análise de Conceito Formal**

**ACM – Association for Computing Machinery**

**AIDS – Acquired Immunodeficiency Syndrome**

**AIO – Arquitetura de Informação Organizacional**

**AOL – America On Line**

**AOS – Arquitetura Orientada a Serviço**

**ART – Adaptive Resonance Theory**

**ASIST – American Society for Information Science and Technology**

**BIW – Business Insights Workbench**

**CAD – Computer Aided Design**

**CBR – Case-based Reasoning**

**CELM – Customer Equity and Lifetime Management**

**CEO – Chief Executive Officer**

**CommonKADS – Common Knowledge Acquisition and Documentation Structuring**

**CPF – Cadastro de Pessoa Física**

**CPLN – Sistema de Conhecimento e Processamento da Linguagem Natural**

**CRM – Customer Relationship Management**

**CSE – Conventional Software Engineering**

**DNA – Deoxyribonucleic Acid**

**EDGAR – Electronic Data Gathering Analysis and Retrieval**

**ER- Entidade-Relacionamento**

**ERP – Enterprise Resource Planning**

**EUA – Estados Unidos da América**

**FCI/UnB – Faculdade de Ciência da Informação da Universidade de Brasília**

**FRI – Fujitsu Research Institute**

**GB – Giga Byte**

**GC – Gráfico Conceitual**

**GE – Gráfico Existencial**

**HP – Hewlett Packard**

**HTML – Hypertext Markup Language**

**IA – Inteligência Artificial**

**IBM – International Business Machines**

**IC – Inteligência Competitiva**

**ILPES – Instituto Latinoamericano y del Caribe de Planificación Económica y Social**

**KADS – Knowledge Acquisition and Documentation Structuring**

**KBS – Knowledge-based Systems**

**KE – Knowledge Engineering**

**MB – Mega Byte**

**MED – Modelagem Essencial de Dados**

**MER – Modelo Entidade-Relacionamento**

**MIT – Massachusetts Institute of Technology**

**MOODLE – Modular Object-Oriented Dynamic Learning Environment**

**NER – Named Entity Recognition**

**OCR – Optical Caracter Recognition**

**OMG – Object Management Group**

**OO – Object-Oriented**

**OR – Object-Relational**

**PANKOW – Pattern-based Annotation through Knowledge on the Web**

**PDF – Portable Document Format**

**PLN – Processamento da Linguagem Natural**

**POS – Part of Speech**

**RAM – Random Access Memory**

**RCI – Recuperação Conceitual da Informação**

**RFP – Request for Proposals**

**RI – Recuperação da Informação**

**RS – Rede Semântica**

**SCM – Supply Chain Management**

**SEC – Securities and Exchange Commission**

**SIC – Sistema de Informação Computacional**

**SOA – Service Oriented Architecture**

**SRI – Sistema de Recuperação da Informação**

**SWOT – Strength, Weakness, Opportunity, Threat**

**TI – Tecnologia da Informação**

**TIC – Tecnologia da Informação e Comunicação**

**UEFA – Union of European Football Associations**

**UML – Unified Modelling Language**

**URL – Uniform Resource Locator**

**WWW – World Wide Web**

# RELAÇÃO DE TABELAS

# RELAÇÃO DE GRÁFICOS

# RELAÇÃO DE QUADROS

# RELAÇÃO DE FIGURAS

# 1. INTRODUÇÃO

O texto a seguir apresenta a defesa de uma tese na qual será possível o desenvolvimento de um construto epistemológico para mineração e modelagem de informação conceitual relevante, a partir de textos digitais disponíveis em fontes abertas, como *praxis* de gestão do conhecimento *a priori* para Inteligência Competitiva (IC) nas organizações de mercado. Estipula-se, como premissas para esta tese, que o estado-da-arte da Inteligência Artificial (IA) e das ferramentas de Processamento da Linguagem Natural (PLN) suportará, numa abordagem de Engenharia do Conhecimento[2], o desenvolvimento de uma metodologia com este objetivo.

Como vantagens para a gestão do conhecimento nos ambientes das organizações, a Engenharia do Conhecimento proporciona (SCHREIBER *et al*., 2000, p. 7):

- ferramentas para identificação de oportunidades e gargalos nos processos de desenvolvimento, distribuição e aplicação dos recursos de conhecimento;
- métodos para compreensão das estruturas e processos utilizados pelos trabalhadores do conhecimento nas organizações;
- métodos para integração das tecnologias da informação no suporte às atividades dos trabalhadores do conhecimento;
- uma metodologia para a construção de melhores sistemas baseados no conhecimento.

Os métodos de raciocínio científico fundamentais utilizados nesta tese são o da *retrodução* e da *abdução* mencionados por Peirce (2010), o primeiro justificando o experimento de laboratório realizado e o segundo o processo de formação de uma hipótese explanatória acerca do fenômeno informacional estudado.

## 1.1 Competitividade e sistemas de negócios

O projeto de pesquisa multidisciplinar que fundamenta esta tese aborda um tema central na Ciência da Informação aplicada ao ambiente das organizações competitivas: a gestão da informação e do conhecimento. Como pesquisa experimental, o trabalho se concentra no desenvolvimento de uma metodologia, com uso de Inteligência Artificial (IA), para mineração e modelagem de informações textuais com objetivo de suportar processos de desenvolvimento de Inteligência Competitiva em organizações e seus sistemas de negócios. Esses processos de inteligência, no entanto, não são pensados *ad hoc*, mas no escopo de uma proposta mais ampla de Gestão da Informação e do Conhecimento, como apresentada em Capuano *et al*. (2009).

O princípio epistemológico que se defende, neste aspecto, é que a Inteligência Competitiva pode ser estruturada, *a priori*, como um macroprocesso corporativo conectado à Gestão da Informação e do Conhecimento, motivo pelo qual esta tese aborda esses conceitos numa mesma

[2] Conforme Schreiber *et al.* (2000), "Engenharia do Conhecimento" consiste na modelagem de diferentes aspectos do conhecimento humano.

abordagem metodológica integrada. É nessa idéia de um construto de inteligência a partir de componentes mais básicos de gestão da informação aberta e disponível dos ambientes de negócio que se apresenta as inovações propostas na tese.

Os conceitos de “sistemas de negócios” e “sistemas organizacionais” adotados nesta tese são abrangentes e intercambiáveis, referindo-se aos modelos e sistemas de estruturas e processos decisórios e operacionais que moldam uma organização complexa contemporânea, assim como aos objetos que constituem insumos e produtos nesses sistemas, como os produtos e serviços oferecidos aos clientes e os próprios sistemas de informação computacionais (SICs). O problema geral para o qual se busca uma solução é o de transformação semi-automática do crescente volume de informações das organizações e sobre as organizações em modelos conceituais que possam responder, com eficácia e eficiência, às suas necessidades de readaptação contínua ao ambiente global de negócios em constante e cada vez mais acelerada mutação, de modo a contribuir para a Inteligência Competitiva.

Esta abordagem se deve ao duplo imperativo que se apresenta às organizações competitivas: a de prever as mudanças no mercado e a de se preparar para as mudanças necessárias na própria organização para readaptação ao meio mutante. Geralmente, a disciplina de Inteligência Competitiva se ocupa apenas do monitoramento do mercado, sem envolvimento com as questões de gestão interna nas organizações, como se bastasse prever os movimentos da concorrência e acertar a “bússola” corporativa para os novos rumos.

A área de concentração da pesquisa se justifica, também, pela notória dependência das organizações contemporâneas em relação aos seus sistemas de um modo geral, especialmente os sistemas de informação de negócios. Os reflexos dessa dependência são observados, no cotidiano, na questão evolutiva dos sistemas informacionais baseados em computador, onde enormes dificuldades se apresentam em projetos de mudanças em escala corporativa (NEVO e WADE, 2007; LUNA-REYES *et al*., 2005; STONE, 1997).

Embora tenham contribuído, historicamente, para a evolução sociotécnica das organizações e do mercado após a 2ª Guerra Mundial, com ganhos de produtividade, os sistemas organizacionais que moldaram o ambiente corporativo da segunda metade do Século XX, de padrão tipicamente industrial, voltado para a automação de processos produtivos, tornaram-se também um sério obstáculo às mudanças evolutivas (ou, às vezes, revolucionárias) nessas mesmas organizações na medida em que sistemas e estruturas de negócio se confundem numa relação simbiótica de forte acoplamento. Em algumas organizações, observa-se que sistemas e processos de negócio se confundem com o próprio negócio criado pela organização, como no caso de empresas mais inovadoras baseadas na *World Wide Web* (WWW).

Com tal cenário, mudanças reestruturantes nos processos operacionais para melhoria da competitividade exigem mudanças nos sistemas organizacionais na mesma velocidade, tornando

a (re)modelagem desses sistemas um problema de solução não trivial em projetos de mudança (CAPUANO, 2007; HAMMER e CHAMPY, 2001; HEHN, 1999; DAVENPORT, 1993).

Em suma, organizações mais competitivas provavelmente serão, no futuro, as mais flexíveis em suas estruturas e sistemas, reunindo as condições necessárias para readaptação às contínuas mudanças nos ambientes de negócio na velocidade e profundidade necessárias.

Os sistemas de informação computacionais (SICs) corporativos têm sido desenvolvidos, a partir dos anos 1960, ao longo de anos, senão décadas, de investimentos nas organizações complexas com base em modelos de negócio bastante estáveis. E algumas organizações, como bancos, por exemplo, ainda utilizam rotinas de processamento de dados desenvolvidas em linguagens de programação de computador primitivas, como *Assembler*, devido também a essa estabilidade e confiabilidade resultantes de anos de prova em ambiente de produção.[3] As próprias metodologias de desenvolvimento de SICs utilizadas nesse período se adequavam aos requisitos de sistemas estáveis, num mundo onde as mudanças nos ambientes de negócios eram, presumidamente, lentas e pouco sensíveis.

Entretanto, com o avanço da globalização dos mercados, os ventos das mudanças começaram a soprar cada vez mais fortes e com maior freqüência, contestando o requisito da estabilidade nos modelos de sistemas organizacionais. De certo modo, a regra da estabilidade dos requisitos sempre representou uma simplificação da realidade para fins de modelagem, um reducionismo epistemológico que, contemporaneamente, tem-se mostrado cada vez menos útil, por exemplo, como metodologia de desenvolvimento de SICs, especialmente entre organizações inseridas em mercados muito competitivos.

Como exemplos dessa dificuldade, alguns paradoxos observáveis nos processos de desenvolvimento de sistemas organizacionais contemporâneos são:

I. A necessidade de reconhecimento do código dos SICs legados para o planejamento das mudanças. Como esses sistemas foram, geralmente, desenvolvidos e evoluíram ao longo de anos nas organizações, os desenvolvedores de *software* históricos não estão todos mais presentes e a documentação técnica disponível geralmente não é suficiente para que outros desenvolvedores possam conhecer, no curto prazo, todos os detalhes dos códigos de programação de modo a projetarem as mudanças necessárias. Esse tipo de atividade é denominado *refatoração de código* (LINDIG, 1997), com objetivo de transformar SICs monolíticos em conjuntos de componentes encapsuláveis e reaproveitáveis em novos SICs atualizados ou readaptados para novos serviços e processos de negócio.
II. A necessidade de se estimar custos e cronogramas de desenvolvimento de SICs *a priori* da modelagem, sendo que a própria modelagem integra o processo de desenvolvimento

[3] Linguagens de programação de computadores primitivas – mais próximas da linguagem de máquina – costumam apresentar desempenho melhor, em termos de velocidade de processamento, que linguagens mais modernas – mais próximas da linguagem natural.

e custos e cronogramas constituem resultados da modelagem, recaindo-se num paradoxo de modelagem recursiva.

III. A visão de sistemas independente das visões de processos e pessoas, como se essas três dimensões organizacionais não constituíssem um todo sistêmico. Como consequência, necessidades de mudanças nos SICs legados se chocam com anseios naturais de estabilidade nos processos de trabalho e na situação funcional das pessoas.

O problema central abordado na pesquisa se refere, portanto, à complexidade inerente à (re)modelagem de sistemas organizacionais competitivos numa velocidade nunca antes experimentada em tão larga escala e tamanha profundidade. Schreiber *et al*. (2000, ix), por exemplo, situam a Engenharia do Conhecimento como um recurso metodológico para atacar esse tipo de problema nas organizações contemporâneas:

> *(...) Gestão do conhecimento é uma área recente na administração de negócios que se ocupa de como disponibilizar conhecimento como um bem e um recurso chave nas modernas organizações. (...) Gerir o conhecimento numa organização é atualmente pouco viável sem a exploração do vasto potencial dos avançados sistemas de informação e conhecimento. De outro lado, desenvolvedores de sistemas de informação e engenheiros do conhecimento tem chegado à conclusão que o trabalho técnico com êxito é possível somente se ele é adequadamente situado num contexto organizacional mais amplo. Os métodos de engenharia do conhecimento tem gradualmente ampliado seu escopo: eles não são somente utilizados para o desenvolvimento de sistemas baseados no conhecimento mas também têm mostrado seu valor na gestão do conhecimento, engenharia de requisitos, modelagem empresarial e reengenharia de processos de negócio.*

Então, o objetivo da pesquisa será o de avançar, numa abordagem de engenharia do conhecimento, nos meios de suporte tecnológico à gestão da informação e do conhecimento em ambientes de Inteligência Competitiva, elaborando e testando, experimentalmente, uma metodologia que possa propiciar uma base de informação conceitual sobre os contextos de negócios das organizações, de modo a propiciar a modelagem de sistemas organizacionais de modo mais automático, tempestivo e compreensivo.

Em síntese, reconhecendo-se, como premissa, a visão sistêmica das organizações complexas em ambientes competitivos, apresenta-se uma tese na qual organizações mais competitivas são aquelas com melhor base epistemológica e capacidade para reconhecer a necessidade de mudanças em seus sistemas de negócios e maior capacidade de implementá-las de modo eficaz, eficiente e efetivo. A literatura de base sobre este tópico introdutório é vasta, sendo apresentada, em linhas gerais, no Capítulo 4.

## 1.2 Gestão da informação e do conhecimento

A contribuição do presente trabalho de pesquisa à área de estudos denominada Gestão da Informação e do Conhecimento é de natureza instrumental e pragmática, com uma abordagem de Engenharia e Gestão do Conhecimento (SCHREIBER *et al*., 2000), mas algumas reflexões teóricas e teses derivam de seus resultados. O conceito de “gestão do conhecimento” adotado

não é algo inovador, mas baseado em conceitos anteriores mais disseminados como “gestão da informação”, “gestão de pessoas”, “gestão do capital intelectual” e “gestão da tecnologia da informação e comunicação”. Esta concepção parte da premissa que não é possível uma gestão do conhecimento autônoma e direta, com base na “comunicação do conhecimento”, como advogam alguns entusiastas, porque o conhecimento é algo inerente aos modelos e processos mentais de cada indivíduo, portanto sempre tácito. O que se tem como “conhecimento publicado” são informações referentes ao conhecimento e não o conhecimento em si (os conceitos de informação e conhecimento serão discutidos mais detalhadamente na revisão de literatura).

A noção de gestão do conhecimento é útil na medida em que se presume que uma organização poderá prover seus colaboradores das condições necessárias para que eles possam, a partir de bases de informações e capacitação técnica adequadas, elaborar e reelaborar seus processos cognitivos e modelos mentais e construir sua própria “base” de conhecimento de interesse para a gestão do negócio. Contudo, o aspecto mais interessante no contexto de uma organização fundamentada na informação e no conhecimento é que suas bases de informações e de conhecimento útil se encontram, geralmente, ao alcance da própria organização, mas desestruturadas, muitas vezes pouco visíveis, como no mosaico informacional típico desenhado na Figura 1.1.

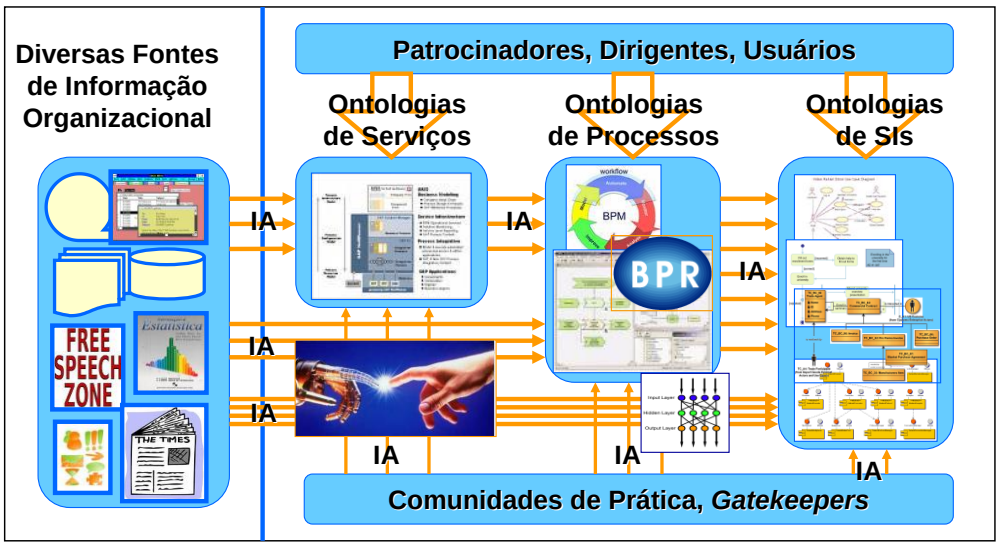



Figura 1.1 Organizações Fundamentadas na Informação e no Conhecimento
(Fonte: do autor da tese)

Embora se atribua à visão epistemológica de gestão do conhecimento importância fundamental no desenvolvimento das organizações do Século XXI, poucas soluções metodológicas e tecnológicas para sua realização se encontram disponíveis, concentrando-se as soluções mais inovadoras em tecnologias de recuperação da informação digital nos grandes repositórios de textos, como os *corpora* idiomáticos, e nos conteúdos da *World Wide Web*

(WWW). O uso do arsenal metodológico e tecnológico da Inteligência Artificial (IA), como sugerido no contexto da Figura 1.1, deve ser encarado como um recurso estratégico, ainda que instrumental, para construção de *pontes linguísticas* (ou uma *interlíngua*) entre o conhecimento dos engenheiros de sistemas e dos operadores do negócio (EKLUND, ELLIS e MANN, 1996).

## 1.3 Mineração conceitual de informações

Como “mineração conceitual de informações” define-se, nesta tese, as atividades de reconhecimento de padrões em textos para aprendizado e população de ontologias de sistemas organizacionais. Os principais antecedentes desse construto epistemológico são: Recuperação da Informação, Processamento da Linguagem Natural, Mineração de Textos, Estruturas Conceituais, Análise de Conceitos Formais, Reticulados de Conceitos e Aprendizado de Ontologias. Essas áreas de desenvolvimento, na Ciência da Informação, têm evoluído isoladamente com o uso de recursos computacionais, mas não têm produzido soluções integradas para solução do problema abordado nesta tese.

Entretanto, mineração da informação (armazenada como dados ou textos) não deve ser confundida com recuperação da informação, pois na mineração da informação não se tem um alvo muito específico nas buscas, em termos de conteúdos. O interesse da mineração se concentra mais nos padrões de informação representados na base e menos em informações indexadas, como na recuperação da informação. A recuperação da informação se concentra em buscas mais específicas, especificadas pelo usuário, utilizando para isso argumentos de busca que são termos-chave ou conteúdos existentes na base subrogada.

Com os resultados da pesquisa apresentada neste volume, procura-se evidenciar o potencial desses construtos integrados num *framework* metodológico para extração de informações textuais úteis para modelagem semi-automática de sistemas organizacionais. Os dados de entrada do modelo podem estar presentes nos crescentes volumes de documentos textuais armazenados nos repositórios digitais das próprias organizações, dos governos e da WWW. Essa abordagem opera como uma *engenharia reversa* do texto, ou seja, como o texto representa o resultado de um processo cognitivo-linguístico do autor, parte-se do texto para descobrir quais são os modelos mentais dos seus autores – e que deram origem ao texto. E, desse modo, tenta-se recuperar o conhecimento tácito do autor (ou mentor) do texto e utilizá-lo como vantagem competitiva no negócio.

Os modelos mentais desenhados com base nessa metodologia são representados, preferencialmente, de modo pictórico, mas embasados em formalismos matemáticos e em linguagens de modelagem de sistemas de informação comumente utilizadas no mercado, buscando-se com isso uma linguagem de modelagem mais genérica, mais próxima da linguagem natural e menos dependente de idiomas artificiais. Com isso, uma das vantagens da metodologia

proposta é o uso da linguagem natural, pois esta é a que mais aproxima as linguagens dos desenvolvedores de sistemas organizacionais da linguagem dos operadores do negócio numa organização.

O processo de elaboração da arquitetura da informação nesse *framework* é formalizado com emprego de entidades lógico-matemáticas de modelagem de conceitos desenvolvidas com a metodologia denominada Análise de Conceito Formal (ACF) e representadas com a Linguagem de Modelagem Unificada (UML). Com essa formalização lógico-matemática, pretende-se construir a base científica para as teses propostas a partir do referencial teórico, dos resultados dos experimentos e dos *insights* decorrentes. Os fundamentos filosóficos e científicos dessa abordagem conceitual são encontrados em obras de pesquisadores europeus da década de 1980 (WILLE, 1982), tendo-se iniciado sua disseminação na Ciência da Informação dos EUA há apenas alguns anos (PRISS, 2005).

As figuras a seguir ilustram o processo básico de descoberta, representação e aprendizado de ontologias[4] para a modelagem de sistemas organizacionais proposto. O modelo parcial de um sistema de informação de uma oficina mecânica apresentado, como exemplo, na Figura 1.3 parte de um recorte de texto representado na Figura 1.2, de onde se extraem alguns objetos e estruturas linguísticas (sintáticas) reconhecidas como padrão no idioma português, com os quais se definem as classes de objetos e seus relacionamentos no contexto. O método de modelagem ontológica, nessa ilustração, é o da UML[5] utilizada no processo de desenvolvimento de sistemas computacionais orientados a objetos.

O reconhecimento de padrões sintáticos no trecho de texto, com uso de um *software* de Processamento de Linguagem Natural (PLN), na corrente linguística estruturalista de Chomsky (1956), é realizado a partir de uma padronização das classes de termos utilizados no respectivo idioma mediante a colocação de uma etiqueta *(tag)*, ou rótulo[6], indicando a natureza sintática desse termo, ao lado de cada respectivo termo na sentença. Esta tarefa, no jargão de PLN, é denominada *Part of Speech (POS) tagging*.[7]

Observe-se, na Figura 1.3, que com apenas uma sentença curta pode-se identificar, inicialmente, quatro classes de objetos (mecânico, motor, automóvel e conserto) e suas relações essenciais de composição (um tipo de associação representado por uma seta partindo da classe de objetos componentes com ponta triangular oca apontando para a classe de objetos que contem esses componentes) e de objeto-atributo (seta simples partindo da classe-objeto para as classes

---

[4] *Ontologia*, na Filosofia, é o *estudo da natureza das coisas e dos seres e suas interrelações no mundo*; na Ciência da Informação, consiste de uma estrutura de conceitos ou entidades, em um domínio, organizada com base em suas interrelações. Sowa (1984, p. 294), de um ponto de vista da análise de conceitos, conceitua *ontologia para um mundo possível* como *um catálogo de tudo que constitui esse mundo, como esse todo se apresenta em conjunto e como ele funciona.*

[5] *Unified Modelling Language* (no original).

[6] Códigos de etiquetas (com base nas etiquetas utilizadas no *software* etiquetador *TreeTagger*): DT – artigo (determinante); NN – substantivos comuns no singular; VB – verbo no tempo presente; IN – preposição.

[7] Etiquetagem de Parte do Texto.

de atributos). Com o padrão sintático representado pela sequência substantivo_preposição_determinante_substantivo (NN_IN_DT_NN) tem-se uma relação instanciada como “motor_do_automóvel”, indicando que a classe “automóvel” possui um atributo denominado “motor”, e com a sequência-padrão substantivo_verbo_determinante_substantivo (NN_VB_DT_NN) tem-se uma relação entre a classe “mecânico” e a classe “motor” onde, com uma operação de transformação do verbo “consertar” no substantivo “conserto”, descobre-se a classe “conserto” com as classes-atributos “mecânico” e “motor”.

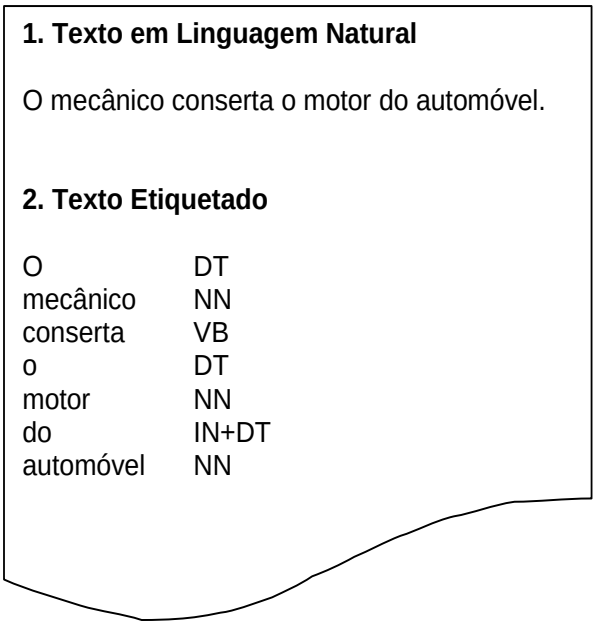



Figura 1.2 Estruturas Sintáticas de Texto em Linguagem Natural
(Fonte: do autor da tese)

As classes representando ações e não objetos, como “conserto”, são importantes porque se referem, geralmente, a serviços executados no interior das organizações, por uma unidade produtiva para outra, e pelas organizações para sua clientela – no caso, o conserto de um componente de um automóvel, como o motor, é uma “transação” (ou *“serviço”*) de negócio da organização no mercado.

Ampliando-se o modelo mediante uma pesquisa ontológica com apoio de um dicionário idiomático com sinônimos, descobre-se, ainda, que a classe de objetos “mecânico” é uma instância de uma classe mais geral, que pode ser denominada “empregado”, e que, continuando a subir na árvore hierárquica dessa ontologia idiomática, essa classe “empregado” pertence a uma outra classe ainda mais primitiva – a classe “pessoa”. A classe “conserto”, com uso de dicionário, também pode representar (instanciar) um tipo de ação mais abstrata, ou uma classe mais genérica denominada “ação”, ou “transação”, no caso.

Outras estruturas sintáticas úteis para esse tipo de aprendizado ontológico são apresentadas como resultados da pesquisa experimental desta tese, sendo acrescentadas a outras descobertas anteriores, de outros pesquisadores, em outros contextos.

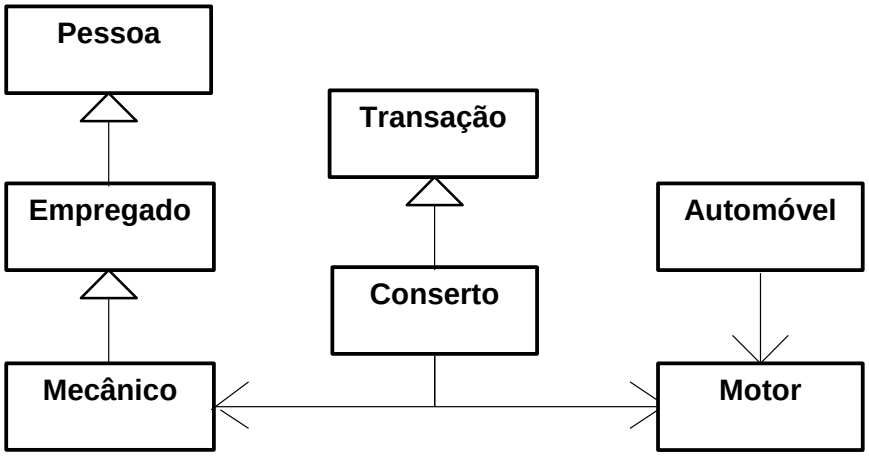



Figura 1.3 Modelo Ontológico de Sistema de Informação
(Fonte: do autor da tese)

Esta abordagem de mineração de texto é diferente das tradicionalmente utilizadas na recuperação da informação e no aprendizado ontológico, e portanto inovadora, porque não necessita de *corpus* idiomático extenso e variado, nem de metodologias de suporte a desambiguação. Com textos relativamente curtos e auxílio de um simples dicionário idiomático pode-se descobrir estruturas sintáticas úteis para a modelagem automática ou semi-automática de sistemas organizacionais.

## 1.4 Modelos como sensores semânticos

Os problemas específicos, em cujas soluções se pretende avançar, parcialmente, com o *framework* metodológico proposto, podem ser sintetizados nas seguintes questões que incomodam o mundo corporativo:

a) como reconhecer formalmente, a partir de fontes textuais disponíveis, os conceitos do negócio de uma organização para uso na Inteligência Competitiva?
b) como reconhecer, a partir desses conceitos, novas necessidades de informação de uma determinada organização em um contexto de competição no mercado?
c) como reconhecer, formalmente, os impactos de mudanças nos ambientes de negócios sobre os sistemas organizacionais legados de uma organização em particular?
d) que tipo de documento digital se mostra mais útil para o aprendizado de ontologias de sistemas organizacionais em contextos de negócios na era da *World Wide Web*?

e) que tipo de tecnologia de Inteligência Artificial (IA) se mostra mais produtiva para implementação da metodologia proposta?

E uma idéia utilitarista é que os modelos de sistemas organizacionais produzidos com a metodologia proposta nesta tese exerçam o papel de sensores semânticos para orientar avaliações e decisões corporativas em situações que exigem inteligência no negócio. Com o acréscimo de mais informações no contexto, pode-se, então, elaborar e reelaborar modelos de sistemas celeremente em escala razoável e compará-los na dimensão tempo, ou seja, pode-se comparar um modelo anterior, representativo de um cenário, com outro posterior a um evento provável ou projetado, e assim avaliar-se os prováveis impactos desse evento sobre os sistemas organizacionais legados.

Embora os modelos de representação do conhecimento empregados no experimento sejam bastante expressivos, em linguagem natural, e com razoável poder de resolução semântica (desambiguação) com base na composição de sintagmas nominais[8] e nas relações entre conceitos, a última palavra nesse tipo de avaliação continuará sendo de responsabilidade dos seres humanos envolvidos, que serão, em última análise, os executivos tomadores de decisões nas organizações. Com este recurso metodológico baseado na psicologia, os engenheiros do conhecimento poderão desencadear processos mentais estruturados de criação de significado e construção de conhecimento para Inteligência Competitiva (WEICK, 1995; CHOO, 2003).

## 1.5 Justificativa

Vários fatores se apresentam como justificativas para o desenvolvimento de uma tese de doutorado, alguns de natureza motivacional essencialmente acadêmica, outros de natureza mais personalística, com base no estoque de conhecimento e no sistema de crenças do estudante-pesquisador. O caso presente não constitui exceção e, por isso, apresenta-se a seguir algo sobre esses fatores essenciais, unindo aspectos motivacionais e de oportunidade.

### 1.5.1 Motivação

O desenvolvimento desta tese teve como motivação acadêmica dois aspectos centrais: o primeiro, numa perspectiva histórica da própria Ciência da Informação, com base nas “grandes questões” enunciadas por Bates (1999); o segundo, numa dimensão contemporânea, com o

[8] *Sintagma* (KOCH e SILVA, 1985; *apud* SANTOS, 2005) *consiste num conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantêm entre si relações de dependência e de ordem*. Organizam-se em torno de um elemento fundamental, denominado núcleo, que pode, por si só, constituir o sintagma. Os sintagmas nominais são sintagmas cujo núcleo é um nome ou substantivo e Perini (2003) denomina de “sintagma complexo” o sintagma que contém oração subordinada, onde geralmente se encontram mais de um substantivo/nome.

advento da *World Wide Web* e o "oceano de informação" aberta e disponível para uma *praxis* de Gestão da Informação e do Conhecimento e inteligência de negócios nas organizações.

Bates (1999, p. 1.048), "seguindo a informação"[9], assim enunciou três grandes questões contemporâneas que se apresentam à Ciência da Informação:

> *As três Grandes Questões podem ser identificadas na seguinte estrutura básica: (1) A questão física: Quais são as características e leis do universo da informação registrada? (2) A questão social: Como as pessoas se relacionam, procuram e usam informação? (3) A questão de projeto: Como pode o acesso à informação registrada ser executado de modo mais rápido e efetivo?*

Compreende-se, na tese apresentada, uma abordagem de solução de problema que se refere, de certo modo, a essas três grandes questões. O experimento mostrou, por exemplo, como se pode, em um mesmo projeto, realizar uma pesquisa de base e desenvolver um artefato de engenharia utilizando ciência e tecnologia da informação.[10]

A questão física de Bates (1999) se reporta às descobertas científicas de base, que nas origens da Ciência da Informação se concentraram na análise estatística e sintática de textos e que, atualmente, é sintetizada numa subdisciplina denominada Processamento da Linguagem Natural (PLN). O experimento que embasou esta tese se tornou viável somente graças à disponibilidade de recursos de *software* de PLN com capacidade para realização de mineração de texto com associação de padrões existentes na linguagem natural (*collocation*[11]*, concgram*[12] e outros). Como restou evidente no experimento, parece que existem, ainda, padrões sintáticos e semânticos a serem descobertos nas estruturas desse universo da informação registrada aguardando esforços de pesquisadores com paciência e talento para desvendá-los.

A segunda grande questão de Bates (1999) se relaciona ao ser humano e seu comportamento informacional, que acaba produzindo os padrões de texto úteis em projetos com uso de PLN para engenharia da informação e do conhecimento. Como discutido no Capítulo 5, as organizações apresentam, em seus *Web* Portais, uma mensagem para o mundo que contem, na essência, o seu próprio "DNA informacional" (FULD, 2007), com atributos que, devidamente analisados, podem levar o Engenheiro do Conhecimento a um aprendizado útil para Inteligência Competitiva.

Esse tipo de pesquisa, sobre o comportamento do usuário da informação nos repositórios digitais, sempre representou uma área de interesse na Ciência da Informação e, atualmente, teve

---

[9] Esse lema de Bates (1999) é similar ao de Fuld (2007) em relação à Inteligência Competitiva com informações da *Web*: "siga as transações do mercado".

[10] É interessante perguntar-se, a propósito, por que a conhecida organização *The American Society for Information Science* (ASIS) mudou, em 2000, seu nome para *American Society for Information Science and Technology* (ASIST), acrescentando o "T" em sua sigla.

[11] *Collocation*: corresponde a um padrão qualquer de posicionamento de palavras observado repetidamente, em linguagem natural, no qual algumas palavras aparecem na vizinhança da palavra utilizada como argumento de busca (SCOTT, 2008).

[12] Conforme Cheng, Greaves & Warren, *apud* Scott (2008, p. 224), *concgram* é o conjunto de permutações da variação constitutiva e posicional gerada pela associação de duas ou mais palavras, significando que um *concgram* em particular pode constituir a fonte para um número de padrões de posicionamento *(collocation).*

sua importância reconhecida além dos limites das bibliotecas, no universo da *World Wide Web* e das grandes empresas fabricantes de *softwares* para estações de trabalho.

A terceira grande questão é um problema de engenharia *par excellence*, que exige o desenvolvimento de tecnologias específicas para solução de problemas. Esta tese, baseada num experimento de laboratório (mas, obviamente, com intuição racional apriorística), reúne características de engenharia que poderão contribuir para o aumento do desempenho dos processos de busca inteligente por informações conceituais em repositórios digitais. O conceito de "busca inteligente", neste caso, se refere à tentativa de igualar o desempenho humano numa dimensão-chave desse empreendimento: utilidade num contexto.

Com o "oceano" de informação digital disponível na atualidade, o modelo de busca ideal será, certamente, aquele composto por recursos de busca ultraveloz e "cirúrgica", retornando apenas um mínimo de informação com o máximo de precisão possível no contexto. Com esses requisitos funcionais, os filtros dos *softwares* de busca precisam ser cada vez mais velozes e precisos, outra importante linha de pesquisa na ciência e tecnologia da informação que consome quantias cada vez mais fantásticas de recursos anualmente nas grandes empresas do mercado.[13]

O problema de engenharia é esmiuçado, de modo mais didático, por Bates (1999, p. 1.048, com grifos nossos) com os seguintes argumentos:

> *Em conseqüência da complexidade lingüística, psicológica, cognitiva, social e técnica da recuperação da informação, cada aumento no tamanho da fonte de informação ou base de dados requer soluções diferentes; escalabilidade é um problema fundamental nesse campo. Eu creio que de um ponto de vista um historiador mostrará que a explosão da informação (conosco desde a invenção da impressão) tem direcionado a maioria das maiores inovações na organização e no acesso à informação. Cada vez que a coleção cresce, em média, para um novo nível, um novo método de acesso precisa ser inventado. Isto tem sido particularmente notável no último século – do desenvolvimento de estruturas de indexação ao desenvolvimento de hyperlinks.*

Como toda motivação para um empreendimento humano também apresenta componentes de natureza pessoal, outros dois trechos seminais de Bates (1999, p. 1.049) contribuíram para o projeto desta tese, algo no conceito de "desafio multidisciplinar" da Ciência da Informação:

> *Em primeiro lugar é importante se reconhecer que a ciência da informação, como a educação e o jornalismo, entre outros, é um campo que atravessa, ou que é ortogonal a, as disciplinas acadêmicas convencionais.*
>
> *(...) a ciência da informação tende a atrair pessoas multitalentosas, pessoas que se comprazem da mistura de tipos de cognição que a natureza do nosso campo exige*

[13] O nível de investimentos em pesquisa e desenvolvimento nessa área pode ser avaliado comparando-o, por exemplo, com o tamanho do interesse da Microsoft pela aquisição de uma empresa de portal de busca na *Web* como o Yahoo!. Gates, da Microsoft, explicou recentemente essa empreitada*: Nós temos uma estratégia para competir no espaço de busca que a Google atualmente domina, que perseguiremos como temos feito antes de termos feito a oferta à Yahoo, e que podemos perseguir sem isso. Envolve engenharia de ponta. Nós pensamos que a combinação com a Yahoo aceleraria as coisas de um modo muito excitante porque eles têm grandes engenheiros, eles têm realizado um bocado de grandes trabalhos. Assim, se você combina o trabalho deles e o nosso trabalho, a velocidade na qual você pode inovar e ter as coisas executadas é dramaticamente mais rápida. Então, é na verdade sobre as pessoas que estão lá e que querem se unir e criar um sistema de busca melhor, um portal melhor para um conjunto muito amplo de clientes. Esta é a visão que se encontra por trás quando dizemos 'hey, esta não seria uma grande combinação?'* (CNET, 2008)

> *para solução de seus problemas de pesquisa. Esta é uma das razões pelas quais nós temos falhado em consolidar-nos como um campo em torno de um paradigma metodológico padrão.*

Outro aspecto dessa motivação para a pesquisa, num plano pessoal do pesquisador, se refere ao interesse e utilidade de outra área multidisciplinar envolvida no contexto: a Inteligência Artificial. Os conceitos de informação, conhecimento e inteligência estão cada vez mais interconectados na era da informação digital e a tese que se apresenta elabora, de um ponto de vista pragmático, uma síntese epistemológica envolvendo esses três aspectos sociotécnicos do comportamento humano individual e coletivo nas organizações. A Inteligência Artificial, pela sua complexidade ao pretender emular a inteligência humana, não tem limites naturais de atuação em termos de objeto de pesquisa, outro atrativo para quem aprecia temas multidisciplinares e transdisciplinares.

Objetivamente, os modelos mentais (ou modelos conceituais) desenvolvidos nesta tese se fundamentam em teses da Inteligência Artificial que se encontram, contemporaneamente, também no discurso de pesquisadores de outras disciplinas, tais como a educação. Com o advento da era da informação digital, percebe-se que a Inteligência Artificial passa a ser tratada como algo natural, sem preconceitos, na medida em que alguns de seus produtos mais evidentes são incorporados ao padrão sociotécnico do dia-a-dia das pessoas, como sistemas de informação computacional embarcados em máquinas e aparelhos eletrônicos diversos (automóveis, microcomputadores, termostatos, telefones celulares, etc).

Ontoria, De Luque e Gómez (2006, p. 41), por exemplo, colocam o tema dos modelos mentais com a seguinte argumentação (grifos nossos):

> *(...) a origem dos mapas mentais (Buzan, 1996:44) é proveniente dos estudos sobre a memória, quando se teve consciência de que associação e ênfase são dois fatores fundamentais para a permanência da lembrança e sua evocação posterior. O agrupamento de conceitos e idéias cria estruturas cognitivas que, na dinâmica do pensamento, relaciona-se entre si ou com outras estruturas novas. Outro fato a ser ressaltado como origem dos mapas mentais é a busca de uma técnica que ajude a memorizar e que, posteriormente, evoluiu para uma técnica do pensamento (Buzan, 1996:168). Igualmente, ressalta-se outro fato como referencial da criação dos mapas mentais, que é a reflexão acerca da poderosa capacidade dos computadores para estabelecer relações entre a palavra e a imagem, e trabalhar em conjunto com ambas: o cérebro tem mais capacidade (Buzan, 1996:87). Finalmente, indica-se como origem dos mapas mentais "o pensamento criativo ou brainstorming", pois o mapa mental é uma manifestação do pensamento criativo.*

As noções de "associação" e "ênfase" (ou "relevância") se encontram na essência da metodologia de mineração e modelagem de conceitos proposto nesta tese. Com esses argumentos, pode-se compreender também, por exemplo, porque o Centro para Sistemas Adaptativos do Departamento de Sistemas Cognitivos e Neurais da Universidade de Boston[14] é uma unidade de pesquisa acadêmica em Inteligência Artificial e Neurociência com interesses que intersectam disciplinas tradicionais tão distintas quanto a Biologia, Ciência da Computação, Engenharia, Matemática e Psicologia. Os principais pesquisadores desse centro desenvolveram

[14] Ver: *http://cns-web.bu.edu/about/cas.html*).

um tipo de redes neurais artificiais emulando o que se conhece sobre os sistemas neurológicos do aprendizado humano – as redes ART (CAPUANO, 2009; CAPUANO e NEHME, 2002) e, atualmente, se dedicam também ao desenvolvimento de sistemas inteligentes para apoio a pesquisas neurais na área biomédica.

### 1.5.2 Oportunidade

Quanto à oportunidade, ela se apresenta na medida em que a Faculdade de Ciência da Informação da Universidade de Brasília (FCI/UnB) realiza sua transição de um modelo acadêmico disciplinar, centrado na Biblioteconomia e na Arquivologia, para um modelo multidisciplinar, com linhas de pesquisa na pós-graduação tão modernas quanto Gestão da Informação e do Conhecimento, Arquitetura da Informação e Comunicação da Informação. Esse ambiente tem se tornado atraente para profissionais com os mais variados perfis acadêmicos e práticos, com bagagens tanto das Ciências Sociais como das Ciências Exatas e Engenharias, constituindo-se, gradativamente, um *locus* privilegiado no Brasil para o desenvolvimento da pesquisa multidisciplinar em Ciência da Informação.

Com esta tese, vislumbrou-se uma oportunidade de acrescentar alguns conhecimentos e experiências ainda não desenvolvidos na FCI/UnB, e talvez ainda não desenvolvidos em nenhum outro centro acadêmico conhecido. E não obstante as inovações em PLN apresentadas na tese, pretende-se com a pesquisa mostrar também a utilidade de algumas disciplinas (ou subdisciplinas) correlatas, por natureza, à Ciência da Informação, como Linguística, Ciência da Computação, Engenharia de Sistemas e Psicologia.

# 2. PROBLEMA, TESE, OBJETIVOS E RELEVÂNCIA DA PESQUISA

## 2.1 Problema geral

Considerando-se, conforme Bates (1999), que se trata de uma tese típica de "engenharia do conhecimento", o problema geral para o qual se busca o desenvolvimento de uma solução, ainda que parcial, na proposta de pesquisa desta tese poderia ser resumido com a seguinte pergunta objetiva: "como o 'arsenal' metodológico e tecnológico da Inteligência Artificial (IA) poderia contribuir para o desenvolvimento de processos automáticos, ou semi-automáticos, de transformação de informação textual das (e sobre as) organizações em conhecimento ontológico útil para a modelagem de seus sistemas de negócios?" O "como", no caso, se reporta a métodos e técnicas de Processamento da Linguagem Natural (PLN) com recursos de Inteligência Artificial (IA), buscando-se rotas de desenvolvimento de uma solução metodológica para este problema-desafio que, na Ciência da Informação, é comentado por Ferneda (2006), Ebecken, Lopes e Costa (2005), Benabdellatif (2002) e outros.

O modelo que define o contexto e a estrutura epistemológica do problema, num recorte de Gestão da Informação e Conhecimento, é apresentado na Figura 2.1, onde o artefato denominado "Árvore de Problemas" representa um construto desenvolvido por pesquisadores da área de planejamento estratégico e adotado como um dos pilares conceituais da Metodologia do Marco Lógico do *Instituto Latinoamericano y del Caribe de Planificación Económica y Social – ILPES*.

De um ponto de vista dos fluxos de informações nas organizações, a baixa capacidade competitiva pode resultar, principalmente, de uma percepção equivocada da cadeia produtiva no mercado em que se insere a organização, problema cujos sintomas mais evidentes são o uso de bases de informações inadequadas (incompletas ou insuficientes).

O uso do modelo mental[15] da Figura 2.1 para definição conceitual e pragmática do problema se justifica pela sua similaridade no contexto. Ortegón, Pacheco e Prieto (2005, p. 15) argumentam, em defesa desse modelo de representação, que "o processo de planejamento nasce com a percepção de uma situação problemática e a motivação para solucioná-la". A Árvore de Problemas da Figura 2.1 é construída com os seguintes procedimentos consecutivos:

I. Analisar e identificar o que se considera como principais problemas da situação.
II. A partir de uma primeira "tempestade" de idéias, estabelecer o problema central que afeta os interessados em sua solução (uma comunidade, no caso do planejamento estratégico governamental), aplicando critérios de prioridade e seletividade.
III. Definir os efeitos mais importantes do problema em questão.
IV. Anotar as causas do problema central detectado.

[15] *Modelo mental* é uma *representação pictórica de um fenômeno do mundo real* utilizado para se obter maior eficiência em processos cognitivos de análise para tomada de decisão em contextos.

V. Construir a árvore de problemas desenhando retângulos com textos identificando causas e efeitos abaixo e acima do retângulo com o texto do problema central.

VI. Verificar a consistência do modelo completo, em relação à identificação das reais causas e efeitos e seus interrelacionamentos.

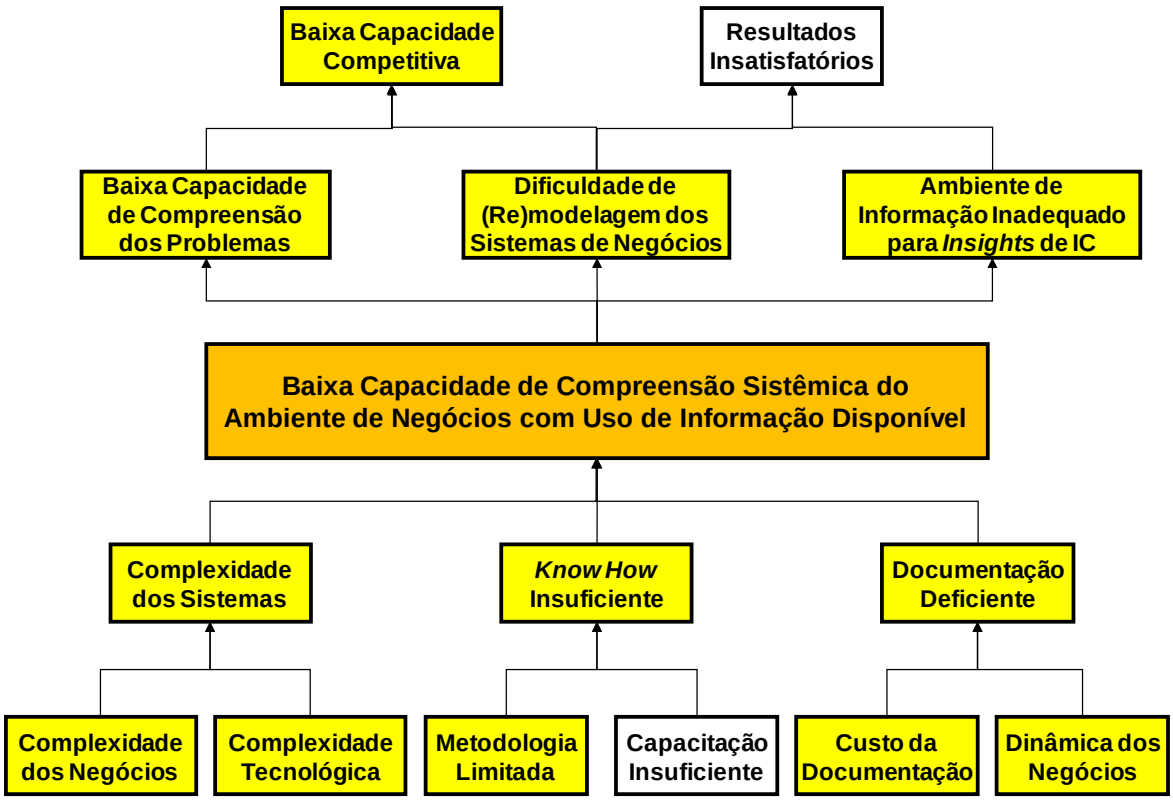



Figura 2.1 Árvore de Problemas

A árvore de problemas deve mostrar uma imagem completa da situação negativa existente, sintetizada na baixa capacidade de compreensão sistêmica do ambiente de negócios como problema central atacado na tese. Essa ausência de visão sistêmica seria a principal causa da dificuldade que as grandes organizações têm encontrado na remodelagem de seus sistemas legados de suporte tecnológico para readaptação contínua ao ambiente de negócios mutante (BELL, 2008; PULIER e TAYLOR, 2008; RITTO, 2005; HAMMER e CHAMPY, 2001; DAVENPORT, 2000; DAVENPORT, 1993; MCGEE e PRUSAK, 1994).

## 2.2 Tese, objetivos gerais e específicos

### 2.2.1 Tese

A tese defendida pode ser resumida na seguinte afirmativa: “é viável estruturar-se uma *praxis* de Gestão do Conhecimento para Inteligência Competitiva com base em informações

conceituais relevantes recuperadas de fontes digitais abertas e modeladas *a priori* das demandas dos usuários, com uso de Inteligência Artificial." E, contrariando outra crença tradicional na Inteligência Competitiva, que pressupõe a relevância como um valor subjetivo de acordo com a necessidade do usuário da informação, esta tese inclui, consequentemetne, uma subtese: "é possível separar-se a informação mais relevante num contexto de modo formal, não subjetivo."

### 2.2.2 Objetivo geral

O objetivo geral da pesquisa é o desenvolvimento de uma metodologia experimental de mineração e modelagem conceitual de informação textual para uma *praxis* de Gestão do Conhecimento em ambientes organizacionais de Inteligência Competitiva. Com essa metodologia, pretende-se defender a tese proposta, apresentando-se para tanto uma experiência de simulação em contexto empírico, que para Peirce (2010) configura um experimento de retrodução (buscando verificar uma hipótese provisória).

O proposto uso da linguagem natural, para representação e comunicação da informação nessa metodologia, se deve à necessidade de modelagem de sistemas organizacionais compreensíveis tanto por tecnólogos de sistemas e informação como por operadores, analistas e executivos de negócios, de modo que os modelos resultantes exerçam o papel de uma "interlíngua" no interior das organizações (EKLUND, ELLIS e MANN, 1996).

### 2.2.3 Objetivos específicos

Como objetivos específicos, a metodologia a ser desenvolvida e testada experimentalmente em laboratório de simulação computacional deverá apontar:

I. Quais recursos do "arsenal" de Inteligência Artificial (IA) são mais promissores na missão de elicitação de conhecimento ontológico de organizações a partir de fontes de informação textuais digitais abertas das organizações e sobre as organizações.
II. Qual a sensibilidade do modelo computacional em relação a mudanças de contexto nas variáveis de entrada. Ou seja, observando-se e avaliando-se as variações nos resultados (saídas do modelo) em função das informações de entrada, exercício conhecido, no jargão da simulação experimental, como questões *what-if* (e-se?), qual será o poder explicativo da metodologia proposta.
III. Qual o nível de interoperabilidade sociotécnica possível das ontologias produzidas pela metodologia. Ou, qual sua capacidade de superação das barreiras de linguagem intra-organizacional e integração de equipes de conhecimentos diversos.

Com isso, a metodologia será testada em sua robustez, buscando-se avaliar sua capacidade de mineração de informações e elicitação de conceitos para modelagem semi-

automática de sistemas nas organizações a partir de fontes de informações textuais disponíveis. Estima-se que, em caso de êxito na defesa da tese principal, restará evidente que o arsenal de IA representa um recurso decisivo para alavancagem de meios de suporte tecnológico a estratégias de gestão da informação e do conhecimento em contextos funcionais de inteligência competitiva nas organizações.

## 2.3 Premissa e hipótese

A premissa na qual se embasa o desenvolvimento da tese se reporta à capacidade do repertório epistemológico da Ciência da Informação para tratamento do problema num modelo explicativo multidisciplinar. Ou seja, parte-se da premissa que se pode desenvolver construtos metodológicos para solução do problema conectados aos temas disciplinares da Ciência da Informação, na visão evolucionária de Bates (1999), com uma abordagem típica de *engenharia*.

E a hipótese de partida para desenvolvimento é que os recursos disponíveis de Inteligência Artificial (IA) implementados em *softwares*, principalmente, de Processamento da Linguagem Natural (PLN) são suficientes para suportar a elaboração de uma metodologia eficaz e eficiente para mineração de conceitos de negócios em textos digitais de fontes abertas, com escalabilidade e robustez adequadas para uso em grandes repositórios de conteúdos em linguagem natural.

Esta abordagem de engenharia na Ciência da Informação, inspirada em Bates (1999), é necessária para se testar a hipótese, ancorando-se também na espécie fundamental de raciocínio da *retrodução* (às vezes denominada *abdução*) mencionada por Aristóteles:

> *Retrodução é a adoção provisória de uma hipótese em virtude de serem passíveis de verificação experimental todas as suas possíveis conseqüências, de tal modo que se pode esperar que a persistência na aplicação do mesmo método acabe por revelar seu desacordo com os fatos, se desacordo houver.* (PEIRCE, 2010, p. 6)

## 2.4 Relevância da pesquisa

A pesquisa é relevante na medida em que não se concebe, no estado-da-arte da Inteligência Competitiva, construtos metodológicos de Engenharia do Conhecimento para uma gestão do conhecimento apriorística nessa área, concentrando-se os esforços em processos de busca da informação sob demanda, em caráter geralmente específico *(ad hoc).* Fuld (2007, p. 25-26), por exemplo, alerta para o que ele considera ser o caráter efêmero dessa inteligência e sua natureza eminentemente pragmática:

> *(...) a inteligência tem um breve instante de vida; ela é muito pessoal, muito customizada, quase única.*
>
> *(...) Você não pode massificar a inteligência. Inteligência, por sua natureza, é válida somente por um curto período de tempo para poucos indivíduos. Uma vez que todos sabem do "insight", não há mais insight; este se torna uma informação comum.*

> *Perde-se a oportunidade de destruir o mercado. Perde-se a oportunidade do elemento surpresa.*
>
> *(...) Inteligência é também um produto muito pessoal, já que ajuda a direcionar a forma como você gerencia e toma decisões.*

Com o desenvolvimento de uma metodologia de Engenharia do Conhecimento para modelagem semi-automática de sistemas organizacionais, espera-se que uma organização complexa possa alcançar maior capacidade competitiva e melhores resultados com base numa maior capacidade de compreensão sistêmica do ambiente de negócios com uso da informação disponível. Essa ampliação cognitiva em nível holístico do ambiente competitivo proporcionará uma maior capacidade de compreensão dos problemas da organização no seu nicho de competição e a construção de um ambiente de informação mais adequado para a geração de *insights* de Inteligência Competitiva *a priori* das demandas dos usuários.

Como resultado, espera-se também o aumento da eficiência operacional (relacionado à remodelagem dos processos e serviços), maior capacidade de mudança (relacionada à readaptação sociotécnica) e a consequente atualização tecnológica (propiciada pelo desenvolvimento dos outros dois fatores correlatos).

O projeto também se fundamenta, em termos de relevância, nas implicações sociotécnicas e econômicas negativas decorrentes da estagnação evolutiva da gestão da informação e do conhecimento "abaixo da linha d'água"[16] nas organizações complexas, com prejuízos econômicos e desperdícios de oportunidades. E, do ponto de vista do mérito acadêmico, no desafio intelectual de desenvolver uma solução tanto genérica quanto possível para um problema considerado complexo na modelagem de recursos informacionais, superando as limitações dos estudos de casos na Ciência da Informação.

O conceito de "conceito" (KEIL e WILSON, 2000) representa, nesta tese, a "cola", a chave epistemológica para integrar a enorme massa de informação digital disponível em construtos gráficos de suporte ao desenvolvimento de inteligência nas organizações. Conceitos são idéias que constituem as maiores unidades mentais dos seres humanos, como os pensamentos, tornando-os centrais em nossa existência do ponto de vista cognitivo.

[16] Este termo, utilizado por Bates (1999), se refere às aplicações concretas e tangíveis da Ciência da Informação que ocorrem nas organizações.

# 3. REVISÃO DE LITERATURA

Considerando tratar-se de um tema multidisciplinar que integra várias áreas clássicas do conhecimento científico, como Linguística, Filosofia, Matemática, Psicologia, Administração e Ciência da Computação, num construto de Recuperação Conceitual da Informação[17] com uso de recursos da Inteligência Artificial (IA), a literatura disponível é vasta e dinâmica, com um volume expressivo de novas e importantes contribuições publicadas a cada dia. A revisão apresentada a seguir é uma fração suficiente desse universo e bastante atual, até porque o conhecimento basilar de métodos e técnicas para automação dos processos de Recuperação da Informação (RI) desenvolvido na segunda metade do Século XX continua sendo útil nos novos contextos de repositórios digitais da *Web*.

Os conteúdos recuperados para estudo nesta tese, no entanto, constituem um conjunto conceitualmente coeso, buscando-se integrar esse conhecimento fragmentado numa nova síntese em Ciência da Informação.

## 3.1 Histórico epistemológico

### 3.1.1 Origens na Ciência da Informação

O tema da pesquisa é algo conceitualmente recorrente, ainda que sem soluções genéricas, inserindo-se no atual cenário corporativo de abundância de informações e pouco êxito dos projetos de sistemas de informação (SIs) baseados em computador em ambientes competitivos (AL NEIMAT, 2007; SIMPSON *et al*., 1996). Este paradoxo se amplia na medida em que quanto mais informações acumulam as organizações, mais complexos se tornam seus ecossistemas informacionais e, necessariamente, seus sistemas de informação computacionais (SICs).

Os problemas conceituais inerentes à área de estudos conhecida como Recuperação da Informação (RI), a partir de fontes não-estruturadas tais como textos em linguagem natural (ou textos livres) e gravações de áudio e vídeo, constituem desafios formidáveis que têm motivado esforços da comunidade da Ciência da Informação desde os seus primórdios nos anos 1950, após a publicação do histórico artigo de Vanevar Bush lançando a idéia do *MEMEX* (BUSH, 1945), considerada uma das precursoras do hipertexto. A quantidade de autores e obras acadêmicas produzidas desde então é vasta, como se pode observar nas publicações patrocinadas pela

[17] O conceito de *informação* adotado nesta tese é discutido, em profundidade, no Capítulo 3 (item 3.5), mas, provisoriamente, pode-se utilizar o conceito filoficamente ingênuo de Drucker, *apud* Davenport (2000, p, 19): *informação é um dado dotado de relevância e propósito*. Entretanto, reconhecendo também a impossibilidade filosófica (fenomenológica) da existência de *dado*, adota-se, nesta tese, um conceito tautológico e pragmático: *dado é o conteúdo de um banco (ou de uma base) de dados* (os "dados" utilizados no experimento de laboratório, analisados no Capítulo 5, foram extraídos da maior base de dados digital aberta existente: a *World Wide Web)*.

*Association for Computing Machinery* (ACM)[18], com fontes de referência bibliográfica de embasamento do presente projeto de pesquisa, confirmando as conclusões de Saracevic (1999) quanto ao perfil desses profissionais da informação. Os artigos seminais de Luhn (1957), Good (1958), Doyle (1965), Minsky (1968), Salton (1970) e outros, até contribuições mais recentes, podem ser encontrados na biblioteca digital da ACM.

A obra de Araújo Jr. (2006) oferece uma visão panorâmica condensada do estado-da-arte das metodologias e tecnologias desenvolvidas para recuperação automática de informação, comparando o desempenho de processos manuais com processos baseados em computação eletrônica. Essa obra mostra que a RI executada com *software* pode não apresentar vantagem qualitativa sobre o processamento manual do ponto de vista da precisão, ainda que com a automação os ganhos em velocidade sejam inquestionáveis.

Entretanto, do ponto de vista da gestão do conhecimento o maior desafio continua sendo o da recuperação inteligente de informação dessas fontes brutas (textos livres, áudio e vídeo), de modo a suportar processos de extração, representação e mapeamento da informação e inspirar a geração do conhecimento individual em um determinado domínio ontológico. Belew (1987, p. 1), por exemplo, cita metáforas de um magistrado do início do Século XX para ilustrar, numa linguagem algo poética, a inerente complexidade epistemológica do tema: "'... (uma) palavra não é um cristal, transparente e imutável, mas é a pele de um pensamento vivo.' Juiz Chefe Holmes, Towne v. Eisner, 191".

Evolutivamente, a terminologia dos pesquisadores que se ocupam de problemas de RI incorporou a expressão "recuperação conceitual de informação" (RCI) para destacar os aspectos mais essenciais relativos a conteúdos extraídos das fontes de informação. Belew (1987) justifica qualquer esforço de pesquisas nessa direção argumentando que existe muito a melhorar na capacidade de recuperação de informações relevantes nos sistemas de recuperação de informação (SRI) baseados nas metodologias e tecnologias tradicionalmente utilizadas – emprestadas da estatística, da lógica e da matemática – e ressalta as possíveis contribuições do campo multidisciplinar da IA para tanto.[19]

Os métodos tradicionais de RI com uso de técnicas estatísticas e lógico-matemáticas são reconhecidamente limitados para a missão de extração de conhecimento de textos em linguagem natural (BELEW, 1987), o que tem motivado pesquisas com uso do arsenal de IA na RI a partir dos anos 1980. Os artigos de Koll (1979), Lebowitz (1983), Zarri (1983), Tong, Askman, Cunningham e Tollander (1985), Breuker e Wielinga (1987), Belew (1987), Brachman e

[18] Considerando-se que, em geral, cada obra citada tem como referências mais de uma dezena de outras obras correlatas, pode-se estimar o volume de publicações nessa área nos últimos 50 anos.
[19] O artigo citado apresenta os resultados da aplicação de um SRI denominado AIR sobre a conhecida base de informações jurídicas WESTLAW, nos EUA.

McGuinness (1988) e Rose e Belew (1989) apenas ilustram essa guinada histórica, que coincide com o ressurgimento do interesse acadêmico sobre a IA após as frustrações do período anterior.[20]

Cooper, um pesquisador da área de biblioteconomia, apresentou o problema, ainda na década de 1980, da seguinte forma:

> *Recuperação da informação [...] inclui uma preocupação com sistemas 'especialistas' ou 'baseados no conhecimento' e seus futuros sucessores. É improvável que sistemas sofisticados desse tipo possam ser desenvolvidos de modo a utilizar uma linguagem inteiramente natural sem assistência de uma teoria avançada e unificada da linguagem e da lógica* (COOPER, 1984, p. 259).

Outro conceito-chave correlato ao trabalho programado nesta tese é o de "mineração de textos", uma variante evolutiva do conceito de "mineração de dados", entendendo-se mineração de dados como o processo heurístico de extração de conhecimento útil a partir de massas de dados preexistentes nas organizações, geralmente armazenadas em bancos de dados de sistemas computacionais. Ebecken, Lopes e Costa (2005) conceituam mineração de textos como um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos, com desdobramentos mais atuais no mundo dos negócios a partir da descoberta de novas dimensões nas informações disponíveis nas empresas. Contudo, observando-se a literatura da Ciência da Informação desde seus primórdios, percebe-se que os tecnólogos da Ciência da Computação trilharam caminhos diferentes para alcançar praticamente o mesmo objetivo da comunidade da Ciência da Informação num período anterior.

De um modo geral, as tecnologias úteis para aquisição de conhecimento em qualquer atividade humana a partir dessas massas de informações brutas (dados e textos, principalmente) derivam de campos da pesquisa lógico-matemática e estatística e da Inteligência Artificial (IA). O corpo de conhecimentos de IA se baseia, conceitualmente, nos conhecimentos de outras disciplinas tais como Biologia, Ciência da Computação, Estatística, Neurologia e Psicologia.[21] E o que, de fato, se produz no campo da IA é a "cola", a engenharia de sistemas para integração e combinação de conceitos dessas disciplinas subjacentes em aplicações complexas que exigem abordagem multidisciplinar.

Essas metodologias e tecnologias podem ser exploradas em todo seu potencial somente porque tecnologias da "Big Science", como redes de comunicação de dados, sistemas de bases de dados e microcomputadores, estão disponíveis no mercado. A *Association for Computing Machinery* (ACM) publicou, recentemente, referências a uma vasta coleção de dissertações versando sobre uso de IA em contextos de IR e extração de conhecimentos de bases de dados e fontes de informações textuais (ACM, 2007).

---

[20] As idéias seminais da IA coincidem com os primórdios do desenvolvimento da ciência da computação, quando se imaginava que com a computação eletrônica se poderia emular o próprio cérebro humano.

[21] Observe-se que *Lógica de Predicados, Matemática, Ciência da Computação e Lingüística* são disciplinas presentes na *Mineração de Textos*, constituindo o campo de pesquisas interdisciplinares denominado *Processamento da Linguagem Natural* (PLN).

### 3.1.2 Cenário sociotécnico multidisciplinar

O cenário sociotécnico onde se encaixa esta pesquisa multidisciplinar em Ciência da Informação é algo parecido com o da época de Bush e de seu artigo seminal na Ciência da Informação – *As We May Think*, que mais tarde incentivaria toda uma comunidade de cientistas no desenvolvimento de métodos e técnicas para recuperação da informação (SARACEVIC, 1999). Simpson *et al.* (1996), contemporaneamente, publicaram um artigo na renomada *Association for Computing Machinery* (ACM), em comemoração aos 50 anos desse texto histórico publicado no *Atlantic Monthly*, que serviu de referência temática num simpósio acontecido em outubro de 1995, em homenagem ao visionário engenheiro do *Massachussetts Institute of Technology* (MIT), creditando-se a Bush, entre outras coisas, o pioneirismo de idéias tão paradigmáticas como o hipertexto, os grupos de trabalho em redes de computadores *(workgroup)* e a teoria das interfaces (IBIBLIO, 2007).

Esses autores apresentam excertos do texto original de Bush que motivou o desenvolvimento da ciência e da tecnologia da informação desde então e observam a similaridade entre o contexto ambiental da gestão da informação em 1945, nos EUA, e na atualidade. Bush ressaltava, por exemplo, que "[...] há uma nova profissão de pioneiros, aqueles que encontram deleite no estabelecimento de caminhos úteis no meio da enorme massa de registros comuns" (SIMPSON *et al*., 1996, p. 50), algo bastante familiar no atual cenário de "explosão da informação" (DAVENPORT, 2000, p. 11).

O atual cenário econômico, entretanto, difere um pouco do período pós-guerra de Bush, pois se os recursos "informação" e "conhecimento" são hoje reificados como armas estratégicas para a competição no mercado global, observa-se que as TIC encontraram seus limites sociotécnicos naturais como a outrora principal alavanca do desenvolvimento das organizações complexas. Este tipo de crítica às TICs é apresentado por McGee e Prusak (1994), Davenport (2000) e vários outros pesquisadores organizacionais. McGee e Prusak (1994) reforçam a tese das limitações das TICs com os seguintes argumentos:

> *Os investimentos em tecnologia da informação eram apregoados por vendedores, consultores e jornalistas como ferramentas que criariam uma revolução no mundo executivo. [...] mas a verdade é que muito pouco desse sonho se realizou. A pesquisa de Stephen Roach sugere que a produtividade dos executivos não cresceu muito nos últimos vinte anos, apesar da magnitude do investimento. [...] Um outro canto da sereia que atraiu muitos nas últimas duas décadas foi a idéia de que os investimentos em tecnologia da informação podem ser estratégicos, capazes de criar uma vantagem competitiva substancial. [...] embora existam evidentes vantagens comerciais a curto prazo, quando se é o primeiro a utilizar um novo sistema, o mercado fará com que tais vantagens continuem a ser de curto prazo* (MCGEE e PRUSAK, 1994, p. 6-7).

McGee e Prusak, além de consultores de empresas, são também profissionais praticantes da Ciência da Informação. Conforme Davenport (2000), Prusak tem experiência em biblioteconomia e McGee atuou com Davenport em pesquisas sobre o hoje difundido conceito de

*ecologia da informação* (DAVENPORT, 2000). Este último relata, como testemunha ocular privilegiada da realidade das empresas em termos de gestão da informação:

> *Na maioria delas, os ambientes informacionais são estarrecedores. Ninguém sabe o que sabe, ou o que precisa saber. Há pouca informação acessível sobre funcionários, clientes e até mesmo sobre os próprios produtos. Mesmo as empresas famosas pela aplicação de sistemas de informação específicos costumam contar com ambientes informacionais internos pobres* (DAVENPORT, 2000, p. 16).

O problema abordado se insere no ecossistema de negócios das organizações complexas[22] e necessita de uma explanação prévia de contexto antes de sua apresentação de modo mais específico. O conceito de "ecologia da informação" de Davenport (2000) e a experiência relatada de McGee e Prusak (1994) como profissionais praticantes da Ciência da Informação oferecem uma visão panorâmica introdutória bastante consistente no tema, apresentando desafios interessantes que motivaram a pesquisa. Davenport (2000, p. 43) assim sintetiza sua visão holística para abordagem dos problemas de informação nas organizações contemporâneas (com grifos nossos):

> *A ecologia da informação inclui uma gama muito mais rica de ferramentas do que aquela empregada pelos engenheiros e arquitetos informacionais. Os ecologistas da informação podem mobilizar não apenas designs arquiteturais e TI, mas também estratégia, política e comportamento ligados à informação, além de suporte a equipes e processos de trabalho para produzir ambientes informacionais melhores.*

Este capítulo busca definir os contornos epistemológicos históricos do problema e oportunidades de inserções temáticas em linhas de pesquisa voltadas para a Gestão da Informação e do Conhecimento na Ciência da Informação.

## 3.2 Contextos e conceitos

### 3.2.1 Inteligência competitiva nas organizações

Tarapanoff (2006a) explica que, na tradição francesa (numa visão assumidamente sistêmica), a inteligência competitiva:

> *(...) desenvolvida a partir da veille technologique[23], é entendida de forma mais ampla, incluindo a busca de qualquer informação na ambiência, de caráter científico, tecnológico, social ou político, sobre os seus competidores e também clientes, fornecedores e parceiros, que possibilite melhor posicionamento da organização na ambiência. [...] Está ligada ao planejamento estratégico da organização, ao seu posicionamento na ambiência e à estratégia* (TARAPANOFF, 2006a).

Conforme o pensamento de Liebowitz (2006), considera-se também que Gestão da Informação e do Conhecimento são construtos epistemológicos multidisciplinares capazes de suportar, metodologicamente, o desenvolvimento de inteligência competitiva em organizações complexas. Nonaka e Takeuchi (1997) colocam esse desafio metodológico estruturante da seguinte forma:

[22] Em geral, entendendo-se organizações complexas como grandes organizações privadas e públicas.
[23] *Veille Technologique*: vigília ou monitoramento do ambiente tecnológico (tradução aproximada).

> *A essência do 'problema da organização' (...) é transformar os agentes que buscam estrategicamente metas mutuamente conflitantes em um sistema cooperativo racional. E, devido à nossa capacidade limitada de processar as informações, o conhecimento é essencial para a garantia da racionalidade cooperativa. Barnard reconheceu a importância da integração dos processos lógicos e não-lógicos da atividade mental humana, do conhecimento científico e comportamental e das funções gerenciais (...)* (NONAKA; TAKEUCHI, 1997, p. 43).

Outros pesquisadores citados por Nonaka e Takeuchi (1997), como March e Simon (1993), também colocam o problema da complexidade e das limitações cognitivas dos seres humanos nas organizações como um desafio teórico na formulação de modelos de administração. Choo (1991), por exemplo, apresenta uma exaustiva revisão de literatura sobre teoria das organizações resgatando Cohen, March e Olsen (1972), que criticam as idéias de March e Simon (1993) e apresentam o "modelo da lata de lixo" e a "teoria da inteligibilidade" para explicar a complexidade das organizações. Eles também argumentam que a realidade é uma realização contínua que surge dos esforços de criar a ordem e tentar entender o que ocorre nas organizações.

Em termos concretos, as conexões entre os fenômenos sociotécnicos que constituem as condições de contorno da ambiência organizacional podem apresentar o macro-problema objeto da pesquisa proposta de modo silogístico[24]: considerando-se (P1) que as organizações contemporâneas operam seus processos de negócio com suporte de processos e sistemas de informação computacionais (SICs) baseados em tecnologia de informação e comunicação (TIC) e que (P2) mudanças cada vez mais freqüentes (e às vezes traumáticas) nos ecossistemas de negócios onde elas se inserem as obrigam a mudanças constantes desses processos para se readaptarem ao meio mutante, pode-se concluir que (C) essas mudanças necessárias de processos de negócios requerem, para se tornarem operacionais, mudanças nos processos decisórios e operacionais e nos SICs legados (ou o desenvolvimento acelerado de novos SICs cada vez mais complexos para suportar novos processos de negócio).

O cenário de mudanças de processos para readaptação das organizações aos seus ecossistemas de negócios é discutido, conceitualmente, por Hammer e Champy (2001) no conceito de "reengenharia". Como as organizações são dependentes de seus sistemas organizacionais para operar o negócio e das TIC que os suportam, as conexões entre mutações ambientais, processos, sistemas de informação e tecnologias são evidentes, problema comum em grandes organizações privadas e públicas em praticamente todo o mundo, inclusive em países em desenvolvimento como África do Sul, Brasil, Chile, Coréia, Sri Lanka e outros (CAPUANO, 2007; CZARNEWSKI, 2007; VAZ, 2007; FERRER e LIMA, 2007; GUTIÉRREZ, 2007; YOON, 2007; HOSKINS, DOBBERNACK e KUPTSCH, 2001).

Observa-se, também, nos casos relatados na literatura, que as soluções tradicionais de modelagem e desenvolvimento de SICs não mais respondem satisfatoriamente aos desafios do ambiente de negócio em constante mutação e McGee e Prusak (1994), citando Waller, assim

[24] P1: Premissa 1; P2: Premissa 2; C: Conclusão.

resumem, com desalento, o problema: “(...) Não há ciência para construção de uma metáfora que permita a representação dos requisitos de informação de uma organização de uma forma compreensível e utilizável por todos” (MCGEE; PRUSAK, 1994, p. 143).

Os processos de mudanças evolutivas no ecossistema organizacional apresentam pelo menos três pontos de dificuldade metodológica considerável para a concretização das soluções propostas pela reengenharia. O primeiro deles se refere à limitada capacidade do ser humano em lidar com a complexidade de modo imediato, problema apresentado por McGee e Prusak (1994) citando uma argumentação de Waller a respeito das relações complexas entre os objetos de informação das organizações:

> *Quando os seres humanos enfrentam a complexidade, precisam descobrir essas relações e (na vida organizacional) comunicá-las entre si de forma inteligível. A manipulação consciente de elementos e relações, entretanto, é uma função da memória imediata. (...) essa parte do sistema cognitivo dos seres humanos possui graves limitações. Isso nos leva a um problema de projeto (...) é preciso encontrar um meio que reconheça as características centrais da complexidade e ainda assim considere pontos fortes e fracos da cognição humana (...). Uma vez que nem a capacidade humana nem a complexidade podem ser alteradas, é preciso encontrar uma forma de estabelecer a ligação entre as duas sem, no entanto, alterá-las. Em linguagem de engenharia, é preciso encontrar um dispositivo de interface que estabeleça a ligação entre os seres humanos e a complexidade, ao mesmo tempo em que preserva as propriedades originais de ambos* (MCGEE; PRUSAK, 1994, p. 141-142).

O segundo ponto tem implicações econômicas imediatas de projeto de sistemas de informação (SIs), onde os profissionais da informação necessitam estimar, muito precocemente nos projetos, os custos inerentes aos recursos de desenvolvimento para subsidiar os estudos de viabilidade e o planejamento dos investimentos. Evidencia-se, assim, o retromencionado paradoxo temporal no ciclo de produção e consumo de informações e conhecimento na medida em que a atividade de modelagem de SIs, ainda que numa fase preliminar, pode consumir expressivo montante de recursos que deverão ser orçados antes da modelagem; e sem um certo nível de modelagem não será possível elaborar estimativas orçamentárias realistas desses recursos para execução do projeto. Questões correlatas são abordadas por Al Neimat (2007), Paula Filho (2005) e Pressman (1995), entre outros, ainda que sem soluções propostas no contexto que se apresenta.

E um terceiro ponto, de natureza estrutural nas organizações complexas, é que os processos de modelagem de SIs geralmente requerem informações e conhecimento dos operadores e gerentes de média hierarquia que são utilizados no dia-a-dia das equipes de “chão-de-fábrica” e de gerência, consistindo o trabalho de modelagem, em grande parte, na elicitação[25] de requisitos a partir do conhecimento tácito desse estamento e de conteúdos documentais disponíveis, mas desestruturados.

De pontos de vista mais instrumentais sobre mineração de dados e mineração de textos, Fayyad, Piatetsky-Shapiro, Smyth e Ramasamy (1996) observam que a busca de padrões em

[25] Anglicismo elaborado para transliteração do termo *elicitation.*

dados brutos tem sido identificada com terminologias variadas, tais como “descoberta de conhecimento em bases de dados”, “descoberta de informações”, “coleta de informações”, “arqueologia de dados e processamento de padrões de dados”. Ebecken, Lopes e Costa (2005) se concentram na definição e análise dos contextos de uso da mineração de textos nas organizações. Como cenário sociotécnico genérico, esses autores observam que todos os tipos de textos que compõem, atualmente, o dia-a-dia de empresas e pessoas são produzidos e armazenados em meios eletrônicos. E que:

> *Inúmeras novas páginas contendo textos são lançadas diariamente na Web. Outros tipos de documentos, como relatórios de acompanhamento, atas de reuniões e histórias pessoais são periodicamente gerados e atualizados. Entretanto, até pouco tempo atrás, essas informações em formato de textos não eram usadas para significar algum tipo de vantagem competitiva, ou mesmo como suporte à tomada de decisões, ou ainda como indicador de sucesso ou fracasso. Com o advento da mineração de textos, a extração de informações em textos passou a ser possível e o imenso e crescente mundo dos textos está começando a ser explorado. Em virtude desse crescimento contínuo do volume de dados eletrônicos disponíveis, técnicas de extração de conhecimento automáticas tornam-se cada vez mais necessárias para valorizar a gigantesca quantidade de dados armazenados nos sistemas de informação* (EBECKEN; LOPES; COSTA, 2005, p. 337).

O capítulo de livro desses autores apresenta uma introdução conceitual às metodologias mais comuns de mineração de textos, classificando-as em dois grandes grupos com abordagens diferentes: semânticas (que dependem do idioma) e estatísticas. Outros pontos de interesse nessa obra são a discussão das etapas da mineração de textos: (i) preparação de textos para mineração, envolvendo o conceito de recuperação da informação; (ii) processamento de dados textuais, apresentando os conceitos de indexação, extração de características, sumarização, clustering (agrupamento) e categorização; (iii) validação das descobertas; e (iv) comparação de ferramentas de mineração de textos disponíveis no mercado. E concluem com uma observação de Zanasi que fundamenta o problema apresentado neste trabalho de pesquisa:

> *(...) as empresas redescobriram suas informações já armazenadas em textos e estão utilizando essas informações como uma vantagem competitiva em relação aos seus concorrentes. A mineração de textos apresentou os principais recursos para que a inteligência pudesse ser efetivada (ZANASI, 2000)* (EBECKEN; LOPES; COSTA, 2005, p. 370).

O artigo de Silva, Prado e Ferneda (2002) é elucidativo em relação ao desenvolvimento de metodologias e tecnologias para mineração de textos, mostrando os problemas de priorização de esforços e interesses algo discrepantes entre a academia e a indústria.

Embora a análise dos dois grupos de cientistas da informação observados por Saracevic (1999) permita deduzir-se que as conexões lógico-cognitivas entre mineração de textos e inteligência competitiva, passando pelo implícito conceito de gestão do conhecimento, ainda sejam ignoradas por expressiva parcela da comunidade da Ciência da Informação, os exemplos de aplicações práticas, no dia-a-dia das organizações, estão se tornando cada vez mais evidentes. O caso descrito por Ebecken, Lopes e Costa (2005) é ilustrativo sobre o potencial dos

mecanismos de *crawler*[26] na *Web*, onde um grupo de empresários do turismo na Região dos Lagos, no Rio de Janeiro, decidiu montar um sistema de informação baseado em computador *on line* para mantê-los atualizados com informações geo-referenciadas sobre pontos de venda mais promissores para seus produtos, além de potenciais parceiros e concorrentes e respectivos modos de operação.

### 3.2.2 Recuperação da informação

A definição do termo "recuperação da informação" é uma tarefa mais simples do que a definição de "informação" isoladamente. Capurro e HjØrland (2007), depois de tempos pesquisando as raízes etimológicas do termo "informação" ao longo da história, descobriram mais de 500 significados para o mesmo – Capurro desistiu dessa empreitada e hoje se dedica à angelética, que é o estudo da mensagem (MATHEUS, 2005).

O conceito de recuperação da informação foi criado por Mooers em 1951, assim definindo os problemas endereçados por essa nova disciplina (GOTTSCHALG-DUQUE, 2005, p. 8):

> *(...) a recuperação da informação trata dos aspectos intelectuais da descrição da informação e sua especificação para busca e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação (MOOERS, 1951).*

Gottschalg-Duque (2005, p. 11) apresenta a *recuperação da informação* como:

> (...) *o clássico problema da recuperação efetiva e eficiente de documentos pertinentes extraídos de uma grande coleção (que nos dias de hoje pode ser entendida como um armazém de informação ou uma base de dados digital) de acordo com uma necessidade de informação específica, consistindo de três processos: coleta, indexação e ordenação*.

A coleta é um processo de identificação, avaliação e armazenamento; a indexação um processo de categorização com uso de palavras-chave para representação de um documento; e a ordenação um processo de disponibilização de documentos aos usuários segundo critérios de representação que satisfaçam suas necessidades.

A construção de um SRI[27] que possa extrair, consistentemente, informações semânticas acuradas de todo tipo de texto nem sempre é possível (KONCHADY, 2006), constituindo um dos maiores problemas conceituais da RI a ambigüidade de termos lingüísticos. Meadow, Boyce, Kraft e Barry (2007) observam que:

> (...) *enquanto identificadores únicos são utilizados para algumas aplicações e indicadores de classificação, em outros casos há incerteza sobre valores ou significados, causando confusão ao leitor humano ou a um programa de computador, ou ambos. E uma fonte de ambigüidade é a semântica – o significado dos símbolos.*

Gottschalg-Duque (2005, p. 29) apresenta uma solução metodológica para o problema propondo que se construam proposições (ou enunciados, na matriz aristotélica) para a interpretação semântica. Objetivamente, este autor define em seu *framework* de recuperação da

[26] *Crawler*: programa de computador (robô) que recupera e categoriza informações da *Web*.
[27] Utilizando-se, no caso, "recuperação" como um termo mais genérico, hiperônimo de "extração".

informação que “uma proposição é uma unidade constituída de sentido e que é maior que o significado de uma palavra e menor que uma narrativa ou uma teoria”.

### 3.2.3 Processamento da linguagem natural

O processamento da linguagem natural se ocupa da extração de representações de significados a partir de textos livres, para se descobrir quem faz o que, para quem, quando, onde, como e porque (KAO; POTEET, 2007). Esta definição pragmática é útil para a contextualização do problema de mineração de informações abordado na pesquisa e as metodologias selecionadas para o experimento empírico. A extração de relações sintáticas e semânticas entre entidades (termos ou estruturas) de um texto é considerada uma tarefa importante e desafiadora no campo do PLN, presente na literatura tanto em obras mais antigas como mais recentes (WINOGRAD, 1972; BORKO, 1981; WILKS, 1975; WINOGRAD, 1983; BALLARD e TINKHAM, 1984; BITTLESTONE, 1985; DOSZKOCS, 1986; BARNETT, KNIGHT, MANI e RICH, 1990; HOBBS, APPELT, BEAR e TYSON, 1991; DAELEMANS, GAZDAR e DE SMEDT, 1992; CHURCH e RAU, 1995; KING, 1996; BERRY, 2004; KONCHADY, 2006; KAO e POTEET, 2007; MEADOW, BOYCE, KRAFT e BARRY, 2007; PRADO e FERNEDA, 2008).

As relações lingüísticas relevantes que podem ser encontradas num texto são abundantes, podendo ocorrer de várias formas, de acordo com as estruturas lingüísticas possíveis segundo as regras de uso do idioma do texto (estruturais ou funcionais), constituindo uma área do conhecimento que tem atraído interesse de pesquisadores desde meados da década de 1950. O artigo seminal de Chomsky (1956), por exemplo, resultou de um esforço de pesquisa patrocinado pelas três armas de defesa dos EUA e pela empresa *Eastman Kodak*, interessados nos aspectos computáveis da linguagem natural, revelando a natureza pragmática desse tipo de pesquisa em ambientes organizacionais competitivos.

Embora o *framework* metodológico proposto na tese utilize várias abordagens de processamento da linguagem natural, assim como análise funcional de textos no processo seletivo de conteúdos como dados de entrada no modelo, os formalismos elaborados se fundamentam em experimentos de análise sintática estruturalista, com as estruturas frasais (CHOMSKI, 1956), e na pragmática. Delbecque, por exemplo, define “sintaxe” como “o estudo da combinação dos elementos da frase segundo esquemas de construção regulares” e “pragmática” como “o estudo daquilo que as pessoas fazem quando utilizam a língua” (DELBECQUE, 2008, p. 114).

A análise sintática é útil no contexto porque, dentro de domínios específicos da atividade humana, como o mundo dos negócios e das organizações, ela apresenta uma capacidade promissora de revelação de certas estruturas gramaticais canônicas (SOWA, 1984) úteis para modelagem de sistemas. Como discutido no capítulo seguinte, sobre o referencial teórico da tese, propõe-se um recorte específico de uma teoria da estrutura do idioma inglês, com alguma

similitude com o idioma português, aplicável à linguagem natural típica do texto organizacional, que geralmente é um texto estruturado. Com esse recorte, propõe-se um método de modelagem automática, ou semi-automática, de sistemas organizacionais com base na premissa de equivalência cognitiva entre pensamento e linguagem. As estruturas frasais dos textos organizacionais seriam estruturas do mundo dos negócios e das organizações, com objetos e ações (eventos e relacionamentos).

Historicamente, alguns autores desenvolveram idéias que, de certo modo, são atualmente agregadas a construtos epistemológicos e artefatos computacionais de PLN. Wilks (1975), por exemplo, situa a provável simbiose entre Inteligência Artificial (IA) e uma "compreensão da linguagem natural" numa abordagem pragmática, argumentando que as limitações metodológicas de ambas as áreas do conhecimento não devem desencorajar o desenvolvimento de tecnologias para solução de problemas correlatos. O autor previra que essa compreensão da linguagem natural, um conceito da época que, embora mais amplo, pode ser considerado precursor do conceito de PLN, deveria ocupar um lugar menos periférico na IA, graças ao pioneirismo de Winograd (1972), e relacionara pelo menos quatro benefícios dessa simbiose (WILKS, 1975, p. 130-131):

> *(i) ênfase em estruturas de armazenamento complexas em sistemas de compreensão da linguagem natural: frames, se o termo agradar (MINSKY, 1974);*
>
> *(ii) ênfase na importância do mundo real, conhecimento indutivo, expresso nas estruturas mencionadas em (i);*
>
> *(iii) ênfase na função comunicativa de sentenças em um contexto, como encontrar a leitura correta-num-contexto para uma sentença, em oposição à visão lingüística padrão, para a qual a tarefa é encontrar um leque de possíveis leituras independentes do contexto;*
>
> *(iv) ênfase na expressão de regras, estruturas e informação dentro de um ambiente operacional/procedimental/computacional.*

A crítica de Wilks à obsessão da linguística tradicional por uma Ciência da Linguagem, ou uma teoria da linguagem natural, se fundamenta no argumento segundo o qual não se pode divisar os limites metafísicos da linguagem natural. O alento que o autor oferece para superação desse impasse epistemológico é eminentemente pragmático, numa abordagem de linguística próxima à do critério de Popper para delimitação de uma área ou de um ramo da ciência: concentrar a atenção não em seu objeto de estudo, mas nos problemas que essa área demarcada da ciência poderá resolver. O autor sugere considerar-se a compreensão da linguagem natural como uma engenharia e não como uma atividade científica de base.

Com a IA, Wilks (1975, p. 131) vislumbrara um casamento multidisciplinar promissor para a área de PLN:

> *Dado um contexto e explicação suficientes, qualquer coisa pode ser acomodada e compreendida: é essa competência básica da linguagem humana que os linguistas gerativos têm sistematicamente ignorado e que uma visão de IA da linguagem poderia se habilitar a tratar.*

O traço marcante dessa visão de Wilks é a crença numa compreensão epistemológica e tecnológica da linguagem natural sem as limitações inerentes a teorias científicas apriorísticas sonhadas pela corrente racionalista da linguística, adotando uma abordagem experimentalista e empirista. Concentra seu ataque nos formalismos, ou no que ele chama "falácia da regra 100%"[28], que não admite exceções às regras e formalismos dos modelos desenvolvidos para o tratamento científico da linguagem natural (WILKS, 1975, p. 132). Ou seja, ele defende uma abordagem multidisciplinar de PLN com IA útil em determinados contextos, onde não se exige modelos explicativos amplamente generalizáveis do ponto de vista de uma Ciência da Linguagem, mas útil para a solução de determinados problemas do mundo real.

Borko (1981), pesquisador seminal da Ciência da Informação pré-*Internet*, defendera idéias pragmáticas como as de Wilks em construtos parecidos na recuperação da informação. As inovações que ele propunha na época, para o desenvolvimento de sistemas de recuperação da informação tecnologicamente mais aptos para enfrentamento dos desafios do mundo da comunicação em redes de computadores digitais que se avizinhava, podem ser hoje observadas em *softwares* de análise sintática *(parsing)* e de mineração de textos, tais como o uso do recurso de agrupamento (clusterização) por atributo para indexação e recuperação automática de documentos, a indexação sintagmática (com uso de trechos de frases e não apenas termos isolados), o abandono da pré-indexação de documentos numa base textual (como medida de economia), e a pesquisa em texto completo de modo completamente automático.

Em termos de metodologia para a indexação automática, Borko (1981) observa que "(...) o computador não lida com conceitos, mas somente com palavras. A tarefa é projetar um algoritmo para selecionar palavras significativas do texto e, como consequência, essas representarão conceitos significativos." Obviamente, Borko propõe essas inovações no escopo de um contexto tecnológico mais avançado de indexação e recuperação automática da informação, como o da *Internet* e dos aplicativos de busca da *World Wide Web*, que se tornaria realidade pouco depois. O termo automático, para ele, se aplica a um processo implementado por um conjunto de programas de computador ao invés de executado com o esforço intelectual de pessoal qualificado. Esse autor, tal como Wilks, também criticara conceitos sedimentados de avaliação do desempenho de construtos epistemológicos e respectivos artefatos tecnológicos da era pré-*Internet*, e o faz mesmo em relação aos sistemas de recuperação da informação mais avançados da época, baseados em redes de computadores com acesso *on line*. Ele questionara, por exemplo, se os parâmetros matemáticos de revocação e precisão seriam, ainda, os melhores critérios para avaliação da efetividade de sistemas de recuperação da informação em ambiente computacional com acessos em rede.

[28] Essa regra estabelece que na ciência é necessário se ter uma teoria completa antes de se ter alguma teoria útil.

O pragmatismo[29] em torno do processamento da linguagem natural cada vez mais presente continuou alavancado, na década de 1980, principalmente pelos avanços na indústria de processadores eletrônicos para computadores. Ballard e Tinkham (1984), cientistas com um pé na academia e outro na indústria eletrônica, trilham o "caminho do meio" entre a corrente científica racionalista e a corrente experimentalista, advogando que "(...) o projeto de formalismos para serem usados em processadores da linguagem natural transportáveis se refere aos aspectos científicos, tanto quanto aos de engenharia, da linguística computacional." Os autores desenvolvem e implementam experimentalmente um formalismo gramatical baseado no que denominam "regras de estrutura frasal expandida", que permite a um analisador sintático "tomar decisões em domínio específico reportando-se a um dicionário e outros arquivos auxiliares produzidos durante uma sessão de aprendizado inicial com o usuário" (BALLARD e TINKHAM, 1984, p. 82).

Outros aspectos interessantes do trabalho de Ballard e Tinkham (1984) se referem à reminiscência das estruturas sistêmicas de Winograd (1972) e à fonte (parcial) de financiamento de sua pesquisa: o Escritório de Pesquisas Científicas da Força Aérea dos Estados Unidos. Essa constatação reforça a natureza estratégica das pesquisas com IA e PLN num contexto específico de inteligência competitiva – o poderio militar.

O ensaio de Bittelstone (1985) representa uma crítica ao senso comum dos sistemas especialistas da época, com uma abordagem que poderia ser denominada "lógico-linguística". O autor critica, especialmente, o uso do termo Inteligência Artificial (IA) como atributo de tais sistemas e sustenta que IA não será implementada em computadores digitais. Contudo, sua contribuição é realçada na medida em que mostra a utilidade de *softwares* de processamento de sentenças de lógica formal, como o PROLOG, na análise sintática e inferencial de textos em linguagem natural.

Doszkocs (1986), numa linha de publicação mais tradicional da Ciência da Informação, avança numa abordagem explicitamente multidisciplinar da recuperação da informação com processamento da linguagem natural, que é assim definida (DOSZKOCS, 1986, p. 191):

> *(...) processamento da linguagem natural (PLN) abrange todas as abordagens baseadas em computador para o manuseio da linguagem irrestritamente escrita ou falada, a partir de procedimentos puramente "mecânicos", como os empregados em muitos editores de textos, processadores de palavras e abordagens de indexação automática na recuperação da informação (RI), até análise "inteligente", compreensão e expressão de "significado", como exemplificado na compreensão da linguagem natural, resposta a perguntas e sistemas especialistas na inteligência artificial (IA)*.

[29] O conceito de pragmatismo adotado nesta tese é o da doutrina filosófica de Peirce, que assim argumenta em relação à escolha desse termo para sua filosofia: *(...) uma concepção, isto é, o teor racional de uma palavra ou outra expressão reside, exclusivamente, em sua concebível influência sobre a conduta da vida; de modo que, como obviamente nada que não pudesse resultar de um experimento pode exercer influência direta sobre a conduta, se se puder definir acuradamente todos os fenômenos experimentais concebíveis que a afirmação ou negação de um conceito poderia implicar, ter-se-á uma definição completa do conceito, e nele não há absolutamente nada mais. Para esta doutrina o presente autor inventou o nome de pragmatismo*. (PEIRCE, 2010, p. 284)

O autor observa que as vantagens do casamento entre RI e IA para o desenvolvimento de sistemas de recuperação da informação eram ignoradas na época por razões históricas e pragmáticas. E que o interesse das pesquisas em RI estava se movendo para a formulação de uma visão unificada dos modelos de recuperação probabilísticos-combinatórios e lógicos num modelo baseado na teoria dos conjuntos nebulosos. O texto de Doszkocs (1986) é didático, esclarecendo a origem siamesa da IA e da PLN nos primórdios da computação eletrônica nos anos 1960. Como os demais autores da corrente linguística pragmática, ressalta que os tecnólogos praticantes de IA descobriram que a captura do conhecimento humano especializado em objetos materiais é absolutamente essencial para o sofisticado processamento da linguagem natural; e que tal processamento parece ser exequível em áreas bem restritas, de objeto bem definido, a fim de se alcançar eficiência e utilidade.

Barnett, Knight, Mani e Rich (1990) apresentam um artefato computacional, que denominam sistema de Conhecimento e Processamento da Linguagem Natural (CPLN)[30], para aplicações em interfaces inteligentes, recuperação de textos e tradução de máquina. Os autores ressaltam questões importantes, como a naturalidade (e facilidade, portanto), para os humanos, do uso da linguagem natural para a comunicação com máquinas, uma tautologia (naturalmente) ignorada pelos projetistas de linguagens de programação de computadores. O embasamento filosófico dessa obra se reporta, perceptivamente, a conceitos kantianos de objeto e referência, argumentando os autores que:

> *(...) existem relações naturais entre construtos linguísticos e estruturas do mundo as quais utilizamos a linguagem para descrever. (...) no mundo existem objetos e ações; correspondentemente, na linguagem existem substantivos e verbos* (BARNETT; KNIGHT; MANI; RICH, 1990, p. 52).

Apesar do objetivo ambicioso com seu artefato – o de elaborar "uma série de regras genéricas ligadas a uma abrangente ontologia do conhecimento do mundo", esses autores observam, também na corrente pragmática do estudo linguístico, que apesar do mapeamento entre a linguagem como um todo e o mundo objeto do discurso ser algo natural e efetivo, não há, *a priori*, razão para se acreditar que os mapeamentos entre um domínio restrito e o recorte linguístico de uma linguagem são naturais no todo, citando exemplos de como isso pode ocorrer. Esses experimentalistas assumem que as principais técnicas de análise sintática e pragmática podem ser aprimoradas com métodos que aumentam sua robustez e apresentam as técnicas que os suportam, denominadas: (i) análise sintática baseada em agenda, (ii) recuperação de falhas em análise sintática, e (iii) análise sintática de subsequências de caracteres de termos. A contribuição dos autores é importante como referência para pesquisas ulteriores, tais como a gramática do sistema DIALOGIC que é utilizada no projeto TACITUS, composta de aproximadamente 160 regras de estruturas frasais do idioma inglês com um "construtor" (que expressa as restrições de uso da regra) e um "tradutor" para produção do formalismo lógico.

[30] KBNL: *Knowledge-Based Natural Language Processing*, no idioma original.

Church e Rau (1995) antecipam, num cenário vislumbrado em meados dos anos 1990, algumas aplicações comerciais de sistemas de processamento de linguagem natural, como o processamento de texto, a publicação de conteúdos digitais e a gestão da informação. As aplicações de PLN na gestão da informação mais atrativas para os autores seriam recuperação da informação, categorização de documentos, extração de dados estruturados de textos em linguagem natural e, no caminho inverso, geração de textos em linguagem natural a partir de dados estruturados.

Contemporaneamente, King (1996) apresenta uma revisão das metodologias de avaliação de sistemas de processamento de linguagem natural influenciada, segundo a própria autora, pelo trabalho dos *Expert Advisory Groups for Language Engineering Standards* (EAGLES) iniciado em 1992. O *framework* de avaliação dos grupos EAGLES originou o padrão ISO 9126 de qualidade de *software*. Contudo, as metodologias de avaliação, à época, eram muito variadas, assim como o escopo dos sistemas avaliados, dificultando o desenvolvimento de um padrão genérico de avaliação de *softwares*. Os exemplos citados por King são de sistemas de funções mais comuns nos primórdios da PLN, tais como os de tradução automática (ou tradução de máquina).

O texto de Leidner (2003) complementa, de certo modo, o de King (1996) oferecendo uma visão panorâmica da engenharia de *software* utilizada no desenvolvimento de aplicativos para PLN. O autor explora os conceitos de “componentização” e “reuso” e seus benefícios, assim como as dicotomias utilitaristas entre experimentos e sistemas, e ferramentas e estruturas.

Dale, Aliod e Schwitter (2002) discutem um *curriculum* acadêmico adequado para o estudo de PLN a partir de sua experiência numa universidade. Os autores também consideram PLN como um subcampo de estudos da Inteligência Artificial (IA) e tentam distinguir, com base em outros estudos, os conceitos de processamento da linguagem natural e linguística computacional, definindo o primeiro como a área de estudos com foco nas técnicas que habilitam máquinas a trabalhar com a linguagem humana, e o segundo como um campo de estudos de mesmo objeto, mas com interesse em aspectos mais teóricos.

Lease (2007), da *Brown University* (pioneira na produção de *corpus* de língua inglesa), renova o desafio de se desenvolver melhores sistemas de recuperação da informação com PLN. O autor observa que, na prática, a compreensão da linguagem humana tem permanecido fugaz e métodos superficiais de abordagem, como os de estilo “saco-de-palavras”, continuam a dominar a área de recuperação da informação. E arremata argumentando que o impacto da subdisciplina de PLN nas tarefas de recuperação da informação tem largamente sido mais de promessa que de substância (LEASE, 2007, p. 1).

### 3.2.4 Mineração de textos

Os métodos e processos de descoberta de informação (que alguns chamam, em certos contextos, de conhecimento) relevante em linguagem natural, conhecidos como mineração de textos, constituem variantes do que se conhece como mineração de dados. Chauke-Nehme (2008), como justificativa para os investimentos em pesquisa e desenvolvimento e o enorme interesse do mercado em tecnologias dessa natureza, argumenta que um importante desafio para os próximos dez anos é a preparação de um "exército de Davi" [31]para tirar vantagem do volume de informações disponíveis exponencialmente crescente no mundo em favor de uma sociedade mais sábia, justa e igualitária em termos de oportunidades para todos. Este autor observa que a mineração de textos, que constitui um dos tipos de tecnologias úteis para o tratamento adequado da "inundação" de informações do mundo contemporâneo, tem apresentado uma significativa evolução na última década, desde o simples processamento de palavras na segunda metade da década de 1990 até hoje, quando o processamento de conceitos (termos ontológicos), ou mesmo a extração de conhecimento de estruturas lingüísticas, tem se tornado possível.

Outros autores, como Prado e Ferneda (2008), definem mineração de textos como a aplicação de métodos e técnicas computacionais sobre dados textuais com a finalidade de encontrar informações intrinsecamente relevantes e conhecimento. Em relação às origens da mineração de textos na modernidade, Penteado e Boutin (2008) recordam o estudo de textos estruturados para mensuração da publicação científica, que surgiu e se desenvolveu a partir dos esforços de pioneiros como Solla Price, Small, van Raan, Swanson, Dou e Porter. Os autores mencionam que a mineração de textos estruturados é encontrada em campos do conhecimento tais como bibliometria, cientometria, informetria, midiametria, museometria e webmetria, esclarecendo que nesses campos se estuda os diferentes aspectos da informação, inclusive sua qualidade, sendo a principal matéria-prima para esses estudos as palavras nos textos.

Araújo Jr. (2007), expondo a recuperação da informação em contextos de mineração de textos, ressalta que "a partir do momento em que as bases de dados são formadas, faz-se necessário o desenvolvimento de mecanismos que permitam a mineração e a identificação de conhecimento perdido nos textos". O autor descreve a tipologia de mineração de textos com nada menos que catorze abordagens metodológicas, onde as que interessam, no contexto de engenharia do conhecimento desta tese, são as de descobertas baseadas na extração de passagens, análise linguística, associação entre passagens, estruturas de textos e agrupamento ou generalização (ARAÚJO JR., 2007, p. 55-61).

Quanto às funções que podem ser desempenhadas por sistemas de mineração de textos, Konchady (2006) observa que não há um conjunto padrão definido das funções desse tipo de

[31] Como metáfora relacionada ao confronto bíblico entre Davi (representando os usuários de informações) e o gigante Golias (representando o enorme volume de informações disponíveis).

tecnologia. O autor, no entanto, apresenta uma lista de soluções com uso (inclusive) de mineração de dados para problemas comuns de gestão da informação: busca *(search)*, extração de informação (busca de padrões de uso semântico), formação de *clusters*, categorização, construção de resumos (sumarização), monitoramento de informação, organização de perguntas e respostas (em aplicações de *Frequent Asked Questions*, por exemplo).

### 3.2.5 Análise de conceito formal

As noções de "contexto formal", "conceito formal" e "análise de conceito formal" são apresentadas nesta revisão de literatura porque representam formalismos matemáticos que embasam a tese, do ponto de vista filosófico e científico, e alguns dos principais pilares metodológicos da pesquisa experimental executada. Contextos formais e conceitos formais geralmente são desenvolvidos, contemporaneamente, em temas correlatos ao aprendizado e população de ontologias a partir de *corpora* idiomáticos de domínio específico (CIMIANO, 2006).

Conforme Priss (2005), a Análise de Conceito Formal (ACF)[32] é um método da análise de dados, representação do conhecimento e gestão da informação ainda amplamente desconhecido entre os cientistas da informação nos EUA, apesar de sua potencial utilidade em aplicações tecnológicas. O método foi criado por Wille (1982) no início dos anos 1980 e desenvolvido, na década seguinte, por um pequeno grupo de pesquisadores e estudantes na Alemanha, talvez devido à necessidade de certo conhecimento matemático necessário para sua utilização. A primeira aplicação industrial em larga escala se deu no projeto de um sistema de exploração de conhecimento para a Engenharia Civil e, nos últimos 15 anos, ocorreu um crescimento do interesse no método na comunidade de pesquisas internacional, observando-se aplicações em várias disciplinas, tais como Linguística, Engenharia de *Software*, Psicologia, Inteligência Artificial e Recuperação da Informação.

A utilidade da ACF na Linguística, com apelo mais explícito para a formalização de processos de recuperação da informação textual baseada em análise sintática, não se restringe a textos em linguagem natural, mas se aplica também a códigos-fonte de programas de computador. A experiência de Lindig e Snelting (1997), por exemplo, representa o estado-da-arte em metodologias de análise e refatoração de *software* legado para reuso em componentes, abrindo novo alento para o desenvolvimento de arquiteturas de sistemas computacionais orientadas a serviço *(Service-Oriented Architecture,* ou SOA, no jargão da indústria do *software)*. Stumme (2002) esclarece, a respeito dessa tendência, que esse vetor de desenvolvimento tecnológico nasceu de uma união de esforços entre as comunidades de Análise de Conceito Formal e de Gráficos Conceituais (SOWA, 1984; MINEAU; STUMME; WILLE, 1999).

[32] FCA: *Formal Concept Analysis*.

As estruturas da ACF, que são fundamentais para a representação da informação em contextos como o da presente tese, têm sido descobertas por diferentes pesquisadores. Contemporaneamente ao desenvolvimento da Análise de Conceito Formal, reaparecem as teorias matemáticas do reticulado *(lattice)*, também denominado "Reticulado de Galois", cuja integração num construto de recuperação da informação com ACF é atribuída a Godin, Gecsei e Pichet (1989), com base nos estudos seminais de Barbut e Monjardet (1970). O uso de ACF em metodologias de IA como aprendizado de máquina, descoberta de conhecimento e mineração de dados aparece, mais recentemente, nos artigos de Kuznetsov (2004) e de Valtchev, Missaoui e Godin (2004). Contudo, o texto introdutório mais completo à ACF é o de Wille (2005), que ressalta a ACF como um subcampo da matemática aplicada baseado na matematização dos termos "conceito" e "hierarquia de conceitos".

O método teve sua origem epistêmica em atividades de reestruturação da teoria matemática da ordem e do reticulado. Wille (2005) observa que mais de uma década depois de sua criação, as conexões da ACF com a lógica filosófica do pensamento humano se tornaram claras, principalmente em relação aos últimos desenvolvimentos da corrente pragmática da filosofia de Peirce. O autor esclarece que com ACF a própria matemática, em si, se beneficiou, citando como exemplo a seguinte tese de sua autoria, publicada num artigo recente: "o objetivo e o significado da matemática, no final das contas, se reportam ao fato da matemática ser habilitada para suportar efetivamente a comunicação racional dos humanos" (WILLE, 2005, p. 2).

E, partindo da premissa que os conceitos são também pré-requisitos para a formação de julgamentos e conclusões sobre os fenômenos, Wille (2005, p. 2) adapta a tese anterior para a ACF do seguinte modo:

> *O objetivo e o significado da Análise de Conceito Formal como teoria matemática de conceitos e de hierarquia de conceitos é suportar a comunicação racional entre humanos mediante o desenvolvimento matemático de estruturas conceituais apropriadas que podem ser ativadas logicamente.*

"Conceitos", para Wille (2005, p. 2), podem ser filosoficamente entendidos como "unidades básicas do pensamento formadas em processos dinâmicos dentro de ambientes sociais e culturais". A tradição filosófica dominante define que um conceito é constituído por sua "extensão", abrangendo todos os objetos que pertencem ao conceito, e sua "intensão"[33], incluindo todos os atributos (propriedades e significados) que se aplicam a todos os objetos da extensão. Outra noção epistemológica basilar é que conceitos somente podem se apresentar relacionados a outros conceitos, onde a relação entre "superconceitos" e "subconceitos" exerce um importante papel, especialmente nas metodologias de classificações de seres e objetos no mundo. O subconceito de um superconceito é um conceito cuja extensão está contida na extensão do superconceito, e cuja intensão contém a intensão do superconceito.

[33] Como não há tradução para o termo *intension*, do idioma original, adotou-se o neologismo *intensão* como uma referência útil no contexto.

O filósofo Aristóteles, em sua noção de *diferenciação (differentiae)*, estabeleceu o fundamento lógico do método de Análise de Conceito Formal (ACF) ao observar que existe uma conexão inversa entre quantidades de objetos e quantidades de atributos, pois quanto mais características se requerem de um objeto, menos instâncias dessa classe de objetos serão encontradas no mundo, e vice-versa, ou seja, quanto menos características se requerem de um objeto, mais instâncias desse objeto genérico serão encontradas (pois um número menor de características comuns é exigido). Galois formalizou a conexão inversa de Aristóteles no que se conhece por "Conexão de Galois" (WILLE, 1982; GANTER, STUMME e WILLE, 2005).

Conforme a teoria da Análise de Conceito Formal (ACF)[34], os objetos são representados como objetos formais e suas características como atributos formais (CIMIANO, 2006). Objetos e características extraídos do texto em linguagem natural são dispostos numa matriz de incidência, cruzando objetos com as respectivas características incidentes nas colunas de atributos, gerando uma matriz com dados binários de relações de incidência entre objetos e características, conforme a Tabela 3.1. Entretanto, essa matriz de relações de incidência ainda não apresenta conceitos formais, conforme o método, mas apenas um contexto formal e a razão disso é que existem atributos que não apresentam relação de incidência com objetos.

Como exemplo de aplicação em um contexto, Cimiano (2006, p. 57) observa que nos atributos e relações de incidência da Tabela 3.1 os atributos são construídos a partir de verbos que acompanham, sintaticamente, os objetos nos textos estudados (relações de pares de termos com estrutura verbo-objeto), definindo *bookable* como o atributo comum a todos os objetos (interpretado como o atributo "reservável"),[35] *rentable* (alugável) como atributo comum aos objetos *apartamento*, *carro* e *bicicleta, driveable* (dirigível) comum a *carro* e *bicicleta*, *rideable* encontrável apenas em *bicicleta* e *joinable* como atributo de *excursão* e *viagem*.

O próximo passo do método, em relação ao contexto formal representado na Tabela 3.1, seria a descoberta de conjuntos fechados entre si, que são obtidos com um conjunto de objetos *O* e um conjunto de atributos *A* quando os atributos em *A* são exatamente os mesmos que são comuns a todos os objetos em *O* e, vice-versa, todos os objetos em *O* são exatamente os mesmos que têm todos os atributos em *A*. Os exemplos de objetos que representam conjuntos fechados na Tabela 3.1 são "excursão" e "viagem" {excursão, viagem}, pois ambos têm os mesmos atributos "reserva" e "adesão" {reserva, adesão}, e vice-versa. Considerando-se, então, que os contextos formais são dados estruturados em unidades que representam abstrações formais de conceitos do pensamento humano, com interpretação compreensiva e significativa, Análise de Conceito Formal é uma técnica de clusterização conceitual que também provê descrições genéricas para os conceitos abstratos ou unidades de dados produzidos no método.

[34] *Formal Concept Analysis* (FCA), no original.
[35] Optou-se pelo uso de termos (atributos) interpretados para o idioma português pela dificuldade e desconforto lingüístico na tradução de termos como *driveable*, *rideable* e *joinable* no original.

Tabela 3.1 Domínio “Turismo” como Contexto Formal

| Objeto | | Atributo e Relação de Incidência | | | | |
|---|---|---|---|---|---|---|
| Nº | Denominação | Reserva | Aluguel | Direção | Selim | Adesão |
| 1 | Hotel | X | | | | |
| 2 | Apartamento | X | X | | | |
| 3 | Carro | X | X | X | | |
| 4 | Bicicleta | X | X | X | X | |
| 5 | Excursão | X | | | | X |
| 6 | Viagem | X | | | | X |

Cimiano (2006, p. 57-58) define, formalmente, “contexto formal” e “conceito formal” do seguinte modo, com uso de lógica e teoria dos conjuntos:

**Definição 1 (Contexto Formal):**[36] uma tupla *(G, M, I)* é chamada um *contexto formal* se *G* e *M* são conjuntos e $I \subseteq G \times M$ é uma relação binária entre *G* e *M*. Os elementos de *G* são chamados *objetos*, os de *M atributos* e *I* é a *relação de incidência* do contexto.

Então, para $O \subseteq G$, define-se:

$O' := \{m \in M \mid \forall g \in O: (g,m) \in I\}$

E, dualmente, para $A \subseteq M$:

$A' := \{g \in G \mid \forall m \in A: (g,m) \in I\}$

$O'$ é o conjunto de todos os atributos comuns aos objetos em *O* e $A'$ é o conjunto de todos os objetos que tem todos os atributos em *A*.

**Definição 2 (Conceito Formal):** um par *(O, A)* é um conceito formal de *(G, M, I)* se e somente se $O \subseteq G$, $A \subseteq M$, $O' = A$ e $O = A'$. Então, *O* é denominado a *extensão*[37] e *A* a *intensão*[38] do conceito

[36] Os símbolos e operações lógicas empregados têm o seguinte significado: $A \subseteq B$, A é um subconjunto de B; $A \leq B$, A é um subgrupo de B; $A \in B$, A pertence a B; $\forall$ B, para todo B; $A \Leftrightarrow B$, equivalência material entre A e B.

[37] Como extensão *(extension)*, entende-se o conceito de “instanciação”, ou representação de um conceito com objetos do mundo. O método adota esse termo com significado oposto ao de intensão *(intension)*, com objetivo de relacionar e exemplificar os subconceitos de um conceito numa árvore de conceitos.

[38] *Intension*, no original. Este é um termo utilizado em lingüística, lógica, filosofia e outras áreas do conhecimento com um significado não encontrado no idioma português e facilmente confundido, foneticamente, no inglês com *intention* (substantivo derivado do verbo *to intend*, no sentido de pretender). O significado de “intensão”, neste caso, é o de “definição” num nível mais geral, ou seja, é alguma propriedade ou conjunto de propriedades de um termo que o torna representante de um conjunto de coisas, um conceito mais geral (como sinônimo de “compreensão”, no sentido de “abrangência”).

formal *(O, A)*. Com isso, pode-se também definir uma ordem entre os conceitos formais, como segue: $(O_1, A_1) \leq (O_2, A_2) \Leftrightarrow O_1 \subseteq O_2 (\Leftrightarrow A_2 \subseteq A_1)$

Conceitos formais são parcialmente ordenados em relação à inclusão de suas *extensões* ou, de modo equivalente, à inclusão inversa de suas *intensões*. Cimiano (2006) argumenta que a relação subconceito-superconceito forma uma rede completa *(β(G, M, I), ≤)*. No exemplo da Tabela 3.1, ({excursão, viagem}, {reserva, adesão}) representa um conceito formal, mas outros conceitos são, por exemplo, ({bicicleta}, {reserva, aluguel, direção, selim}), ({carro, bicicleta}, {reserva, aluguel, direção}). E o conceito formal ({bicicleta}, {reserva, aluguel, direção, selim}) é um subconceito de ({carro, bicicleta}, {reserva, aluguel, direção}).

As unidades de busca nos *corpus* de textos, que são insumos para o modelo de aprendizado ontológico, são os termos, que podem ser palavras simples ou compostos de várias palavras com um significado conjunto específico, possivelmente com um significado técnico num determinado contexto ou domínio. Em tarefas de descoberta de termos para o aprendizado de ontologias consideram-se, geralmente, como termos quaisquer palavras ou conjuntos de palavras relevantes no domínio estudado. Os insumos (entradas) nesse tipo de processamento de dados são coleções de documentos em linguagem natural ou computacional representando o domínio de interesse e os produtos (saídas) são conjuntos de símbolos lingüísticos *(strings)* $S_C$ e $S_R$ representando termos que serão utilizados como sinais de representação de conceitos e suas interrelações, respectivamente. E, tal como na modelagem com UML, é importante estabelecer-se relações de sinonímia nesse processo, que também pode ser entendida como de co-hiponímia (palavras “descendentes” num tesauro, ou classes de conceitos hierarquizadas sob uma mesma superclasse), tendo como resultado uma simplificação do modelo – Cimiano (2006, p. 24) argumenta que os sinônimos, neste caso, correspondem aos *synsets* da *WordNet*.

Como observado anteriormente, os conceitos são definidos, formalmente, como uma tupla <i(c), [c], $Ref_C$(c)>, onde i(c) é a intensão do conceito, [c] a extensão desse conceito e $Ref_C$ descreve a realização léxica desse conceito num *corpus*. Deste modo, a formação de conceitos deve prover (i) uma definição intensional de conceitos, (ii) sua extensão e (iii) os sinais léxicos que são usados como referência aos mesmos (BUITELAAR *et al*., 2004, *apud* CIMIANO, 2006).

Quanto às relações, o método considera apenas incidências binárias entre objetos e atributos, entendendo-se como aprendizado de relações a tarefa de descoberta de identificadores ou rótulos *r* de relações tanto como de seu domínio *dom(r)* e amplitude *amp(r)*, distinguindo-se quatro sub-tarefas (CIMIANO, 2006, p. 25):

- encontrar conceitos (conjunto *C*) nos documentos com alguma relação ontológica não-taxonômica;
- especificar as relações (conjunto *R*) encontrando rótulos apropriados e identificadores de relações com base nos *corpora* utilizados ou outros *corpora* de referência úteis;

- dada uma determinada relação $r \in R$, determinar o nível apropriado de abstração com base na hierarquia de conceitos para o domínio e a amplitude da relação;
- aprender uma ordem hierárquica $\leq_R$ entre as relações em $R$.

Outro exemplo mais simples de aplicação da ACF é o de Priss (2005), apresentado na Tabela 3.2, com o reticulado de conceitos da Figura 3.1, ilustrando as noções de conceitos hierárquicos e de contextos e conceitos formais. A tabela mostra as relações de incidência entre os objetos, que são os nomes de alguns animais famosos, e alguns de seus atributos.

Tabela 3.2 Contexto Formal de Animais Famosos

| Objetos (animais) | Atributos | | | | | |
|---|---|---|---|---|---|---|
| | desenho | real | tartaruga | cão | gato | mamífero |
| *Garfield* | X | | | | X | X |
| *Snoopy* | X | | | X | | X |
| *Socks* | | X | | | X | X |
| *Greyfriar's Bobby* | | X | | X | | X |
| *Harriet* | | X | X | | | |

A Conexão de Galois é uma noção central para a ACF, implicando que objetos com poucos atributos serão sempre mais abundantes no mundo que objetos com muitos atributos. Essa noção é observada comumente quando se trata de documentos e termos contidos nos mesmos, algo familiar na indexação e recuperação da informação: quando se deseja documentos com muitos conteúdos específicos (termos, palavras-chave ou sintagmas), o número de documentos recuperado (a revocação) é menor que quando se deseja documentos com menos conteúdos específicos (documentos mais genéricos).

Observa-se, na Tabela 3.2, objetos na coluna à esquerda e atributos nas demais colunas, com as respectivas marcações (em *X*) de incidência de atributos em objetos. Os animais famosos têm atributos coincidentes ou não entre si: *Garfield* e *Snoopy* apresentam o atributo "desenho" *(cartoon)*, ou seja, não são animais reais, mas criados por cartunistas, e são ambos mamíferos; contudo, enquanto *Garfield* é um gato, *Snoopy* é um cão. Os animais com nomes *Socks*, *Greyfriar's Bobby* e *Harriet* são reais: o primeiro era o gato da família Clinton na Casabranca; o segundo, um cão lendário de Edinburgo (Escócia); e o terceiro é considerado um dos animais que Darwin trouxe das ilhas Galápagos para casa e que viveu até 2006.

Os objetos da Tabela 3.2, em ACF, são denominados "objetos formais", e os atributos são denominados "atributos formais", sendo que os objetos formais, os atributos formais e as respectivas "relações de incidência" entre eles constituem um construto denominado "contexto formal", base para a análise de conceitos formais. Observa-se, na tabela, outra propriedade interessante da Conexão de Galois: analisando-se um grupo de animais como *Socks*, *Greyfriar's*

*Bobby* e *Harriet*, por exemplo, descobre-se um atributo comum aos três, isto é, o fato de serem animais reais. Essa propriedade comum aos três mostra que esse subconjunto de objetos formais constitui um conceito formal, pois existe uma relação fechada apenas entre eles no contexto formal – o atributo de serem animais reais. E, como consequência, nesse conjunto não se pode nem aumentar a quantidade de objetos nem a quantidade de atributos.

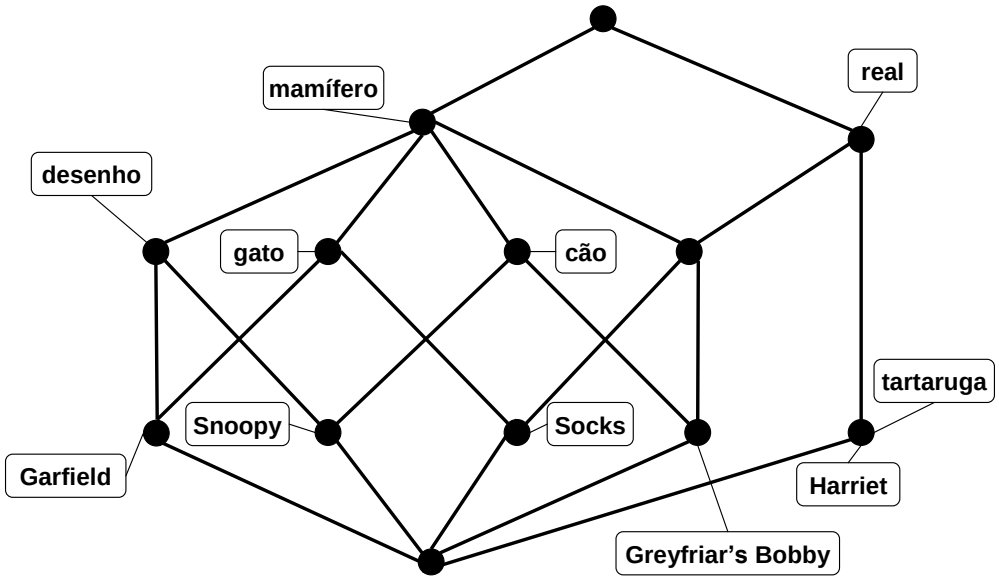



Figura 3.1 Reticulado de Conceitos de Animais Famosos
(Fonte: PRISS, 2005)

Outros conceitos formais podem ser destacados da Tabela 3.2, como os objetos *Socks* e *Bobby* com os atributos "real" e "mamífero". A Figura 3.1 (PRISS, 2005), que representa um reticulado de conceitos, mostra como se pode organizar e interpretar todos os conceitos formais possíveis com os dados da Tabela 3.2, decorrente de um formalismo matemático que é inerente ao método de Análise de Conceito Formal e que o torna generalizável.

A leitura da Figura 3.1 deve ser executada da seguinte maneira: o objeto *Garfield* tem os atributos de ser um gato e ser um desenho, atributos estes que são comuns também aos animais com o atributo de mamífero (por isso, os nós dos atributos "desenho" e "gato", no reticulado, são ligados ao nó do atributo "mamífero*"* logo acima); o objeto *Snoopy* tem os atributos "cão" e "desenho", sendo que ambos atributos são comuns também aos animais com o atributo "mamífero"; *Socks* tem o atributo "gato", que implica ser um 'mamífero", mas tem também o atributo de ser um animal do mundo real; situação similar ocorre com o objeto *Greyfriar's Bobby*, que tem o atributo "cão" e, por consequência, "mamífero", mas que também é "real". Com *Harriet*, ocorre que ela tem apenas dois atributos (ver Tabela 3.2): ser uma "tartaruga" e ser "real".

É importante ressaltar-se que os nós da Figura 3.1 representam, por si mesmos, conceitos formais, e que para um contexto formal, os conceitos formais, suas extensões e intensões são

definidas de modo único e fixo (PRISS, 2005). A Figura 3.2 mostra como funciona o modelo da relação entre um subconceito e um superconceito na Análise de Conceito Formal.

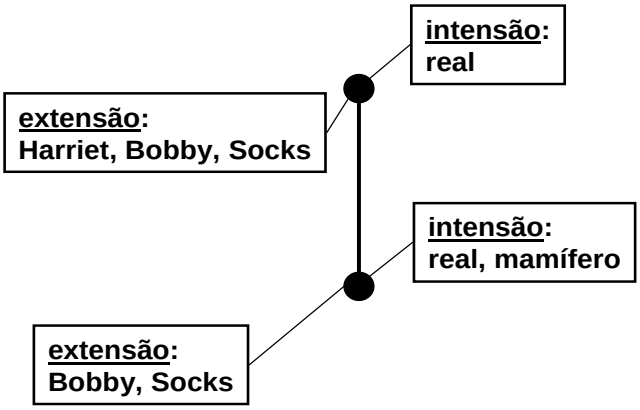



Figura 3.2 Relação Subconceito-Superconceito
(Fonte: PRISS, 2005)

Em termos de compreensão dos agrupamentos de objetos no reticulado de conceitos, aqueles que se situam nos nós mais acima representam um conjunto maior que os que se situam nos nós mais abaixo, constituindo a noção de "extensão". Ou seja, pela dualidade da Conexão de Galois conclui-se que objetos com menos atributos (menor número de itens intensionais), são mais abundantes no mundo, formando conjuntos com mais elementos, como nos casos da extensão *Harriet, Bobby, Socks*, com o atributo *real* (intensão), e a extensão do nó mais abaixo com somente *Bobb*y, *Socks*, mas com os atributos (intensão) "real", "mamífero". E do ponto de vista dos atributos, ou da intensionalidade, objetos (nós) mais acima no reticulado têm um número menor de atributos, ou seja, são mais genéricos que os que estão em nós mais abaixo no reticulado. Os objetos situados mais abaixo têm mais atributos e são, portanto, mais específicos que os acima (têm intensionalidade mais detalhada).

Priss (2005) apresenta alguns exemplos da variedade de aplicações possíveis da ACF mencionando: os reticulados cruzando documentos e termos de Salton (1970) – um dos construtos basilares da recuperação da informação, formalismos de representação do conhecimento da Inteligência Artificial, recuperação da informação de componentes de *softwares* legados em linguagens de programação de computador e sistemas de descoberta de conhecimento. Como estado-da-arte, a autora menciona os esforços empreendidos desde 1996 nas comunidades de Análise de Conceito Formal e de Gráficos Conceituais para a construção de formalismos de representação do conhecimento:

> *Uma combinação de ACF e Gráficos Conceituais não é apenas outro formalismo ontológico porque ela tem fundamento filosófico. Wille (2000) percebe a lógica (e o raciocínio e argumentação humana) como uma tríade Kantiana de conceitos, juízos e conclusões. A meta é utilizar ACF para matematização dessas três doutrinas filosóficas num framework de Lógica Contextual (PREDIGER, 1998). Enquanto ACF é utilizada para matematização de conceitos, Gráficos Conceituais são utilizados para matematização de juízos. Uma combinação de Gráficos Conceituais, ACF e Lógica*

> *Descritiva poderá, então, ser utilizada para matematizar conclusões. Wille encara isso como uma continuação da lógica de signos e classes de Boole (WILLE, 2000 e 2004). Mas, em contraste a Boole, que vislumbrara um conjunto ou classe Universal, a ACF é focada em contextos formais, que são finitos na maioria das aplicações (GANTER e WILLE, 1999) e que evitam algumas das confusões causadas pela assunção da universalidade* (PRISS, 2005, p. 16).

## 4.3 Modelagem de informações organizacionais

O uso de modelos para explicar os fenômenos do mundo com a percepção científica representa algo inerente ao próprio método científico, com as vantagens de representação simplificada de uma realidade complexa e o estabelecimento de conexões (ou relações) entre os conceitos (ou ideias) modelados. Algumas das abordagens de modelagem apresentadas a seguir são as mais comuns no desenho de sistemas de informação baseados em computador e outras são adequadas aos construtos de representação da informação conceitual nesta tese.

É comum a busca de definições para conceitos que, por sua vez, servirão de base para o desenvolvimento de novos conceitos, construindo-se tramas de conceitos que delimitam a Ciência. Contudo, alguns temas apresentam dificuldade, como *Gestão do Conhecimento*. O que é o *conhecimento,* afinal?

*Conhecimento*, tal como *dado* e *informação*, é um conceito que apresenta muita dificuldade para definições precisas e Hessen (1976, p. 87), discorrendo filosoficamente sobre a "essência do conhecimento", define *conhecimento* como *uma relação entre um sujeito e um objeto* (no caso, o sujeito é quem tenta ter acesso ao conhecimento sobre um objeto). Esse autor alerta que o problema do conhecimento consiste na complexidade dessa relação entre o sujeito e o objeto, quando se busca a concordância (verdadeira) do conteúdo do pensamento com o objeto.

O dilema da filosofia, quanto ao problema do conhecimento, se reporta às duas possibilidades de caminhos epistemológicos fundamentais para se definir o conhecimento: ou se admite que todos os objetos possuem uma imagem correlata ideal, mental, *a priori* de sua percepção pelos sentidos humanos – tese do *idealismo*, ou se parte da premissa que além dos objetos ideais há, também, objetos reais, independentes do pensamento humano – tese do *realismo*. Hessen (1976), ao final, conclui que a tese do *realismo* é filosoficamente mais defensável, do ponto de vista da fenomenologia (uma corrente mais moderna da Filosofia), mas não descarta a utilidade operacional do *idealismo* em áreas de conhecimento mais abstrato como a Lógica e a Matemática.

Outro aspecto importante na tentativa de conceituação de conhecimento se refere ao conceito de verdade, que se conecta, necessariamente, ao de conhecimento – o *conhecimento*, portanto, deverá ser o *conhecimento verdadeiro*. O conhecimento sem conexão com a verdade não seria conhecimento e, portanto, não teria a necessária credibilidade para generalização nas ciências em geral. Contudo, isso leva a outra questão: o que é (ou qual é) a verdade?

Os critérios de verdade que sustentam a Ciência também são problemáticos e provisórios, como as noções de *substancialidade* (sobre as propriedades das coisas) e *causalidade* (sobre a relação de causa e efeito), o que implica a necessidade de adoção de conceitos (e critérios) de verdade suficientes em contextos.

Assim, como os problemas centrais na definição rigorosa de conhecimento são filosoficamente insolúveis, parte-se então para conceituações pragmáticas[39]. E, assim, Hessen (1976, p. 50-51) apresenta a definição de verdade a seguir (grifos nossos):

> *Como o cepticismo, também o pragmatismo abandona o conceito da verdade no sentido da concordância entre o pensamento e o ser. Porém, o pragmatismo não se detém nesta negação, mas substitui o conceito abandonado por um novo conceito de verdade. Segundo ele, verdadeiro significa útil, valioso, fomentador da vida.*
> *(...) O intelecto é dado ao homem não para investigar e conhecer a verdade, mas sim para poder orientar-se na realidade.*

O conhecimento verdadeiro, portanto, estaria na concordância do conteúdo do pensamento com o objeto numa visão utilitarista, que agregue valor à existência humana. Esta é uma definição suficiente que será desenvolvida mais adiante, no item 3.5 (de um ponto de vista cognitivo), com exemplificações empíricas (experimentais) no Capítulo 6.

### 3.3.1 Modelos e conhecimento

Schreiber, Wielinga e Breuker (1993) são autores historicamente vinculados ao desenvolvimento de uma conhecida metodologia para modelagem de sistemas baseados no conhecimento (*Knowledge-Based Systems*, acrônimo KBS) denominada *Knowledge Acquisiton and Documentation Structuring*, ou KADS, atualmente atualizada no *framework* de gestão e engenharia do conhecimento denominado *CommonKADS*. Essa metodologia foi desenvolvida na Universidade de Amsterdã, constituindo o padrão europeu para o desenho de KBS, fundamentando-se em dois princípios:

I. A introdução de múltiplos modelos como meio de se lidar com a complexidade de processos de engenharia do conhecimento.
II. O uso de descrições, no assim chamado “nível de conhecimento” dessa metodologia, como um modelo intermediário entre os dados de especialistas e o projeto do sistema.

Contudo, o aspecto mais importante da obra desses autores, no contexto da tese que se defende neste trabalho de pesquisa, é o reconhecimento de que os modernos sistemas de informação transacionais corporativos estão se tornando tão complexos quanto os KBSs pensados no passado, o que sugere, também, uma oportunidade de utilização das mesmas técnicas de modelagem. Schreiber, Wielinga e Breuker (1993) declaram, a propósito:

[39] Hessen (1976) considera que o fundador da doutrina filosófica do *pragmatismo* é William James e não Charles Sanders Peirce. E que outro notório representante do pragmatismo seria Schiller (filósofo inglês), que chamou essa doutrina de *humanismo*.

> *Engenharia do conhecimento (KE) e engenharia de software convencional (CSE) são campos intimamente relacionados. Apesar de ambos campos enfatizarem diferentes aspectos do processo de desenvolvimento de sistemas, não há uma precisa fronteira entre sistemas de software convencional e sistemas baseados no conhecimento. [...] com a crescente complexidade de sistemas convencionais, a fronteira é, na maioria das vezes, muito tênue. E, também, a tendência é utilizar-se aplicações KBS não como aplicações de usuários individuais, mas em combinação com outras aplicações mais convencionais* (SCHREIBER, WIELINGA e BREUKER, 1993, p. 151).

Esse depoimento é importante para se evidenciar, no contexto da tese, o problema da Ciência da Informação apresentado por Saracevic (1999), com base nas pesquisas bibliométricas de White e McCain:

> *(...) há duas áreas principais ou sub-disciplinas. Em uma das seções estão autores que atuaram no 'estudo analítico de literatura', suas estruturas, estudos de textos como objetos de conteúdos nascentes, comunicação em várias populações, particularmente comunicação científica, contexto social da informação, usos da informação, busca e comportamento da informação, várias teorias da informação e tópicos correlatos. (...) denomine-se este grupo de 'cluster dominante', como fizeram White e McCain, apesar que 'básico' poderia ser um bom rótulo. Na outra seção (...) estão autores que se concentraram em teoria de IR e algoritmos de recuperação de informação, processos e sistemas práticos de IR, interação humano-computador, estudos de usuários, sistemas de bibliotecas, OPAC e tópicos correlatos. Chamemos este cluster de cluster de recuperação da informação, apesar que 'aplicado' poderia ser também um bom rótulo. (...) Infelizmente, estes dois clusters principais são muito desconectados. Apenas um número muito pequeno de autores atuou nos dois clusters. Em outras palavras, há muito poucos trabalhos integrados* (SARACEVIC, 1999, p. 1.055).

Ou seja, na prática a multidisciplinar Ciência da Informação é cisalhada, resultando em dois grandes feudos de cientistas, um se ocupando predominantemente dos aspectos epistemológicos de base e outro de aplicações – uma ciência da informação básica e uma aplicada (que poderia levar a uma nova "engenharia da informação"). Conforme Saracevic (1999), Solla Price classifica o primeiro grupo como cientistas da *Small Science* e o segundo como da *Big Science*; ou seja, o primeiro grupo lida com um campo de conhecimento ainda não consolidado, enquanto o segundo se ocupa de aplicações práticas interdisciplinares consolidando conhecimentos sedimentados em outras áreas do conhecimento, tais como a computação eletrônica, a matemática, a estatística, a psicologia, a lingüística, a biologia e a neurosciência.

Em relação aos modelos, seu potencial como método de representação do conhecimento é ressaltado (ou revivido) por Michaud (2006b) na obra organizada por Tarapanoff (2006a). Michaud define que:

> *(...) um modelo é uma representação externa e explícita de parte da realidade vista pela pessoa que deseja usar aquele modelo para entender, mudar, gerenciar e controlar parte daquela realidade – navega nos domínios do conhecimento atual e do conhecimento futuro (...)* (MICHAUD, 2006b, p. 221).

Michaud (2006b) sugere uma abordagem conceitual de modelagem do conhecimento onde o conhecimento deve operar como um elo para se conectar o entendimento do passado com projeções de futuro (muito conhecida na estatística). Essa abordagem, de fato, é similar ao conceito de simulação com base em modelos utilizados tanto nas ciências exatas, como nas engenharias e nas ciências sociais (LUNA-REYES *et al*., 2005) – também prevista no trabalho de

pesquisas em “laboratório” proposto na tese. Os pontos de sinergia cognitiva do presente projeto de pesquisa com as idéias de Michaud (2006b) aparecem no respectivo artigo referenciado:

> *Muito foi escrito sobre dados, informações, conhecimento e sua gestão, modelos e modelagem. (...) É preciso instigar a reflexão sobre a relação estreita e permanente dos conceitos de sistema, modelo e conhecimento, mostrando que um não vive sem os outros* (MICHAUD, 2006b, p. 211).

Campos (2004) também apresenta uma visão sobre modelização de domínios de conhecimento útil no plano conceitual da metodologia proposta, ainda que a autora revele, no texto referenciado, desconhecimento das metodologias de desenvolvimento de sistemas orientada a objetos (OO) e seus fundamentos teóricos. Ela afirma que a classificação de um domínio não cabe numa OO porque seus princípios foram estabelecidos para tratar da modelagem de dados e não de unidades de conhecimento, chocando-se, nesse posicionamento conceitual, com os próprios fundamentos da metodologia OO.

Paula Filho (2005), por exemplo, esclarece que o processo unificado (*Unified Modeling Language* – UML) de modelagem OO, independente dos processos de desenvolvimento, apresenta como características centrais: (i) direcionamento a casos de uso; (ii) a concentração na arquitetura; e (iii) a interatividade e o incrementalismo. Outro autor importante, Pressman (1995), explica que é importante reconhecer que a OO e a modelagem de dados são abordagens diferentes, com um ponto de vista bastante diferente, opiniões similares às de Armstrong (2006) e Giguette (2006). E toda a complexidade cognitiva das metodologias OO para o desenvolvimento de sistemas de informação computacionais (SICs) é abordada, com profundidade filosófica, na obra de Pesonen (2002).

Outra obra interessante para a compreensão das conexões naturais entre modelos e conhecimento é a de Lima-Marques e Macedo (2006). Esses autores definem que:

> *A arquitetura da informação fornece suporte às ações de gestão do conhecimento à medida que visa promover a acessibilidade à informação armazenada para garantir a eficácia do processo decisório nas organizações. (...) para que a gestão seja eficaz, é preciso considerar a interação entre os ambientes informacional, organizacional e externo, compreendendo a dinâmica dos fluxos de informação, e mapear os recursos informacionais para direcioná-los às necessidades de uso. Neste contexto, a função da arquitetura da informação seria a de estruturação do ambiente informacional para viabilizar os processos de gestão*” (LIMA-MARQUES e MACEDO, 2006b, p. 250).

Os modelos organizacionais mencionados, que deverão ser abordados na tese, são: modelo de organização, modelo de processos (com instâncias relativas a tarefas, agentes, insumos e produtos) e modelo de ambiente externo à organização.

O problema com os projetos de sistemas de informação computacionais (SICs) evidenciado nas pesquisas do *Standish Group*, relatados, por exemplo, por Nevo e Wade (2007) e Luna-Reyes *et al*. (2005), e as frustrações com as TIC relatadas por Davenport (2000), McGee e Prusak (1994), revelam apenas a “ponta do *iceberg*” de um cenário adverso para as organizações informatizadas que, para Bates (1999), poderia ser o paradigma “acima da linha d’água*”*. O que está “abaixo da linha d’água” são as práticas de campo da Ciência da Informação, ou seja, as

metodologias e as TIC responsáveis pelo êxito ou fracasso da gestão da informação e do conhecimento nas organizações. E uma definição de modelo organizacional bem "abaixo da linha d'água" é a de Schreiber, Wielinga e Breuker (1993): "Um modelo organizacional provê uma análise do ambiente sócio-organizacional no qual o KBS terá que funcionar. Ele inclui uma descrição das funções, tarefas e gargalos na organização" (SCHREIBER, WIELINGA e BREUKER, 1993, p. 6).

O modelo de serviços, conforme Spohrer e Riecken (2006), ainda não tem uma ciência para si, ao passo que os processos estão hoje no centro das atenções dos desenvolvedores de pacotes de *softwares* corporativos no mercado (suítes de gestão de conteúdos eletrônicos, de fluxos de trabalho, ERP[40], entre outros), pelas suas relações muito próximas com os sistemas de informação baseados em computador. O renovado entusiasmo com a reengenharia também tem contribuído para esse interesse, haja vista a proliferação de iniciativas de desenvolvimento de novos modelos de gestão nas organizações, tanto na iniciativa privada como no setor público.[41]

Maglio (2006) evidencia lacunas epistemológicas no que ele denomina "sistemas de serviços"; Natis e Schulte (2004) esboçam conexões entre serviços e eventos na arquitetura de componentes de *software* de suporte a processos de negócio; e Robredo (2000) aborda o tema "processos" como preliminar ao desenvolvimento de SICs, temas correlatos à conhecida obra de Prahalad e Krishnan (1999) sobre qualidade na sociedade da informação.

A obra de Oliveira (2004) também é interessante, no contexto, pela abordagem inédita de modelos de gestão com foco em processos e na certificação de qualidade. E Bitner e Brown (2006) lançam uma espécie de provocação à comunidade acadêmica a respeito da "Ciência dos Serviços" nas escolas de Administração.

O livro de Havey (2005) é útil para a compreensão da noção de "padrões" no desenho de processos, uma característica de modernas linguagens de notações de processos de negócio que pode facilitar, sensivelmente, o desenvolvimento e popularização de linguagens de representação biunívocas e interoperáveis do conhecimento entre usuários de estamentos diversos nas organizações.

Em relação aos SICs, tanto os processos de desenvolvimento de *software* quanto os processos de produção têm evoluído continuamente, mas ainda aquém da complexidade inerente. O número de novidades publicadas recentemente na ACM, por exemplo, é sintomático, como se pode observar em vários artigos publicados em sua biblioteca digital. Dietz (2006) apresenta uma análise de estruturas essenciais de processos de negócio "abaixo da superfície" nas organizações. Outro artigo interessante, no contexto, é o de Gregg e Walczak (2006) sobre extrações de informações da *Web* para uso em aplicações de inteligência de negócios.

[40] ERP: *Enterprise Resource Planning* (Planejamento de Recursos Empresariais).
[41] No Poder Executivo federal do Brasil, como ilustração, órgãos como Instituto Nacional do Seguro Social, Ministério da Fazenda, Ministério das Minas e Energia e Agência Nacional de Vigilância Sanitária têm projetos dessa natureza com escopo corporativo.

O entendimento da modelagem da complexidade no contexto de projetos de SICs é apresentado, de modo bastante abrangente, no artigo de Silva *et alii* (2005). O texto é propositivo, apresentando um *framework* conceitual de Gestão da Informação e do Conhecimento para o desenvolvimento de sistemas complexos com inspiração nas metodologias *KADS* e *CommonKADS*.

Em relação ao instrumental metodológico para gestão do conhecimento no contexto da pesquisa, também existe uma vasta coleção de obras publicadas na ACM e outros repositórios eletrônicos. Os textos abordam, com freqüência, temas com enfoques ontológicos relacionados com a produção de mapas de informação em domínios específicos ou na WWW (tema popularizado como *Web Semântica*). Ou, ainda, temas correlatos à representação do conhecimento, outro campo de pesquisas tradicional, mas ainda atraente na academia. Os artigos de Auer (2006), Cantele *et al*. (2006), Lee (2005), Gomez-Pérez e Corcho (2002), De Ville (2001) e Brachman e McGuinness (1988) apenas ilustram a proliferação de publicações correlatas.

### 3.3.2 Modelo entidade-relacionamento

O denominado Modelo Entidade-Relacionamento (MER) integra uma metodologia de modelagem de dados para o desenvolvimento de sistemas de *software* que começou a evoluir desde a década de 1970, representando atualmente uma corrente doutrinária da análise de dados ainda bastante robusta mesmo tendo como concorrentes as metodologias de desenvolvimento de sistemas orientadas a objetos (OO). E prova disso é que o que mais se observa, na indústria do *software*, são metodologias de mapeamento entre classes e entidades de dados culminando com abordagens de modelagem lógica de sistemas de bancos de dados denominadas “objeto-relacional” (OR), numa tentativa de aliar as vantagens da modelagem de dados Entidade-Relacionamento (ER) com as vantagens de encapsulamento ontológico e produtividade de código da modelagem OO.

Chen (1976, p. 10), autor seminal da abordagem MER, define uma entidade como “uma coisa que pode ser identificada de modo distinto das demais e um relacionamento como uma associação entre entidades” como, por exemplo, “pai-filho” é um relacionamento entre duas entidades “pessoa”. Outro conceito fundamental no MER é o conjunto de entidades, que esse autor apresenta com um exemplo do seguinte modo (CHEN, 1976, p. 11):

> *Denote-se ‘e’ uma entidade que existe em nossas mentes. Entidades são classificadas em diferentes conjuntos de entidades tais como EMPREGADO, PROJETO e DEPARTAMENTO. Há um predicado associado a cada conjunto de entidades para testar se uma entidade pertence a ele. Por exemplo, se conhecemos uma entidade no conjunto de entidades EMPREGADO, então sabemos que ela tem as propriedades comuns às outras entidades no conjunto de entidades EMPREGADO. Entre essas propriedades se encontra o predicado de teste anteriormente mencionado. Denote-se ‘$E_i$’ conjuntos de entidades. Observe que conjuntos de entidades podem não ser mutuamente disjuntivos. Por exemplo, uma entidade que pertence ao conjunto de entidades MASCULINO-PESSOA também*

> *pertence ao conjunto de entidades PESSOA. Neste caso, MASCULINO-PESSOA é um subconjunto de PESSOA.*

O autor observa que o MER é uma síntese dos modelos de dados em rede, relacional e de entidades, com propósito de superar os pontos fracos e combinar os pontos fortes de cada um desses modelos anteriores. Esse pioneiro argumenta que o MER adota uma visão mais unificada e natural na qual o mundo real consiste de entidades e relacionamentos, incorporando algumas das importantes informações semânticas da realidade. E que, do ponto de vista formal, o modelo é baseado na teoria dos conjuntos e na teoria das relações.

O conjunto de componentes primários para a diagramação Entidade-Relacionamento (ER) é composto de: objetos de dados, atributos, relações e vários indicadores de tipos. O principal propósito dessa modelagem, nos projetos de sistemas de informação, é representar os objetos de dados e as relações entre si. A notação de um diagrama ER é relativamente simples, e numa das primeiras versões, conforme Pressman (1995), os objetos de dados eram representados por retângulos rotulados e os relacionamentos por losangos unindo linhas de relacionamentos entre os retângulos das entidades. Atualmente, as relações entre objetos são representadas de modo mais simples ainda, sem uso do losango, como na Figura 3.3.

O modelo da Figura 3.3, com uso da notação “pé-de-galinha”, se refere a objetos e relacionamentos da estrutura de cargos dos órgãos públicos no Brasil, conforme a respectiva legislação, e o que se pretende resolver são os conhecidos problemas do relacionamento múltiplo entre *órgãos* e *cargos* e entre *pessoas* e *vagas* nos *cargos*. Os objetos mais visíveis no mundo real, nesse caso, são os órgãos públicos, os tipos de cargos existentes e, obviamente, as pessoas que os ocupam. Ocorre, entretanto, que a relação entre um órgão e um tipo de cargo é “N-para-N”, ou seja, um órgão pode ter vários tipos de cargos e um tipo de cargo pode existir em vários órgãos, tornando o modelo inviável caso não se identifique nenhuma instância de relacionamento entre órgãos e cargos. E a solução é justamente uma instância de relacionamento entre ambos denominada “vaga”, algo previsto na legislação (que constitui o conjunto de regras do negócio).

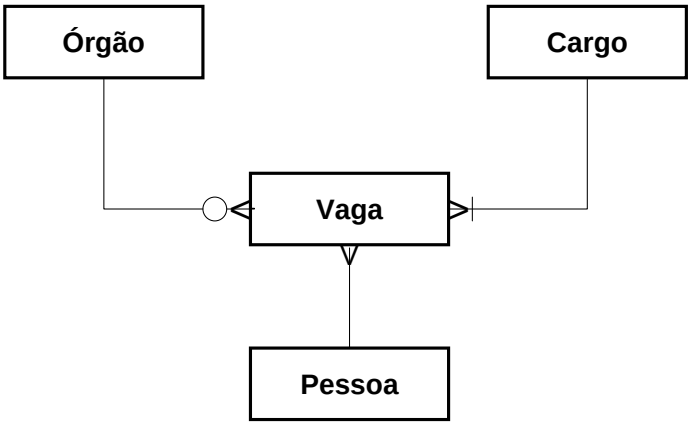



Figura 3.3 Exemplo de Modelo de Dados Entidade-Relacionamento

(Fonte: do autor da tese)

Com a solução apresentada no modelo, reduz-se o relacionamento entre os objetos (ou entidades) “Órgão” e “Cargo” a “um-para-N” mediante a introdução de um objeto de mediação denominado “Vaga”, pois no mundo real os cargos têm vagas a serem preenchidas e os órgãos alocam vagas para cada tipo de cargo em sua estrutura. Esse tipo de estrutura no modelo ER pode ser útil, em ambientes de mineração e modelagem de conceitos para Inteligência Competitiva, para identificação e representação de serviços nas organizações (internos ou externos), estado resultante da interação entre duas ou mais entidades formando uma terceira mais abstrata – como é o próprio conceito de “serviço” (o conserto do motor de um automóvel) ilustrado, anteriormente, na Figura 1.3.

### 3.3.3 Gráficos conceituais

Gráficos Conceituais (GCs) são estruturas conceituais, baseadas na Linguística, Psicologia e Filosofia, que constituem um formalismo para representação do conhecimento (SOWA, 1984; LUKOSE *et al*., 1997). O interesse nesse tipo de estrutura de representação cognitiva, no contexto da tese, se deve aos recursos de visualização pictórica com uso de uma linguagem próxima à linguagem natural, requisito de modelagem para propiciar sua compreensão tanto por cientistas e tecnólogos da informação como por usuários das áreas de negócio nas organizações. Eklund, Ellis e Mann (1996) apostam nas estruturas conceituais como uma possível “interlíngua” para representação do conhecimento, ou seja, uma linguagem comum, compreensível, manipulável em qualquer domínio do conhecimento.

Os GCs, desenvolvidos por Sowa (1984), têm origem nos Gráficos Existenciais (GEs) de Peirce (1909) e nas Redes Semânticas da Inteligência Artificial (QUILLIAN, 1968). Peirce, talvez devido aos seus estudos de Kant, observou, em 1866, que o único modo de se conduzir uma pesquisa metafísica seria “(...) adotando-se nossa lógica como nossa metafísica” (SEARLE *et al*., 1997, p. 1). Embora Peirce (1909) depositasse no seu construto metodológico de modelagem da informação e conhecimento a capacidade de execução de operações com base em diagramas, externas ou internas, tomando o lugar de experimentos com coisas reais, sua obra começou a ser reconhecida somente na década de 1960, com Zeman (1964), e de 1970, com Roberts (1973).

Searle *et al.* (1997, p. 2) argumentam que os GEs de Peirce representam o primeiro modelo articulado de processamento da informação e conhecimento implementando a idéia de representação do “curso geral do pensamento”, com exatidão, *vis a vis* o cálculo de predicados de primeira ordem – o que, para Peirce, seria “a lógica do futuro”. É importante ressaltar-se o conceito de experimentalismo de Peirce, que é primordialmente racionalista, elegendo a lógica como uma forma primária de experimentação sobre a estrutura da realidade. Conforme comentam Lukose *et al.* (1997, p. 3):

> *Peirce entendeu, tanto como qualquer dos seus ancestrais filosóficos, que nunca é fácil ser um metafísico realista, particularmente se o mesmo abranger, como Peirce o*

> *fez, a necessidade de experimentação e as modalidades do possível, do provável, e suas chances de serem constituintes da realidade.*

Outro importante aspecto da obra seminal de Peirce, no contexto, é seu pragmatismo. Lukose *et al*. (1997, p. 4), ponderando sobre críticas contemporâneas[42] à obra de Peirce, assim contra-argumentam:

> *O que é mais importante, certamente tanto na visão de Peirce como na nossa, é que aqui há evidência de uma emergente e crescente comunidade de pesquisa na qual o potencial heurístico e explanatório que Peirce viu em sua lógica gráfica caminha de mãos dadas com sua visão de Pragmatismo. Não é o instrumentalismo do tipo 'qualquer-coisa-que-funcione' que levou Max Horkheimer e Theodor Adorno a abandonar o pragmatismo (...) como quase um segundo pensamento (...), nem é o suave relativismo que Richard Rorty tem às vezes reclamado como o ponto do pragmatismo. É, ao contrário, a visão onde o significado de qualquer proposição é a soma das consequências que seguem de sua aceitação (para abreviar a famosa máxima do pragmatismo de Peirce).*

Obviamente, os GCs de Sowa surgiram num contexto diferente dos GEs de Peirce e foram desenvolvidos de modo independente das Redes Semânticas (RSs), embora esses construtos tenham como objetivo o suporte metodológico do raciocínio e a construção de artefatos de Inteligência Artificial e Linguística Computacional. Sowa (1984) apresenta sua fundamentação teórica dos GCs a partir de uma visão fenomenológica de mundo, definindo conceitos a partir de algo mais fundamental – as próprias percepções humanas acerca das coisas.

O conceito de "percepção", para Sowa 1984), é definido como o processo de construção de um "modelo operacional" que represente e interprete entradas de sinais sensoriais neurais, tendo dois componentes: uma parte sensorial formada por um mosaico de percepções de coisas mais atômicas, as quais ele chama de *percepts*, cada uma das quais encontrando algum aspecto das entradas; e uma parte mais abstrata chamada "gráfico conceitual", que descreve como os *percepts* se organizam em conjunto para formar o mosaico. A percepção seria, então, baseada nos seguintes mecanismos (SOWA, 1984, p. 69):

> *A estimulação é gravada por uma fração de segundo na forma de um ícone sensorial. O comparador associativo procura, na memória de longo prazo, por percepts que representem todo ou parte de um ícone. O montador dispõe os percepts em conjunto num modelo operacional que forma uma montagem aproximação da entrada. E um registro dessa montagem é armazenado como um gráfico conceitual. Mecanismos conceituais processam conceitos concretos que têm associados tanto percepts como conceitos abstratos que não têm percepts associados.*

O processo estruturante da percepção humana de um objeto, para Sowa, se iniciaria com o uso de um dos órgãos sensoriais – como a visão, por exemplo, que excitaria ícones sensoriais e os associaria (no comparador associativo) a partículas de percepção (os *percepts*) estocadas em seu cérebro, resultado de suas experiências sensoriais anteriores; os pares "ícone-*percept"* que se revelarem similares são, então, selecionados e transformados, por um mecanismo de montagem lógica, em gráficos conceituais. O conceito de "montagem" dos Gráficos Conceituais é

[42] Essas críticas, basicamente, contrapõem a necessidade do uso da Teoria de Conjuntos e da Álgebra, para o desenvolvimento de aplicações computacionais e de estratégicas de pesquisa em Inteligência Artificial, ao realismo filosófico (com sua dimensão especulativa) de Peirce.

fundamental, do ponto de vista científico, na medida em que implementa os três princípios elementares da Inteligência Artificial baseada na natureza (FREEDMAN, 1995):

I. A melhor maneira de se compreender como funciona a inteligência humana é estudar, antes, como funciona um modelo de inteligência mais elementar, como a inteligência animal;
II. A inteligência pode ser emergente, uma propriedade da interação complexa de elementos mais simples;
III. A inteligência é demasiado complexa para que possa ser projetada a partir do zero.

Entendendo a Inteligência Artificial (IA) como uma área de estudos de representação do conhecimento e seu uso na linguagem, raciocínio, aprendizado e solução de problemas (SOWA, 1984), algo como uma disciplina de engenharia do conhecimento, tornando operacionais construtos teóricos que imitam a capacidade de aprendizado e raciocínio humanos, a montagem de um Gráfico Conceitual (GC) parte de elementos mais simples, os *percepts*, para chegar a estruturas conceituais mais complexas, implementando o primeiro e o segundo princípios da IA baseada na natureza. Os *percepts*, como partículas elementares da percepção humana, são o ponto de partida necessário para implementação de composições mais complexas de conceitos, implementando o terceiro princípio mencionado.

Sowa combinou, para desenvolvimento dos GCs, recursos das RSs de Quillian com os Gráficos de Dependência de Tesniere-Hays para formar uma representação semântica para a linguagem natural (QUILLIAN, 1968; HAYS, 1964; TESNIERE, 1959) . A notação gráfica com caixas (ou retângulos) e círculos (ou elipses), considerados elementos pictóricos primitivos adequados do ponto de vista pragmático pretendido, foi influenciado pelos *templates* plásticos adotados no desenho de modelos computacionais de *softwares*. E alguns termos, como *join* (ajuntamento) e *projection* (projeção) foram adaptados da Teoria da Base de Dados Relacional de Codd, motivo pelo qual os GCs podem ser considerados descritores de relações extensionais de bases de dados (LUKOSE *et al*., 1997).

Em suma, os GEs de Peirce propiciaram os ingredientes anteriormente ausentes, mas necessários para uma versão rica e flexível de redes semânticas, que culminaram no desenvolvimento de uma notação simples e elegante, juntamente com regras de inferência baseadas em gráficos ao invés das regras de substituição da notação algébrica. A Figura 3.4 ilustra essas propriedades dos GCs no mesmo contexto semântico anteriormente modelado com uso da UML na Figura 1.2, para a sentença: “O mecânico conserta o motor do carro”.

Observe-se, na Figura 3.4, que os retângulos podem representar tanto classes de objetos, identificáveis em substantivos (nomes frasais) numa sentença de texto, como verbos e relações nominais e verbo-nominais. Os círculos com abreviaturas, nos GCs, significam que o objeto central da sentença, que é “consertar” (um verbo), tem um agente (AGNT: AGENT) que é um “mecânico”, e “consertar” tem um paciente (PTNT: PATIENT) que é um “motor”, parte (PART)

integrante de um “carro” (ou “o carro possui um componente que é um motor”). Com isso, resta evidente o funcionamento do modelo de GC da Figura 3.4, que expressa o próprio modelo geral de tradução (gráfica para texto) de um GC: “uma linha direta com rótulo *R* apontando do objeto *A* para o objeto *B* deve ser lida, geralmente, como *‘A tem um R que é B’;* ou *‘B é um R de A’*” (WILLE; 1997, p. 293).

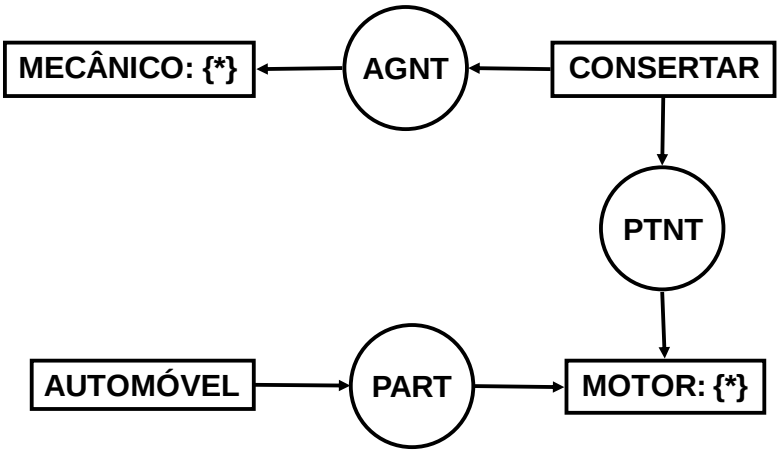



Figura 3.4 Exemplo de Gráfico Conceitual Representando uma Sentença
(Fonte: do autor da tese)

O termo gramatical “paciente” vem do inglês *patient*, significando uma pessoa ou coisa que é afetada pela ação de um verbo; “agente”, de *agent* (AGNT: AGENT), seria a pessoa ou coisa que realiza uma ação expressa como o sujeito de um verbo ativo, ou em uma frase de um verbo passivo com uso da partícula apassivadora “por” (ou *by*, no idioma inglês). Outros elos sintáticos entre objetos aparecem, com certa frequência, nas elipses de transição dos GCs, como “estado” (STAT: STATE), que seria a condição mental, emocional ou física em que se encontra uma pessoa ou coisa, que também pode ser interpretada como um atributo, e “recipiente”, do inglês *recipient* (RCPT: RECIPIENT), significando uma pessoa ou coisa que recebe algo.

O uso dessas relações nos GCs é um tanto intuitivo e derivado das próprias regras do léxico, especialmente quanto à regência e transitividade verbal, não apresentando dificuldade para os conhecedores das regras sintáticas do idioma natural de onde se extraem os conceitos.

Zeman (1997, p. 15) argumenta que, considerando-se a semiótica de Peirce, pode-se também afirmar que os GCs desempenham a função de representação simbólica da informação e não uma representação iconográfica ou de indexação. Ou seja, nos GCs “o símbolo é conectado ao seu objeto em virtude da idéia de mente-que-utiliza-símbolos, sem o que tal conexão não existiria”. Os símbolos primordiais processados pela mente humana, no caso, se referem aos da linguagem natural, no construto basilar “pensamento-linguagem” – eis, portanto, a força representacional dos GCs com base na filosofia.

Outro exemplo é o da Figura 3.5, de Cao e Creasy (1997, p. 418), que ilustra a capacidade de generalização de conceitos com GCs na representação da sentença: “Todo curso é oferecido por um professor universitário, que designa um estudante de doutorado para ser o tutor”. O modelo de representação da informação textual com uso de GC da Figura 3.5 mostra uma variante onde as elipses contêm, elas mesmas, a relação entre os objetos dos retângulos, decorrente do uso dos verbos na sentença, com exceção dos círculos com a indicação de um atributo para o objeto “professor” e um atributo para o objeto “estudante” (ATTR: ATRIBUTE). Cao e Creasy resgatam, com esse exemplo, os conceitos de “marcador universal” (com símbolo lógico ∀) e de “marcador genérico” (com símbolo ‘*’), chamando-os de “conceito universal” e “conceito existencial”, respectivamente.

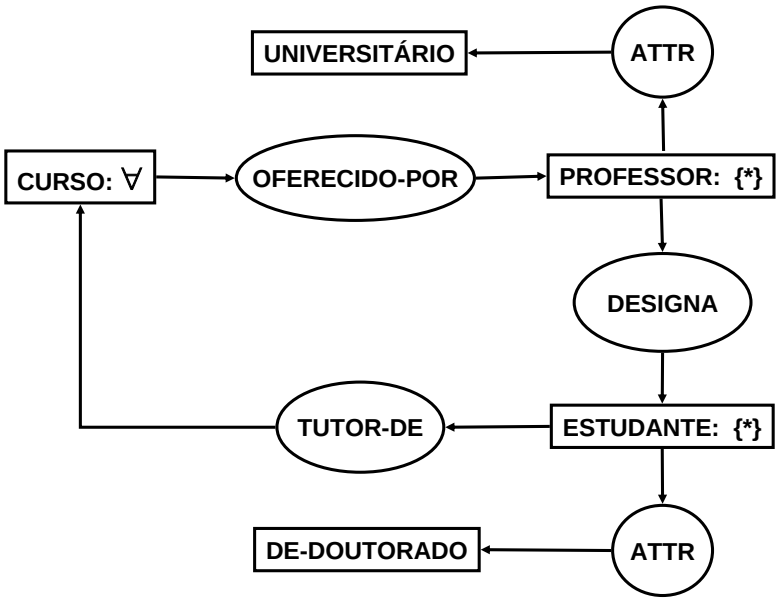



Figura 3.5 Generalização com Gráficos Conceituais
(Fonte: CAO e CREASY, 1997)

Esses autores argumentam que a força do GC é que sua linguagem gráfica tem um formalismo lógico que a maioria das outras linguagens gráficas não tem. E afirmam que isto não somente garante às representações de um GC significados precisos, mas também permite o raciocínio direto sobre essas representações (CAO e CREASY, 1997, p. 416). Essas observações são corroboradas pelas de Sowa, *apud* Wille (1997, p. 292):

> *(...) Com seu mapeamento direto para linguagem natural, os gráficos conceituais podem servir de linguagem intermediária para tradução de formalismos computacionais para e de linguagens naturais. Com sua representação gráfica, eles podem servir de linguagem imediatamente legível, mas também como uma linguagem formal de projeto e especificação.*

Outro aspecto interessante do GC como ferramenta analítica de processamento sintático e semântico da informação textual é sua capacidade de suportar raciocínios com conjuntos de conceitos. Cao e Creasy (1997, p. 416) comentam, sobre esse recurso, que “dados dois GCs,

pode-se determinar se um subsume o outro por uma projeção de GC executada diretamente nos dois GCs" (projeção e ajuntamento são duas operações de GCs).

Aplicações de GCs têm se multiplicado em várias áreas e domínios do conhecimento desde sua criação nos anos 1980, com entusiasmo crescente na última década. Lukose *et al*. (1997) comentam, a propósito, que tais esforços têm se expandido em áreas como processamento da linguagem natural, aquisição de conhecimento, recuperação da informação e fundamentos filosóficos da lógica. Gerbé (1997) apresenta um formalismo gráfico para representação da memória corporativa baseado em GCs, algo parecido com os elementos primitivos da UML. Eklund, Ellis e Mann (1996) publicaram, tal como Lukose *et al.* (1997), uma coletânea de artigos inclusive com aplicações organizacionais e aplicativos de *software* com GCs.

### 3.3.4 Linguagem de modelagem unificada

A Linguagem de Modelagem Unificada (UML) é uma linguagem gráfica para visualização, especificação, construção e documentação de artefatos de sistemas intensivos em *software* (BOOCH, RUMBAUGH e JACOBSON, 2005). O potencial da UML, no contexto, é avaliado em função de sua utilização na indústria do *software*, sendo a linguagem padrão de modelagem orientada a objetos mais disseminada nas organizações.

As origens da UML remontam à década de 1980, quando metodologistas de modelagem de sistemas computacionais assumiram o desafio de desenvolver uma nova linguagem de modelagem para complexos projetos de *software* com linguagens de programação orientadas a objetos tais como *Smalltalk, Objetive C, C++* e *Eiffel*. O desenvolvimento de uma linguagem de modelagem de sistemas envolve dois aspectos fundamentais: o escopo e alcance da linguagem, que se refere às suas funcionalidades, e o equilíbrio entre expressividade e simplicidade. Booch, Rumbaugh e Jacobson (2005), com relação a este dualismo, observam que um ponto de equilíbrio é crucial, pois uma linguagem de modelagem muito simples poderá limitar a amplitude de problemas que ela pode resolver e uma linguagem muito complexa poderá se tornar inatingível por desenvolvedores "mortais".

A primeira versão da UML foi liberada ao público em 1995 e desde então a linguagem vem sendo aprimorada com versões mais atuais (a atual é a 2.0, liberada em 2003). Ela se tornou uma linguagem padrão de modelagem de sistemas computacionais a partir das chamadas de propostas de padronização *(Request for Proposals – RFP)* do *Object Management Group* (OMG) e de sua aceitação no final de 1997.

Corroborando teses defendidas no escopo da Arquitetura Orientada a Serviços (*Service Oriented Architecture* – SOA), Booch, Rumbaugh e Jacobson argumentam que embora o principal produto de uma equipe de desenvolvimento de sistemas seja o código-fonte de um *software*, o desenvolvimento de um *software* de qualidade elevada requer uma sólida fundação de arquitetura

que seja resiliente a mudanças. O uso de modelos, nesse contexto, se justifica a partir do próprio conceito de modelagem e sua finalidade:

> *Um modelo é uma simplificação da realidade. Nós construímos modelos de modo que possamos melhor entender o sistema que estamos desenvolvendo. Nós construímos modelos de sistemas complexos porque não podemos compreender tal sistema em sua totalidade. Modelagem é uma técnica de engenharia comprovada e bem aceita. Nós construímos modelos de arquitetura de casas e edifícios para auxiliar seus usuários visualizarem o produto final. Nós podemos até construir modelos matemáticos para analisar os efeitos de ventos ou terremotos em nossas edificações. (...) Nos campos da sociologia, economia e gestão de negócios, nós construímos modelos de modo que possamos validar teorias existentes ou testar novas com o mínimo de risco e custo* (BOOCH, RUMBAUGH e JACOBSON, 2005, p. 5-6).

Outra vantagem da modelagem de um sistema de *software*, para os criadores da UML, é que isto facilita a comunicação na organização. A atividade de modelagem de um complexo sistema intensivo em *software* depende, para esses metodologistas, de uma modelagem adequada como a orientada a objetos, mas eles também reconhecem, de modo bem humorado, a prática bem diversa nas organizações e os contextos mais comuns onde os *softwares* corporativos são desenvolvidos (BOOCH, RUMBAUGH e JACOBSON, 2005, p. 5):

> *Curiosamente, um bocado de organizações que desenvolvem software começa querendo construir edifícios mas adota abordagens do problema como se estivesse construindo uma casinha de cachorro. Às vezes, tem-se sorte. Se você tem as pessoas certas no momento certo e todos os planetas se alinham adequadamente, então você pode, apenas pode, conseguir que sua equipe produza um software que impressione seus usuários. Tipicamente, no entanto, você não pode ter as pessoas certas (as pessoas certas estão, frequentemente, assoberbadas), nunca é o momento certo (ontem teria sido melhor), e os planetas nunca parecem alinhar-se (ao invés, eles se mantêm em movimento e fora de seu controle)*.

Essas observações são importantes, no contexto, porque uma metodologia de modelagem automática, ou pelo menos semi-automática, de sistemas de negócios a partir de informações preexistentes em algum repositório digital poderia amenizar essa dificuldade de se ter as pessoas certas no momento certo para se iniciar um projeto. Com o fator “tempo” sendo, também, um recurso cada vez mais escasso em ambientes de mudança contínua para adaptação e readaptação das organizações aos ambientes de mercados competitivos, presume-se que alguma modelagem geral das ontologias de sistemas organizacionais *a priori* dos projetos específicos poderia significar um “atalho” importante para se alcançar novos patamares de desempenho tecnológico com sistemas de informação, especialmente em projetos de reengenharia e Arquitetura Orientada a Serviços.

A visão contemporânea do desenvolvimento de *software* deposita na orientação a objetos (OO) todo seu embasamento epistemológico, significando que o principal bloco de construção dos sistemas de *software* é a classe de objetos. Conforme Booch, Rumbaugh e Jacobson (2005), como “objeto” entende-se uma “coisa” (real ou abstrata), geralmente extraída do vocabulário do domínio do problema ou do domínio da solução; e como “classe” uma descrição de um conjunto de objetos que são suficientemente similares, do ponto de vista do responsável pela modelagem, para compartilharem uma especificação comum. Outra definição de classes e objetos é que

classes são abstrações e objetos são manifestações concretas dessas abstrações. E todo objeto tem uma identidade, para que possa ser distinto de outros objetos, um estado (geralmente, dados associados a ele), e um comportamento (pode-se realizar ações com o objeto, e o mesmo pode, também, realizar interações com outros objetos).

Como uma linguagem, a UML provê um vocabulário e as regras de combinação de palavras desse vocabulário para o propósito da comunicação. E como uma linguagem de modelagem, seu foco se concentra na representação conceitual e física de um sistema. Os criadores da linguagem argumentam que a UML permite tanto a codificação automática, ou seja, o mapeamento direto de modelos gráficos em códigos-fonte, como a engenharia reversa, com mapeamento direto de códigos de programação em modelos gráficos.

O vocabulário da UML contém três tipos de blocos de construção de modelos: “coisas”, “relações” e “diagramas”. As coisas são as mais importantes abstrações dos modelos; as relações ligam essas coisas em conjunto; e os diagramas agrupam coleções interessantes de coisas. Os quatro tipos de coisas na UML são classificados em estruturais, comportamentais, de agrupamento e anotações. Os aspectos (ou coisas) estruturais da UML são os substantivos (nomes), que são também as partes mais estáticas num modelo, e envolvem classes de objetos, interfaces, colaborações, casos de uso, classes ativas, componentes, artefatos e nodos. Os aspectos comportamentais representam a parte dinâmica dos modelos da UML, extraídos de verbos da linguagem natural, prevendo interações, máquinas de estado e atividades. Os agrupamentos são a parte organizacional da UML, onde o modelo pode ser decomposto em pacotes, que são mecanismos puramente conceituais de organização tanto de coisas como de comportamentos – eles são, também, reflexivos, empacotando os próprios agrupamentos. Os artefatos anotacionais representam a parte explanatória, em linguagem natural, dos modelos com UML (comentários, marcações e notas).

As relações entre classes e objetos na UML, conforme a representação pictográfica da Figura 3.6, são de três tipos: (1) dependência, (2) associação, e (3) generalização e realização. A dependência é uma relação semântica entre dois elementos de um modelo na qual a mudança de um elemento (o independente) pode afetar a semântica do outro elemento (o dependente); uma associação é uma relação estrutural entre classes que descreve um conjunto de ligações, uma ligação sendo uma conexão entre objetos, que são instâncias das classes; uma generalização, de fato, é uma relação de especialização e generalização (dependendo do ponto de vista da análise) na qual o elemento especializado (elemento “filho”) incorpora a especificação do elemento generalizado (elemento “pai”) e compartilha a estrutura e seu comportamento; e uma realização é uma relação semântica entre classificadores, onde um classificador especifica um contrato que outro classificador garante cumprir (isso ocorre entre interfaces, classes de implementação de interfaces e casos de uso e colaborações).

O padrão atual da UML prevê até treze diagramas num modelo de sistema de *software*, representando: classes, objetos, componentes, estruturas compostas, casos de uso, sequências, comunicação, estados, atividades, implantação, empacotamento, *timing* e supervisão de interação. Como a essência da modelagem com UML se encontra nas classes e objetos, a presente pesquisa concentrar-se-á na representação do conhecimento com uso desses dois diagramas e do diagrama de atividades. Os diagramas de classes e objetos mostram a estrutura estática das relações entre as unidades do modelo e o diagrama de atividades as interações e fluxos de informações e decisões entre os mesmos.

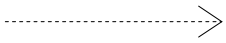
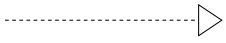
| Relação | Representação Pictórica |
|---|---|
| **Dependência** |  |
| **Associação** |   |
| **Generalização** |  |
| **Realização** |  |

Figura 3.6 Representação de Relações na UML
(Fonte: BOOCH, RUMBAUGH e JACOBSON, 2005)

A Figura 3.7 ilustra um modelo com diagrama de classes e relações de generalização entre as mesmas. As subclasses específicas “Retângulo”, “Círculo” e “Polígono” encontram na superclasse abstrata “Forma Geométrica” seu gênero ontológico. A superclasse, no caso, apresenta um único atributo denominado “origem”, cujos dados são do tipo “Ponto”; as subclasses, por serem entidades mais específicas, têm mais atributos, tais como “largura” e “altura” (na classe “Retângulo”), “raio” (na classe “Círculo”) e “posição de vértice” (“posVértice”, na classe “Polígono”). Os tipos de dados multivaloráveis desses atributos são típicos da geometria, indicando distâncias. O tipo de dados denominado “Lista de Pontos” seria derivado de um sistema de coordenadas geométricas.

Observe-se também, na Figura 4.7 (BOOCH, RUMBAUGH e JACOBSON, 2005, p. 65), que as classes têm métodos (ou funções computacionais) semanticamente algo parecidos, tais como “mover”, “redimensionar” e “exibir” (mostrar). Entre as três subclasses os métodos são os mesmos, diferindo as mesmas somente nos atributos. Os métodos são funções que deverão ser codificadas, com algoritmos computacionais adequados, na linguagem de programação adotada pelos desenvolvedores do *software*.

E uma classe pode ter nenhum (zero), um ou mais de um “pai” (ou superclasse de generalização). Com nenhuma superclasse relacionada uma classe é denominada “classe-raiz” ou “classe-base”; com apenas uma superclasse a classe utiliza “herança simples” e com mais de uma superclasse (classe-pai) uma classe utiliza “herança múltipla”. As classes que não têm filhos (subclasses) numa relação de generalização são denominadas “classes-folha” (BOOCH, RUMBAUGH e JACOBSON, 2005).

O recurso da herança é considerado um dos pontos fortes na metodologia de modelagem de sistemas orientada a objetos (OO), permitindo que uma subclasse utilize uma função prevista na superclasse ao qual se relaciona por generalização, resultando em economia de esforço de programação.

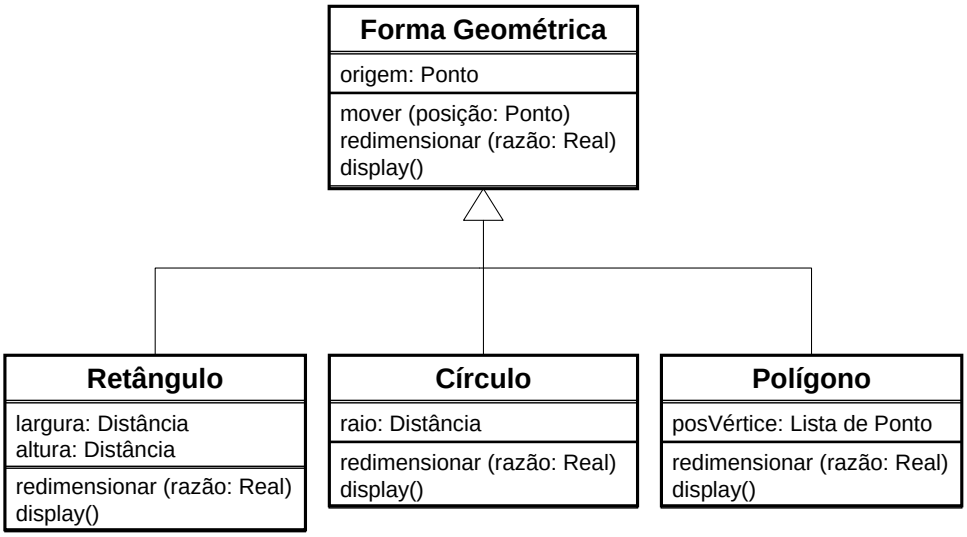



Figura 3.7 Generalização entre Classes
(Fonte: BOOCH, RUMBAUGH e JACOBSON, 2005)

A Figura 3.8 mostra uma representação pictórica dos relacionamentos estruturais típicos de um modelo ontológico de sistema computacional com uso da UML. O exemplo apresenta as classes “Escola”, “Departamento”, “Aluno”, “Curso” e “Instrutor” e dois tipos básicos de relacionamentos entre si: agregação e associação. A classe “Escola” é uma entidade maior que é integrada pelas classes “Aluno” e “Departamento”, caracterizando-se esse tipo de agregação com um “diamante” (paralelogramo) na extremidade da linha de associação próxima à classe “Escola”. O modelo também indica que a classe “Instrutor” integra a classe “Departamento” (ou seja, um instrutor é vinculado a um ou mais departamentos da escola).

As associações lembram o modelo entidade-relacionamento dos bancos de dados, indicando a cardinalidade dos relacionamentos entre as classes. O modelo da Figura 3.8, de Booch, Rumbaugh e Jacobson (2005, p. 73), indica que: uma escola pode ter um ou mais departamentos; um departamento pode ter zero ou um diretor e um ou mais instrutores; um instrutor pode estar vinculado a um ou mais departamentos; um instrutor pode lecionar em muitos

cursos; cursos podem ter um ou mais instrutores; cada departamento pode ter um ou mais cursos; alunos frequentam cursos; uma escola pode ter muitos alunos e alunos podem ser membros de uma ou mais escolas.

Deve-se destacar, na Figura 3.8, uma diferença fundamental entre a relação associativa de dependência existencial (ou de composição) entre as classes “Departamento” e “Escola”, representada pelo losango cheio na extremidade da linha próxima a “Escola”, e a relação associativa de “agregação” entre “Instrutor” e “Departamento”, representada pela extremidade com losango vazio. Weilkiens e Oestereich (2007, p. 48) assim definem essas duas modalidades de associação na UML:

> *Agregação é uma associação expandida pelo comentário semanticamente não restrito que as classes participantes não têm nenhum relacionamento de comparação equitativa; ao invés, elas representam uma hierarquia todo-partes. Agregação é utilizada para descrever como algo que é um todo é logicamente composto pelas suas partes.*
>
> *Composição é uma forma restrita de agregação onde a existência de suas partes depende do todo. O todo é o proprietário das partes. Descreve como algo que é um todo é composto de partes individuais.*

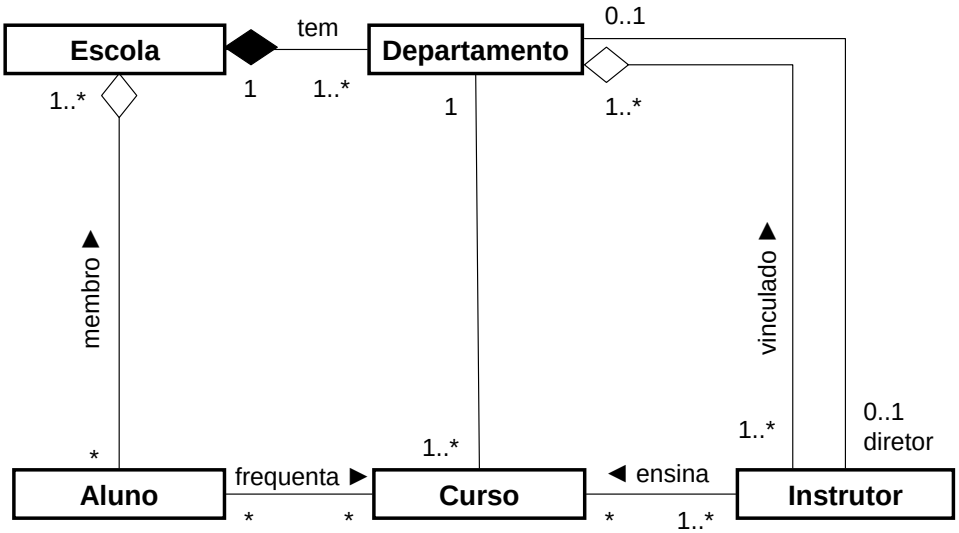



Figura 3.8 Relacionamentos Estruturais

(Fonte: WEILKIENS e OESTEREICH, 2007)

Embora pareça que a distinção entre uma agregação e uma composição não seja muito clara na UML, um critério conceitual deve ser utilizado para tanto: uma composição, ao contrário de uma agregação, não admite instanciação. Ou seja, uma classe que representa uma parte no todo de uma composição não pode pertencer a mais de uma composição simultaneamente, como numa associação entre as classes “Carro”, “Motor” e “Barco”: um objeto “motor” real, como uma instância da classe “Motor”, não pode ser agregada a um “carro” e a um “barco” simultaneamente.

Outro aspecto importante da UML, que tem conexões profundas com as estruturas sintáticas da linguagem natural, se refere às propriedades e atributos dos objetos. Conforme Weilkiens e Oestereich (2007, p. 42), “uma propriedade é uma característica estrutural especial que se

pertence a uma classe é um atributo*"*, mas que também pode pertencer a uma associação. E "um atributo é um elemento (dado) que é igualmente contido em cada objeto de uma classe e é representado por um valor individual em cada objeto*"*, não tendo identidade externamente à sua relação de pertencimento ao respectivo objeto.

A Figura 3.9 ilustra uma relação de associação ou notação (opcionais) entre um objeto da classe "Cliente" e seu atributo "reservas" de duas maneiras e três modos de representação: na primeira opção, tem-se uma relação de associação entre uma classe "Cliente" e uma classe-atributo "Reservas", onde a classe "Cliente" tem uma propriedade denominada "reservas" (sobreescrita à seta). A seta indica a direção da navegação do objeto para seu atributo e a notação de cardinalidade *0..** indica que cada cliente pode ter nenhuma reserva ou qualquer número de reservas na organização. Com a segunda opção, o atributo "reservas" aparece apenas como uma anotação de atributo dentro do retângulo da classe "Cliente".

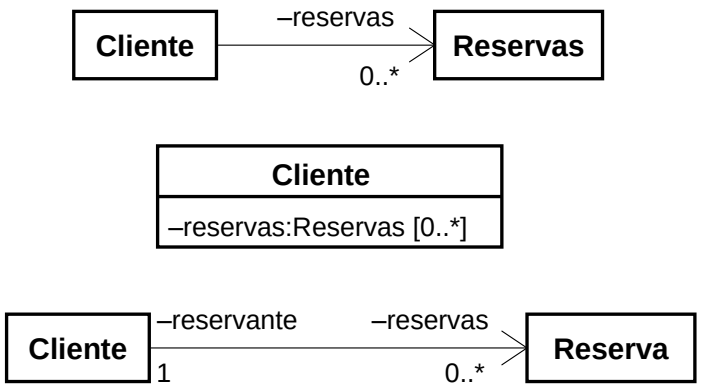



Figura 3.9 Atributo como Associação ou Notação
(Fonte: WEILKIENS e OESTEREICH, 2007)

O terceiro modo de representação dessa relação, na Figura 3.9, esclarece melhor a questão sobre a quem (a quais classes) pertencem o atributo "reservas" e seu simétrico "reservante": a classe "Cliente" se relaciona com a classe "Reserva" com cardinalidade um-para-zero ou um-para-muitas e tem o atributo "reservas"; simetricamente, seria de se esperar que o modelo indicasse que a classe "Reserva" tivesse o atributo "reservante", mas "reservante", no caso, é um atributo da relação em si (com a anotação *possExtrem* sob o retângulo da propriedade "reservante") e não de uma das classes envolvidas ("Reserva" nada conhece sobre "reservante").

Com a UML pode-se, ainda, esclarecer a estrutura dessa relação com uma representação denominada "repositório" modelada conforme a Figura 3.10 (WEILKIENS e OESTEREICH, 2007, p. 44). O metamodelo da Figura 3.10 mostra que "Cliente" e "Reserva" são objetos de "Classe" e que as propriedades "reservante" e "reservas" são objetos da classe "Propriedade", observando-se também um objeto anônimo da classe "Associação".

As classes são os tipos de propriedades e essa associação reconhece cada uma das duas propriedades como uma extremidade da linha de representação da associação *(assocExtrem)*. A associação descreve um conjunto de tuplas cujos valores se referem às instâncias da tipologia, observando-se que a propriedade “reservas” pertence à classe “Cliente” (é um atributo desta, representado pela anotação *possAtributo*). Em suma, na UML o próprio sentido da navegação na linha de notação da associação mostra a relação de propriedade do atributo pela classe/objeto.

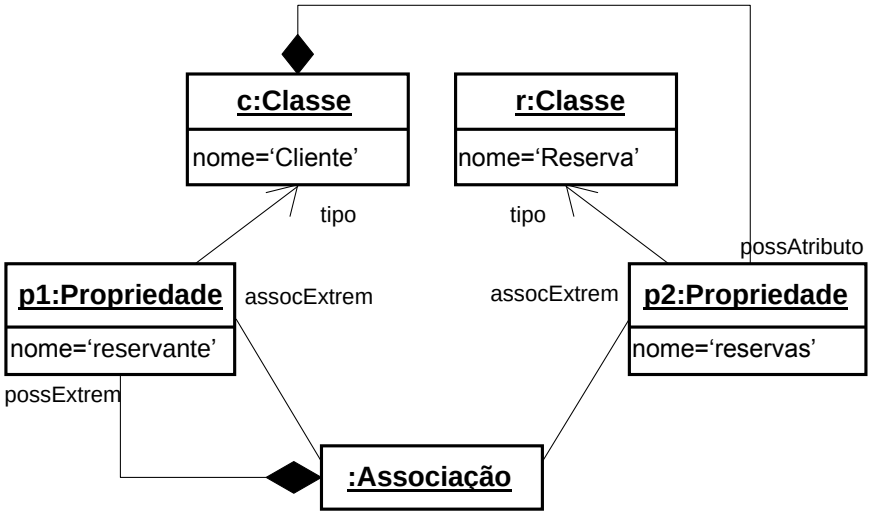



Figura 3.10 Exemplo de Repositório na UML
(Fonte: WEILKIENS e OESTEREICH, 2007)

As associações entre classes, e entre classes e atributos, também podem ser “n-árias” na UML, ou seja, uma associação pode ter mais que duas extremidades, como na relação entre as classes “Avião”, “Passageiro” e “Assento*”*. Este tipo de relação de associação teria três extremidades, sendo uma para cada classe, com todas as três classes se relacionando entre si. O ponto de encontro das linhas de associação, nesses casos, prevê um losango (sem preenchimento).

Outra representação da informação com UML considerada útil no contexto é o de atividades, que se relacionam com processos de negócio numa organização. O padrão da UML adota, neste caso, representações pictóricas derivadas de Redes Petri, com uso de retângulos com vértices arredondados, retângulos com vértices de ângulos retos, setas e textos indicativos de ações.

A Figura 3.11 apresenta um exemplo de modelo de atividade com UML ilustrado por Weilkiens e Oestereich (2007, p. 78). O retângulo maior representa as bordas (limites) do modelo pictórico da atividade como um todo, os retângulos menores (com vértices retos) nas bordas de entrada e saída do modelo representam a entrada (garrafas produzidas) e a saída (novo pacote de seis garrafas) do processo, os retângulos com vértices arredondados representam as ações que compõem a atividade, os retângulos menores ainda situados na ponta das setas

(denominados *pins*) representam os parâmetros de entrada e saída para a execução das ações, o losango vazio representa um ponto de decisão no processo e os dois círculos concêntricos um ponto de parada no processo.

Quando tanto o parâmetro de entrada como de saída de uma ação é o mesmo, utiliza-se um retângulo maior, com vértices retos, com os parâmetros escritos internamente (no caso, o parâmetro é “garrafa”, mas com o atributo de “rotulada”). Os parâmetros são também escritos abaixo do nome da atividade, que é escrita no canto superior esquerdo do retângulo do modelo. O modelo também declara as pré-condições e pós-condições da execução da atividade com expressões colocadas no canto superior direito do retângulo da atividade.

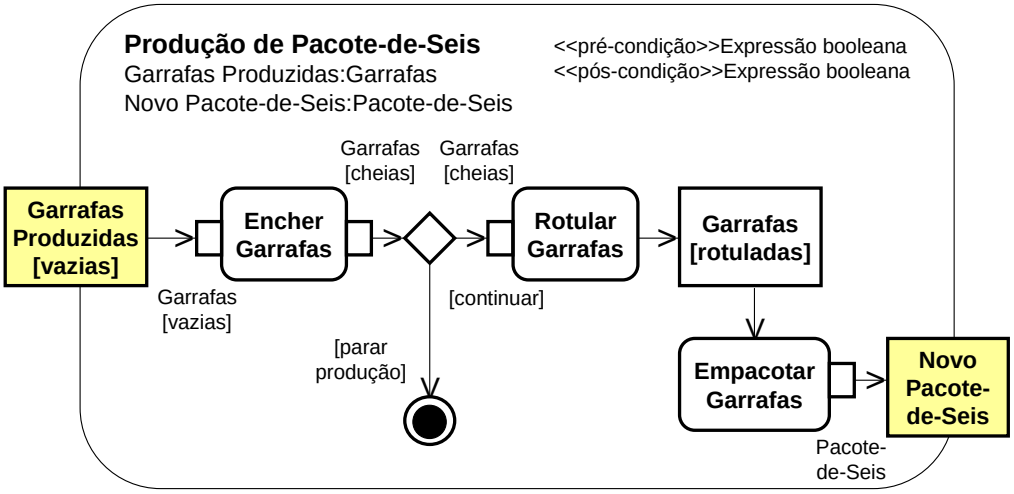



Figura 3.11 Modelo de Atividade na UML
(Fonte: WEILKIENS e OESTEREICH, 2007)

O exemplo da Figura 3.11, de Weilkiens e Oestereich (2007), mostra uma atividade industrial de produção de pacotes de seis garrafas (saídas) de uma determinada bebida a partir de garrafas vazias (entradas). Conforme o modelo, garrafas vazias passam por uma ação inicial de enchimento (“Encher Garrafas”), quando então se decide se a produção continua no período ou se é parada para intervalo (que pode ser ao final do turno de um dia de produção); caso o processo não seja interrompido, as garrafas cheias são então rotuladas (ação “Rotular Garrafas”) e, finalmente, empacotadas em pacotes de seis unidades (ação “Empacotar Garrafas”), quando a atividade é concluída com a apresentação das saídas (novos pacotes-de-seis unidades).

Gonçalves (2000) argumenta que uma organização é uma coleção de processos, um conceito que se encontra na essência da epistemologia da arquitetura de sistemas organizacionais, às vezes mencionados, simplesmente, como fluxos e comportamentos (ou regras). Portanto, qualquer avanço tecnológico no sentido da automação do desenho de processos organizacionais a partir de informações disponíveis (como entradas) pode ser

considerado útil em contextos competitivos, pois a velocidade com que as organizações se adaptam às mudanças no ambiente de negócios representa um diferencial importante.

Em relação à metodologia de mineração e modelagem de conceitos desta tese, os diagramas de atividades previstos na UML poderão ser criados como detalhamento de classes e objetos do domínio do negócio, após a modelagem ontológica. Como num exemplo de competição por preço apresentado por Fuld (2007, p. 95-102), muitas vezes o detalhamento de um processo operacional de produção, mostrando-se suas atividades e tarefas, pode revelar os segregos da competitividade empresarial. O autor utiliza uma metáfora muito sugestiva para sensibilizar os gestores sobre a importância de se conhecer os processos da concorrência e compará-los com os da sua própria organização: “enxergar as árvores para compreender a floresta”.

## 3.4 Aprendizado de ontologias

### 3.4.1 Origens

O que se conhece como *aprendizado de ontologias* é um conceito útil para a construção de Arquitetura da Informação Organizacional (AIO) e sua *praxis*, a Arquitetura Orientada a Serviço (AOS), explorando a Teoria de Conjuntos em sua essência. Como exemplo, apresenta-se a Análise de Conceito Formal (WILLE, 1982) como um método de classificação de ontologias baseado na noção de diferenciação *(differentiae)* de Aristóteles, formalizado na Conexão de Galois, para o aprendizado e população de ontologias no processo de construção de AOS. O método emprega abordagens lógico-matemáticas, lingüísticas, estatísticas e de Inteligência Artificial para extração de objetos e relações de documentos textuais e o aprendizado de ontologias, culminando com o povoamento de árvores hierárquicas de conceitos.

Conforme Cimiano (2006), o termo “aprendizado de ontologias” é uma criação de Mädche e Staab (2001) historicamente correlata à idéia de *Web Semântica*, que pode ser descrita como a aquisição de um modelo de domínio a partir de dados. O aprendizado ontológico pode ser entendido, grosso modo, como um processo de engenharia reversa onde, inicialmente, o autor do texto teria formalizado, por meio de linguagem natural, suas idéias acerca de um certo domínio epistemológico e, no aprendizado, tenta-se resgatar essas idéias e mostrá-las como uma representação de conceitos – em suma, essa reengenharia seria a reconstrução do modelo de mundo do autor do texto.

O processo pode ser compreendido a partir do “triângulo de significação” de Sowa (2000), com base nas noções distintas (em alemão) de sentido (*Sinn*) e referência (*Bedeutung*) de Frege, conforme a representação da Figura 3.12, onde se tem um exemplar de gato (objeto) como uma instância do conceito desse animal doméstico, exemplar que é conhecido como *Yojo* (nome que é um conjunto de símbolos lingüísticos utilizado como referência ao animal). Em outras palavras, o

autor de um texto, conhecendo um conceito de coisas no mundo, descreve um objeto que se refere a esse conceito com uso de símbolos lingüísticos. O objetivo da engenharia reversa seria redescobrir esse conceito indiretamente, sem acesso ao autor do texto, realizando o processo inverso, partindo dos símbolos para os objetos e seus conceitos.

O processo de aprendizado das extensões dos conceitos e das relações resultantes é conhecido como "população de ontologias", suportado por uma coleção de métodos desenvolvidos ou resgatados nos últimos anos. Cimiano (2006) apresenta alguns experimentos com o método Análise de Conceito Formal na indução de árvores de conceitos a partir de textos, demonstrando que é possível separar-se os termos extraídos dos textos em camadas ontológicas de significância (Figura 3.13). O modelo de "bolo" para o aprendizado ontológico dessa figura apresenta as fases desse tipo de atividade a partir da análise de textos em linguagem natural. A primeira fase se refere à aquisição de termos relevantes no texto, passando-se depois para a identificação de sinônimos ou variantes lingüísticas de termos, avançando-se depois para a formação de conceitos e sua hierarquização arbórea, aprendizado das relações entre os conceitos extraídos do texto (com informações do próprio texto ou textos auxiliares), hierarquização dessas relações, instanciação de axiomas e, finalmente, a definição arbitrária de axiomas úteis para o aprendizado de ontologias do domínio.

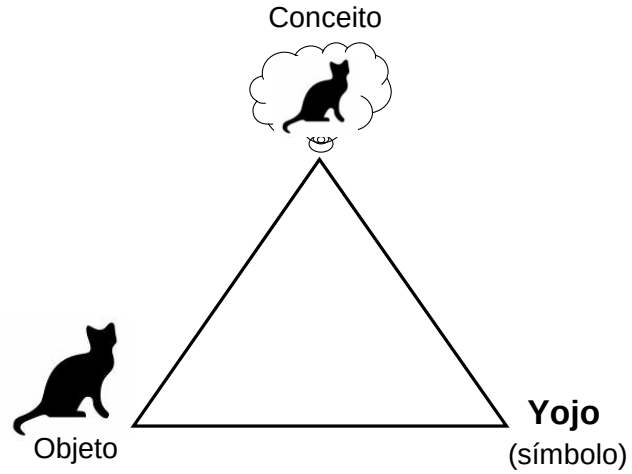



Figura 3.12 Triângulo de Significados de Sowa
(Fonte: CIMIANO, 2006)

"Árvores de conceitos" são construtos úteis para a classificação de objetos e seres na natureza geralmente empregados por taxonomistas na Biologia, Química, Lingüística e outras áreas do conhecimento científico e também nas organizações. O recurso metodológico da classificação de objetos constitui o núcleo conceitual dos processos de desenvolvimento de *software* na atualidade (GIGUETTE, 2006), com tendência de extensão de seus conceitos também para os domínios de processos de negócio e seus fluxos de atividades nas organizações.

O método, portanto, é apresentado como ferramenta auxiliar dos arquitetos de informação na construção de árvores de conceitos para desenvolvimento de AOS nas organizações, com objetivo de suprir uma lacuna existente nos métodos tradicionais de modelagem de processos e *softwares*.

Os métodos de modelagem tradicionalmente utilizados por engenheiros de sistemas, tais como os baseados na *Unified Modeling Language* (UML), dependem de interações pessoais entre analistas de sistemas e usuários, com pouco ou nenhum uso da informação documental existente nas organizações, ao passo que com ACF pode-se extrair, do acervo de informações legadas (textos com suporte papel ou digital) de uma ou sobre uma organização, termos lingüísticos que poderão ser classificados como conceitos úteis para modelagem do negócio e seus sistemas.

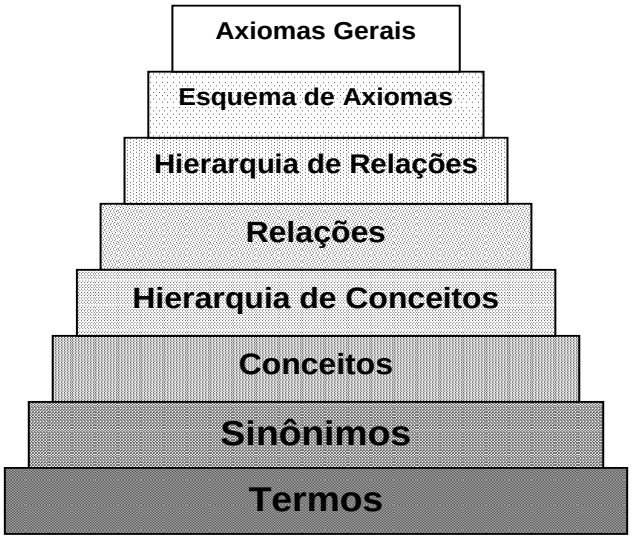



Figura 3.13 Camadas de Aprendizado Ontológico
(Fonte: CIMIANO, 2006)

O conceito de "Arquitetura Orientada a Serviço*"*, neste contexto, subsume o conceito de "serviço*"* como um modelo padronizado de entrega de algo de valor, que no caso do *software* corporativo poderá assumir uma variedade de situações. O conceito de serviço se refere à função de um componente da arquitetura de *software* corporativo, podendo significar uma simples rotina de verificação computacional da validade de um número de Cadastro de Pessoa Física (CPF) de um cliente até uma aplicação maior e mais complexa, como uma folha de pagamentos com milhares de beneficiários. Com a ACF, pode-se elicitar serviços, processos e classes de objetos e atributos de *software*, além de relações entre classes e instanciação (definição de objetos de processamento).

Outra utilidade da ACF, nesse cenário, atende à necessidade de demarcação dos limites ontológicos dos componentes de *softwares* legados para fins de refatoração, encapsulamento e reaproveitamento (reutilização) numa AOS. As principais vantagens do método são: (i) automação

do processo de modelagem de sistemas de informações computacionais (SICs) nas clássicas fases de elaboração e análise dos métodos orientados a objetos (OO); (ii) uso de fontes de informação documental textual em linguagem natural para elicitação de objetos de modelagem para AOS; e (iii) possibilidade de apresentação pictórica de modelos resultantes, facilitando o compartilhamento de modelos mentais e o processo de comunicação de conteúdos entre analistas de sistemas e usuários gestores do negócio. As experiências de Lindig e Snelting (1997) ilustram o potencial e as limitações desse construto na análise ontológica de sistemas de *software* legados em contextos de reengenharia de sistemas de informação.

Estima-se sua utilidade também em ambientes organizacionais que necessitam de integração de dados de sistemas heterogêneos, como os governos em geral, para prestação de serviços baseados em informação que necessitam do acesso a vários sistemas para sua composição e apresentação ao usuário.

### 3.4.2 Construção de árvores de conceitos

Embora nas camadas do aprendizado de ontologias da Figura 3.13 se preveja instanciação de axiomas e sua generalização, o método proposto não emprega tamanho rigor lógico (CIMIANO, 2006, p. 27). A tarefa de população de ontologias, para os fins a que se destinam neste caso, ocorrem em domínios específicos, culminando na definição e extensão de conceitos úteis para a modelagem de sistemas de negócio. Com isso, não se corre o risco de cometer erros que justifiquem a análise axiomática, dispensando-se, em princípio, a descoberta e discussão ontológica de objetos do negócio nesse nível.

A estrutura geral *(framework)* do modelo completo de aplicação do método proposto por Cimiano (2006) pode ser analisada na Figura 3.14. O processo se inicia com a análise sintática *(parsing)* de um texto ou conjunto de textos (documentos) sobre um determinado domínio, no caso produzidos pela organização ou sobre a organização (produzidos por terceiros), na qual se busca extrair, conforme Grefenstette (1994), Lin (1998), Gamallo *et al*. (2005) e outros autores, estruturas sintáticas como “verbo-preposição-complemento” e dependências de “verbo-objeto” e “verbo-sujeito”. Considera-se, para o ciclo de análise posterior, um par de termos composto por um verbo e um sujeito, um objeto ou uma frase preposicional. Então, para cada substantivo ou nome aparecendo como cabeçalho desses argumentos de busca, utiliza-se os verbos correspondentes como atributos para construção do contexto formal e calcular, no final, a rede de conceitos formais *(concept lattice)*.

Em seguida, buscando-se consolidar números menores de contextos a partir da eliminação de variações sintático-morfológicas, realiza-se a lematização ou redução dos termos a suas raízes ou bases morfológicas no respectivo idioma (como norma do método, se a palavra não pertence ao léxico de referência, a mesma não é lematizada).

A próxima fase envolve a comparação estatística dos pares de termos encontrados nas fases anteriores com pares de termos idênticos encontrados em *corpora* de referência adotados para o estudo, geralmente compostos de textos do domínio em questão (por exemplo, se a organização que estão sendo modelada é da área médica, provavelmente será recomendável utilizar-se um *corpus* como o MEDLINE). Esse requisito do método se justifica pelo fato, comumente observado nesse tipo de busca, que se extrai um grande número de objetos e relações que ofuscam a compreensão do domínio e levam à proliferação de dados para análise, muitas vezes sem correspondência dos pares nos *corpora* de referência para análise de freqüência. Assim, deve-se procurar estabelecer correspondência entre as freqüências de pares conhecidas (presentes nos *corpora*) e desconhecidas (presentes nos documentos de entrada no processo, mas desconhecidas nos *corpora*), a fim de se reduzir o número de pares no modelo de análise, processo que deve se apoiar em estimativas de freqüência de pares algo parecidos no contexto. Os pares de termos, a seguir, são ponderados, estatisticamente, de modo a estabelecer-se algum grau de importância respectiva ao domínio e ao contexto. Com isso, considera-se para a próxima fase apenas os pares com freqüência maior que o limiar definido no processo, que comporão o contexto formal no qual será aplicada a ACF.[43]

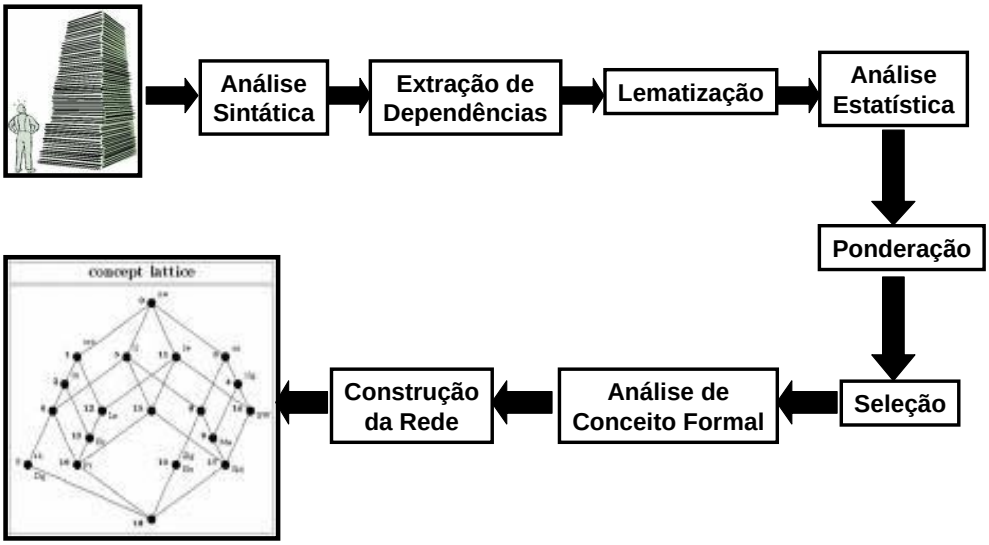



Figura 3.14 Processo de Indução de Hierarquia de Conceitos
(Fonte: CIMIANO, 2006)

A rede resultante desse processo é transformada, então, numa estrutura de ordem parcial que é próxima de uma hierarquia de conceitos tradicional. Como a ACF induz, naturalmente, a uma proliferação de conceitos, essa ordem parcial é compactada com a remoção dos conceitos abstratos comumente produzidos, obtendo-se uma ordem parcial compactada que é a hierarquia

[43] Conforme se descobriu no experimento que embasa esta tese, cada estrutura sintática utilizada como padrão de mineração nos textos, para se tornar útil, deve atender a propósitos específicos do projeto e refletir estruturas presentes em quantidade adequada nas fontes de informação.

de conceitos resultante. Cimiano (2006, p. 96) apresenta o processo de aprendizado e população de ontologias a partir de textos no algoritmo a seguir:

**Algoritmo**: Construir Hierarquia de Conceitos (D, T)
/* construir uma hierarquia para termos T com base em documentos D */
AnáliseSintática = analisar(*POS-tag*(D));[44]
DependSintática = represtags(AnáliseSintática);
DependSintáticaLematizada = lematizar(DependSintática);
DependSintáticaEqualizada = equalizar(DependSintáticaLematizada);
DependSintáticaPonderada = ponderar(DependSintáticaEqualizada);
DependSintática' = aplicarLimiar(DependSintáticaPonderada);
*K* = obtenhaContextoFormal(T,DependSintática');
*(β,≤)* = computarRede(*K*);
(*C',≤')* = transformar*(β,≤)*;
*(C",≤")* = compactar*(C',≤')*;
retorne*(C",≤")*;

A formalização do processo apresentada no algoritmo é importante para se demonstrar o cálculo lógico-matemático do método. Então, partindo do contexto formal *K* = (*G, M, I)* até a rede compactada *C",≤"*, onde *G* se refere a objetos, *M* a atributos e *I* às incidências, a demonstração a seguir, conforme Cimiano (2006, p. 97-98), é desenvolvida em duas definições formais:

**Definição 3 (Transformação de *(β,≤)* em (*C',≤')*):** em primeiro lugar, a ordem parcial *C'* contém tanto objetos como conjuntos de atributos:

$C' := G \cup \{B \mid (A,B) \in \beta\}$

$\leq' := \{(g,B_1) \mid \gamma(g) = (A_1,B_1)\} \cup \{(B_1,B_2) \mid (A_1,B_1) \leq (A_2,B_2)\}$

Como ACF tipicamente produz um grande número de conceitos, comprime-se a hierarquia de conceitos ontológicos resultante removendo-se qualquer nodo interno no qual a extensão é a mesma que a de um conceito-filho em termos de nodos-folha subsumidos. O resultado é uma ordem parcial $(C'',\leq''_c)$, conforme formalizada a seguir:

**Definição 4 (Hierarquia de Conceitos Compactada *(C",≤")*)**: assumindo-se que *lext(c)* é o conjunto de nodos-folha dominado por *c*, de acordo com $\leq'_c$, então:

$C'' := \{c_2 \in C' \mid \_c_1 \in C'\ c_2 \leq' c\ c_1 \rightarrow lext(c_2) \neq lext(c_1)\}$

$\leq''_c := \leq'_c \mid c'' \times c''$

Isto é, $\leq''_c$ é a relação $\leq'_c$ restrita aos pares de elementos de *C"*.

[44] POS: *Part-of-Speech* (parte da fala); tag: etiqueta de metadados que define a função sintática de cada palavra (exemplos: <VB> indica um verbo, <PP> indica uma preposição, <NP> indica um substantivo (nome) e <ADJ> indica um adjetivo). As *tags* acompanham as palavras no texto etiquetado, geralmente colocado ao seu lado posterior, como em *corpora* existentes em cada idioma.

O desenvolvimento de métodos e técnicas para implementação do *framework* de indução de hierarquias de conceitos de Cimiano (2006) tem duas correntes originais: os métodos de recuperação da informação para indexação de termos, baseado nas idéias de Salton e outros (SALTON e BUCKLEY, 1988), e as pesquisas em processamento da linguagem natural (PLN). Em suma, a extração de termos a partir de textos em linguagem natural implica o uso de níveis mais ou menos avançados de processamento lingüístico para identificação de frases nominais complexas que podem expressar termos interessantes num contexto e análise sintática de dependência de termos para identificação de suas estruturas internas. Geralmente, utiliza-se *softwares* etiquetadores *(taggers)* de termos para identificação de palavras, quando disponíveis *corpora* de comparação adequados (o *Treetagger*, da Universidade de Stuttgart, desenvolvido conceitualmente também na Universidade da Pennsylvannia, é um exemplo), ou se realiza etiquetação manual de *corpus*, aplicando depois algum método de reconhecimento de padrões sintáticos manualmente ou com uso de *softwares* de análise sintática *(parsing).*

Os passos seguintes desse método são mais ou menos automatizáveis, com uso de *softwares* estatísticos (alguns *parsers* implementam análise estatística também, como o *WordSmith*, desenvolvido na Universidade de Oxford) e aplicativos de construção de redes *(lattice builders)* e de organização de ontologias (*Protegè, ToscanaJ* e outros). Contudo, ainda não se conhece nenhum aplicativo integrado disponível no mercado, como *software* proprietário ou livre, para implementação de todas as fases do processo previsto no *framework* discutido neste capítulo. Como literatura de apoio para a fase seguinte, de identificação de termos sinônimos, Cimiano (2006) recomenda as técnicas de Harris (1968), baseada na hipótese que todos os termos que são extensões de algum conceito, num determinado contexto, são semanticamente similares. Essa abordagem tem sido explorada por Grefenstette (1994), Turcato *et al*. (2000), Buitelaar e Sacaleanu (2002) e Navigli e Velardi (2004). Outras técnicas são a *Latent Semantic Indexing (LSI)*, com Landauer e Dumais (1997) e a *Probabilistic Latent Semantic Indexing (PLSI)*, de Hoffman (1999), e suas variantes. Existe, no entanto, uma tendência mais atual de utilização da própria WWW como um imenso *corpus* lingüístico para pesquisa de sinônimos baseada em técnicas estatísticas.

A extração de conceitos de textos é uma tarefa um tanto polêmica exatamente em sua essência, ou seja, em relação à definição dessa atividade. Enquanto alguns autores abordam essa atividade na perspectiva de formação de agrupamentos *(clusters)* de termos correlatos considerados como conceitos, tais como Hindle (1990), Lin e Pantel (2001) e Reinberger e Spyns (2004), ou mediante a exploração de conexões entre palavras a partir da redução das dimensões de busca, como Schütze (1993) e Landauer e Dumais (1997), para formação de *clusters*, outros desenvolvem técnicas de formação e população de conceitos com a extensão de outros conceitos, como Evans (2003) e Etzioni *et al*. (2004). Existem, ainda, técnicas de aprendizado de

conceitos de modo “intensional”, como no caso do sistema *OntoLearn*, de Velardi *et al*. (2005), que deriva destaques para conceitos de domínios específicos (como na *WordNet*), com base em interpretação composta do significado dos conjuntos de termos (CIMIANO, 2006).

As estratégias mais conhecidas para indução de hierarquia de conceitos a partir de textos em linguagem natural são, basicamente, três:

I. Recuperação de termos com base em padrões léxico-sintáticos: com a aplicação de *templates* de palavras representando padrões estruturais do idioma, tais como os de Hearst (1992), descobre-se relações de interesse para formação e hierarquização de conceitos; entretanto, os pontos fracos dessa estratégia são que certos padrões ocorrem raramente em *corpora* e, por serem padrões sintáticos, são de precisão razoável, mas de baixa taxa de revocação *(recall)*; outros autores que desenvolvem essa estratégia são Buitelaar *et al*. (2004), com padrões sintáticos além dos de Hearst (1992), explorando as estruturas internas de frases nominais para derivar relações taxonômicas.
II. Hipótese distributiva de Harris: utiliza abordagens de *clustering* para derivar, automaticamente, hierarquias de conceitos a partir de textos em linguagem natural, aproveitando-se a dupla vantagem da clusterização: (i) formação de conceitos e (ii) indução de hierarquia de conceitos. A potência das técnicas de clusterização se revela na medida em que ao formarem os *clusters* de palavras similares, que são, naturalmente, extensões de um conceito, elas permitem a generalização dessas extensões para formação de conceitos que representam essas extensões, ordenando os *clusters* hierarquicamente. Os autores dessa estratégia são Faure e Nedellec (1998), Bisson *et al*. (2000) e Cimiano *et al.* (2004).
III. Análise de co-ocorrência de termos na mesma sentença, parágrafo ou documento: conforme Cimiano (2006), a técnica de Sanderson e Croft (1999), por exemplo, apresenta uma noção de subsunção baseada em documento onde um termo $t_1$ é mais específico que um termo $t_2$ se $t_2$ aparece em todos os documentos em que $t_1$ ocorre.

Existem poucas abordagens de aprendizado de relações ontológicas de textos publicados na literatura, aparecendo o trabalho seminal de Mädche e Staab (2000) como o mais conhecido. O método desses autores é uma variante do algoritmo de extração de regras de associação baseado na co-ocorrência de termos em sentenças. Conforme Cimiano (2006, p. 29), outros autores produtivos nessa linha de pesquisa são: Ciaramita *et al*. (2005), Gamallo *et al*. (2002) e Schutz e Buitelaar (2005), ainda que essas iniciativas tenham apenas “arranhado a superfície do problema”.

Quanto ao problema da população de ontologias, existem, nos métodos e técnicas mais conhecidos, restrições de escala. Em geral, o problema é atacado na abordagem conhecida como Reconhecimento de Entidades Nominadas (NER)[45] e Cimiano (2006) apresenta um método

[45] *NER: Named Entity Recognition.*

alternativo, com poder de processamento escalar, baseado na *Web*, que ele denomina *Learning by Googling*. A implementação desse método é apresentada no paradigma denominado *Pattern-based Annotation through Knowledge on the Web (PANKOW)*, cuja vantagem reside na simplicidade de se utilizar um motor de busca na *Web* para encontrar padrões léxico-sintáticos. Os resultados dessa busca são, então, agregados para se descobrir um conceito apropriado para uma determinada instância conceitual e as anotações semânticas são aproximadas analisando-se, estatisticamente, suas respectivas ocorrências na *Web* em face de determinadas estruturas indicando alguma relação de interesse para a tarefa de população de ontologias.

## 3.5 Criação de significado

O último ato de um processo de mineração de dados ou de textos para o aprendizado de conceitos num contexto é aquilo que Choo (2003) denomina, nos ambientes organizacionais, *construção do conhecimento*. É o ponto em que os engenheiros do conhecimento precisam mostrar o resultado do processo – o produto elaborado de inteligência que denominamos *conhecimento*. Esse conhecimento também precisa ser útil para tomada de decisão no contexto de negócio onde se insere a organização que patrocina os investimentos em Inteligência Competitiva.

Considerando-se o modelo de inteligência proposto, onde o conhecimento é representado por conceitos elicitados a partir de uma base de informação textual, deve-se separar, dos conceitos elicitados, aqueles que representam a informação mais relevante para o desenvolvimento de *insights*. É nesse aspecto que a relevância da informação, no caso, é associada com oportunidades de geração de nova *informação sobre informação* – ou, em síntese, de novas elaborações mentais que propiciem a geração do conhecimento útil para a organização.

Os construtos epistemológicos da literatura discutida a seguir apontam, cada um a seu modo, mas com notável convergência, para o conceito de *diferença* (ou *contraste*) no reconhecimento de padrões em modelos mentais como a chave para a percepção da relevância da informação. Esses construtos, apesar de desenvolvidos, em sua maioria, a partir do estudo da complexidade do ser humano como indivíduo, podem ser compreendidos também numa versão organizacional, como a síntese de Choo (2003), baseada nos ensaios cognitivistas de Weick (1995), ambos talvez os construtos mais completos acerca dos processos de aprendizado e tomada de decisão no contexto das organizações.

O conceito de *diferença* é tão fundamental para a compreensão da informação relevante que aparece, originalmente, nos tratados metafísicos de Aristóteles sobre categorias de coisas que existem no mundo, e, no Século XX, num construto basilar da engenharia eletrônica utilizado em artefatos tecnológicos de transmissão de dados.

Kneller (1980) argumenta que o homem primitivo estava em grande parte à mercê da natureza e que, talvez, o seu motivo mais forte para a investigação natural fosse atingir a paz de espírito por meio de alguma explicação plausível para os desastres da natureza, fenômenos que perturbavam seu sistema cognitivo. Ele observa que os cientistas, ao longo da história, parecem ter sido inspirados menos por motivos éticos e práticos que por duas emoções primordiais: o assombro e o medo. Essas emoções se encontram relatadas nos escritos de Epicuro:

> *(...) se não fôssemos perturbados por apreensões acerca de fenômenos no céu e a respeito da morte, se nada disso nos afetasse de um modo ou de outro, e também se não fôssemos perturbados por nosso fracasso em perceber os limites das dores e dos desejos, não teríamos necessidade alguma de estudar a natureza.* (*apud* KNELLER, 1980, p. 12)

Esse conceito primitivo de *diferença* como uma *perturbação cognitiva*, como discutido a seguir, representa a base antropológica e psicológica da própria noção de *informação* no sentido de *notícia*, aquilo que acrescenta algo novo na *equação do conhecimento* de Brooks (1980).

### 3.5.1 Conceito, sentido e referente

O uso da linguagem natural para representação das coisas do mundo para nós mesmos e para interpretar o discurso de outros seres humanos é um antigo tema de interesse da filosofia. O problema se reporta ao significado *(meaning)* das palavras, sentenças e discursos, que é tratado em disciplinas como lingüística e semiótica (ou semiologia). Os últimos cem anos de produção filosófica sobre o problema do significado são reapresentados por Richard (2003), com artigos seminais de Davidson, Frege, Kripke, Horwich, Putnam, Quine, Soames, Wilson e outros. Em particular, interessa ao desenvolvimento das idéias desta tese a obra de Frege (2003) sobre os três aspectos cognitivos básicos de significância das coisas, que são denominados “conceito”, “sentido” (*Sinn*, no original em alemão) e “referente” *(Bedeutung).*

Frege (2003) inicia sua argumentação questionando se o termo “igualdade”, comumente utilizado, também, em linguagem matemática, é uma relação ou não. E se, em caso positivo, é uma relação entre objetos ou entre nomes ou símbolos[46] de objetos. Ele observa que as igualdades “a = a” e “a = b” têm valores cognitivos diferentes porque a primeira, de acordo com Kant, tem um significado analítico *a priori*, mas a segunda contém, frequentemente, extensões muito valiosas de nosso conhecimento e não pode sempre ser estabelecida *a priori*. O que Frege (2003, p. 36) questiona é um ponto chave que interessa ao estudo semântico do discurso:

> *(...) se entendêssemos a igualdade como uma relação entre as coisas que os nomes “a” e “b” designam, pareceria que “a = b” não poderia ser diferente de “a = a” (assumindo-se que a = b é uma verdade). (...) O que se quer dizer com “a = b” parece ser que os símbolos ou nomes “a” e “b” designam a mesma coisa, de modo que esses próprios símbolos estariam sob discussão; uma relação entre eles seria estabelecida.*

O filósofo alerta sobre o risco que se corre ao se atribuir símbolos às coisas, pois como essa

[46] Assumiu-se *sign* como “símbolo”, ao invés de “sinal”, por considerá-lo mais apropriado no contexto.

simbolização pode ser arbitrária, pode-se também, consequentemente, perder a veracidade factual (ou conhecimento) expressa pela relação "a = b". O problema reside no modo de representação, que pode ser diferente apesar de se referir ao mesmo objeto, e Frege (2003, p. 37) oferece o seguinte exemplo como recurso para esclarecimento desse aspecto:

> *Sejam 'a', 'b', 'c' os segmentos de reta conectando os vértices de um triângulo com os pontos médios dos lados opostos. O ponto de intersecção de 'a' e 'b' é, então, o mesmo que o ponto de intersecção de 'b' e 'c'. Deste modo temos diferentes designações para o mesmo ponto e esses nomes ('ponto de intersecção de a e b', 'ponto de intersecção de b e c'), do mesmo modo, indicam o modo de representação; e, ainda, a declaração contém o conhecimento no caso.*

As noções de "sentido" e "referente" aparecem no argumento de Frege (2003) para distinguir o objeto (ou "referente"), que é o ponto de intersecção dos dois segmentos de reta no interior do triângulo, de seu modo de representação geométrica (ou "sentido"). O modo de representação estaria contido no sentido do símbolo e o objeto como uma referência desse símbolo. Complementando o exemplo do triângulo com outro, o da "estrela da manhã" (que é o planeta Vênus)[47], Frege (2003, p. 37) argumenta que *o referente da "estrela da tarde" seria o mesmo que o da "estrela da manhã", mas não o sentido*.

Em termos práticos, com Frege (2003) conclui-se que um referente é um objeto identificado e definido no mundo (uma instância de uma classe de objetos, portanto), mas os sentidos que se referem a esse objeto podem envolver mais de um conceito ou relação. Considerando-se que textos publicados de organizações, processos e produtos envolvem conceitos expressos, muitas vezes, por termos abstratos tais como "sistema de saúde", "gestão da cadeia de suprimento", "inteligência de negócios" e outros, a distinção de referentes e sentidos é um requisito na mineração de conceitos, algo que se relaciona, no jargão técnico da recuperação da informação para população de ontologias, como "desambiguação" – ou resolução semântica.

Essa trilogia lingüístico-cognitiva de Frege (2003, p. 38-39) é completada com a noção de "ideia" ou "pensamento" *(thought)*:

> *O referente e o sentido de um símbolo devem ser diferenciados da ideia associada. Se o referente de um símbolo é um objeto perceptível pelos sentidos, minha ideia do mesmo é uma imagem interna, surgindo de memórias de impressões dos sentidos que tenho tido e de atos, internos e externos, que tenho desempenhado. (...) Isto resulta, em decorrência, numa variedade de diferenças nas ideias associadas com o mesmo sentido.*

Este tipo de preocupação com ideias, sentidos e referentes, no uso e na interpretação de textos, numa etapa prévia ao da criação de significado, aparece naturalmente na mineração de conceitos, como será mostrado no experimento que embasa esta tese. O destaque de informação relevante resultante da mineração, no caso, deve decidir se a informação relevante, que se revela por contraste ou diferença com o estoque de informação disponível ao engenheiro do conhecimento (tácita ou explícita), se diferencia no nível das ideias, dos sentidos ou dos referentes. Frege (2003, p. 43), a propósito, advoga que o conteúdo-verdade se encontrará

[47] O planeta Vênus pode ser observado, a leste, antes do amanhecer ou logo após o pôr-do-sol.

sempre na conexão verdadeira do sentido com seu referente:

> *Nós nunca podemos nos preocupar somente com o referente de uma sentença; mas, novamente, o mero pensamento, isoladamente, não produz nenhum conhecimento, mas somente o pensamento em conjunto com seu referente, isto é, seu valor-verdade.*

Em síntese, com base no argumento filosófico de Frege (2003) tem-se um requisito de mineração conceitual de textos em função do qual se deve tentar conectar os conceitos mais complexos com conceitos cada vez menos complexos, num processo recursivo de resolução semântica, com nova mineração mais orientada ou supervisionada. O modelo proposto de mineração recursiva ilustra a utilidade desse argumento.

### 3.5.2 Relevância da informação

#### 3.5.2.1 Abordagem da engenharia: auto-informação

O conceito de "auto-informação" *(self-information)* é uma contribuição das mais antigas para o próprio conceito de "informação relevante", remontando aos primórdios da comunicação eletrônica. É uma medida quantitativa de informação formulada por Shannon (1948) que se relaciona com o conceito de "diferença" ou "discrepância" de Bateson (2002) e Weick (1995), constituindo a base científica da compressão de dados utilizada na engenharia eletrônica. Obviamente, como recurso de engenharia, a auto-informação mede, fundamentalmente, a eficiência na comunicação de sinais úteis (discrepantes em relação a um padrão anteriormente conhecido), que se revela em situações tais como o índice de eficiência de um algoritmo compactador de sinais eletrônicos num processo de transmissão de dados.

Sayood (2000, p. 14) apresenta esse conceito da seguinte forma:

> *Shannon definiu uma quantidade chamada 'auto-informação'. Suponha que temos um evento A, que é um conjunto de resultados de algum experimento randômico. Se P(A) é a probabilidade do evento A ocorrer, então a auto-informação associada ao evento A é determinada, matematicamente, pela equação:*
>
> $$i(A) = \log_b 1 / P(A) = - \log_b P(A)$$

Onde i(A) é o valor da auto-informação relacionada à ocorrência de um evento "A" e o parâmetro "b" depende de cada caso, representando o universo de resultados possíveis em relação a esse evento. Quando se tem um evento com apenas dois resultados possíveis, como no caso do lançamento de uma moeda, tem-se b = 2 (cara e coroa); no caso de lançamento de um dado, b = 6; e assim por diante. Assim, P(A) poderia ser a probabilidade de resultar "cara" no lançamento de uma moeda.

Intuitivamente, pode-se compreender esse modelo paramétrico de informação relevante com o argumento lógico-matemático de Sayood (2000, p. 14):

> *O uso do logaritmo para se obter uma medida de informação não foi uma escolha arbitrária (…). Lembre-se que log(1) = 0 e que – log(x) cresce à medida que x*

> *decresce de um para zero. Assim, se a probabilidade de um evento é baixa, a quantidade de auto-informação associada é alta; se a probabilidade de um evento é alta, a informação associada é baixa. E mesmo se ignoramos a definição matemática de informação e simplesmente usamos a definição que usamos na linguagem do dia a dia, isto faz algum sentido intuitivo. O latido de um cão durante um assalto é um evento de alta probabilidade e, por isso, não contém muita informação. Entretanto, se o cão não late durante o assalto, este é um evento de baixa probabilidade e contém um bocado de informação.*

Contudo, a associação entre as definições de informação sob os pontos de vista da auto-informação e da semântica deve ser estabelecida conhecendo-se o contexto e suas implicações epistemológicas. Como exemplo, pode-se questionar, numa comparação entre uma série de símbolos da linguagem natural distribuídos randomicamente e um texto com discurso estruturado, qual dos dois conjuntos de símbolos apresenta valor mais alto de auto-informação. A resposta a essa questão deve seguir, logicamente, a resposta de uma questão mais básica no modelo matemático: "Qual dos dois conjuntos de símbolos é mais provável ocorrer na comunicação entre humanos no mundo real?" Obviamente, o evento mais provável é o de um texto estruturado e legível para os humanos e, portanto, será este o conjunto de sinais com menor quantidade de auto-informação e, desse modo, menos relevante.

Outro exemplo evidente é o de eventos compostos: Quais seriam os valores da auto-informação, numa sequëncia de três lançamentos de uma mesma moeda, do resultado combinado "três caras" (ou "três coroas")? E de "duas caras e uma coroa" (ou vice-versa)? Em qual situação se teria um valor mais alto de auto-informação e, portanto, de "surpresa"? O cálculo, nesse caso, seria:

M = evento de três lançamentos de uma moeda;

$P(M_{HHH \vee TTT})$ = probabilidade de ocorrência de "cara-cara-cara" ou "coroa-coroa-coroa";

i(M) = valor da auto-informação de M;

b = universo de resultados combinados possíveis.

Onde "b" poderá ser:

cara-cara-cara, cara-cara-coroa, cara-coroa-cara, coroa-cara-cara,
coroa-coroa-cara, coroa-cara-coroa, cara-coroa-coroa, coroa-coroa-coroa

Logo, b = 8 (combinações possíveis).

Considere-se, ainda, $M_{HHH}$ como o resultado "cara-cara-cara" (H de *head*), $M_{TTT}$ como o resultado "coroa-coroa-coroa" (T de *tail*) e $M_{XXX}$ como outro resultado qualquer.

Como se tem $P(M_{HHH}) = P(M_{TTT}) = P(M_{XXX}) = 1/8$,

então: $P(M_{HHH\&TTT}) = 2 \times 1/8 = 1/4$, e

$P(M_{HHT\&HTH\&THH\&...\&TTT}) = 6 \times 1/8 = 3/4$

Finalmente, tem-se os seguintes valores de auto-informação:

$i(M_{HHH\&TTT}) = - \log_b P(M_{HHH\&TTT}) = - \log_8 (1/4) = 0{,}667$

$$i(M_{HHT\&HTH\&THH\&...\&TTT}) = - \log_b P(M_{HHT\&HTH\&THH\&...\&TTT}) = - \log_8 (3/4) = 0{,}138$$

Conclui-se, com base no cálculo de auto-informação, que os resultados iguais (“cara-cara-cara” ou “coroa-coroa-coroa”) contêm muito mais auto-informação que os resultados de combinações outras, mais comuns, nos três lançamentos de uma moeda. Ou seja, três resultados iguais em sequência causam mais surpresa que três resultados mesclando “cara” e “coroa”.

Outra propriedade do modelo matemático de auto-informação é o cálculo com a composição de eventos independentes e simultâneos. Com dois eventos independentes A e B, teríamos (SAYOOD, 2000, p. 14):

$$i(AB) = \log_b 1 / P(AB),$$

$$P(AB) = P(A).P(B),$$

$$\text{logo:} \quad i(AB) = \log_b 1 / ((P(A).P(B)),$$

$$i(AB) = \log_b 1 / P(A) + \log_b 1 / P(B),$$

$$i(AB) = i(A) + i(B)$$

O impacto cognitivo esperado com três resultados iguais em três lançamentos de uma moeda é intuitivo, mas quando utilizamos esse adjetivo assumimos, explicitamente ou implicitamente, um determinado contexto que condiciona nosso raciocínio, no qual desenvolvemos essa intuição. É o conhecimento do contexto do evento que nos permite calcular (ou estimar) as probabilidades de cada resultado possível e inferir sobre sua plausibilidade. É o contexto, também, que nos permite avaliar o “tamanho” da surpresa em relação à informação (ou “auto-informação”) sobre os resultados do evento.

O conceito de auto-informação se encontra associado, por exemplo, ao desenvolvimento de algoritmos de *softwares* etiquetadores de textos baseados em Redes de Markov e compressores de sinal eletrônico para comunicação de dados, onde a probabilidade de uma próxima letra, numa palavra (para reconhecimento da função sintática), é fortemente influenciada pelas letras precedentes.

Em contextos menos “comportados”, a surpresa com uma nova informação num fluxo de eventos pode não ser tão grande como em contextos onde existem crenças arraigadas. O processo indutivo utilizado na ciência para generalização do conhecimento empírico, por exemplo, é bastante suscetível a surpresas, podendo apresentar o que Taleb (2007) chama de *o impacto do altamente improvável*, algo na linha de raciocínio de Popper com a descoberta de cisnes negros quando, pela série de observações anteriores, pressupunha-se a existência de apenas cisnes brancos. É com a exploração dessas forças contraditórias, por um lado a previsibilidade (psicológica) das informações num contexto e, por outro lado, as incertezas do mundo real das organizações, que se mostra bastante promissora a abordagem do processo de criação de

significado e construção do conhecimento proposto nesta tese.

### 3.5.2.2 Abordagem ecológica de Bateson

Bateson (2002, p. 27), desenvolvendo uma abordagem alternativa da ecologia como ciência, argumenta que o conceito de *diferença* se encontra na essência do próprio conceito de *ciência*:

> *(...) ciência é um modo de perceber e de fazer o que nós chamamos de "sentido" de nossas percepções. Mas a percepção opera somente sobre "diferença". Toda recepção de informação é necessariamente a recepção de notícias sobre diferença, e toda percepção de diferença é limitada por limiar. Diferenças que são muito pequenas ou que se apresentam muito vagarosamente não são perceptíveis. Elas não são alimento para percepção. (...) o que nós, como cientistas, podemos perceber é sempre limitado por um limiar. (...) Conhecimento em um determinado momento será uma função dos limiares de nossos meios disponíveis para percepção. (...) Como um método de percepção – e isto é tudo o que a ciência pode querer ser – a ciência, como todos os outros métodos de percepção, é limitada em sua capacidade de coletar os sinais externos e visíveis daquilo que pode ser a verdade. A ciência explora; ela não prova.*

Ele observa, de um ponto de vista filosófico, que para existir diferença é necessária a existência simultânea de três entidades reais ou imaginárias: dois objetos e um relacionamento entre os mesmos. O modelo de relacionamento entre esses objetos deve ser capaz de evidenciar essa diferença entre os mesmos, ou de mostrar a *notícia dessa diferença* – que é *informação*. Com esse argumento, Bateson (2002) define *informação* como *a diferença que faz diferença* e assume que essa *notícia da diferença* entre dois objetos pode ser representada como uma diferença dentro de uma unidade de processamento de informação tal como um cérebro ou, talvez, um computador.

O conceito de informação de Bateson (2002) é de compreensão bastante simples no dia-a-dia dos seres humanos. Como exemplo, tome-se o conceito de notícia de um jornal: a informação publicada por um jornal, para ser valiosa, precisa ser novidade (ou notícia), o que, do ponto de vista de conteúdo cognitivo, precisa ser diferente de tudo o que já foi publicado anteriormente. Ou seja, para o leitor uma informação relevante é uma informação diferente – não faria sentido para ninguém a publicação de uma mesma informação em números subseqüentes de um mesmo jornal, a não ser nas seções de anúncios e classificados.

É interessante também o conceito de *mente* de Bateson (2002), numa tentativa de solução do conhecido problema filosófico basilar denominado *mente-corpo*[48], um dos temas centrais em

[48] O problema *mente-corpo* remonta à antiguidade clássica e se refere à possível diferenciação essencial entre a natureza do corpo físico e da mente humana. A corrente dualista, com origem mais provável em Platão, advoga que corpo e mente são coisas essencialmente distintas, onde o corpo pode ser definido em termos de grandezas físicas e a mente não. A corrente monista, esposada por filósofos como Parmênides e Espinoza, e, contemporaneamente, por neurocientistas, assumem que mente e corpo são da mesma natureza física e podem ser explicadas como partes de um todo complexo.

sua obra seminal. Os critérios para se classificar algo como uma mente, para Bateson (2002, pp. 85-86), devem ser os seguintes:[49]

> 1. *Uma mente é um agregado de partes ou componentes que interagem entre si.*
> 2. *A interação entre partes de uma mente é disparada por diferença e a diferença é um fenômeno não-substancial e não localizado no tempo ou espaço; diferença é relativa a negentropia e entropia mais que a energia.*
> 3. *Processo mental requer energia colateral.*
> 4. *Processo mental requer cadeias circulares (ou mais complexas) de determinação.*
> 5. *Em processos mentais, os efeitos de diferença devem ser vistos como transformações (isto é, versões codificadas) de eventos que os precederam. As regras de tais transformações devem ser comparativamente estáveis (isto é, mais estáveis que os conteúdos), mas elas próprias são sujeitas a transformações.*
> 6. *A descrição e classificação desses processos de transformação revelam uma hierarquia de tipos lógicos imanentes no fenômeno.*

A mente, para Bateson (2002, p. 90), é provida de um sistema sensorial (que, para ele, talvez exista também em outras criaturas e plantas) que *pode operar somente com "eventos", que podemos chamar de "mudanças". A não mudança é imperceptível, a não ser que estejamos desejando nos mover no evento.*

O pragmatismo proposto no modelo de engenharia do conhecimento desenvolvido nesta tese encontra suporte filosófico também em Kant, *apud* Bateson (2002), argumentando sobre o que é relevante quando uma mente se depara com uma *diferença que faz diferença*. Bateson (2002, p. 92) sugere, como exemplo, a reflexão sobre um ponto marcado com giz sobre um quadro negro, perguntando-se onde estaria a diferença entre o ponto e o quadro:[50]

> *Kant argumentou, há muito tempo, que essa peça de giz contém um milhão de fatos potenciais (Tatsachen), mas que somente muito poucos deles se tornam verdadeiramente fatos ao afetarem o comportamento de entidades capazes de responder a fatos. Eu substituiria os Tatsachen de Kant por diferenças e esclareceria que o número de potenciais diferenças nesse giz é infinito, mas que muito poucos deles se tornam diferenças efetivas (isto é, itens de informação) no processo mental de qualquer entidade maior.*

Ou seja, as diferenças relevantes estão vinculadas a contextos, ou são diferenças em contextos, onde os seres humanos podem perceber o que lhes é familiar e o que não é, concentrando sua atenção (e estudos mais aprofundados) sobre o que é uma *diferença que faz diferença* (em termos mais comuns, uma novidade que desperte interesse).

#### 3.5.2.3 Abordagem psicológica de Weick e Choo

Weick (1995) declara que a *criação de significado (sensemaking)* é mais um conjunto de ideias em desenvolvimento, com possibilidades explanatórias, do que um corpo de conhecimento.

---

[49] O termo *negentropia*, uma contração de *entropia negativa* (*negative entropy*, ou *negentropy*), também conhecido como *sintropia (syntropy),* é atribuído ao físico Erwin Schrödinger em seu livro *What is Life?* (de 1943), sendo aplicado a contextos de sistemas vivos como a entropia que o sistema exporta para manter sua própria entropia em nível baixo.

[50] Outro exemplo de Bateson (2002, p. 92) se deve ao argumento metafórico de Berkeley: *o fato que acontecer na floresta será sem sentido se o indivíduo não estiver lá para ser afetado pelo mesmo*.

E a idéia central, de fundo psicológico, é apresentada de modo bastante intuitivo: *A criação de significado é testada ao extremo quando as pessoas encontram um evento cuja ocorrência é tão implausível que elas hesitam em reportá-la com medo de não terem crédito* (WEICK, 1995, p. 1).

Essa idéia central permite uma modelagem de Engenharia do Conhecimento para torná-la operacional num ambiente de mineração conceitual da informação, como veremos no Capítulo 7. O psicologismo de Weick (1995) é acessível a partir de nossa experiência mental do dia-a-dia como seres humanos. Ele oferece uma fundamentação bastante sólida para o uso de modelos mentais e linguagem natural para representação de conceitos relevantes nas organizações:

> *As organizações também têm suas próprias linguagens e símbolos que produzem importantes efeitos na criação de significado. A sua relevância no exemplo da BCS é a impressionante diferença entre a frase 'mau trato intencional' e a frase 'criança espancada'. Esta última frase evoca uma imagem gráfica de pais batendo e matando suas crianças. A imagem pode mobilizar ultraje e ação.*
> *O ponto mais geral é que palavras vívidas chamam a atenção para novas possibilidades (...) sugerindo que as organizações com acesso a mais variadas imagens se engajarão na criação do significado mais adaptável que as organizações com vocabulários mais limitados.* (WEICK, 1995, p. 3-4)[51]

É notável a convergência das idéias de Weick (1995) e Bateson (2002) sobre a centralidade do conceito de *diferença* (ou *discrepância*) entre os sinais cognitivos (informação) esperados e observados no processo de criação de significado. Citando a obra de Louis, Weick (1995) argumenta que:

> *Eventos discrepantes, ou surpresas, disparam uma necessidade de explicação, ou de pós-dicção, e, de modo correspondente, uma necessidade de um processo pelo qual interpretações de discrepâncias são desenvolvidas. Interpretação, ou significado, é atribuída a surpresas (...). É crucial notar-se que significado é associado à surpresa como uma saída do processo de criação de sentido mais que emergindo concorrentemente com a percepção ou detecção de diferenças.* (*apud* WEICK, 1995, p. 4-5)

Como eventos discrepantes pode-se deparar, também, com a não ocorrência de algo que era esperado. Quando algo previsto não acontece, o ser humano se encontra diante de algo diferente do padrão em seu modelo mental, que impacta a criação de significado na medida em que exige uma explicação. Weick (1995, p. 45-46) se refere, a propósito, à interrupção de fluxos:

> *(...) embora as pessoas estejam imersas em fluxos, elas raramente são indiferentes ao que passa por elas. Isto é particularmente verdade no caso de interrupção de projetos. A realidade dos fluxos se torna mais aparente quando o fluxo é interrompido. Uma interrupção de um fluxo tipicamente induz uma resposta emocional, que então pavimenta o caminho para a emoção influenciar a criação de conhecimento. (...) Interrupção é um sinal que mudanças importantes têm ocorrido no ambiente. Assim, um evento-chave para a emoção é a 'interrupção de uma expectativa'. E faz um bom sentido evolucionário construir-se um organismo que reage significativamente quando o mundo não é mais do jeito que era.*

Weick (1995, p. 45), citando Berscheid e Mandler, explicam um fenômeno emocional

[51] BCS: *Battered Child Syndrome* (Síndrome da Criança Espancada). Weick (1995) reporta o conhecido processo de criação de significado a partir da observação contínua de certos padrões de lesões corporais em crianças, que levaram um médico nos EUA, na década de 1950, interagindo com técnicos de radiologia, a estudar e teorizar sobre as causas dessas lesões tão similares, mas com explicações insatisfatórias apresentadas pelos pais.

bastante discutido na neurociência em contextos de Inteligência Artificial: o conceito de *disparo* de comportamentos em determinadas situações. Eles argumentam que:

> *(...) uma condição necessária para a emoção é o 'estímulo' ou descarga no sistema nervoso autônomo. E o estímulo é disparado por interrupções da atividade em curso. O estímulo tem um significado fisiológico porque ele prepara as pessoas para reações lute-ou-corra. (...) A percepção de estímulo dispara um ato rudimentar de criação de significado. Ela providencia um aviso que há algum estímulo para o qual deve-se prestar atenção a fim de iniciar uma ação apropriada. Esse sinal sugere que o bem-estar de alguém pode estar em risco.*

Outra situação propícia para criação de significado ocorre quando se foca num detalhe marcante de um todo e esse detalhe sugere implicações importantes para esse todo. Weick (1995, p. 49) cita, neste ponto, outro autor:

> *James (1890/1950, vol. 2, p. 340-343) aponta para a importância dos sinais extraídos para criação de conhecimento em sua discussão acerca dos 'dois grandes pontos de raciocínio'. Os pontos são, 'primeiro, um caracter [sinal] extraído é considerado como equivalente ao inteiro dado do qual ele provem.' (...) O segundo ponto de raciocínio é que o 'caracter extraído' assim considerado sugere uma determinada conseqüência mais obviamente que a sugerida pelo dado total originalmente. (...) O sinal extraído realçou uma implicação distinta que era invisível no objeto indiferenciado.*

O exemplo, nesse caso, é o da pessoa que no processo de avaliação de uma peça de roupa, para comprá-la, argumenta, com base no conhecimento da tintura usada nessa roupa, que a roupa poderá desbotar. Essa peça de roupa (objeto indiferenciado), considerada no todo, não apresenta a informação sobre a tintura problemática, mas a experiência do comprador com essa tintura (sinal extraído) em outra roupa permite um raciocínio com implicações para toda a peça: o provável desbotamento da tintura dessa peça de roupa.

O contexto, portanto, é determinante para o processo de "extração de sinais": primeiro, porque o contexto afeta o que é extraído como um sinal; e, segundo, porque o contexto também afeta o modo como o sinal extraído é interpretado. É importante, também, o modo como os eventos são apresentados no contexto: *se os eventos são noticiados, as pessoas elaboram sentidos para os mesmos; e se eventos não são noticiados, eles não são disponíveis para a criação de significado* (WEICK, 1995, p. 51).

Essas observações empíricas de base psicológica representam uma fundamentação metodológica para mineração de conceitos com base no repertório tradicional da disciplina Recuperação da Informação, na Ciência da Informação, que considera mais relevantes os termos mais freqüentes no texto. Considerando-se *eventos noticiados* como os mais freqüentes (ou "mais noticiados") no texto, na mineração de conceitos pode-se utilizar a estatística como recurso metodológico. E pode-se, também, observar na composição de sintagmas com mais de um substantivo (ou nome) a presença de termos que são novidades no contexto, como sugerido por Koch (1997, p. 125, com grifos nossos):

> *Além de ser uma forma de aprendizagem (e de aprendizagem de línguas em particular), a repetição constitui (...) um meio de criar categorias: itens novos, desconhecidos, podem ser agrupados em categorias lingüísticas e culturais subjacentes, ao lado de itens conhecidos, familiares, quando aparecem em frames repetidos no discurso. Em outras palavras, a repetição permite assimilar o que é novo*

*ao que é já conhecido.*

A obra de Choo (2003) constitui uma síntese contemporânea sobre os processos de criação de significado e construção do conhecimento para tomada de decisão nas organizações, apoiada nos argumentos psicológicos dos autores que o precederam, inclusive Weick (1995). Ele argumenta que *a criação de significado é provocada por uma mudança no ambiente, que produz descontinuidade no fluxo da experiência, envolvendo as pessoas e atividades de uma organização* (CHOO, 2003, p. 21).

O mérito de Choo (2003) se encontra na elaboração de um modelo mental que integra os três processos cognitivos presentes na utilização estratégica da informação num construto holístico unificado. Esses processos, ou "arenas" de uso da informação nas organizações, se referem a: criação de significado, construção de conhecimento e tomada de decisões. Esse construto holístico de Choo é denominado "organização do conhecimento", que confere às organizações uma *especial vantagem, permitindo-lhe agir com inteligência, criatividade e, ocasionalmente, esperteza* (CHOO, 2003, p. 31).

Como se pode observar no trecho a seguir, Choo se apoia nos mesmos fundamentos de Weick (1995) e na ideia central do pensamento de Bateson (2002) sobre o impacto cognitivo das *diferenças* no fluxo da informação:

> *A criação de significado começa quando ocorre alguma mudança no ambiente da organização, provocando perturbações ou variações nos fluxos de experiência e afetando os participantes da empresa. Essa mudança ecológica exige que os membros da organização tentem entender essas diferenças e determinar seu significado. Ao tentar entender o sentido dessas mudanças, um agente dentro da organização pode isolar uma parte das mudanças para um exame mais detalhado.* (CHOO, 2003, p. 33, grifos nossos)

É importante destacar-se, também, as elaborações de Weick (1995) sobre interpretação e criação de significado, que para ele não se confundem. Weick (1995, p. 7-13) advoga que:

> (...) *interpretação literalmente significa uma tradução na qual uma palavra é explicada por outra palavra. (…) O processo de criação de significado é voltado tanto para a inclusão da construção e destaque dos sinais textuais que são interpretados quanto para a revisão dessas interpretações baseada na ação e suas consequências. Criação de significado é como a elaboração de conteúdo em relação à interpretação, e como a criação em relação à descoberta. (…) A distinção chave é que a criação de significado trata dos modos como as pessoas geram o que elas interpretam. (…) Criação de significado se refere claramente a uma atividade ou um processo, enquanto interpretação pode ser um processo mas, provavelmente, apenas para descrever um produto.*

Choo (2003, p. 49), citando Nonaka e Takeuchi (1997), entende a dimensão cognitiva da construção do conhecimento como representável por *esquemas, modelos mentais, crenças e percepções que refletem nossa imagem de realidade (o que é) e nossa visão de futuro (o que deve ser).* Em suma, com Choo (2003, p. 66) tem-se uma elaboração mais completa do processo de construção do conhecimento nas organizações. O autor reconhece, de modo algo poético, que *a informação e o insight nascem no coração e na mente dos indivíduos* e que *a busca e o uso da informação são um processo dinâmico e socialmente desordenado que se desdobra em camadas*

*de contingências cognitivas, emocionais e situacionais.*

Outro aspecto interessante na obra de Choo, pela sua utilidade no contexto desta tese, é que ele reforça os argumentos de Bateson (2002) e Weick (1995) sobre o impacto da informação nova num fluxo (CHOO, 2003, p. 241, grifos nossos):

> *(…) a informação redundante é aquela que transmite algo que o indivíduo já sabe ou reconhece, e é fácil avaliar sua relevância e utilidade. Portanto, a informação redundante pode dar confiança e reduzir a incerteza. A informação nova pode ampliar o conhecimento, mas pode não corresponder à atual estrutura cognitiva do indivíduo, exigindo que ele reconstrua significado e significância.*

Sem perda de conteúdo epistemológico para os objetivos desta tese, pode-se consolidar os aspectos emocionais e situacionais do processo de criação de significado de Choo (2003) no psicologismo cognitivo de Weick (1995). Essa síntese será discutida no Capítulo 6.

Conclui-se, nesta revisão de literatura, que apesar da abundância de abordagens metodológicas para tratamento da informação desenvolvidas na academia e na indústria, não há publicação sobre um construto metodológico não supervisionado[52] para mineração e modelagem semi-automática de conceitos relevantes em textos digitais (em linguagem natural) abertos, com uma abordagem integrada de Inteligência Artificial, para suportar uma *praxis* de Gestão do Conhecimento em contextos de Inteligência Competitiva nas organizações.

Entretanto, os clássicos métodos e técnicas de Recuperação da Informação (RI) continuam sendo úteis nesse tipo de missão crítica, contribuindo para a automação de processos de busca da informação digital em ambientes tão complexos e volumosos como a *Web*.

Os métodos que mais se aproximam da proposta desta tese, como o de Cimiano (2006), apresentam abordagens pouco eficientes para tratamento da ambigüidade e não completam a última fase de um processo padrão de mineração de textos – a da criação de significado. E tampouco se mostram úteis no apoio a processos de construção do conhecimento com a necessária base cognitivista.

Confirmando a visão contemporânea de Ciência da Informação de Bates (1999), é com uma visão integrada do processo de mineração de conceitos que se propõe a metodologia desta tese, articulando-se peças de conhecimento (mono)disciplinar num construto multidisciplinar de *Engenharia do Conhecimento*, como num lego.

[52] O modo de mineração de dados ou textos *supervisionado* pressupõe uma atuação mais intensa do usuário no sentido da indicação dos caminhos a seguir a cada passo na mineração, controlando e "calibrando" o processo de aprendizado do mecanismo de busca. E no modo *não supervisionado* pressupõe-se que o usuário não interage com o mecanismo durante o processo de busca; esse mecanismo deverá, no entanto, com algoritmos apropriados de Inteligência Artificial, aprender a se "autocalibrar" durante o processamento para recuperação dos padrões de informação presentes na base.

# 4. REFERENCIAL TEÓRICO E METODOLOGIA DA PESQUISA

## 4.1 Fundamentação metodológica

A metodologia desenvolvida na tese consiste num processo de reengenharia reversa, como mencionado por Cimiano (2006), sendo proposta sua aplicação num contexto específico – o da modelagem de sistemas de negócios nas organizações competitivas. Em termos filosóficos, essa engenharia reversa se fundamenta na premissa da existência de elos epistêmicos entre o pensamento e a linguagem. O pensamento seria o mundo dos conceitos e modelos e a linguagem o mundo da informação primordial disponível em meio digital.

O construto proposto, além da fenomenologia, também se fundamenta na análise sintática (estruturalista) de Chomsky (1956), em teorias lógico-matemáticas de conjuntos, em teorias de conceitos de Aristóteles (1969; 1995), Peirce (1909; 2010), Sowa (1984) e Wille (1982; 1997; 2005) e, quanto à relevância da informação e criação de significados, na Teoria da Informação de Shannon (SAYOOD, 2000) e nas teses psicológicas de Bateson (2002), Weick (1995) e Choo (2003). Peirce (2010), por exemplo, considera que ciência da Semiótica (ou "ciência dos signos"), que abrange as questões de criação de significado centrais nesta tese, tem três ramos: *gramática especulativa* (ou *gramática pura*), *lógica* e *retórica pura*. Com os experimentos executados, exercita-se este último ramo da Semiótica, sobre o qual Peirce (2010, p. 46) argumenta:

> *O terceiro ramo, imitando a maneira de Kant de preservar velhas associações de palavras ao procurar nomenclatura para novas concepções, denomino 'retórica pura'. Seu objetivo é o de determinar as leis pelas quais, em toda inteligência científica, um signo dá origem a outro signo e, especialmente, um pensamento acarreta outro.*

Como premissa da tese, assume-se a possibilidade de automação, ainda que parcial, do processo de representação gráfica da informação extraída de documentos elaborados em linguagem natural, com base em modelos conceituais de sistemas. Esta premissa comporta, também, a hipótese do poder multidisciplinar da Ciência da Informação ser suficiente para validar a solução metodológica desenvolvida.

É importante, também, esclarecer-se o método de se fazer ciência adotado nesta tese. Como método e técnica, a Engenharia do Conhecimento orienta a execução do experimento e a retrodução aristotélica (teste de uma hipótese). A Teoria da Informação de Shannon (SAYOOD, 2000), apoiada pelas teses de Bateson (2002), Weick (1995) e Choo (2003), confirmam a possibilidade de integração dos métodos dedutivo, indutivo e abdutivo de raciocínio de Peirce (2010, p. 220) no construto elaborado para explicar o processo de criação de significado a partir da mineração de informações conceituais:

> *Abdução é o processo de formação de uma hipótese explanatória. (...) A dedução prova que algo deve ser; a indução mostra que alguma coisa é realmente operativa; a abdução simplesmente sugere que alguma coisa pode ser. Sua única justificativa é que a partir de suas sugestões a dedução pode extrair uma predição que pode ser*

*verificada por indução, e isso, se é que nos é dado aprender algo ou compreender os fenômenos, deve ser realizado através da abdução.*

O início do processo de aplicação da metodologia, conforme a fases elencadas no Quadro 4.1, deve ocorrer com a definição da organização, ou grupo de organizações com alguns atributos em comum, num contexto de estudo em Inteligência Competitiva, passando, na fase seguinte, à seleção e recuperação de documentos textuais digitais que melhor representem a identidade, estrutura, processos de negócio, sistemas e produtos dessa organização ou grupo de organizações.

<table>
<tr><td>Natureza<br>da Atividade e<br>Produto</td><td>Fase 1</td><td>Fase 2</td><td>Fase 3</td><td>Fase 4</td><td>Fase 5</td><td>Fase 6</td></tr>
<tr><td>Análise<br>Sintática</td><td></td><td rowspan="2">Recuperação<br>de Documentos</td><td>Mineração<br>de Textos</td><td rowspan="2">Análise de<br>Conceito<br>Formal</td><td rowspan="2">Modelagem<br>de Sistemas de<br>Negócio</td><td rowspan="2">Inferência<br>&<br>Aprendizado</td></tr>
<tr><td>Análise<br>Semântica</td><td>Definição do<br>Contexto<br>Organizacional</td><td></td></tr>
<tr><td>Produto<br>Parcial</td><td>Objeto de<br>Estudo</td><td>Textos</td><td>Objetos<br>Conceituais</td><td>Contextos<br>e Conceitos<br>Formais</td><td>Modelos de<br>Informação<br>Conceitual</td><td>Insights de<br>Inteligência<br>Competitiva</td></tr>
</table>

Quadro 4.1 Fluxo Operacional Experimental

Na terceira fase do fluxo, submetem-se os textos à mineração de informações, com ênfase no reconhecimento de padrões sintáticos com utilidade semântica no contexto (no caso, sintagmas complexos), e na quarta fase executa-se a análise de conceitos formais no contexto elaborado com os objetos identificados na mineração de texto, obtendo-se contextos e conceitos formais representados no reticulado de conceitos.

A quinta fase é dedicada à seleção e modelagem gráfica dos conceitos que interessam para a modelagem de sistemas de negócio, onde se utilizam técnicas de Análise de Conceito Formal (ACF) e de representação da informação com a Linguagem de Modelagem Unificada (UML). E a sexta e última fase do fluxo de atividades é dedicada ao desenvolvimento de inferências típicas de ambientes de aprendizado com os conceitos modelados na fase anterior, contrastando-os com modelos mentais e conhecimento tácito existente na organização. Esse aprendizado deve ser eminentemente hermenêutico e pragmático, como são os processos de tomada de decisão nas organizações complexas, buscando-se *insights* úteis em contextos de Inteligência Competitiva.

O mecanismo central desse construto de Gestão da Informação e do Conhecimento é a estruturação de conceitos com uso de modelos gráficos a partir de padrões sintáticos textuais. Como observa Winograd (1983), os padrões e classes gramaticais são comuns a qualquer linguagem natural e a liberdade dos falantes de um idioma na criação de sentenças de conteúdo

útil ocorre dentro de limites ou restrições da sintaxe do idioma, restrições que, intuitivamente, poderão ser ainda maiores num contexto.

Com esse elo de representação gráfica da informação, tem-se uma espécie de “interlingua” entre os modelos mentais expressos nos documentos textuais em linguagem natural e os modelos de sistemas de negócio produzidos com a metodologia. As vantagens dessa síntese metodológica se encontram no uso da linguagem natural para modelagem gráfica da informação, sem perda do rigor científico proporcionado pela lógica de primeira ordem e pela teoria matemática de conjuntos, e no tratamento semântico da informação por meio das teorias de conceitos formais e de modelagem ontológica. O próprio método de mineração de conceito proposto contribui para uma redução dramática, senão a eliminação, de problemas semânticos da informação recuperada.

O contexto das organizações competitivas apresenta certos padrões de documentos em linguagem natural que propicia o reconhecimento de estruturas canônicas, que são úteis para a modelagem e padronização de sistemas organizacionais para missões de análise de inteligência e até de sistemas de informação corporativos.

Outros recursos de representação de conceitos proporcionados pelos Gráficos Conceituais, de utilidade para o tratamento semântico da informação, são a “abstração” e a “definição” de objetos e idéias do mundo a partir de “modelos de tipificação”. Sowa (1984) argumenta que verbos como “pensar”, “conhecer” e “acreditar” tomam sentenças completas como seus objetos, conectando-se, necessariamente, a estruturas de gráficos nos nodos de outros gráficos, e que um dos verbos mais importantes nesse sentido é “definir”, na medida em que uma definição pode envolver um gráfico inteiro como um nome (coisa) ou um rótulo de tipo. O autor explica que:

> *Definições podem especificar um tipo de dois modos: estabelecendo condições necessárias e suficientes para o tipo, ou apresentando uns poucos exemplos e afirmando que tudo que for similar a esses exemplos pertencem ao tipo. O primeiro método deriva do método de definição de Aristóteles mediante ‘genus’ e ‘differentiae’, e o segundo é mais próximo de Wittgenstein (1953)* (SOWA, 1984, p. 104).

Sowa (1984) menciona alguns artefatos computacionais desenvolvidos para o tratamento dessas duas abordagens de definição, como construtos de Inteligência Artificial. E ressalta, também, que Gráficos Conceituais suportam tanto definições de tipos por generalização e diferenciação como esquemas e protótipos, que especificam conjuntos ou famílias de estruturas conceituais. Esses métodos são baseados na noção de abstração dos Gráficos Conceituais, que são gráficos canônicos com um ou mais conceitos designados como parâmetros formais.

Observa-se, na Figura 4.1, um exemplo de Gráfico Conceitual representando a definição do tipo conhecido, em linguagem natural, como “beijo” (SOWA, 1984, p. 106). A leitura desse gráfico em linguagem natural, como exercitado anteriormente, é: o tipo BEIJO genérico consiste num tipo específico de TOQUE executado por um agente (AGNT) denominado PESSOA, que tem como parte (PART) constituinte LÁBIOS, onde esse TOQUE é executado de maneira (MODO) SUAVE com o instrumento (INST) LÁBIOS.

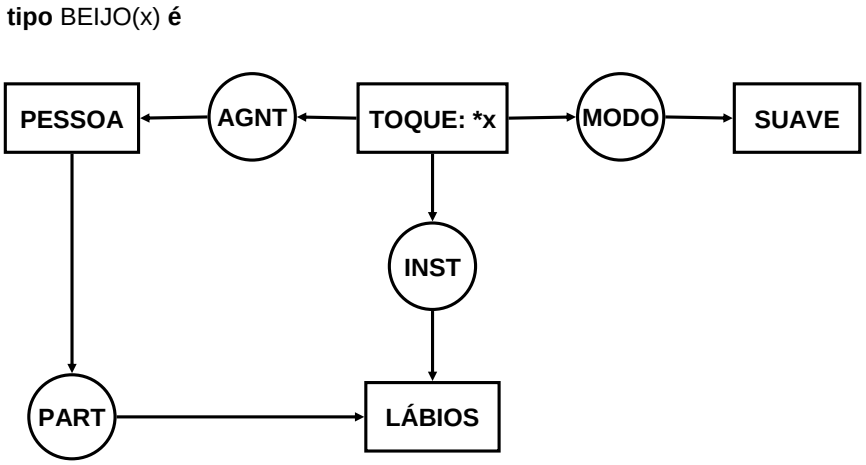



Figura 4.1 Definição de Tipo para BEIJO
(Fonte: SOWA, 1984)

Outros tipos canônicos são ilustrados por Sowa (1984), demonstrando o poder de resolução semântica dos Gráficos Conceituais com estruturas de objetos e relações construídas para representação de abstrações. Esse tipo de recurso é útil para representação de conceitos com referentes mais difíceis de se encontrar em objetos concretos, como é o caso da definição de "serviço". Os Gráficos Conceituais também apresentam recursos importantes para a generalização, especificação, agregação e individualização de conceitos, fundamentais para a teorização científica, motivos pelos quais foram adotados na metodologia.

Os Gráficos Conceituais, no entanto, serão utilizados com parcimônia na metodologia de mineração e modelagem de conceitos propostas nesta tese, com aplicação em situações especiais onde se necessita detalhar, com predicados mais completos, conceitos modelados anteriormente com a UML. Estima-se que sua aplicação se apresente útil no esclarecimento de modelos mentais de conceitos abstratos ou para se elicitar processos e dados para modelagem de sistemas de informação, como no exemplo da Figura 1.3.

O construto epistemológico sintetizado no processo do Quadro 4.1 propõe, portanto, um processo de Gestão da Informação útil para Gestão do Conhecimento em contextos organizacionais de Inteligência Competitiva.

## 4.2 Metodologia experimental

O objetivo da experimentação é simular um processo de extração de informações conceituais de documentos textuais digitais, como ilustrado no Quadro 4.1, para modelagem de sistemas organizacionais, numa abordagem de engenharia reversa da linguagem natural para modelos mentais e estruturas conceituais que possam representar o conhecimento sistêmico das organizações. Com essa abordagem, pretende-se mostrar, experimentalmente, como se poderá

utilizar o crescente volume de informações digitais disponíveis das e sobre as organizações para se avançar na automação dos processos de remodelagem de sistemas e reengenharia, com uso intensivo da informação e suas tecnologias.

### 4.2.1 Simulação computacional

Os métodos e técnicas de extração e mapeamento da informação e do conhecimento organizacional que emulam características do cérebro humano – Inteligência Artificial (IA) – para instrumentação eficiente das atividades voltadas para a gestão da informação e do conhecimento são objeto de investigação na pesquisa proposta. Contudo, será necessária a definição de um método de modelagem operacional do problema e das possíveis soluções no contexto laboratorial da pesquisa, assunto discutido a seguir.

Os dados a serem coletados no desenvolvimento da pesquisa deverão resultar do processamento do modelo de mineração de informações em ambiente de simulações baseado em computador. Com simulações baseadas em computador, pretende-se uma aproximação experimental do mundo real das organizações e suas fontes de informações.

Os referenciais teóricos apresentados em publicações sobre uso de técnicas de simulação em pesquisas na área de ciências sociais são um tanto escassos na literatura, caracterizando-se uma tendência histórica de produção acadêmica mais intensa desse tipo de recurso metodológico nas ciências exatas, mais especificamente nas engenharias. Encontra-se, contudo, organizações acadêmicas onde esse método é mais popular, com linhas de pesquisa dedicadas, por exemplo, na sociologia, tais como o da Universidade de Surrey, no Reino Unido (GILBERT, 2007). E Pidd (1988) argumenta que métodos de simulação computacional têm se desenvolvido desde o início dos anos 1960, onde se incluem, talvez como as ferramentas mais comumente utilizadas para tanto, as utilizadas na ciência da Administração.

Em suma, o engenheiro do conhecimento que utiliza simulação computacional desenvolve um modelo do sistema de interesse, que imita a complexidade do problema real, escreve um programa de computador que implementa esse modelo e utiliza um computador para emular o comportamento do sistema quando submetido a uma variedade de políticas (ou situações) operacionais (PIDD, 1988). Como resultados observados das simulações, o engenheiro do conhecimento terá então dados para análise e condições de compor um modelo completo do problema e da melhor solução possível, comparando, inclusive, as implicações de adoção de uma solução ou outra em termos de desempenho.

Conforme Pidd, o método que utiliza simulação pode ser considerado um método experimental na medida em que testa um modelo conceitual de problema e solução. Esse autor argumenta, no contexto da simulação computacional na Administração, que:

> *O modelo é utilizado como um veículo para experimentação, freqüentemente com "tentativa e erro", para demonstrar os prováveis efeitos de várias políticas. Então,*

> *aquelas que produzem os melhores resultados no modelo seriam implementadas no sistema real* (Pidd, 1988, p. 5-6).

A Figura 4.2 apresenta a idéia básica da simulação computacional como método experimental de pesquisa científica, semelhante ao modelo primitivo "caixa-preta" da Teoria Geral de Sistemas.

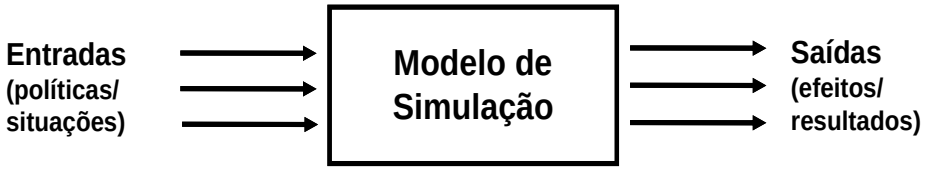



Figura 4.2 Simulação como Método Experimental
(Fonte: do autor da tese)

Os experimentos com simulação computacional podem ser sofisticados, envolvendo uso de técnicas estatísticas se diferentes efeitos podem ser produzidos, por exemplo, como resultado da interação de várias políticas, quando então se deseja analisar os impactos cruzados entre elas. E podem também ser bastante simples, tomando a forma de questões "e se?", onde as simulações se destinam a produzir efeitos perceptíveis no modelo com base em interações conceitualmente mecanicistas de causa e efeito.

Pidd (1988) observa, contudo, que a simulação computacional como método experimental de pesquisa consome tempo e, geralmente, é uma tarefa bastante complexa, mas apresenta também as seguintes vantagens:

a) custo: em geral é menor que um experimento real nas organizações;
b) tempo: após a modelagem e a construção do artefato computacional, as possibilidades de emulação de situações reais com simulação são ilimitadas, podendo-se também simular períodos maiores de operações reais em questão de segundos no computador;
c) possibilidade de replicação: um experimento simulado pode ser replicado por outros pesquisadores, uma vez que as condições ambientais são modeladas e estruturadas de modo inteligível, algo difícil de ocorrer num experimento real, pois as organizações e os ambientes reais geralmente não se repetem naturalmente de modo muito similar;
d) segurança: uma simulação geralmente não apresenta riscos para as organizações ou pessoas, o que nem sempre ocorre em experimentos reais (experiências reais mal-sucedidas podem acarretar danos ambientais e/ou prejuízos para as pessoas e organizações).

### 4.2.2 Aplicação do método na pesquisa

O uso de simulação computacional na pesquisa que embasa a tese apresenta todas essas vantagens, superando-se o problema do desenvolvimento de *softwares* complexos para modelagem do ambiente de simulação utilizando-se *softwares* existentes e disponíveis, especialmente produtos *shareware* e *freeware*, com custos muito reduzidos ou isentos de custos de licenciamento. Com tais produtos, o tempo para implementação do modelo no ambiente computacional é reduzido, ajustando-se às necessidades do projeto.

A Figura 4.3 mostra, num modelo de atividade da UML, as ações da pesquisa, partindo da concepção do projeto, revisão de literatura e avaliação de teorias, e em que ponto do trabalho de pesquisa é integrado o ambiente de simulação computacional. Os experimentos laboratoriais previstos no fluxo de atividades para execução (ação "Executar Experimentos") com uso de simulação computacional aparecem após a seleção e montagem do modelo experimental na fase anterior (na ação de "Selecionar Estratégias").

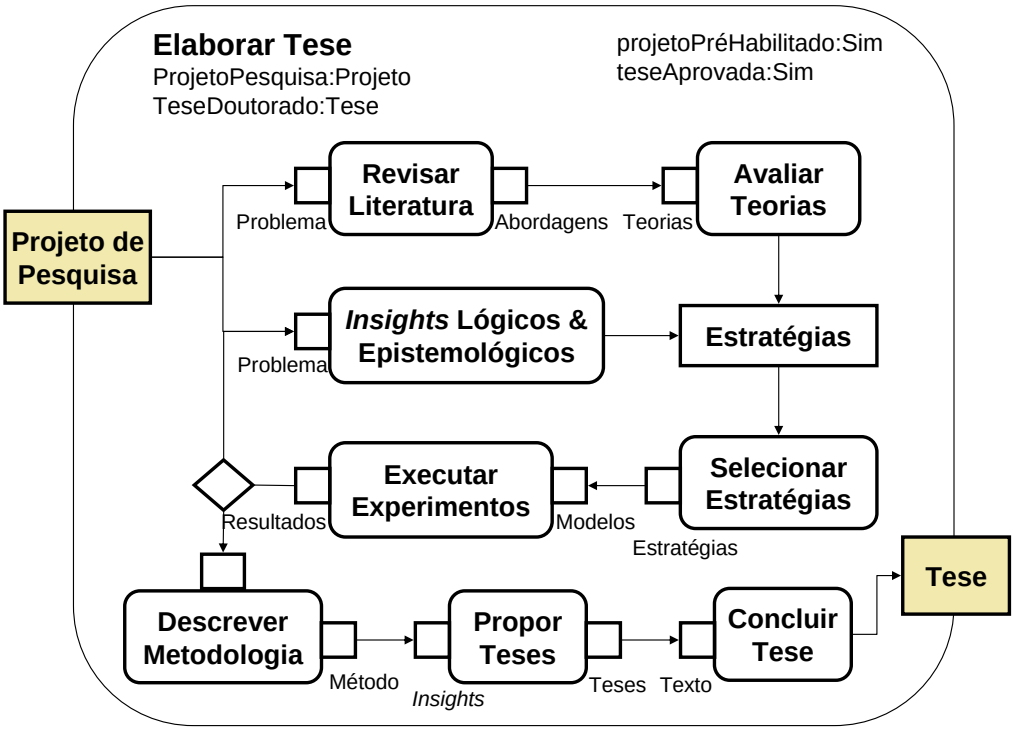



Figura 4.3 Fluxo Lógico da Pesquisa
(Fonte: do autor da tese)

Como sugere o fluxograma da Figura 4.3, as fases de desenvolvimento conceitual do modelo experimental resultam dos estudos nas fases iniciais do projeto, com avaliação de teorias e elicitação da melhor (ou das melhores) estratégias e modelos de experimentação.

A ação “Executar Experimentos*”* corresponde à pesquisa laboratorial com simulação computacional, onde os textos selecionados serão submetidos à mineração de informações para reconhecimento de padrões sintáticos. A mineração de textos será processada com *softwares* aplicativos de Processamento da Linguagem Natural disponíveis.

A ação seguinte se concentrará na descrição da metodologia considerada exitosa, que será apresentada como suporte das teses defendidas. Em “Propor Teses”, com base nos resultados da pesquisa, serão propostas teorias que sintetizam a estrutura *(framework)* conceitual do processo de Gestão da Informação e do Conhecimento implícito, concluindo-se o trabalho com uma análise de desempenho da metodologia experimentada face aos objetivos da tese.

### 4.3 Procedimentos gerais

Os procedimentos de simulação detalham as fases de modelagem, programação e experimentação categorizadas por Pidd (1988). Os procedimentos gerais executados em cada fase são apresentados a seguir.

A fase de modelagem da simulação exigiu pesquisa na literatura em busca de estudos de caso em que se utilizaram técnicas promissoras no contexto, além do reestudo de alguns experimentos com metodologias e tecnologias conhecidas pelo pesquisador. O uso de estratégias e metodologias mistas, e até do uso de dicionários eletrônicos de sinônimos e tesauros ontológicos, foram avaliados nessa fase, mas descartados da fase laboratorial do desenvolvimento da pesquisa, pela sua inadequação (assumiu-se que não se exigiria o concurso de estruturas ontológicas *a priori* da mineração de texto).

Como ressaltado anteriormente, na fase de laboratório selecionou-se alguns *softwares* para implementação do modelo de experimentação. Esse processo seletivo teve como critérios básicos sua adequação aos requisitos do projeto experimental e os custos de instalação e operação envolvidos, com base no conhecimento acumulado pelo pesquisador.

Na fase de experimentação foram executadas as operações computacionais de apoio aos mecanismos de testes da metodologia elaborada. É nessa fase que foram testados os limites da capacidade explicativa do modelo computacional simulado, buscando-se respostas para as questões que motivaram o desenvolvimento da pesquisa. Como exemplo, utilizando-se a vantagem da replicação do método de simulação computacional para repetir o processamento do modelo com instâncias variáveis dos atributos das organizações simuladas, testou-se uma situação de simulação do tipo “e-se?”, buscando-se conhecer o impacto conceitual de um incremento de informação na base textual em tempo futuro – ou seja, alterando-se os valores das variáveis de entrada do modelo para se observar quais impactos produzem essas alterações nos resultados (saídas) do modelo.[53]

[53] Experiências de simulação para responder questões estilo *e-se*? *(what-if?)*.

Com o modelo experimental proposto viabilizou-se, portanto, a simulação de uma organização com suas fontes de informação digital não-estruturadas sobre o negócio e a descoberta de conceitos ocultos nos conteúdos dessas fontes de informação, além de testar os efeitos das mudanças nos ambientes dessa organização com base nas mudanças de contexto e conteúdos nas fontes de informação. E assim evidenciou-se, ainda que numa base experimental, o potencial do arsenal de Inteligência Artificial (IA) embarcado em *softwares* de Processamento da Linguagem Natural (PLN), mostrando sua utilidade nos processos de Recuperação da Informação (RI) em contextos instrumentais de gestão da informação e do conhecimento para suporte ao desenvolvimento de inteligência competitiva nas organizações.

O tema conceitual é atual e Choo (2007) o aborda, do ponto de vista dos aspectos epistemológicos inerentes à recuperação de informação em ambientes organizacionais, num texto bastante recente. Brachman e McGuinness (1988) também ilustram como conectar os conceitos de conexionismo (relativo às redes neurais) como instrumento técnico para recuperação conceitual de informações em contextos de gestão do conhecimento, idéia alinhada com o desenvolvimento de ferramentas tecnológicas para apoiar processos de inteligência competitiva nas organizações.

# 5. DADOS, EXPERIMENTOS E RESULTADOS

*Carnudos substantivos; verbos esguios*.[54]
(LISPECTOR, 1998)

Com este capítulo, são apresentados os dados utilizados na pesquisa[55], os experimentos de laboratório executados e os resultados observados do ponto de vista da mineração conceitual da informação. Entretanto, como comentado na introdução a esta tese, defende-se que o modelo-padrão de apresentação de um texto de tese desta natureza deve ser flexibilizado para que não se caia na tentação de tentar condensar todos os conteúdos científicos de base, que abrangem várias disciplinas em vários níveis de profundidade e utilidade, antes de sua contextualização empírica. Caso os comentários mais especializados de contextos de Inteligência Competitiva fossem inseridos nos capítulos introdutórios do texto, existiria o risco de tornar a compreensão de sua importância *a priori* algo confuso para o leitor, motivo pelo qual optou-se por apresentá-los neste capítulo – do contrário, poderia se chegar à conclusão inicial que os tópicos científicos abordados constituiriam uma mera "sopa de disciplinas" sem muito sentido prático.

O texto, deste ponto em diante, mostra os dados e sua interpretação em contextos, ou seja, tenta associar seus significados em ambientes de negócios caracterizados pela competição. Esta abordagem, tão genérica quanto possível, é também utilizada na otimização de modelos com variáveis quantitativas, onde se busca descobrir o melhor modelo explicativo possível de um fenômeno mensurável a partir de dados de contorno para solução do problema, ou de "variáveis-alvo" adotadas como parâmetros para otimização do modelo.

Entretanto, uma questão primordial deve ser explicitada antes dessa sondagem no mundo real das informações digitais abertas *das* e *sobre as* organizações: "como reconhecer as informações essenciais para Gestão do Conhecimento em contextos de Inteligência Competitiva?"

Fuld (2007, p. 15-16) apresenta a seguinte metáfora como pista para uma abordagem metodológica dessa questão:

> *Não existe melhor maneira de compreender a perspectiva competitiva do que visitar um museu de arte. Preste atenção nas pinturas dos impressionistas. A escola dos impressionistas, conhecida como pontilhismo, vigorosa no final do Século XIX, reproduz cenas completas com milhares de pontos coloridos em vez de linhas contínuas. Olhe de perto para qualquer uma dessas pinturas e verá somente pontos. Afaste-se uns dez passos e verá um campo de flores ou pessoas passeando em um parque com guarda-sóis abertos.*
>
> *(...) Quando você vê toda a pintura, você fica maravilhado como milhares de pontos formaram uma figura coerente e de fácil compreensão. Você percebe o barco a vela*

[54] "É claro que, como todo escritor, tenho a tentação de usar termos suculentos: conheço adjetivos esplendorosos, carnudos substantivos e verbos tão esguios que atravessam agudos o ar em vias de ação, já que palavra é ação, concordais?" (LISPECTOR, 1998, p. 15) Esses comentários aparecem na tela de interface de comando do *software* de análise sintática *WordSmith*.
[55] Coletados nos meses de junho e julho de 2009.

> *balançar com a brisa, o remador, árvores pendendo sobre a margem do rio. Se você examinar a imagem novamente, você começará a perceber mais aspectos sutis, os contornos, as formas geométricas, as proporções, uma torre escondida atrás de algumas árvores ao longe. Você consegue ter perspectiva.*[56]

Essa abordagem de Fuld (2007) apresenta certa semelhança conceitual com um dos três princípios da Inteligência Artificial baseada na natureza: o da "inteligência emergente*"*. Conforme Freedman (1995), esse princípio assume que a inteligência pode emergir da interação complexa entre elementos mais simples, como no caso dos pontinhos da pintura que, de uma perspectiva adequada, mostra ao observador uma rica paisagem impressionista. Outros exemplos similares desse fenômeno cognitivo podem ser: o mestre de xadrez que, numa partida, enxerga um cenário vantajoso do seu jogo numa certa seqüência de jogadas à frente (enquanto, no presente, nada mais se pode observar que uma posição estática pouco alentadora no tabuleiro); o astrônomo que "lê" um mapa celeste a partir do conjunto das posições de algumas estrelas-chave isoladas; uma seqüência de *bits* que codifica uma palavra em linguagem natural.[57]

Outra parte da argumentação de Fuld (2007, p. 16) esclarece algo sobre o que procurar nas fontes de informação digitais e, talvez, também sobre o "como" procurar as informações essenciais em textos:

> *Desenvolver inteligência competitiva é semelhante a criar uma pintura pontilhada. Seu objetivo não é criar uma imagem perfeita, mas uma imagem representativa da realidade (como Seurat fez ao criar sua cena serena com o mínimo de pontilhados). Afaste-se alguns passos da pintura de Seurat e você sentirá como se estivesse sentado na margem do rio naquele dia quente e vagaroso de verão.*

## 5.1 Estratégia de recuperação da informação

Com inspiração na metáfora impressionista de Fuld (2007) e na estratégia de uso de sintagmas nominais como indexadores e palavras-chave para recuperação da informação textual desenvolvida por Capuano (2009), baseada nas idéias de Gottschalg-Duque (2005), adotou-se uma estrutura padrão textual de busca, nos textos digitais selecionados, composta por sintagmas plurinominais. Esses sintagmas complexos são formados pela composição de mais de um substantivo (ou nome) que se encontram próximos numa sentença sem intercalação de verbos, indicando uma expressão com maior teor semântico que um substantivo/nome isolado.

A busca seletiva por sintagmas com esse padrão de composição léxica se justifica na medida em que o objetivo da mineração textual, no contexto epistemológico desta tese, é primeiro conhecer-se os conteúdos informacionais disponíveis sobre as organizações e seus negócios, ou seja, conhecer-se "sobre o que" estão falando. E, em segundo plano, "como" estão falando sobre

[56] O autor se refere, como exemplo, a uma pintura do artista Georges Seurat denominada *The Seine at the Grande Jatte*, exposta no *Royal Museum of Fine Arts* de Bruxelas, na Bélgica.
[57] As imagens digitais codificadas em formato de *bitmaps* (mapas de *bits*) também são exemplos correlatos.

esses conteúdos, questão que se insere numa segunda etapa de aplicação da metodologia, quando se busca detalhes dos temas apresentados como essenciais nas fontes disponíveis.[58]

Esta estratégia de mineração de sintagmas complexos apresenta três vantagens técnicas evidentes, motivos de sua escolha para mineração de textos nesta tese:

I. Elimina, quase completamente, o problema da ambigüidade na recuperação da informação, pois uma expressão sintagmática tende a apresentar, naturalmente, um maior poder de resolução semântica que um termo isolado (Figura 5.1).

II. Utiliza como termos de busca os substantivos, que são os componentes sintáticos representativos das "coisas do mundo", numa visão ontológica de contexto (ou "a expressão mais simples dos conceitos em linguagem natural").

III. A mineração de texto com sintagmas compostos desse modo são estruturas-padrão da linguagem natural que podem ser facilmente modeladas em como argumentos de *queries* (perguntas) com *softwares* especializados em funções de busca sintática (como a busca de estruturas sintáticas em modo *collocation*[59]).

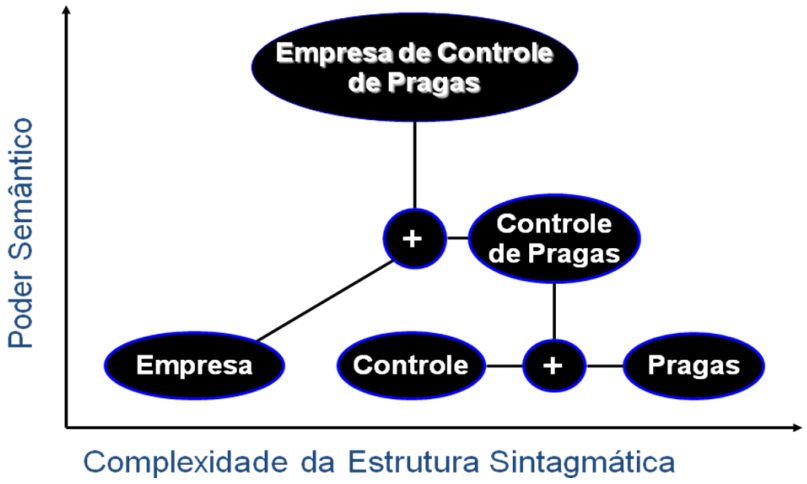



Figura 5.1 Poder Semântico dos Sintagmas Plurinominais
(Fonte: do autor da tese)

Quando se utiliza como termo de busca, ou se recupera, no exemplo da Figura 5.1, a palavra "empresa", pode restar a necessidade de qualificação dessa empresa ou de sua área de negócio no mercado. O mesmo ocorre com as palavras "controle" e "pragas", que unidas pela preposição "de" formam a expressão "controle de pragas", que qualifica esse "controle" como "de

[58] É interessante observar-se, neste ponto, o conceito de "informação" encontrado em Peirce (2010, p. 107): *(...) "a soma das proposições sintéticas na quais o símbolo é sujeito ou predicado", antecedente ou consequente.*

[59] "*Collocates* são palavras que ocorrem na vizinhança de sua palavra utilizada como argumento de busca. *Collocates* de carta podem incluir correio, selo, envelope, etc. No entanto, palavras muito comuns como 'a' podem também 'colocar' com carta (SCOTT, 2008, p. 94). *Collocation*, portanto, é um modo de mineração de texto em que se define *a priori*, como argumentos de busca, um conjunto de palavras ou padrões sintáticos e as respectivas posições (colocações) de seus componentes na frase.

pragas", um tipo específico de atividade de controle, constituindo aquilo que na gramática se denomina "locução com substantivo adjetivado". E quando, finalmente, compõe-se a expressão sintagmática mais completa "empresa de controle de pragas", resta pouca ambigüidade quanto ao tipo de empresa da qual se trata, ou aos significados das palavras "controle" e "pragas" no contexto da informação.

O problema da ambigüidade na recuperação da informação textual, desdobrado na polissemia e na sinonímia, tem atormentado os estudiosos da recuperação da informação por décadas, sem soluções triviais (KONCHADY, 2006). Meadows *et alii* (2007), por exemplo, colocam o problema da seguinte forma:

> *(...) enquanto identificadores únicos são utilizados para algumas aplicações e indicadores de classificação, em outros casos há incerteza sobre valores ou significados, causando confusão ao leitor humano ou a um programa de computador, ou ambos. E uma fonte de ambigüidade é a semântica – o significado dos símbolos.*

Outro importante fator que contribui para a redução dos problemas semânticos em textos "de" e "sobre uma" organização é o contexto de negócio em que ela se insere, que geralmente tem ontologias e metadados conhecidos, ainda que tacitamente, pelos seus colaboradores, parceiros e clientes.

Os sintagmas nominais serão, portanto, os "pontilhados impressionistas" de Fuld (2007), que organizados de modo inteligente poderão auxiliar os práticos da Inteligência Competitiva a reconhecer conceitos-chave de negócio em questão e compor cenários para o desenvolvimento de *insights* úteis para sua organização. Como demonstrado com os resultados de laboratório, no Capítulo 5, outra vantagem decisiva para a busca de sintagmas com mais de um nome, ou plurinominais, como estrutura sintática padrão na mineração de textos, é sua eficiência no sentido de representar os substantivos (ou nomes) mais freqüentes em textos desse tipo de fonte.

O que se pretende, com isso, é propor uma metodologia de organização de conceitos expressos em linguagem natural, com uso de Análise de Conceito Formal, Linguagem de Modelagem Unificada e Gráficos Conceituais para representação da informação, de modo a propiciar a identificação de significados nos próprios conceitos lingüísticos resultantes da mineração textual e nas relações entre os mesmos (Figura 5.2).

O ser humano elabora conceitos mais complexos apoiados em conceitos mais simples e concretos (no sentido físico), mais próximos da "substância" primordial e seus atributos da classificação ontológica (metafísica) de Aristóteles. Obviamente, a pilha de conceitos sugerida na Figura 5.2 é uma abstração da realidade, sugerindo que essa construção de conceitos ocorre como uma pirâmide; na realidade, o que se observam são conceitos que se relacionam em rede, ainda que exista uma hierarquia de significados, onde conceitos mais complexos envolvem conceitos mais simples na sua composição etimológica e epistemológica.

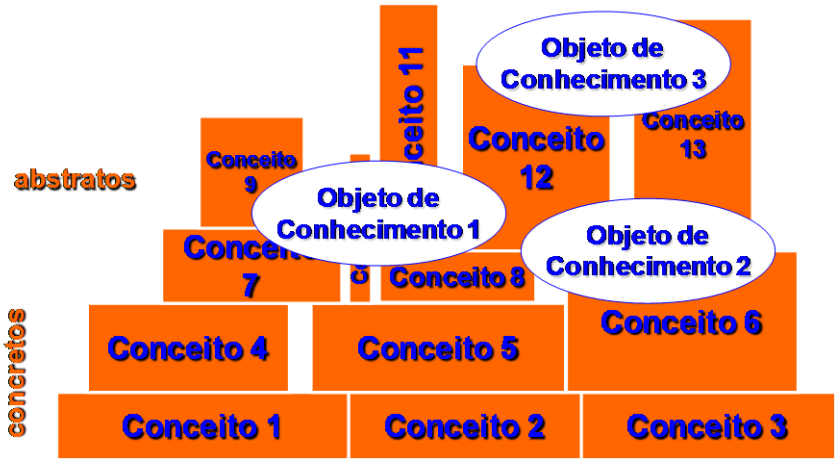



Figura 5.2 Pilha de Conceitos

(Fonte: do autor da tese)

Com esse modelo ontológico, espera-se que os objetos de conhecimento, ou "bens de conhecimento" (SCHREIBER *et alli*, 2000) mais preciosos se encontrem nas conexões (ou relacionamentos) entre os conceitos. Sowa (1984, p. 76) argumenta, a respeito, que:

> *Conceitos abstratos adquirem significados não mediante associações diretas com percepções, mas mediante uma vasta rede de relacionamentos que, no final, os conectam a conceitos concretos.*

Outras estratégicas de busca da informação, com mineração textual, para construção e população de ontologias, como a proposta por Cimiano (2006), utilizam sintagmas compostos por substantivos e verbos correlatos, com a vantagem de facilitar sua classificação posterior em contextos, mas com a desvantagem da necessidade de formação de atributos adjetivados a partir dos verbos. Considera-se essa estratégia particularmente complexa tanto no processo de transformação de verbos em adjetivos como em sua desambiguação posterior, pois pode-se chegar à conclusão que alguns atributos têm mais de um significado – problema de polissemia, ou que alguns atributos podem ser fundidos devidos à sinonímia.

Estratégias dessa natureza, com uso de verbos, apresentam outras desvantagens, pois sua representação posterior também será, naturalmente, mais complexa, pelos seus aspectos de estrutura predicativa. O uso de Gráficos Conceituais (SOWA, 1984) seria uma solução, mas alguns experimentos de laboratório nessa direção realizados no desenvolvimento desta tese têm mostrado a inviabilidade de representação visual de conteúdos de textos mais extensos que uma ou duas páginas de um sítio na *World Wide Web*, por exemplo.

Contudo, esse tipo de representação da informação textual com substantivos/nomes e verbos será discutido mais adiante, nesta tese, nos resultados da mineração conceitual (Capítulo 6), como recurso adicional ao método da Análise de Conceito Formal.

## 5.2 Fontes de informação

Considerando que a Inteligência Competitiva se propõe ao estudo e utilização de informações abertas, de acesso público (portanto lícito), para alcançar seus objetivos nos ambientes das organizações, a identificação, avaliação e seleção de fontes de informação são atividades primordiais a serem desenvolvidas em seus processos constitutivos. O tema "fontes de informação", em particular, representa uma área de pesquisa com significativa representatividade na Ciência da Informação, reportando-se aos fundamentos epistemológicos dessa área do conhecimento.

### 5.2.1 Era pré-*Internet*

Fuld (2007, p. 127-128) relata sua experiência de busca da informação de negócios na era pré-*Internet* mostrando como esse tipo de informação se encontrava disponível. Em relação à sua particular necessidade de informação:

> *(...) o conceito de Internet não foi criado nos anos 90 com o lançamento da World Wide Web e dos browsers que tornaram a rede acessível. Esta nasceu em 4 de junho de 1927, em Boston, Massachussetts, com a inauguração da Baker Library no novo campus da Harvard Business School, localizado ao lado do rio de Cambridge.*
>
> *George Fisher Baker, o banqueiro de Nova York de 87 anos, doou US$ 5 milhões para sua construção. Ele acreditava que essa biblioteca de negócios, que foi uma das primeiras no ramo, estava destinada a tornar-se a biblioteca de negócios mais proeminente, uma fonte central de informações. No momento da dedicatória, ele explicou por que decidiu doar US$ 5 milhões, enquanto os representantes de Harvard haviam solicitado US$ 1 milhão: 'Minha vida tem sido dedicada aos negócios e gostaria de fornecer um novo começo para melhores padrões de negócios'.*

Essa noção de "padrões de negócios" de Baker constitui um dos fundamentos dos modelos analíticos construtivos da Inteligência Competitiva de Fuld (2007) e da própria mineração de dados e de textos. O reconhecimento desses padrões, com efeito, compõe o que se pode chamar "o coração de diversas metodologias de mineração", tais como as do raciocínio baseado em casos (*Case-based Reasoning* – CBR) e de redes neurais artificiais. E pode-se afirmar, também, que o reconhecimento de padrões se encontra no cerne da pesquisa científica, uma vez que a elaboração de modelos para explicação dos fenômenos necessita do conhecimento dos padrões observados de manifestação desses fenômenos.[60]

O autor relata, sobre sua experiência na *Baker Library* (no início dos anos 1980), que abundavam informações de negócios de todo tipo nessa biblioteca, como pesquisas de investimentos, manuais *Moody* e conteúdos da *Encyclopedia of Associations*, *Standard & Poor's* e do *Register's*. Outras fontes pré-*Internet* citadas por Fuld (2007) são *Dow Jones & Reuters, Dialog, Lexis-Nexis* e o *Wall Street Journal*.

[60] A noção de "padrão", na Ciência, está associada à necessidade de "generalização".

Outra das mais conhecidas fontes pré-*Internet* é a *Bloomberg*, empresa fundada em 1981 com a seguinte visão de negócios:

> *(...) criar um serviço de informação e uma empresa de mídia e notícias que disponibilizasse aos profissionais de negócios e de finanças as ferramentas e os dados que eles necessitam, numa única e abrangente plataforma.* (ver: *http://about.bloomberg.com/company.html*)

Deve-se ressaltar, no entanto, que mesmo antes do advento da *Internet* e da *World Wide Web* existiam fontes de informação de qualidade acessíveis por redes de computadores dedicadas. Fuld (2007, p. 156) contextualiza essas fontes e sua evolução mais recente esclarecendo que:

> *Embora os serviços de informações on-line cobrados tenham sido marginalizados nessa era de informação grátis, alguns têm durado desde o final dos anos 60 e ainda são superiores para informar na Web de graça, particularmente em relação ao índice de notícias tradicionais e de bases de dados científicas. Sistemas de serviço de informação como o Dow Jones & Reuters Factiva, o Dialog ou o Lexis-Nexis têm supervisão editorial e permitem ao pesquisador mais controle e confiança de que ele ou ela encontrará o correto pedaço de informação. No caso do Factiva, todo o sistema é indexado, o que significa que os editores literalmente atribuem índice de termos para cada e todo artigo ou registro.*
>
> *Factiva, Lexis-Nexis, Dialog e outros competidores on-line têm continuamente improvisado a capacidade analítica e de busca de seus arquivos. O Factiva inclui as informações eletrônicas da Reuters contidas no Wall Street Journal, tanto quanto uma grande variedade de notícias, artigos e outros conteúdos, contendo aproximadamente nove mil fontes com centenas de milhões de artigos. O Lexis-Nexis também declara possuir um grande arquivo: nesse caso, de 4,1 bilhões de documentos armazenados on-line.*

Este autor cita, ainda, declarações de Hart, CEO[61] da Factiva, sobre o produto da empresa (FULD, 2007, p. 157):

> *A Factiva fornece tecnologia que o ajuda a rastrear uma grande massa de dados utilizando o inteligente índice da Factiva, tanto quanto seus 300 mil códigos de empresas, seus 372 códigos geográficos, seus 426 códigos de assunto, 735 códigos industriais, (...). Podemos trabalhar com empresas para adicionar códigos à taxonomia, obtendo acesso para os artigos mais relevantes, o que significa que você deve prestar atenção à taxonomia.*

Em países de economia mais desenvolvida como os EUA, existe também um grande número de *Web* Portais governamentais, como o EDGAR *(Electronic Data Gathering Analysis and Retrieval) Online*, da *Securities and Exchange Commission (SEC)*[62], destinados a publicarem informações de interesse público. O EDGAR surgiu na década de 1970, nessa época utilizando papel e microfilmes, com objetivo de publicar informações sobre as empresas de capital aberto nos EUA, para evitar um novo *crash* financeiro como o da Bolsa de Valores de 1929.[63]

[61] CEO: *Chief Executive Officer* (Executivo-Chefe da Empresa).
[62] Tradução aproximada: EDGAR – Analise e Recuperação de Coleções de Dados Eletrônicos; SEC – Comissão de Seguros e de Câmbio.
[63] Obviamente, com a nova quebra do sistema financeiro em 2009, questiona-se a eficácia dessa medida.

### 5.2.2 Evolução da *Web*

Com o advento da *Internet* e da *World Wide Web* (conhecida pelo acrônimo WWW ou, simplesmente, *Web*), no início dos anos 1990, o cenário desse "mercado de informação" de negócios mudou completamente, com impactos espetaculares tanto do lado da demanda como da oferta. Em termos de processos produtivos, o uso cada vez mais intensivo de TIC (Tecnologias de Informação e Comunicação) resultou num novo paradigma de mercado da informação, onde os custos são dramaticamente reduzidos em relação ao paradigma sociotécnico imediatamente anterior, baseado preponderantemente no papel, como mídia de comunicação escrita, e na apresentação oral, como mídia de massa (nas campanhas de *marketing* pela televisão).

O primeiro resultado sensível dessa revolução sociotécnica é a "sobrecarga"[64] ou "inundação" da informação em todo o mundo ocidental, ainda atraente objeto de estudo em relação aos seus efeitos sociais e econômicos de longo prazo. O uso cada vez mais intensivo de avançados sistemas de comunicação digital pelas populações em geral, como correio eletrônico e telefone sem fio (com tecnologia *celular*), transformou o computador doméstico e o telefone portátil em equipamentos eletrodomésticos incorporados aos hábitos de consumo de pessoas comuns (trabalhadores em geral, estudantes, crianças e adolescentes, etc). Concomitantemente, as organizações acompanharam esses movimentos construindo sua "imagem digital" na WWW, disponibilizando cada vez mais completos e sofisticados sítios ou portais de informação corporativa acessíveis por sua clientela e parceiros por meio da *Internet*.

Autores como Davydov (2001, p. 25), por exemplo, definem esse inovador ambiente de informação em rede como "ciberespaço", com os seguintes significados:

> *Eu defino ciberespaço como um ecossistema conceitual baseado na informação onde todos os dados ("informação potencial") são armazenados. É também o ambiente de infra-estrutura onde todos os trabalhadores da informação existem e operam. Se aceitarmos a noção que ciberespaço é um ecossistema, então para melhor compreender seus princípios podemos nos voltar para os modelos sugeridos pela Teoria Geral de Sistemas, que é a cibernética.*
>
> *A cibernética nos fala que informação é "energia" que necessita ser consumida e processada por um sistema para que ele sobreviva, reproduza seus elementos chave e, como parte de um ecossistema maior, contribua com aquele sistema de modo a ampliar outros sistemas ou membros do ecossistema em geral.*

Estima-se que atualmente a *Web* com informação aberta congregue um bilhão de usuários no mundo e 250 milhões de sítios, crescendo a um ritmo anual de 30%. E que a *Web* invisível, não acessível ao público em geral pela *Internet*, contenha 500 vezes mais informação que a *Web* aberta (QUONIAM, 2010). Essa percepção de tendência é corroborada, aproximadamente, pelos dados de Fuld (2007, p. 156) de anos anteriores: utilizando-se o *Google* como referência, este indexou mais de 8 bilhões de páginas da *Web* até 2004, e a *Internet Society of Virginia* publicou um artigo estimando em 100 milhões o número de sítios nessa época.

[64] *Overload*, no original em inglês.

Considerando-se que esses sítios pertencem, na sua esmagadora maioria, a organizações, pode-se concluir que a *Web* é o maior repositório de informação aberta sobre as organizações que existe. Em termos de qualidade, deve-se ter em mente que a informação aberta e grátis não rivaliza com a informação editada das fontes especializadas pré-*Internet*, mas apresenta algumas propriedades que estas não têm: abrangência de escopo sem limite, alta velocidade de fluxo e atualização (significando que sua variedade de conteúdos e o desempenho de seu processo de produção, comunicação, disponibilização, discussão e socialização não têm rivais).

Fuld (2007, p. 131) assim resume esse cenário:

> *(...) a Internet tem um padrão, uma forma de pesquisa. Esse é o catálogo das transações de negócios. Ao saber onde o dinheiro é trocado, você pode encontrar a inteligência. Essa é a vastidão da rede, que a torna uma ferramenta de inteligência tão poderosa – ainda que frustrante. (...) É uma estranha combinação de diversas coisas da Baker Library – tais como arquivos – com a dinâmica informacional e a confusão de um bazar aberto no Oriente Médio. Isto tudo está lá – conhecimento, insight e até mesmo a fofoca maliciosa e o rumor destrutivo. Você precisa encontrar um caminho para enxergar além das distrações e distorções e selecionar a informação poderosa que pode estar escondida nos sites das empresas, em grupos de discussão. (...) As bibliotecas por ora possuem barreiras artificiais à informação. Elas não podem refletir a natureza da dinâmica de como a informação flui. (...) As bibliotecas tradicionais podem somente captar o que os editores colocam em seus documentos. A maioria dessa informação é limitada pelo que o editor considera incluir (ou excluir). Além disso, a maioria dessa informação já está ultrapassada no momento em que chega às prateleiras da biblioteca. A informação da biblioteca, por essa razão, não lhe dá vantagem competitiva. É somente informação (informação antiga), não é inteligência. A Internet reflete mais precisamente a forma como a informação flui no mundo real do que qualquer arquivo de biblioteca.*

A metáfora adotada por Fuld (2007, p. 127) para que as vantagens e desvantagens da informação da *Web* sejam corretamente avaliadas é a de que esse ambiente é uma “casa de espelhos*”*. Contudo, mesmo que essa fonte tenha a tendência de desviar e confundir o senso de realidade das pessoas, ela é considerada uma das fontes mais valiosas de inteligência. Com métodos adequados, esse autor sugere que se possa filtrar o enorme volume de informação da *Web* com objetivo de se obter informação útil para *insights* de Inteligência Competitiva (como discutido no Capítulo 6), alguns dos métodos apresentados pelo autor podem ser suportados pelo mosaico de informação conceitual proposto nesta tese).

Outro aspecto importante a se considerar, em relação à *Web*, é a sua evolução conceitual, com impactos sociotécnicos evidentes. Ding e Xu (2007) defendem a tese que a *Web* evolui de modo não supervisionado, com base em leis naturais, independente do direcionamento político que alguma entidade supervisora internacional, tal como o consórcio *World Wide Web Council* – WWWC, pretenda imprimir à mesma. Esses autores argumentam que uma evolução natural ocorreu entre a denominada *Web* 1.0 (o modelo primitivo) e a *Web* 2.0 (o modelo atual), cujos postulados fundamentais são:

I. A evolução da *Web* é um processo direcional baseado em estágios.
II. A evolução da *Web* é um processo de progressiva “clonagem” comportamental da sociedade humana.

Davis (2008) apresenta uma versão dessa evolução de um ponto de vista semântico da informação, argumentando que:

> *A onda semântica abrange quatro estágios do crescimento da Internet. O primeiro estágio, Web 1.0, se referia à conexão de informação e disponibilização na rede. Web 2.0 trata da conexão de pessoas – colocando o "eu" em interface de usuário e o "nós" em redes de participação social. O próximo estágio, Web 3.0, está iniciando agora. É sobre representação de significados, conexão de conhecimento e utilização disso tudo de modo a tornar nossa experiência de Internet mais relevante, útil e prazerosa. A Web 4.0 virá mais tarde. E será sobre a conexão de inteligências numa Web onipresente, onde tanto as pessoas como as coisas raciocinam e se comunicam em conjunto.*

Os estudos prospectivos da *Web* 2.0 e além indicam, portanto, uma tendência de aumento contínuo tanto do volume de informações abertas disponíveis na rede como da diversidade de formas utilizadas para sua representação. A característica marcante desse estágio, em termos de funcionalidade, é a presença de ambientes de colaboração, onde os usuários promovem interação entre si de modo "*N* para *N"*, em modelos para troca de experiências e aprendizado coletivo.

A Figura 5.3 apresenta um modelo evolutivo da WWW com base em observação empírica e extrapolação. Esse modelo é apresentado por Davis (2008), apontando os prováveis caminhos evolutivos da *Web* até 2020, denominados de "onda semântica" a caminho da *Web* 3.0 e além (esse artigo tem suscitado debates interessantes na comunidade de entusiastas e acadêmicos da *arquitetura da informação*).

O ponto de origem nessa evolução é o modelo da "*Web* primitiva", com aplicações mais simples voltadas para publicação e busca de informação em sítios digitais pelo mundo afora, cuja tecnologia se destinava a conectar conteúdos com endereços de hipertexto na *Internet* – seria, portanto, a *Web* 1.0. Essa primeira versão da *Web* ainda existe em abundância no mundo todo, sendo a de custos mais acessíveis para pessoas e organizações que não necessitam nada além de suas funções básicas.

A versão seguinte, que denominamos *Web* 2.0, é a "*Web* Social", com objetivo de desenvolver meios mais avançados de conexão de pessoas pela *Internet* em aplicações de correio eletrônico, redes sociais, portais de comunidades e outros recursos. É comum observar-se, inclusive em portais mais avançados, as características desta versão amalgamadas com as da versão anterior.

A "*Web* Semântica" seria o próximo passo nessa evolução histórica, onde a busca de significados, em contextos de Gestão do Conhecimento e Inteligência Artificial, passa a ocupar as maiores atenções em termos de desenvolvimento. Davis (2008), de modo simplista, caracteriza essa versão pela "conexão de conhecimento", algo errôneo, de um ponto de vista filosófico, considerando-se que não se pode conectar o cérebro das pessoas pela *Web*. É este modelo de *Web* que poderá ser adotado, nas redes digitais internas das organizações competitivas, para publicação dos conteúdos estudados nesta tese.

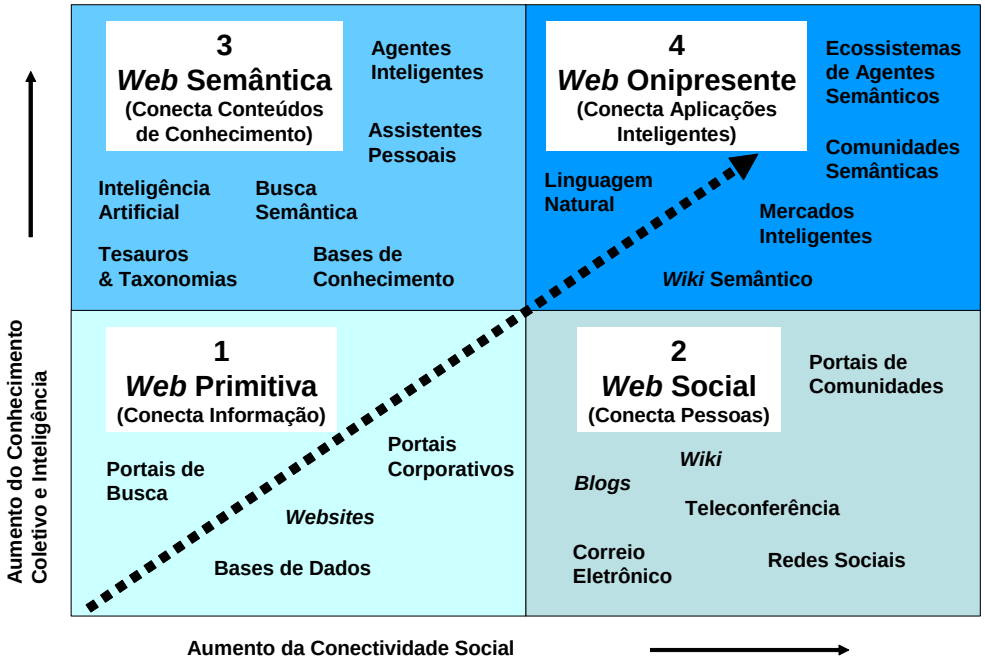



Figura 5.3 Cenário Evolutivo da *Web*
(Fonte: DAVIS, 2008)

O quarto estágio evolutivo da *Web* até 2020, conforme o modelo adaptado de Davis (2008) apresentado na Figura 5.3, seria caracterizado por um ambiente de recursos digitais para solução inteligente de problemas do dia-a-dia das pessoas e organizações, onde o termo "inteligente" pode ser entendido como "comportamento inteligente e automático", isto é, executado por aplicativos de *software* (conforme a necessidade do usuário).

Como exemplo desse admirável mundo novo da *Web* 3.0 nesse último estágio até 2020, o autor prevê o uso disseminado de ambientes virtuais com aplicações de "mercado inteligente" (*smart market)*. De fato, esse tipo de aplicação já existe na *Web* atual, com objetivo de otimização das transações, num determinado momento, entre *N* consumidores e *M* potenciais fornecedores de um determinado produto, como ocorre, por exemplo, na distribuição de gás natural. Esse tipo de aplicação busca satisfazer vários critérios de otimização, tanto do ponto de vista do consumidor como do fornecedor, com uso de lógica e sofisticados métodos matemáticos.

Com esse cenário, embora se reconheça a possibilidade de obtenção de informação digital sobre organizações em outros repositórios, tais como algumas bibliotecas digitais especializadas (algumas de livre acesso[65]), considera-se nesta tese a WWW como a principal fonte de informação para Inteligência Competitiva. E pode-se, nesse cenário, destacar três grandes agrupamentos de informações mais relevantes para a função de compartilhamento da informação, no modelo da *Web* 2.0, tendo em vista uma política de Gestão do Conhecimento para Inteligência Competitiva:

[65] "Livre" no sentido de acesso gratuito pela *Internet*.

I. Informação para a clientela: consumidores.
II. Informação para os parceiros: internos (colaboradores no quadro da organização) e externos (acionistas e outros colaboradores e organizações parceiras).
III. Informação para outros interessados: órgãos fiscais e de controle estatal, auditores independentes, observadores externos, etc.

Com base em seus objetivos, a presente pesquisa se concentra na disponibilização de informações corporativas para o aprendizado dos gestores da informação e do conhecimento e gestores do negócio em áreas de inteligência competitiva nas organizações complexas.

### 5.2.3 *Web* portais corporativos

A definição conceitual de um novo objeto identificado no mundo digital da WWW nem sempre é tarefa trivial e alguns autores utilizam, para tanto, metáforas, tais como no caso de *Web Portal Corporativo*:

> *A ampla difusão da Internet e de todos os tipos correlatos de "e"-tecnologias e recursos de informação deu origem ao conceito de portal, ou, de modo mais ampliado, de "estação do ciberespaço". Inicialmente, esse conceito era estruturado em torno das possibilidades de busca e disseminação da informação apresentadas pela World Wide Web. Então, o termo portal significava um ponto de entrada ou sítio Web originário para combinação da fusão de conteúdos e serviços de disseminação da informação, assim como para prover uma "base doméstica" personalizada para seus usuários, a partir da qual eles estariam aptos a lançar "expedições" de exploração no ciberespaço.*
>
> *(...) um portal de e-business é um mecanismo (ou um dispositivo físico) que provê acesso a um conjunto dinâmico de serviços do negócio, ao mesmo tempo provendo acesso a serviços de disseminação de informações específicas que podem ser usados para comunicação ou entrega de informação para um usuário*. (DAVYDOV, 2001, p. 57-58)

A Figura 5.4, de Davydov (2001), mostra uma classificação evolutiva dos portais na *Web* baseada na taxonomia utilizada no mercado. Os três tipos de portais que interessam a esta tese são os dedicados, originalmente (no primeiro nível de classificação), à busca e à disseminação da informação, e o dedicado à Gestão do Conhecimento.

Essa taxonomia arbórea é teórica, ainda que inspirada na experiência, em geral coexistindo, de modo evolutivo, características de mais de uma classe em um mesmo portal. À medida que os *Web* portais evoluem, novas funcionalidades de suporte a serviços são adicionadas a eles, mas algumas funcionalidades "de origem" não são abandonadas, mas aperfeiçoadas, tais como a de busca da informação.

Contudo, como a *Web* é muito vasta, os portais de busca se especializaram em portais de informação para orientação ao consumidor (que são os mais conhecidos) e portais de informação verticais da indústria (*Vertical Industry Portal*, ou *Vortal*). O *Vortal* é um sítio na *Web* para provimento de caminhos para acesso à informação específica correlata a um determinado ramo de negócios do mercado – são, portanto, portais de busca de informação mais especializados.

Os portais para disseminação da informação apresentam características de *Intranets*, ou redes de conteúdos digitais com tecnologia da *Internet* para uso interno nas organizações. Davydov (2001, p. 131) apresenta esse tipo de *Web* portal como uma mídia de massa para colaboração e interação entre indivíduos e seus computadores, acrescentando, ainda, que:

> *No campo emergente dos portais corporativos, a principal meta é expor e entregar informação relevante e específica do negócio num contexto de auxílio aos modernos empregados, para que os mesmos sejam produtivos e competitivos. E para ser produtivo e competitivo requer-se não somente acesso a informação, mas também uma habilidade para interagir (comunicar-se) com os outros utilizando a informação obtida como base.*

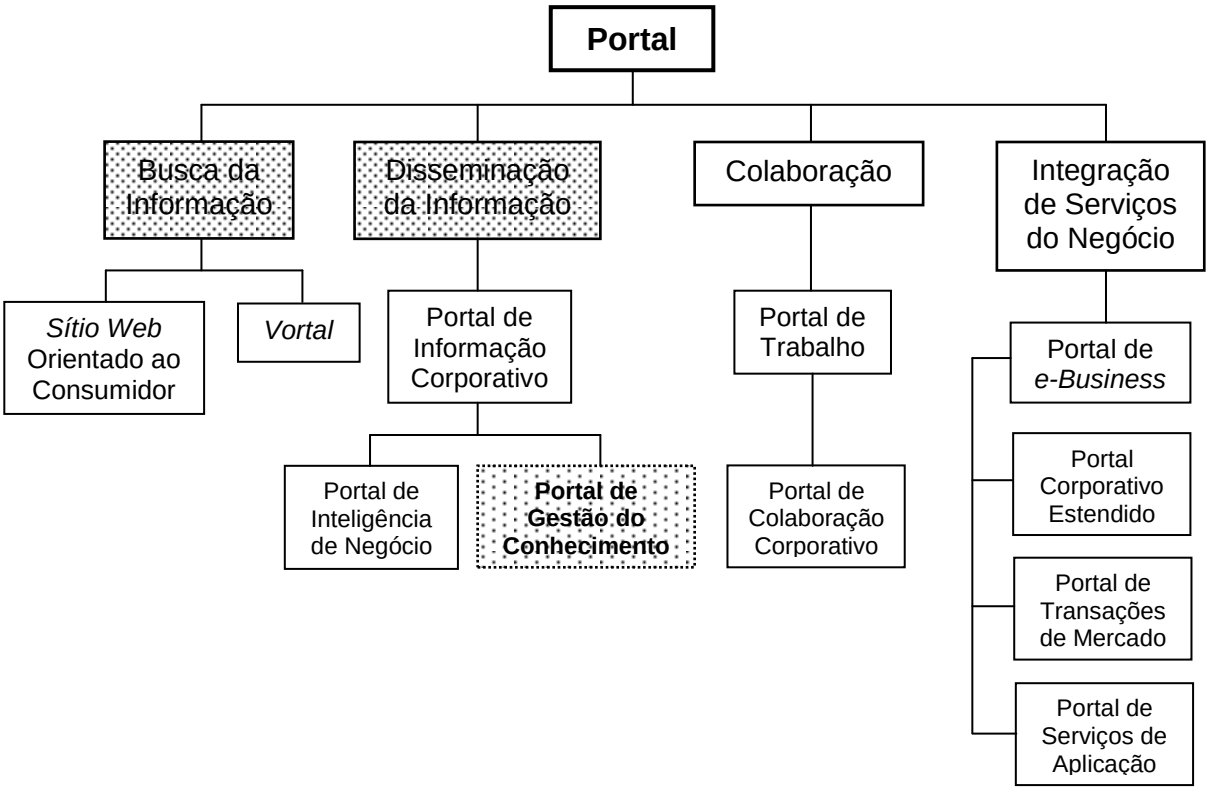



Figura 5.4 Classificação Evolutiva dos *Web* Portais
(Fonte: DAVYDOV, 2001)

O objetivo de entrega de informação útil, de qualidade, para os colaboradores e parceiros na organização é endereçado a dois subtipos de portais especializados: os portais de *Inteligência de Negócios* (*Business Intelligence*, com acrônimo BI, no idioma original) e os portais de Gestão do Conhecimento. Considerando-se que "conhecimento" pode ser resumidamente definido, de modo prático, como "informação sobre informação" (SCHREIBER *et alli*, 2000), esses portais podem ser considerados uma evolução natural dos primitivos portais dessa classe. E, no contexto desta tese, é importante destacar-se as características chave desses portais referentes às duas "categorias críticas" de Davydov (2001, p. 132-133):

1. *Identificação e categorização de recursos de informação corporativa e produção e entrega de conteúdo relevante.*
2. *Processamento de informação orientada ao conhecimento.*

Quanto aos portais de colaboração, eles destinam-se não somente a conectar pessoas a fontes de informação, mas também, e talvez principalmente, pessoas a pessoas em contextos de realização de tarefas (por isso são classificados como "portais de trabalho"[66]). Eles se destinam à solução de problemas em equipe, sendo, portanto, focados na disponibilização de ambientes virtuais onde grupos de pessoas podem interagir trocando informação ou desenvolvendo tarefas em conjunto com uso de aplicativos de *software* em rede. Exemplos desse tipo de ambiente abundam, podendo-se citar, na academia, o MOODLE *(Modular Object-Oriented Dynamic Learning Environment)* e, no mercado os ambientes de gerenciamento de projetos, o AutoCAD (para projetos de engenharia) e o Microsoft Project (para projetos em geral) e seus similares com código aberto (um primitivo sistema de colaboração é o próprio ambiente de correio eletrônico, onde os usuários podem trocar uma série de mensagens interativas até se chegar a um consenso sobre determinado assunto).

A WWW, atualmente, oferece uma variedade de ambientes de colaboração sem custos diretos para os usuários, tais como o *Google Docs*. Esses ambientes são utilizados, geralmente, por comunidades de prática, que constituem uma categoria de fonte de informação para Gestão do Conhecimento nas organizações.

E a última classe de *Web* portais apresentada na Figura 5.4, a de *portal de negócios (e-Business Portal)*, representa o último estágio na automação das transações de uma organização de mercado com seus clientes e parceiros externos. É nesse ambiente tecnológico que uma organização no estado-da-arte do comércio eletrônico realiza tanto as transações de vendas de produtos como de aquisição de insumos para seus processos produtivos em geral. O conceito de "Gestão da Cadeia de Suprimento" (*Supply Chain Management*, ou SCM) é realizado nesse ambiente, conectando, pela *Internet*, conteúdos de informação e serviços de todos os elos – ou organizações e pessoas – da cadeia produtiva relacionada aos produtos da empresa proprietária do *Web* portal.

Outro conceito, o de "organização em rede", se aplica bem a essa categoria de portal, que pode se desdobrar, ainda, em três subtipos: Portal de Empresa Estendida, Portal de Transações de Mercado e Portal de Serviços de Aplicação. O primeiro, comentado anteriormente, dispensa esclarecimentos complementares; o segundo pode ser observado em abundância na atual *Web* 2.0, onde muitas empresas dispõem, para os internautas, ambientes de transações eletrônicas onde o cliente pode escolher o produto que deseja adquirir, às vezes até com possibilidade de composição de partes – como no caso dos portais de vendas de computadores, e realiza, mediante cadastro pessoal prévio, o pagamento via cartão de crédito. Quanto ao terceiro subtipo,

[66] Esta expressão é uma tradução aproximada (ou interpretação) de *Workplace Portal*, no original.

trata-se de um portal especializado na disponibilização de serviços baseados em *softwares*, onde o cliente busca um recurso tecnológico para uma aplicação específica de seu interesse – um cliente poderá, por exemplo, mediante contrato, utilizar um sistema de processamento remoto de folha de pagamento disponibilizado nesse portal.

### 5.2.4 Identificação e seleção de fontes

Os dados selecionados na amostra são coleções de textos de 15 (quinze) organizações de vários portes e ramos de negócio disponíveis na *Web*, no idioma inglês. A opção por textos neste idioma se deve à indisponibilidade de *softwares* etiquetadores *(taggers)* no idioma português para uso em escala de textos e ao fato dele ser o idioma mais comumente utilizado pelas organizações na *Web* – podendo-se afirmar que o inglês é o idioma dos negócios na rede mundial de computadores.

A amostra é composta de conteúdos de informação em linguagem natural de Portais Corporativos de organizações consideradas, na maioria, entre as 500 maiores do mundo em 2009 (13 delas na lista *Global 500*), parte delas classificadas, também, na lista *Fortune 100* (seis delas)[67]. Essas classificações ordenam as organizações pelo volume monetário de receitas anuais (em US$ milhões), mostrando, também, o volume de lucros anuais.[68]

Os ramos de negócio representados na amostra são: livraria e conteúdos digitais (uma empresa), banco (duas empresas), indústria química e farmacêutica (uma empresa), advocacia (uma empresa), indústria de equipamentos fotográficos (uma), pesquisa em *hardware* e *software* (três), comércio de alimentos (uma), cervejaria (uma), siderurgia e construção naval (uma), indústria petrolífera (uma), indústria de *hardware* para computadores digitais (uma), hipermercado (uma). Dessa amostra, as quatro organizações que não constaram da lista *Fortune 500* em 2009 são as de advocacia, comércio de alimentos e indústria de *hardware* para computadores (caso todas as organizações selecionadas estivessem na lista *Fortune 500*, a amostra corresponderia a 3% desse universo).

O critério de seleção dessa amostra se deve, em parte, à liderança e representatividade dessas organizações em suas áreas de negócio e, em parte, aos volumes de informação aberta apresentados em seus portais na *Web*. O número de coletâneas adotado buscou compatibilizar a necessidade de uma amostragem minimamente representativa do universo das organizações com a disponibilidade de tempo e de recursos humanos e materiais no horizonte da pesquisa, para o esforço de laboratório – considerou-se, portanto, esse número satisfatório para uma avaliação dos dados das organizações mais representativas para o desenvolvimento da tese.

[67] A lista *Fortune 100* contempla apenas empresas norte-americanas.
[68] Dados disponíveis em: *http://money.cnn.com/magazines/fortune/global500/2009/full_list/...*.

Os testes de laboratório com textos de artigos científicos, mais presentes na "*Web* invisível", tanto corroboram os resultados observados nos textos de portais corporativos na *Web* como suscitam algumas indagações ulteriores interessantes, voltadas para uma possível classificação tipológica de conteúdos de textos em linguagem natural a partir dos parâmetros quantitativos de análise adotados nos experimentos. Questões como "Que tipo de texto apresenta, naturalmente, maior conteúdo semântico?", ou "Quais parâmetros sintáticos estatísticos indicam o nível semântico de um texto?", talvez possam ser respondidas com a metodologia de análise conceitual proposta nesta tese.

## 5.3 Laboratório

Como proposto na metodologia de desenvolvimento da pesquisa, as atividades de coleta, armazenamento, organização, processamento e análise dos dados contaram com suporte tecnológico de um mini-laboratório de *Processamento da Linguagem Natural* (PLN) composto por um microcomputador e alguns *softwares* com as seguintes especificações básicas:

I. Microcomputador: notebook *HP Pavilion Entertainment PC*, com microprocessador Intel Core 2 Duo T 6600, barramento de 2,2 GHz, 4 GB (gigabytes) de memória RAM e 500 GB de capacidade de armazenamento de dados no disco rígido.
II. Sistema Operacional: *Microsoft Windows 7 Home Premium,* de 64 bits.
III. Rede de Acesso à *Internet*: no cabo, da última milha externa até o *hub* local, e *wireless,* do *hub* até a estação de trabalho (microcomputador pessoal).
IV. Editor de Textos: *Microsoft Word 2003*.
V. Planilha Eletrônica: *Microsoft Excel 2003*.
VI. Aplicativos de PLN:
    - Etiquetador: *TreeTagger*, desenvolvido por Helmut Schmid no *Institute for Computational Linguistics of the University of Stuttgart* (Alemanha), com parametrização para o idioma inglês desenvolvida na *University of Pennsylvannia* (EUA) e interface gráfica para Sistema Operacional *Microsoft Windows XP* desenvolvida por Ciarán Ó Duibhín; é um *software* livre disponível em *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.*
    - Analisador Sintático Estatístico e Aplicativo de Mineração de Textos: *WordSmith*, desenvolvido por Mike Scott e comercializado[69] na *Web* em: *http://www.lexically.net/wordsmith/*.
    - Reticulador para Análise de Conceito Formal: *ToscanaJ*, desenvolvido em parceria na *University of Queensland* (Austrália) e na *Technical University of Darmstadt* (Alemanha), *software* livre disponível no endereço: *http://toscanaj.sourceforge.net/.*

[69] O preço de licenciamento do *WordSmith* é bastante acessível.

- Editor de Gráficos Conceituais: *CharGer*, desenvolvido na *University of Alabama at Huntsville* (EUA), *software* livre disponível no endereço: *http://sourceforge.net/projects/charger/.*

VII. Impressora: *HP Laser Jet 1020*.

O ambiente computacional assim composto atendeu, em geral, às expectativas tanto em termos de requisitos funcionais como não-funcionais, executando as tarefas de PLN com precisão aceitável e tempos de resposta bastante satisfatórios. O *TreeTagger* e o *WordSmith* se revelaram bastante velozes para o volume de textos utilizado, com tempos de resposta de no máximo poucos segundos a cada transação, ainda que suas interfaces gráficas não estejam no nível de usabilidade dos melhores *softwares* aplicativos comerciais – o *TreeTagger*, em particular, desenvolvido originalmente para Sistema Operacional *Linux* e interface de linha de comando, tem uma interface gráfica bastante limitada em termos de recursos.

O autor do manual do usuário do *WordSmith* apresenta o produto como "um conjunto integrado de programas de computador para observação de como as palavras se comportam em textos" (SCOTT, 2008, p. 2). Esse aplicativo tem capacidade de processamento de textos em linguagem natural *(plain texts)* para produção de listas e estatísticas de frequência de palavras, busca textual com palavra-chave, busca de segmentos de textos com base em padrões sintáticos predefinidos (implementando a função *concordance*) ou com base em horizontes de posicionamento de termos contíguos ou não nas sentenças (função *collocation*), descoberta de agrupamentos (*clusters)* mais frequentes de palavras, descoberta de padrões de combinação de conjuntos de palavras predefinidas *(concgram)* e outros recursos de PLN. Os dados processados resultantes podem ser obtidos tanto com a interface gráfica do aplicativo como em formato de planilha eletrônica.

O *ToscanaJ* representa uma evolução, em linguagem de programação Java, de uma conhecida plataforma computacional para Análise de Conceito Formal originalmente desenvolvida em linguagem C utilizada na academia. As interfaces de entrada de dados são, também, bastante limitadas, com pouca usabilidade, mas permitem a elaboração de contextos com milhares de objetos de entrada e atributos. Contudo, os reticulados *(lattice)* produzidos se tornam pouco úteis quando os contextos apresentam um número relativamente elevado de atributos: os experimentos de laboratório mostraram que para mais de uma dezena de atributos, em contextos com centenas de objetos, o número de conexões de conceitos pode ser elevado o bastante para tornar sua leitura e interpretação bastante penosas para o analista usuário numa tela de computador. A experiência de Lindig e Snelting (1997) corrobora esta percepção, mostrando um reticulado resultante que se parece com um feixe de espaguete, com análise visual praticamente impossível.

O *CharGer* é um aplicativo útil para edição de Gráficos Conceituais de Sowa (1983), com base nos estudos de Peirce sobre Gráficos Existenciais, e suas limitações são as comuns em editores de textos, com pouca inteligência associada.

## 5.4 Coleta de dados

As estruturas textuais de *Web* Portais Corporativos apresentam, geralmente, uma composição de documentos que se complementam de modo a cobrir todos os conteúdos da "mensagem para o mundo*"* que as organizações pretendem propagar. Com os primeiros contatos com as fontes selecionadas, observou-se que os textos da primeira página geralmente servem de chamada para textos mais aprofundados nos respectivos temas nas páginas posteriores, os da segunda página muitas vezes remetem o leitor internauta para as páginas seguintes, e assim por diante, sugerindo-se assim uma estrutura de encapsulamento de hipertexto que pode ser modelada como uma cebola (Figura 5.5).

Os textos das camadas mais externas são mais curtos e menos densos, mas apresentam mensagens mais objetivas e "agressivas" para os leitores, com notícias e chamadas para produtos e vantagens oferecidas pela empresa. Outra característica mais atual dessas páginas frontais é a nuvem de *tags* (etiquetas) com palavras-chave para indexação de conteúdos mais recentes (geralmente notícias) do *Web* Portal. Esses textos também apresentam mais sinais de marcação gráfica da própria linguagem (*Hypertext Markup Language*, HTML e suas derivações evolutivas), que constituem "ruídos" nos sinais semânticos das mensagens necessitando ser eliminados na mineração de conceitos.

As camadas mais profundas de um *Web* Portal apresentam textos mais longos, alguns parecidos com pequenos artigos, apresentando detalhes sobre produtos e sobre a organização em si, no sentido de convencimento do leitor (atual ou potencial cliente) sobre as vantagens de um futuro relacionamento com a mesma. E nas camadas centrais, as mais profundas possíveis, tende-se a encontrar formulários para cadastro de clientes, aplicações de *software* com serviços úteis para os usuários (para cotação de produtos, por exemplo) e textos mais longos, com conteúdos apresentados em formatos mais próximos de textos em papel (tais como relatórios sobre produtos, balanços de negócios e outros, geralmente em formato PDF).

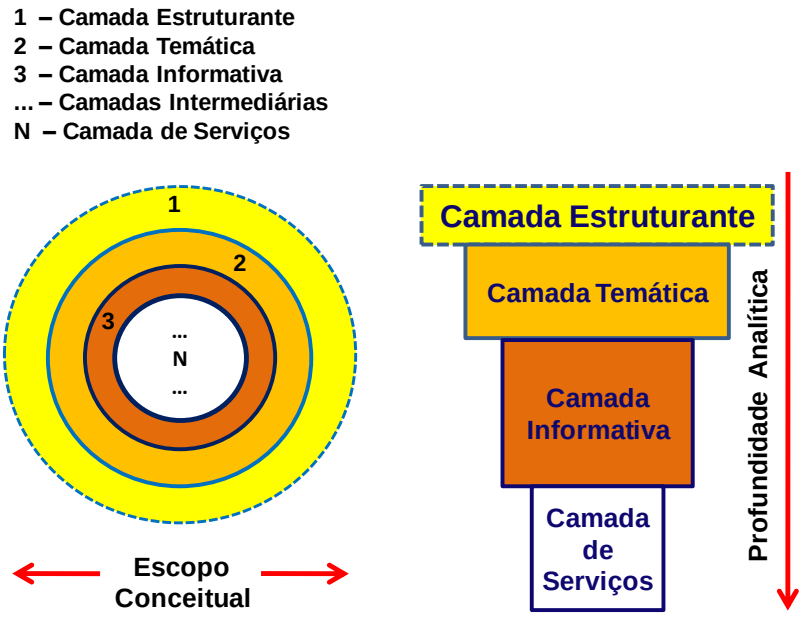



Figura 5.5 Camadas de Conteúdos em *Web* Portais Corporativos

(Fonte: do autor da tese)

Em suma, observam-se as seguintes composições de tipos de conteúdos de textos e sinais gráficos nessas camadas de um *Web* Portal Corporativo do porte pesquisado:

I. Camada de 1ª Página:
   - sinais de linguagem de marcação para controle de acesso ao sítio;
   - informações e botões de navegação no sítio;
   - lemas (frases) de visão empresarial;
   - notícias corporativas;
   - nuvem de *tags*.

II. Camada de 2ª Página:
   - informações e botões de navegação no sítio;
   - lemas de visão empresarial;
   - notícias corporativas;
   - informações organizacionais;
   - detalhes sobre os temas da 1ª página;
   - conexões de hipertexto para baixar documentos *(download)*;
   - *blogs* corporativos;
   - outros.

III. Camada de 3ª Página:
   - informações e botões de navegação no sítio;
   - lemas (frases) de visão empresarial;
   - notícias corporativas;

- textos com desenvolvimento dos temas da 1ª e da 2ª páginas;
- conexões de hipertexto para baixar documentos *(download);*
- outros.

Os textos da 1ª camada são úteis para a presente pesquisa porque apresentam sintagmas abstratos, com alto poder semântico, para definição de conceitos (antigos e novos) sobre o negócio e a organização, tais como: *credit card, investment services, Bayer Healthcare, creditor's rights, Heineken Experience, business solutions*, etc. Observam-se poucos verbos nesta camada, que é mais caracterizada por substantivos indicando conceitos. Assim, denominamos esta camada dos portais corporativos, para os fins desta tese, de "camada estruturante*"* de conteúdos.

Os conteúdos da 2ª camada são úteis, por sua vez, porque utilizam sintagmas para chamada de temas e serviços de informação a serem desenvolvidos nas páginas mais profundas do portal, tais como: *Video Podcast: Bayer's Perspective on Innovation, Use Mobile Banking, Bankrupticy Discharge, Research Areas*, etc. É nessa camada dos portais que começam a aparecer verbos nos textos, sendo mais caracterizada, no entanto, pela apresentação dos temas do *Web* Portal, motivo pelo qual a denominamos "camada temática*"*.

Com a 3ª camada e além obtêm-se textos explorando mais profundamente conceitos (ou temas) do negócio com uso de sintagmas complexos de alto poder semântico, muito específicos do negócio (geralmente endereçados a públicos mais especializados), tais como: *Protein Hunters in the Brain, Full Faith and Credit Clause, Operations Strategy, UEFA Champions League, Computational linguistics*, etc. Esta camada de conteúdos, que pode se estender por mais alguns níveis de produndidade no *Web Portal*, é caracterizada pela disponibilização de informação nos formatos clássicos da era do papel, a qual denominamos de "camada informativa*"*.

Os dados paramétricos das coletâneas extraídas dos *Web* Portais das organizações selecionadas são apresentados no quadro do "APENSO I – Densidade Sintagmática Plurinominal: Amostra de Textos da *World Wide Web"*. Coletou-se páginas dos portais em pelo menos três camadas, desde a camada estruturante até a camada informativa, para constituição de arquivos consolidados com todos os conteúdos de cada organização. A operação de acesso aos *Web* Portais e cópia dos conteúdos se deu com uso do portal de buscas *Google*, montando-se arquivos de textos sem formatação, codificados no padrão *Unicode*, para redução dos sinais de marcação HTML e outros indesejados (de navegação, figuras, etc), operação realizada com o *software* de edição de textos *Microsoft Word*.

Esta abordagem de coleta de dados brutos dos *Web* Portais Corporativos, sem nenhum tipo de filtragem prévia, teve como objetivo aproximar o experimento o máximo possível das condições de trabalho do mundo real, com informações e ruídos coexistindo nas fontes. Com isso, pretende-se alcançar maior robustez para operação nas "trincheiras" da Inteligência Competitiva, evitando-se custosas e tediosas operações de preparação (geralmente, seleção e reformatação) de textos para mineração encontradas em outras metodologias.

O primeiro bloco de dados à esquerda do quadro, no APENSO I, denominado “Texto”, apresenta uma coluna com uma numeração das organizações com portais pesquisados, em ordem alfabética; outra coluna, com o título da coletânea, se refere ao nome da organização mais conhecido; e uma terceira coluna informa o tamanho de cada coletânea em *bytes* (coluna “A”). A maior coletânea digital resultante é a da empresa *Sun Microsystems*, com 559,5 MB (megabytes), e a menor a do *Bank of America*, com 16,1 MB.

O bloco seguinte mais à direita, denominado “Palavras”, contém dados sobre: (coluna “B”) a quantidade de palavras encontradas em cada coletânea, considerando-se inclusive as repetições de palavras, (coluna “C”) a quantidade de palavras distintas encontradas (ou “tipos” sem repetição) e (coluna “D”) a relação de proporção entre a quantidade de palavras distintas e a quantidade total de palavras encontradas em cada coletânea (em percentual) – parâmetro denominado, em PLN, relação “tipo/palavra” *(type/token).*

As colunas do bloco de dados “Substantivos” mostram a quantidade de substantivos (ou nomes) com repetição em cada coletânea (coluna “E”), a quantidade de substantivos distintos (coluna “F”), a quantidade de substantivos distintos que aparecem na composição de sintagmas plurinominais (coluna “G”), a relação entre o número de substantivos distintos e o número de palavras distintas (coluna “H”), e a relação entre o número de substantivos distintos compondo sintagmas e o número de substantivos distintos.

O último bloco de dados à direita, nesse quadro, mostra a quantidade de sintagmas plurinominais distintos (coluna “J”) e a relação entre esse número e a quantidade de substantivos distintos (coluna “K”). O quadro ainda mostra, nas quatro últimas linhas, a soma total das quantidades de cada coluna, as médias, os desvios-padrões e as relações entre os desvios-padrões e as médias.

Os procedimentos para obtenção dos dados e cálculos dos parâmetros estatísticos de Processamento da Linguagem Natural – PLN das coletâneas do APENSO I são apresentados e discutidos, metodologicamente, a seguir.

## 5.5 Análise sintática e mineração de textos

### 5.5.1 Metodologia de recuperação da informação

O processo de coleta, organização, armazenamento, processamento e análise dos dados (textos em linguagem natural) do método adotado na experiência de laboratório da pesquisa é apresentado na Figura 5.6. As atividades expressas nos retângulos de cor mais clara são automáticas, com uso de *softwares*, enquanto as atividades dos retângulos de cor mais escura são apenas semi-automáticas, envolvendo penosas e morosas atividades manuais com suporte de um editor de textos.

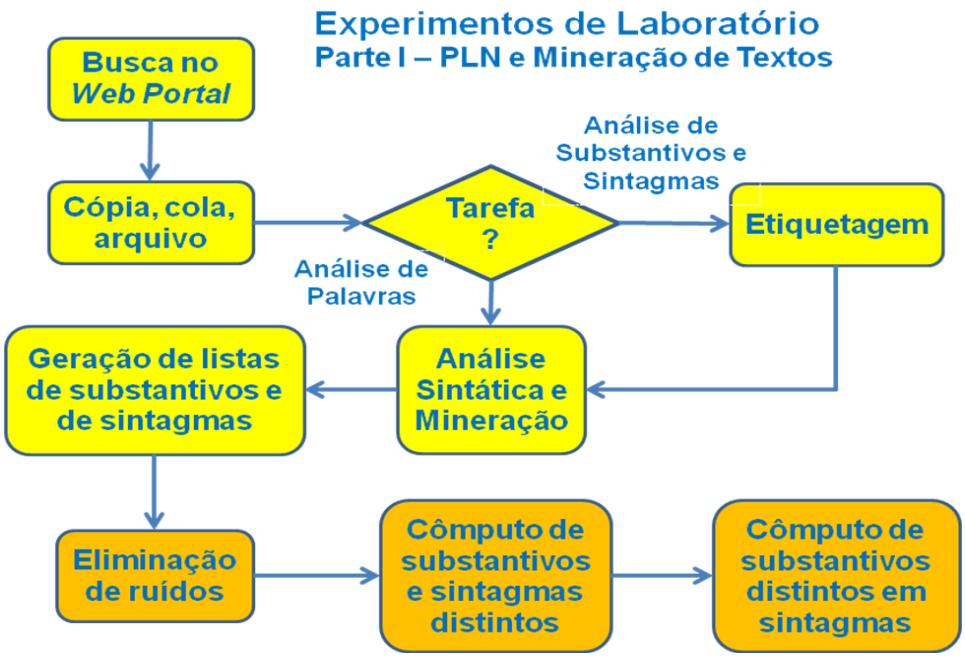



Figura 5.6 Processo Experimental – Parte I

(Fonte: do autor da tese)

O processo se inicia após a seleção da organização e seu respectivo *Web* Portal Corporativo como fonte de informação, tendo como primeira atividade a própria busca dos conteúdos das páginas com uso dos recursos de um *Web* Portal de busca como o *Google*. O passo seguinte envolve atividades de "folheamento" *(browsing)*, cópia, "cola" *(paste)* e arquivo de textos das páginas pesquisadas ("colando-se" os textos copiados de modo sequencial no arquivo, partindo-se da camada estruturante para as camadas mais profundas do *Web* Portal Corporativo). Os textos são colados em arquivos gerados pelo *software* editor de textos, em formato *.txt padrão *Unicode*.

O próximo passo requer uma decisão prévia sobre o tipo de atividade a ser desenvolvida na sequência: caso tratar-se da análise estatística de palavras *(tokens)* do texto, sem preocupação com sua classificação sintática, deve-se partir direto para a análise sintática com o *software* WordSmith; caso contrário, ou seja, necessitando-se da classificação prévia das palavras para se analisar a frequência de substantivos e de sintagmas plurinominais, deve-se submeter o texto à etiquetação antes da análise sintática e mineração.

A atividade seguinte consiste na geração das listas de substantivos e de sintagmas plurinominais, com base em listas de palavras *(tokens)* etiquetadas anteriormente. Essas listas devem ser processadas com o analisador sintático introduzindo-se, como argumentos de busca em modo *collocation* de mineração de textos (no *software* WordSmith), os símbolos de substantivos utilizados pelo etiquetador *(TreeTagger)*: NN (substantivo comum no singular), NNS (substantivo comum no plural), NP (nomes próprios no singular) e NPS (nomes próprios no plural).

Deve-se, também, estipular previamente o que se denomina “horizonte de busca” no *software* analisador sintático, que são os números de palavras à frente e atrás do padrão sintático utilizado presentes nos segmentos de texto, processando-se então uma consulta para cada tipo padrão de substantivo ou nome, obtendo-se um relatório de saída para cada horizonte. Como exemplo, numa das sessões de mineração de texto utilizou-se o par de padrões sintáticos “NN” e “NNS” com objetivo de extração de segmentos de textos com 100 letras (incluindo-se os espaços em branco), estipulando-se a captura de todos os tokens com etiquetas “NNS” encontrados até 10 *tokens* à frente ou atrás do *token* “NN”.

As saídas do WordSmith (Figura 5.7) se apresentam como relatórios em formato próprio desse *software* e precisam, por isso, ser exportadas para planilha eletrônica ou outro formato mais adequado para processamento de listas textuais. A quantidade de substantivos com repetição, para preenchimento da coluna “E” do APENSO I, é obtida mediante a observação, nos relatórios de saída do WordSmith, e soma das quantidades de substantivos comuns no singular (com etiqueta “NN”) e plural (com etiqueta “NNS”) e nomes próprios (“NP” e “NPS”).

Entretanto, as listas de saída do analisador sintático contêm um volume expressivo de “ruídos”[70], necessitando de uma limpeza antes da análise de conteúdos. Os ruídos são oriundos de sinais da própria etiquetagem (exemplo na Figura 5.8), geralmente símbolos de etiquetas e lemas das palavras etiquetadas e de *tokens* que não interessam no contexto. Isso ocorre porque a busca em modo *collocation* recupera trechos de sentenças em horizontes predefinidos pelos usuários e não palavras etiquetadas isoladas (o minerador de texto garante que nesses trechos se encontram as estruturas sintáticas de interesse, mas não realiza a filtragem de ruídos). Deve-se também, nesta etapa, eliminar as presenças excessivas de um mesmo substantivo ou sintagma, inclusive suas variações em número (em princípio, para a análise de conceitos tanto faz deixar-se um substantivo no singular como no plural).

O penúltimo passo do processo se refere às atividades de contagem de substantivos e sintagmas plurinominais, sem repetições, com uso de um *software* editor de textos comum, como o Microsoft Word.

O último passo do processo de recuperação da informação, que denominamos, na Figura 5.6, “Computar substantivos distintos em sintagmas”, consiste na comparação entre a lista de substantivos e a lista de sintagmas plurinominais, computando-se o número de substantivos que aparecem, também, na lista de sintagmas. Com isso, obtêm-se os dados necessários para completar as colunas “F”, “G” e “J” da planilha do APENSO I.

[70] Ruídos: símbolos sintáticos de palavras que não são de interesse no contexto de análise de conceitos que vêem a seguir na metodologia.

| | |
|---|---|
| 1 | ---------- NN ---------- |
| 2 | ---------- NN ---------- |
| 3 | ove NN move into IN into the DT the right JJ right lane NN lane and CC and follow VV follow the DT the Veteran N |
| 4 | owards IN towards Pittsburgh NP Pittsburgh . SENT . Exit NN exit at IN at the DT the Boulevard NP Boulevard of IN |
| 5 | of IN of the DT the Allies NP . SENT . Move NN move into IN into the DT the right JJ right lane NN l |
| 6 | is VBZ be located VVN locate at IN at the DT the corner NN corner of IN of Grant NP Grant and CC and Seventh NP |
| 7 | toward IN toward Pittsburgh NP Pittsburgh . SENT . Exit NN exit at IN at the DT the Veterans NPS Veterans Bridge |
| 8 | through IN through the DT the traffic NN traffic light NN light , , , crossing VVG cross over IN over Grant NP |
| 9 | P North NP North signs NNS sign . SENT . Exit NN exit to TO to your PP$ your left VVN leave at IN at |
| 10 | raight RB straight through IN through the DT the traffic NN traffic light NN light , , , crossing VVG cross over |
| 11 | is VBZ be located VVN locate at IN at the DT the corner NN corner of IN of Grant NP Grant and CC and Seventh NP |
| 12 | ffic NN traffic light NN light , , , turn NN turn right NN right , , , passing VVG pass the DT the Convention NP |
| 13 | the DT the Omni NP Omni Hotel NP Hotel . SENT . Drive NN drive across IN across Liberty NP Liberty Avenue NP Av |
| 14 | at the DT the second JJ second traffic NN traffic light NN light , , , turn NN turn right NN right , , , passin |
| 15 | JJ second traffic NN traffic light NN light , , , turn NN turn right NN right , , , passing VVG pass the DT the |
| 16 | is VBZ be located VVN locate at IN at the DT the corner NN corner of IN of Grant NP Grant and CC and Seventh NP |
| 17 | raight RB straight through IN through the DT the traffic NN traffic light NN light , , , crossing VVG cross over |
| 18 | through IN through the DT the traffic NN traffic light NN light , , , crossing VVG cross over IN over Grant NP |
| 19 | : : : Take VV take the DT the middle JJ middle lane NN lane , , , following VVG follow signs NNS sign to TO |
| 20 | CC and I-579 NP North NP North . SENT . Exit NN exit to TO to your PP$ your left VVN leave at IN at |
| 21 | Local Counsel NP Counsel to TO to debtor JJ debtor Sold NN assests NNS assest thereby RB thereby saving |
| 22 | S creditor , , , successful JJ successful reorganization NN reorganization Medical NP Medical Malpractice NP |
| 23 | of IN of our PP$ our people NNS people . SENT . CLIENT NN client REPRESENTATION NP DESCRIPTION NP |
| 24 | S creditor , , , successful JJ successful reorganization NN reorganization Brownsville NP Brownsville General NP Ge |
| 25 | diated VVD mediate a DT a medical JJ medical malpractice NN malpractice case NN case to TO to the DT the benefit |

Figura 5.7 Resultados da Mineração de Texto com Analisador Sintático

(Fonte: do autor da tese)

```
Bernstein   NP    Bernstein
-     :     -
Pittsburgh  NP    Pittsburgh
Law   NP    Law
Firm  NN    firm
When  WRB   when
Joe   NP    Joe
Bernstein   NP    Bernstein
started     VVD   start
his   PP$   his
law   NN    law
practice    NN    practice
more  JJR   more
than  IN    than
40    CD    @card@
years NNS   year
ago   RB    ago
```

Figura 5.8 Exemplo de Resultado da Etiquetagem

(Fonte: do autor da tese)

É importante observar-se, na coluna “I” do APENSO I, o percentual de substantivos que integram sintagmas plurinominais em cada texto analisado, que denominamos de “Índice de Representação Substantiva em Sintagmas Plurinominais (IRSP)”, com valor experimental médio de 63,5%, nesta pesquisa, para listas completas com todos os substantivos dos textos. O nível de

eficiência do método na recuperação de conceitos (representados, nesta etapa, pelos substantivos e nomes próprios), com base nesse indicador, aumenta sensivelmente quando se utiliza apenas os substantivos e nomes mais freqüentes num texto, aspecto paramétrico (estatístico) fundamental que será discutido a seguir.

Esse indicador é importante porque mostra o nível de eficiência do método de recuperação da informação proposto e testado em laboratório, podendo ser comparado ao índice de *revocação* da recuperação da informação tradicional com uso de palavras-chave.

### 5.5.2 Análise estatística

É importante ressaltar-se, neste ponto, que a análise estatística apresentada não deve ser encarada como uma análise similar à de um *corpus* lingüístico, quando se pretende conhecer padrões de uso da linguagem natural em estudo. O que se pretende, nesta análise estatística de sintagmas conceituais dos textos corporativos publicados na *Web*, é conhecer a mensagem que uma organização está transmitindo para o mundo por meio dos componentes textuais constitutivos dessa mensagem. Esses componentes característicos do léxico que tornam possível o conhecimento de uma organização, na *Web*, constituem o seu DNA[71], que pode ser filtrado com algoritmos inteligentes de busca denominados de "ímãs" por Fuld (2007, p. 156, grifos nossos):[72]

> *Você precisa compreender a linguagem e as nuances culturais do conteúdo da Internet, mas você não está equipado para racionalmente classificar a informação crítica. (...) O truque é identificar somente a informação correta da Web e enxergar claramente por meio das milhares de distrações, por meio e ao redor dos muitos espelhos, refletindo os muitos pedaços de informações – alguns verdadeiros outros não, alguns relevantes outros pura distração. Ao fazer isso, você precisa criar filtros de inteligência. (...) Esses filtros inteligentes começam com o conceito que denomino 'ímãs'. Um ímã é uma série de palavras ou frases que você coloca junto em uma busca, cada um representando um objetivo inteligente. Os bibliotecários denominam a conexão de palavras booleana como 'operadores' e incluem palavras como 'e', 'ou', e 'não'. O mundo on-line tradicional e as ferramentas de busca da Web, como Google, empregam estes ou outros termos similares ou frases para excluir ou incluir conceitos de busca. Os ímãs têm papel estratégico nessa idéia. Um ímã é uma frase que combina os elementos da idéia que você está tentando capturar.*

Os dados do APENSO I mostram, inicialmente, que na amostra de textos estudada o número de palavras distintas (não repetidas, ou "tipos") num texto representa apenas uma fração de todas as palavras utilizadas nesse texto – algo em torno de 21,8%, na média – e que à medida que os textos aumentam de tamanho (em quantidade de palavras ou *tokens*) a proporção de palavras não-repetidas *(types)* diminui. Este é um fenômeno conhecido na Ciência da Informação, indicando que quanto maior o texto maior a tendência de uso repetido de *tipos*, algo que contribui

[71] DNA: *Deoxyribonucleic Acid*. É um tipo de molécula que codifica a informação genética de um organismo humano e apresenta algumas características que o identificam de modo único, daí o uso da metáfora para qualificar certos traços informacionais que identificam uma organização que publica na *World Wide Web*.
[72] O autor considera essa *linguagem de busca* desenvolvida num contexto de Inteligência Competitiva como um segredo empresarial *per se*.

para os objetivos desta tese – a busca de conceitos expressos com sintagmas complexos, resultantes da combinação de substantivos ou nomes.

O mesmo fenômeno acontece em relação aos nomes/substantivos, obtendo-se, na amostra de textos estudada, proporções quase idênticas, com uma média de 20,2 % de nomes/substantivos não-repetidos em relação ao número total de nomes/substantivos (ver resultados no Gráfico 5.1).

Este resultado apenas confirma, no ambiente de linguagem natural da *Web* corporativa, observações anteriores de pesquisadores em textos diversos. Entretanto, mostra a utilidade de outros padrões sintáticos de textos comuns nesse tipo de ambiente, abrindo as portas da *Web* para a aplicação do conhecimento clássico da Ciência da Informação.

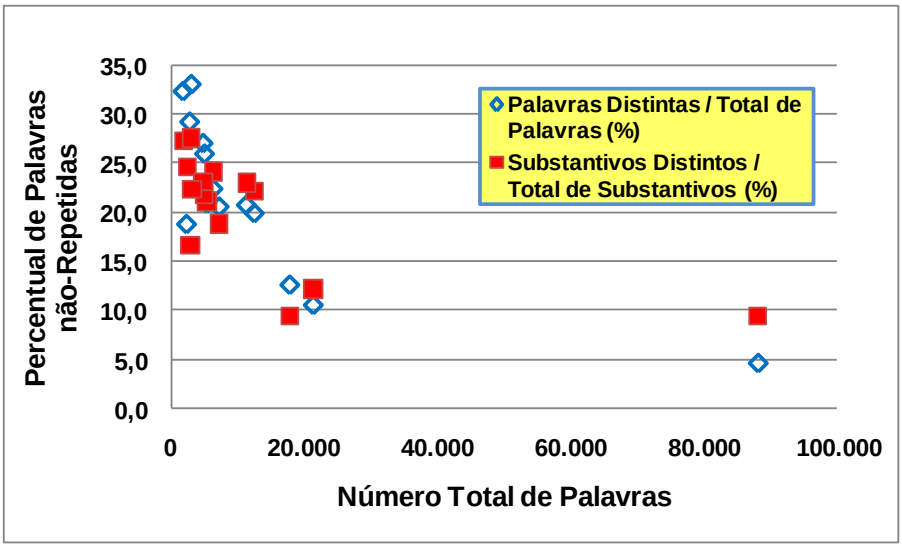



Gráfico 5.1 Curvas de Repetição de Termos na Amostra Textual

O Gráfico 5.2 mostra a razão (em percentual) entre o número de substantivos e o número de palavras (ou termos gerais) que aparecem nos textos da amostra, com e sem repetição. Observem-se as concentrações de pontos em torno da marca de 50% tanto no que se refere à razão entre substantivos e palavras distintas como entre substantivos e palavras em geral – as médias obtidas na amostra de textos, para ambos os casos, são de 50,8% e 48,6%, respectivamente.

Outro aspecto interessante nos Gráficos 5.1 e 5.2 são as posições dos pontos do texto da Sun Microsystems – um *outlier* devido ao seu tamanho bem maior que os demais na amostra.

Os resultados mais importantes obtidos no experimento, do ponto de vista de PLN, podem ser observados no "APENSO II – Frequência de Substantivos em Sintagmas Plurinominais". Os dados apresentados na planilha do APENSO II resultaram de contagens de nomes/substantivos mais freqüentes em sintagmas plurinominais, diferindo da abordagem experimental que resultou os dados do APENSO I apenas por este detalhe. Os quatro primeiros blocos de dados dessa

planilha se referem às quantidades de substantivos consideradas nesta análise, que se referem aos 2,5%, 5,0%, 7,5% e 10,0% mais freqüentes (os dados do último bloco, intitulado “100,0%”, se referem aos mesmos do APENSO I, com todos os substantivos utilizados pelas organizações em cada coletânea de *Web* Portal).

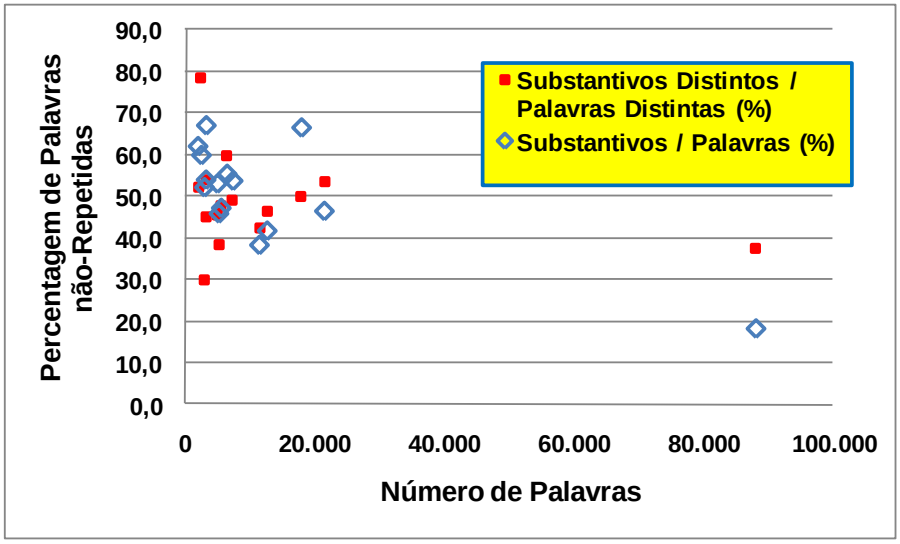



Gráfico 5.2 Razão entre Substantivos e Palavras na Amostra de Textos

Analisando-se os dados do primeiro bloco à esquerda no APENSO II (após as colunas de numeração e títulos dos textos), com os 2,5% de substantivos mais freqüentes, observam-se na primeira coluna à esquerda as quantidades absolutas de substantivos (QAbs) nessa frequência, na segunda coluna as quantidades desses substantivos mais freqüentes que também compõem sintagmas plurinominais (“Quantidade em Sintagmas”, ou QeS) e na terceira coluna a relação (percentual) entre os substantivos que compõem sintagmas e o total de substantivos nessa freqüência. As colunas dos blocos seguintes seguem essa mesma lógica de apresentação, com os 5,0%, 7,5% e 10,0% de substantivos mais freqüentes em cada texto.

Os cálculos das relações entre as quantidades de substantivos mais freqüentes que compõem sintagmas (QeS) e as quantidades absolutas desses mesmos substantivos mais freqüentes (QAbs) revelam que quando se consideram apenas os substantivos mais freqüentes, ao invés de todos os substantivos distintos utilizados nos textos, obtém-se um poder de revocação de nomes/substantivos significativamente maior que no modo “todos os nomes/substantivos”. Ou seja, utilizando-se a conhecida abordagem de indexação de textos com os termos mais freqüentes pode-se atingir níveis de desempenho mais elevados na representação de nomes/substantivos em sintagmas plurinominais, que denominamos “conexão sintagmática de substantivos mais freqüentes”, com uma proporção (ou percentual de revocação) de 80% a 90% – ressaltando-se que esses sintagmas comporão os conceitos da base de informação para Inteligência Competitiva no modelo de Gestão do Conhecimento proposto.

Em média, o percentual de conexão sintagmática dos 2,5% substantivos mais freqüentes nas 15 coletâneas da amostra é de 90,2%, significando que em cada dez substantivos/nomes dessa lista de freqüência nove deles compõem sintagmas compostos por mais de um substantivo/nome, que se denominam, nesta tese, “objetos” ou “instâncias” de conceitos. Esse desempenho médio, no pior caso estudado – quando se utiliza os 10% de substantivos/nomes mais freqüentes nas coletâneas amostradas, cai para 81,7%, ainda um percentual de revocação de substantivos/nomes bastante interessante para os fins propostos (mais de quatro objetos recuperados a cada cinco existentes).

O Gráfico 5.3, baseado nos dados do APENSO II, mostra os resultados do experimento, em termos de conexão sintagmática, para todos os textos da amostra e os substantivos mais freqüentes, nas escalas discretas adotadas.

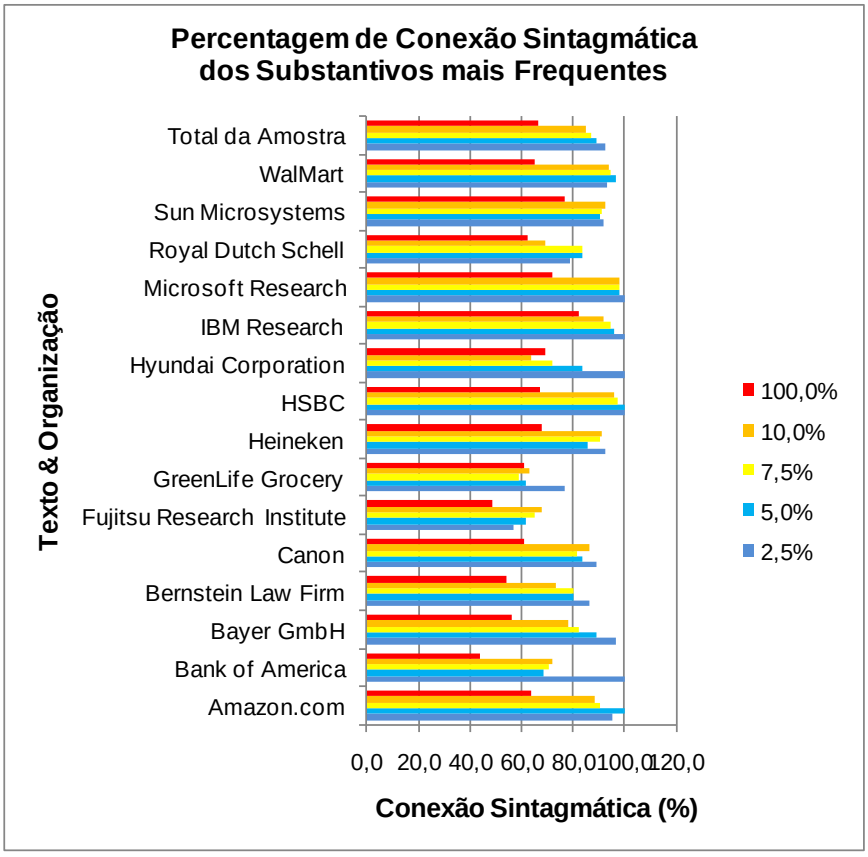



Gráfico 5.3 Percentuais de Conexão Sintagmática de Substantivos

A conclusão geral, conforme as curvas do Gráfico 5.4, é que quanto mais freqüentes os nomes/substantivos considerados, melhor o desempenho da revocação desses nomes/substantivos com o uso de sintagmas plurinominais na mineração conceitual de textos – ou melhor o percentual de conexão sintagmática dos nomes/substantivos. Ou ainda, numa linguagem mais adequada ao objeto desta tese, quanto mais freqüentes os nomes/substantivos constitutivos dos sintagmas, mais representativos, em termos conceituais, se tornam os sintagmas

plurinominais. Considerando-se até 10% dos nomes/substantivos mais freqüentes, o percentual de revocação com mineração de sintagmas plurinominais varia de 80% a 90%, em média, podendo até chegar, em casos específicos, a níveis de desempenho próximos de 100%.

As duas curvas do gráfico da Figura 5.4 mostram o espaço de variação padrão dos percentuais de conexão sintagmática dos nomes/substantivos num texto comum de *Web* Portal Corporativo. Esse espaço varia de um valor calculado como a média menos o desvio-padrão, para um determinado percentual de substantivos mais freqüentes, até um valor calculado como a média mais o desvio-padrão nessa freqüência.

É importante observar-se, no APENSO II, que quando se tomam apenas os 2,5% de nomes/substantivos mais freqüentes nas coletâneas da amostra, obtém-se 100,0% de conexão sintagmática em cinco dos quinze textos, ou em 1/3 deles. E que esse nível máximo de conexão sintagmática pode ocorrer em textos de organizações de nichos de mercado tão variados como bancos (Bank of America e HSBC), siderurgia (Hyundai) e pesquisa em TIC (IBM Research e Microsoft Research).

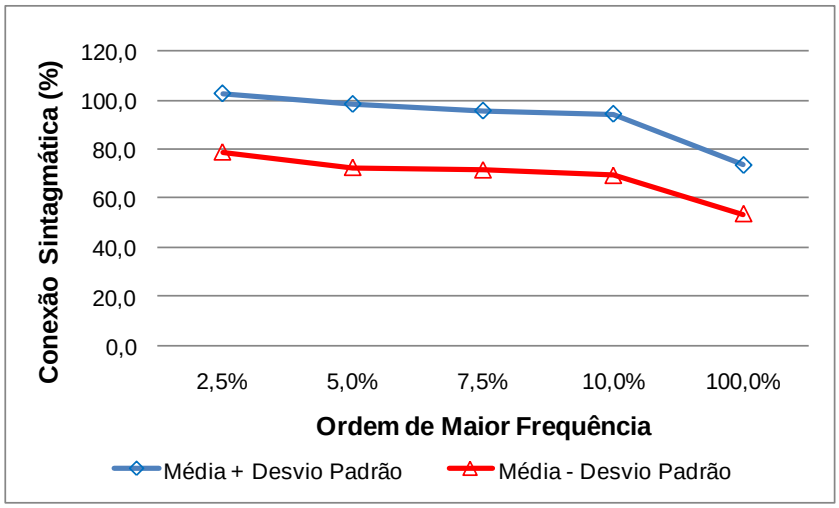



Gráfico 5.4 Curvas de Conexão Sintagmática dos Substantivos mais Frequentes

Com esta primeira parte do experimento concluída, restou evidente que o método de mineração de conceitos elaborado pode ser bastante eficiente, de modo a abranger a maior parte dos substantivos/nomes mais freqüentes de um texto. Ou, ainda, razoavelmente eficiente caso se queira analisar todos os nomes/substantivos que aparecem num texto da *Web* corporativa sem uso de recursos adicionais para desambiguação, como evidenciado no próximo capítulo.

Os "carnudos substantivos" de Lispector (1998) podem ser, portanto, assim recuperados com uso das metodologias e tecnologias de PLN utilizadas. Com isso, tem-se acesso à informação sobre os temas tratados com mais ênfase pelas organizações em seus *Web* Portais corporativos, numa abordagem de mineração de textos para posterior modelagem de conceitos de negócio, com missão algo parecida à da Modelagem Essencial de Dados (MED).

As propriedades estatísticas de textos digitais em linguagem natural descobertas no experimento também reforçam o modelo estatístico de relevância da informação de Zipf (1949), baseado na curva logarítmica que apresenta, no eixo horizontal, a posição *(rank)*, em termos de ordem (da maior freqüência para a menor), e no eixo vertical o número de palavras repetidas em cada ordem de freqüência. Como se observou no experimento desta tese, os substantivos que exprimem os conceitos em sintagmas complexos se situam no meio dessa escala de ordem de freqüência, entre as palavras mais freqüentes (que são *stop words*[73]), e as menos freqüentes.

A propósito, Meadow *et al.* (2007, p. 208) assim relatam uma experiência histórica com essa propriedade dos textos:

> *Luhn (1958) reconheceu que as palavras de freqüências mais altas tendiam a ser as comuns ou as que não geravam informação. Ele também sentiu que apenas uma ou duas ocorrências de uma palavra num grande texto não poderiam ser tomadas como significativas para definição do assunto temático. Por isso, ele sugeriu utilizar-se as palavras no meio da amplitude de freqüência. As palavras comuns melhor seriam eliminadas mediante o uso de uma lista de stop words e, em relação às palavras com baixa freqüência, pelo simples descarte daquelas com freqüência 1, 2, ou mais, dependendo do tamanho do texto.*

A segunda parte do experimento, apresentada a seguir, se concentra no problema da composição das redes de conceitos como suporte metodológico à Gestão da Informação e do Conhecimento em ambientes de Inteligência Competitiva.

## 5.6 Contextos e conceitos formais

O próximo estágio do experimento de laboratório consiste na organização e apresentação dos dados coletados na mineração de textos, para elaboração dos modelos gráficos de informação conceitual que deverão suportar as análises e geração de *insights* para desenvolvimento de Inteligência Competitiva. Esta segunda parte do experimento se reporta, portanto, à questão da representação do conhecimento, um tema desenvolvido em várias disciplinas correlatas à Ciência da Informação, tais como a Gestão do Conhecimento.

Como mencionado no Capítulo 4, os principais métodos e técnicas adotados nesta tese para representação do conhecimento – ou, de um ponto de vista filosófico basilar, para representação da "informação sobre informação" – são a Análise de Conceito Formal e a Linguagem de Modelagem Unificada (utilizando-se, eventualmente, Gráficos Conceituais como recurso de modelagem recursiva). Com efeito, a Análise de Conceito Formal é um método de classificação de objetos com base na incidência de atributos em um predeterminado contexto, que depende da necessidade e da abordagem do usuário. E, para sua consecução, o primeiro passo consiste na elaboração de uma lista de objetos e seu conjunto de atributos gerais parcialmente compartilhados, obtida do experimento.

[73] *Stop words*: palavras que compõem a estrutura sintática, mas sem importância semântica para a compreensão da essência do tema tratado no texto (em inglês, *the, of, a* e outras).

Esses sintagmas-objetos compõem, essencialmente, a mensagem da organização para o resto do mundo através da *Web*, revelando-se como os *cromossomos* de seu DNA corporativo (FULD, 2007), componentes de informação que a distinguem entre todas as outras organizações com presença na rede mundial de computadores. Esse material bruto necessita, no entanto, ser "lapidado" para que possa mostrar aos engenheiros do conhecimento seu potencial de "alavancagem" cognitiva e geração de *insights*.

Conforme a metodologia proposta, os objetos dessa lista, classificados mediante Análise de Conceito Formal, comporão os conceitos agrupados no reticulado de conceitos. Esses agrupamentos de objetos em conceitos e os relacionamentos entre conceitos constituirão a base de informação, o ponto de partida conceitual para o desenvolvimento de análises e *insights* posteriores, num ambiente de Gestão de Conhecimento apoiando a Inteligência Competitiva numa organização. Contudo, uma questão precisa ser resolvida inicialmente, antes da elaboração do contexto para Análise de Conceito Formal: qual o modelo de classificação de objetos mais adequado a ser adotado? Ou seja, com quais atributos deverão ser classificados todos os objetos sintagmáticos da lista?

### 5.6.1 Seleção de categorias

A Figura 5.9 apresenta as atividades do processo experimental da "Parte II – Análise de Conceito Formal", onde se busca conhecer os conceitos mais promissores de e sobre uma organização publicados em seu *Web* Portal Corporativo[74]. E, a partir desses conceitos esparsos, pretende-se desenvolver um modelo conceitual integrado que possa servir de plataforma de Gestão do Conhecimento tendo como alvo a Inteligência Competitiva.

O método da Análise de Conceito Formal apresenta propriedades bastante adequadas para a identificação e estudos de conceitos em quaisquer ambientes de informação. Contudo, o modelo de classificação dos objetos a ser adotado é intencional, isto é, vincula-se a algum critério de utilidade para o engenheiro do conhecimento. E neste ponto aparecem opções e dilemas que precisam ser devidamente ponderados, previamente, para aplicação da metodologia proposta.

#### 5.6.1.1 Categorização sintática

Experimentalmente, propôs-se testar alguns desses modelos de classificação (ou de categorização) de objetos textuais até encontrar-se um modelo adequado para os fins da presente tese, representando as primeiras duas atividades do processo experimental representado na Figura 5.9. O primeiro modelo de categorização testado consistiu na utilização dos próprios

[74] Obviamente, outras fontes de informação textual digital também poderiam ser utilizadas, como se mostrou no caso do contexto "Microsoft Research".

termos (palavras) dos sintagmas como objetos e atributos – considerando-se o primeiro substantivo/nome do sintagma como objeto e os demais como atributos, montando-se assim um contexto puramente sintático. Esse modelo de categorização, no entanto, revelou duas desvantagens avassaladoras para qualquer experimento analítico:

- distribuição de atributos e, consequentemente, de conceitos, excessivamente esparsa, resultando em matrizes de "objeto-atributo" de contexto quase quadradas, com "N" objetos e algo próximo a "N" atributos, portanto de pouca utilidade analítica (o teste com os objetos textuais da coletânea "Bayer Company" ilustra essa observação empírica);
- necessidade de se realizar uma mixagem de conceitos, após a construção da matriz de contexto formal (da Análise de Conceito Formal), a fim de reduzir a proliferação de conceitos de ontologias superficiais (extraídas diretamente do léxico).

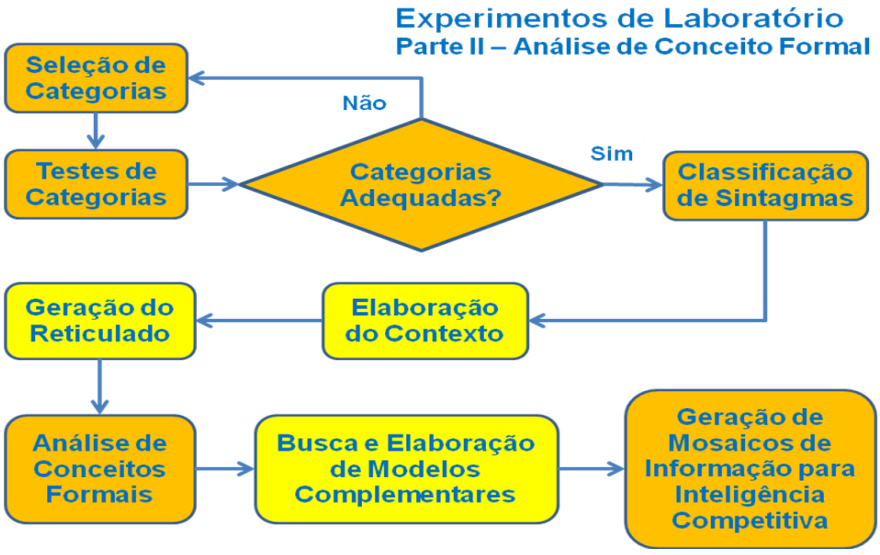



Figura 5.9 Processo Experimental – Parte II

(Fonte: do autor da tese)

5.6.1.2 Categorização aristotélica

Com essa frustração inicial em mente, pensou-se em um modelo de categorização mais compacto e mais básico acerca das coisas do mundo, que suportasse a natural e esparsa distribuição dos atributos sintáticos dos sintagmas sem, no entanto, permitir um modelo muito esparso de formação de conceitos. E o modelo então testado, para atendimento desses requisitos, inspirou-se no modelo de categorias de Aristóteles, assim apresentado pelo filósofo (grifos nossos):

> *Das expressões que são ditas sem qualquer combinação, cada uma significa ou uma substância, ou uma quantidade, ou uma qualificação, ou um relativo, ou onde, ou quando, ou estar numa* posição*, ou ter, ou fazer, ou ser afetado. (...) toda a afirmação parece ser ou verdadeira ou falsa; mas nenhuma das expressões que são ditas sem*

> *qualquer combinação (como, por exemplo, 'homem', 'branco', 'corre', 'vence') é verdadeira ou falsa.* (ARISTÓTELES; SANTOS, 1995, p. 39)

A categoria fundamental de Aristóteles para a classificação das coisas do mundo é a de "substância", que é assim definida pelo filósofo (ARISTÓTELES; SANTOS, 1995, p. 39-40):

> *Substância – aquilo a que chamamos substância de modo mais próprio, primeiro e principal – é aquilo que nem é dito de algum sujeito nem existe em algum sujeito, como, por exemplo, um certo homem ou um certo cavalo. Chamam-se substâncias segundas as espécies a que as coisas primeiramente chamadas substâncias pertencem e também os gêneros dessas espécies. Por exemplo, um certo homem pertence à espécie homem, e animal é o gênero da espécie; por conseguinte, homem e animal são chamados substâncias segundas.*
>
> *(...) Das substâncias segundas, a espécie é mais substância do que o gênero, pois está mais próximo da substância primeira. Pois se tivermos de dizer de uma substância primeira o que ela é, será mais informativo e mais adequado indicar a espécie do que indicar o gênero. Por exemplo, de um certo homem será mais informativo dizer que é um homem do que dizer que é um animal (pois o primeiro é mais próprio de um certo homem, enquanto o segundo é mais comum); e, para dizer o que é uma certa árvore, será mais informativo dizer que é uma árvore do que dizer que é uma planta. Além disso, é porque as substâncias primeiras são sujeitos de todas as outras coisas, e todas as outras coisas ou se predicam delas ou existem nelas, que elas são principalmente chamadas substâncias.*

Esta escolha do modelo Aristotélico para teste deveu-se aos próprios fundamentos dessa filosofia, que aparentemente se mostraram abertos à análise de conceitos para além da análise puramente sintática, como nas seguintes passagens da obra citada (ARISTÓTELES; SANTOS, 1995, p. 84-85, com grifos nossos):

> *Sendo a teoria das categorias um elemento essencial da filosofia aristotélica, são muitas as obras e as passagens onde Aristóteles a apresenta. Do estudo dessas passagens conclui-se que as categorias são os gêneros supremos da realidade.*
>
> *As categorias são então apresentadas, nesta passagem, como os gêneros a que pertencem as coisas significadas pelas expressões lingüísticas simples. Desta classificação dos gêneros pode evidentemente derivar-se uma classificação correspondente das próprias expressões: expressões que significam uma substância, expressões que significam uma quantidade, etc. Mas não é este o objetivo principal de Aristóteles. O que ele pretende classificar são as coisas significadas, e não as expressões que as significam. O que equivale a, tomando a linguagem como guia, efetuar uma classificação dos gêneros mais elevados de toda a realidade.*

Esta relativização das expressões lingüísticas quanto ao conhecimento está presente, também, na obra de Saussure (2002, p. 23), versando sobre a natureza do objeto:

> *Será que a lingüística encontra diante de si, como objeto primeiro e imediato, um objeto dado, um conjunto de coisas evidentes, como é o caso da física, da química, da botânica, da astronomia, etc.? De maneira alguma e em momento algum: ela se situa no extremo oposto das ciências que podem partir do dado dos sentidos. Uma sucessão de sons vocais, por exemplo, mer (m + e + r) é, talvez, uma entidade que regressa ao domínio da acústica, ou da fisiologia; ela não é, de jeito nenhum, nesse estado, uma entidade lingüística. Uma língua existe se a "m + e + r" se vincula uma idéia.*

Essas observações de Saussure também se relacionam com os "jogos de linguagem" de Wittgenstein (2001, p. 7):

> *Será possível dizer-se: Na linguagem (...) temos diferentes "tipos de palavras". Entre as funções da palavra "tijolo" e da palavra "bloco" existem mais semelhanças que entre as funções de "tijolo" e "d". Mas como agrupamos palavras em tipos dependerá do objetivo da classificação – e de nossa própria tendência.*

> *(...) É fácil imaginar-se uma linguagem consistindo somente de ordens e respostas em batalha. Ou uma linguagem consistindo somente de questões e expressões para se responder sim e não. E inumeráveis outras. E imaginar-se uma linguagem significa imaginar-se uma forma de vida.*
>
> *Mas o que dizer sobre isto: é o pedido de "Tijolo!" no exemplo (2) uma sentença ou uma palavra? Se é uma palavra, certamente não tem o mesmo significado da palavra de som parecido de nossa linguagem ordinária (...). Mas se é uma sentença, não é certamente a sentença elíptica: "Tijolo!" de nossa linguagem. (...) Porque se você grita "Tijolo!" você realmente quer dizer: "Apanhe um tijolo para mim."*

Outra passagem, de outra obra da filosofia aristotélica, esclarece ainda mais acerca dos propósitos desse modelo de categorias (ARISTÓTELES, 1969, p. 13-14):

> *A substância é anterior para o conhecimento. Conhecemos melhor uma coisa quando sabemos o que ela é do que quando estamos informados de sua qualidade, quantidade e lugar. E, com efeito, se queremos conhecer alguma coisa pertencente a uma categoria que não a de substância, não devemos indagar das qualidades, etc, que possui, mas sim do que ela é: qual é a sua quase-substância, o que faz com que ela seja o que é. Neste contexto, é evidente que a substância não está sendo concebida como a coisa concreta, mas como a natureza essencial. E esse duplo significado impregna todo o tratamento da substância por Aristóteles.*

Esta última passagem é interessante porque mostra, também, a preocupação do filósofo em esclarecer o papel do conceito de substância na epistemologia das coisas do mundo. O estudo da natureza essencial das coisas, no contexto desta tese, se reporta ao problema da semântica, endereçado, como solução, na composição sintagmática plurinominal das estruturas de textos recuperadas na mineração.

Outro aspecto importante desse esclarecimento das idéias de Aristóteles é o da anterioridade da substância para o conhecimento. Ou seja, o filósofo adota uma abordagem eminentemente empírica em relação ao conhecimento, tendo como base a substância denotada, em sua essência, pela referência lingüística. Com isso, a análise conceitual dos "carnudos substantivos" de Lispector (1998) como ponto de partida nos parece uma decisão acertada em termos de estratégia experimental nesta tese.

O modelo de categorias de Aristóteles, no entanto, também apresentou uma dificuldade considerada excessiva para os fins a que se destinaria: complexidade do processo de classificação dos objetos sintagmáticos. E isso se deve, paradoxalmente, à sua maior vantagem, que é a capacidade de suporte a quaisquer objetos lingüísticos expressos pelo ser humano. Como os objetos textuais (ou "substâncias") minerados nos *Web* Portais corporativos representam, na maior parte, gêneros mais abstratos compostos por objetos mais simples e denotam conceitos complexos elaborados no contexto do negócio, a velocidade de classificação desses objetos complexos se revelou excessiva, tornando essa atividade morosa e tediosa.

Como exemplos, a classificação de objetos como "abstraction for the business artifacts", "aftermarket services", "bandwith availability", "cross-industry Supply Chain Operations Reference Model", "development of service-delivery solutions", "e-government services", "information services hub" e outros, nas categorias de Aristóteles, requer a análise de vários níveis hierárquicos de composição sintática e semântica e até o uso de dicionário, algo que, certamente, consumirá uma

fração de tempo além do razoável em tarefas de classificação manual de milhares de objetos. O desenvolvimento de algoritmos de classificação automática ou semi-automática desse tipo de objeto textual com uso de *softwares* inteligentes parece uma tarefa não trivial, com relação custo/benefício pouco atraente no contexto, motivos pelos quais foi descartada.

Outro aspecto negativo das categorias de Aristóteles como método de classificação, neste contexto, se refere à pouca intuitibilidade no aprendizado sobre os agrupamentos conceituais de objetos, como resultados da Análise de Conceito Formal.

#### 5.6.1.3 Categorização da engenharia do conhecimento

"Engenharia do Conhecimento" é um termo utilizado nas comunidades de Engenharia de Sistemas e Ciência da Computação para contextualizar a disciplina que se ocupa do ambiente epistemológico de desenvolvimento de sistemas computacionais nos quais o raciocínio inteligente e o conhecimento desempenham papéis centrais. Os assim denominados Sistemas Baseados no Conhecimento *(Knowledge-Based Systems)* são considerados o principal produto da Inteligência Artificial como disciplina, em benefício do mundo dos negócios.

Os métodos desenvolvidos no escopo dessa Engenharia do Conhecimento são utilizados, também, em contextos de Gestão do Conhecimento, Engenharia de Requisitos, Modelagem Empresarial e Reengenharia de Processos de Negócios (SCHREIBER *et al*., 2000).

O terceiro modelo de categorização de objetos testado, baseado então na Engenharia do Conhecimento, é o do método denominado *CommonKADS*, mencionado anteriormente no Capítulo 3. Esse método é promissor porque parte de um princípio derivado da abordagem pragmática do modelo de categorização proposto: a de que uma arquitetura de conhecimento de sistema é algo complexo e dependente do contexto. E, também, porque apresenta uma abordagem estruturada, baseada em alguns poucos princípios emanados da experiência de especialistas práticos em Gestão do Conhecimento.

Schreiber *et al.* (2000) ressaltam que a abordagem *CommonKADS* é encarada pelos seus idealizadores como uma atividade de modelagem, entendendo-se um modelo como uma abstração intencional de uma parte da realidade. O conhecimento deve, então, ser modelado em um nível conceitual, de modo independente dos construtos das linguagens computacionais. Os conceitos utilizados na modelagem do conhecimento devem se referir a e refletir o domínio do mundo real e ser expressos com um vocabulário compreensível pelas pessoas envolvidas no projeto do sistema – eis a razão, portanto, pela qual os artefatos do método utilizam a linguagem natural, outro atributo de interesse para o experimento em questão.

Os desenvolvedores do *CommonKADS* também revelam uma preocupação típica de abordagens de mineração de dados: a descoberta de padrões. Schreiber *et al*. (2000, p. 16-17) argumentam que (grifos nossos):

> *Não é preciso dizer que conhecimento, raciocínio e solução de problemas constituem fenômenos extremamente ricos. O conhecimento pode ser complexo, mas não é caótico: conhecimento parece ter, ao contrário, uma estável estrutura interna na qual podemos ver padrões de similaridade repetindo-se diversas vezes. Apesar de uma arquitetura de conhecimento ser claramente mais complicada que uma de sistemas baseados em regras (...), o conhecimento tem uma estrutura compreensível e isto constitui o gancho prático para se realizar análises de sucesso. Conceitualmente, os modelos de conhecimento nos auxiliam na compreensão do universo da solução de problemas humanos com a elaboração de tipagem de conhecimento. E um importante resultado da moderna engenharia do conhecimento é que o conhecimento humano pode ser analisado em termos de categorias, padrões e estruturas de conhecimento estáveis e genéricas.*

Epistemologicamente, a abordagem *CommonKADS* também se mostra adequada ao pragmatismo do experimento na medida em que define seu conceito de conhecimento:

> *Conhecimento é algo intimamente relacionado com "informação". Diríamos que o fato de um paciente apresentar a temperatura de 39,0 ºC é uma peça de informação e que os médicos têm o conhecimento para deduzir desse fato que o paciente tem febre. De um ponto de vista da engenharia de sistemas, conhecimento é provavelmente melhor compreendido como um tipo especial de informação, denominado "informação sobre informação". O conhecimento nos diz algo sobre certos itens de informação. E uma forma simples de conhecimento é incorporada em hierarquia de subclasses, que tem se tornado uma ferramenta comum na modelagem de dados.*
>
> *(...) O conhecimento pode ser utilizado, freqüentemente, para se inferir novas informações. (...) não há uma linha divisória bem definida entre conhecimento e informação. Conhecimento é "apenas" informação complexa, tipicamente nos dizendo alguma coisa sobre outra informação.* (SCHREIBER *et. al*., 2000, p. 85-86)

O método *CommonKADS* é estruturado com base numa modelagem do conhecimento em três níveis conceituais: contexto, conceito e artefato. Entretanto, o nível de modelagem *CommonKADS* que interessa a esta parte do experimento que embasa a tese é o de contexto, onde a informação disponível deve ser classificada em três "categorias" de modelos analíticos: modelo de organização, modelo de tarefa e modelo de agente. Entretanto, considerando-se que a Inteligência Competitiva será sempre voltada para o ambiente externo, ainda que com as forças internas da organização, acrescentou-se, inicialmente, mais uma categoria para o experimento, denominada "modelo de ambiente externo".

O primeiro experimento de categorização se deu com os objetos sintagmáticos extraídos da coletânea da organização IBM Research (APENSO III-B), com resultados ainda um tanto insatisfatórios em termos de velocidade do procedimento de classificação manual. O "gargalo" de classificação apareceu em se deparando com objetos intuitivamente pouco caracterizáveis como correlatos aos modelos adotados, tais como: "accuracy of the solution", "advances in information technology", "analytics capabilities", "bandwith availability", "BlueStar technology", "cell phone", "crisis situations", "holiday seasons", "level of complexity", "performance factors", "project experience" e outros. Esse tipo de percepção também se fez sentir na classificação dos objetos sintagmáticos de outras duas organizações selecionadas para essa experiência preliminar: Fujitsu Research e Microsoft Research.

Com isso, buscou-se um reestudo do modelo de categorização de objetos de contexto do método *CommonKADS* antes de se prosseguir nos experimentos. Então, analisando-se a natureza dos objetos que, intuitivamente, não se encaixavam bem nas categorias definidas até então, optou-se pela inclusão de mais uma categoria, denominada “modelo de insumo e produto”.

O problema que se apresentou, a seguir, se deveu à ausência de metadados acerca da natureza de cada categoria, cuja solução se buscou na própria taxonomia *CommonKADS* e nos conceitos econômicos de insumos e produtos. Com as definições a seguir, prosseguiu-se nas atividades de “Classificação de Sintagmas” e “Elaboração do Contexto” integrantes do processo expresso na Figura 5.9:

- Modelo de Organização: compreende, para fins de classificação, (i) os atributos de objetos sintagmáticos que se refiram às características da própria organização analisada ou de outras organizações, parceiras ou não, como *IBM’s Haifa Research Lab*, e (ii) os atributos que lembrem, em quaisquer circunstâncias, a noção de organização das coisas no mundo dos negócios, tais como *enforcement of a corporate privacy policy, enterprise clients, nuances of the business world*, e outros.
- Modelo de Tarefa: compreende os atributos de objetos que podem ser considerados como partes de processos produtivos ou de gestão, exceto os insumos e produtos; em geral, substantivos derivados de verbos, representando processos, atividades e tarefas, se referem a esse atributo, como os presentes nos objetos sintagmáticos *distribution operations, supply chain relationships, power of information and communication, use of classification technologies*, e outros.
- Modelo de Agente: compreende os atributos de objetos sintagmáticos que podem ser considerados agentes executores de processos, atividades ou tarefas, representados por seres humanos, organizações, sistemas de informação ou outras entidades capazes de executar ações na própria organização ou numa organização externa (parceira ou não); como exemplos, tem-se *end-to-end solution for the client, automobile manufacturer, part to a customer on time, policy makers, sensor data, system design*, e outros.
- Modelo de Insumo e Produto: compreende os atributos de objetos sintagmáticos que se referem a insumos ou produtos dos processos produtivos da própria organização ou de organizações externas (parceiras ou não), excetuando-se aqueles que podem ser disponíveis a qualquer organização, como recursos naturais, pessoas, estatísticas publicadas, etc, podendo se referirem a objetos do mundo físico, como em *conference hall, container logistics optimization, device platforms, research facilities*, ou objetos abstratos, como em *entity analytics, e-science efforts, growth plan, help of IT solutions, product cycles times*, e outros.
- Modelo de Ambiente Externo: compreende quaisquer atributos de objetos sintagmáticos de organizações posicionadas no ambiente externo à própria organização analisada,

excluindo-se aqueles que se referem a objetos genéricos que podem se relacionar a todas as organizações, como conhecimento científico publicado, questões ambientais, questões econômicas, regulamentação governamental, etc; como exemplos nos contextos do experimento, tem-se *access for people, Cambridge Lab, client experience, cutting-edge technologies, energy grids, research institutes in Japan, spare parts business*, e outros.

### 5.6.2 Classificação de objetos sintagmáticos

Os objetos sintagmáticos obtidos na mineração de texto, para modelagem de conceitos, nem sempre representa uma atividade trivial devido ao nível de abstração dos objetos. Embora se possam desenvolver critérios de classificação ao longo de uma etapa de treinamento da equipe de engenharia do conhecimento, com regras estáveis e satisfatórias, inicialmente o problema da classificação se apresenta de modo evidente, exigindo uma solução.

Esse tipo de problema, denominado "processo de elicitação do conhecimento" (SCHREIBER *et. al*., 2000), ou, ainda, de "socialização do conhecimento" (NONAKA e TAKEUCHI, 1997), também se apresenta em outras abordagens de modelagem de conceitos com uso de quaisquer outros métodos de classificação, como relatado por Schreiber *et al*. (2000, p. 187-214):

> *Elicitação do conhecimento é o processo de obtenção dos dados necessários para a modelagem do conhecimento. Existe uma variedade de técnicas de elicitação. (...) entrevistas, análise de protocolo, laddering, ordenação de conceitos e grades de repertório. (...) A elicitação do conhecimento compreende um conjunto de técnicas e métodos que tentam elicitar o conhecimento de um especialista em algum domínio por meio de alguma forma de interação direta com o especialista.*
>
> *É importante ter em mente que não se deve tentar obter reais descrições formais com uso de uma técnica de elicitação. Impor-se representações formais na elicitação tipicamente conduz a forte viés no processo de elicitação e frequentemente resulta em dados ruins. Elicitação deve ser focada e estruturada, mas também tão aberta quanto possível. É a tarefa de modelagem do conhecimento para converter o material elicitado em uma descrição mais formal do processo de solução do problema.*
>
> *As pessoas que conduzem a elicitação e análise do conhecimento, os engenheiros do conhecimento (também chamados "analistas do conhecimento"), tipicamente não são pessoas com um profundo conhecimento do domínio da aplicação. No caso mais simples, o engenheiro do conhecimento pode ser habilitado para extrair informação de uma variedade de recursos não humanos: livros-texto, manuais técnicos, estudos de caso e assim por diante.*

As técnicas que, pela sua natureza, apresentam recursos que poderão ser úteis para apoiar os engenheiros do conhecimento na tarefa de classificação de objetos sintagmáticos, como objetos do negócio, em conceitos e modelos conceituais no estilo *CommonKADS* adotado nesta tese, são *laddering*[75] e ordenação de conceitos *(concept sorting).* Schreiber *et. al*. (2000, p. 199-200) explicam como funcionam essas duas técnicas de elicitação/socialização do conhecimento:

[75] Em uma interpretação baseada no texto de Schreiber *et al.* (2000, p. 199), *laddering* poderia ser traduzida como *negociação de conceitos de domínio com modelagem gráfica*.

- *Laddering:*

> *Laddering é uma técnica um tanto planejada e você deverá explicá-la inteiramente ao especialista antes do início. O especialista e o engenheiro do conhecimento constróem uma representação gráfica do domínio em termos das relações entre o domínio e os elementos de solução do problema. O resultado é um gráfico qualitativo de duas dimensões onde os nodos são conectados por arcos rotulados. O gráfico toma a forma de uma hierarquia de árvores. Nenhum método de elicitação a mais é utilizado neste caso, mas o especialista e o elicitador constróem o gráfico em conjunto mediante negociação.*
>
> *O ponto-chave é que, tendo-se adquirido alguns dos termos-chave no domínio, organizá-los em alguma forma de estrutura é uma coisa natural a se fazer. (...) Os termos "conceito" e "atributo" devem ser interpretados de modo flexível no contexto de laddering. Por exemplo, nenhuma distinção estrita necessita ser feita ainda entre "conceitos" e "instâncias" (algo que é difícil nas fases iniciais de modelagem do conhecimento). (...) Os tipos de objetos podem ser definidos pelo usuário e se postarem disponíveis como marcadores de texto.*

- Ordenação de conceitos:

> *Ordenação de conceitos é uma técnica útil quando se deseja descobrir diferentes formas como um especialista vê os relacionamentos entre um conjunto fixo de conceitos. Na versão mais simples, é apresentado a um especialista um número de cartões escrito, em cada um, uma palavra conceitual. Os cartões são embaralhados e é solicitado ao especialista ordenar os cartões ou em um número fixo de pilhas ou em algum número de pilhas que o especialista considere apropriado. O processo é repetido muitas vezes.*
>
> *Utilizando esta técnica tenta-se obter múltiplas visões da organização estrutural do conhecimento mediante solicitação ao especialista para executar a mesma tarefa várias vezes. Cada vez que o especialista ordena os cartões ele deve criar ao menos uma pilha que difere, de algum modo, das ordenações anteriores. O especialista deve também providenciar um nome ou rótulo de categoria para cada pilha de ordenação.*
>
> *Variantes da ordenação simples são diferentes formas de ordenação hierárquica. Em uma dessas versões solicita-se ao especialista produzir, em primeiro lugar, duas pilhas de cartões; na segunda sessão de ordenação, três; então, quatro, e assim por diante. Finalmente, solicita-se ao especialista informar se existem duas pilhas com algo em comum. Caso existirem, tem-se isolado um conceito de ordem mais alta que pode ser usado como base para futura elicitação.*
>
> *(...) É rápido para se aplicar e fácil de se analisar. Força os construtos subjacentes ao entendimento de um especialista a um formato explícito. E uma ordenação pode conduzir o especialista à visão de estrutura do domínio que ele mesmo não tinha, conscientemente, articulado anteriormente.*
>
> *(...) A ordenação de conceitos pode descobrir novos conceitos e atributos e é, a propósito, particularmente útil na construção de um esquema de domínio em domínios não familiares.*

Com a técnica de *laddering*, os engenheiros do conhecimento poderiam, por exemplo, contar com especialistas no domínio do negócio para dirimirem eventuais dúvidas sobre a classificação mais adequada de um determinado conceito baseado em sintagmas minerados dos textos. E com a ordenação de conceitos, poderiam desenvolver entre os próprios membros da equipe de inteligência, ou envolvendo, também, especialistas no domínio, várias sessões para se obter uma média de decisões sobre classificação das instâncias de conceitos mais problemáticas.

Contudo, entende-se que malgrado não se possa obter precisão absoluta, ou algo próximo disso, em termos genéricos de classificações de objetos, deve-se manter uniformidade de aplicação dos critérios de classificação. As engenharias em geral, como a Engenharia do

Conhecimento, por exigirem níveis de precisão de acordo com a necessidade do contexto, suportam aproximações que devem se coadunar com os seguintes lemas pragmáticos (SCHREIBER *et al*., 2000, p. 112):

> *Conhecimento pode ser frequentemente usado para se inferir nova informação. (...) Conhecimento é "apenas" informação complexa, tipicamente dizendo-nos algo sobre outra informação. (...) O raciocínio sempre tem uma "razão". Em outras palavras, um importante aspecto do conhecimento é o que queremos fazer com ele.*

### 5.6.3 Elaboração do contexto

O passo seguinte do experimento, com base na metodologia de Análise de Conceito Formal, se concentrou na atividade de elaboração dos contextos apresentados nos APENSOS III-A, III-B e III-C, com uso de *software* aplicativo de planilha eletrônica *(Microsoft Excel)* e de Análise de Conceito Formal *(ToscanaJ).* Essa atividade, após o procedimento de classificação dos objetos manualmente, não apresentou dificuldade, mas se revelou morosa e tediosa na medida em que a entrada de dados no aplicativo de Análise de Conceito Formal utilizado teve que ser realizada como numa planilha, marcando-se nas respectivas "células" de interface do usuário as incidências de atributos em objetos (o módulo de operação de entrada de dados em lote, no *ToscanaJ*, apresenta certa complexidade e requer tempo considerado demasiado no contexto, com formatos de dados de tecnologia proprietária).

Observe-se que esses contextos mostram todos os objetos sintagmáticos minerados dos textos de três organizações de um mesmo nicho de mercado, que serão analisados em conjunto, no Capítulo 6, como uma simulação prática de Gestão da Informação e do Conhecimento num ambiente de Inteligência Competitiva.

É importante destacar-se, neste ponto, algumas propriedades estruturais interessantes dos contextos elaborados com essa metodologia:

I. Indexação temática: como os sintagmas plurinominais representam expressiva parte dos substantivos/nomes presentes nas coletâneas, os contextos podem ser utilizados para recuperação de informação do negócio publicada nos *Web* Portais corporativos; em algumas situações os contextos de objetos sintagmáticos podem funcionar, também, como listas invertidas de busca a partir de um conjunto de objetos, como no caso do substantivo *business*, que pode remeter o usuário a dezenas de objetos contendo esse termo em sua composição sintagmática (como *business applications, business decision, business goals, business networks*, etc).
II. Apresentação polissêmica dos substantivos/nomes: com os sintagmas plurinominais, observam-se, com freqüência, os mesmos substantivos compondo mais de um sintagma, mostrando-se as possibilidades semânticas utilizadas nos textos, inclusive as

opções mais freqüentes no contexto (essas possibilidades, como se observou no experimento, geralmente são poucas devido ao próprio contexto do negócio).

III. População de ontologias: como existem muitos substantivos/nomes que aparecem com freqüência maior que um, pode-se observar sintagmas plurinominais que permitem a identificação e população de ontologias; como exemplo, o conceito de *client* pode gerar, numa hierarquia de conceitos, um objeto ontológico "pai" com os "filhos" *client engagements*, *client experience*, *client projects*, *client solutions*, *client's e-commerce*, *client's money*, etc.

IV. Caracterização da organização: a partir da observação dos substantivos/nomes mais freqüentes no contexto, pode-se descrever a organização e seu ambiente de negócios interpretando-se esses substantivos/nomes como seus atributos, algo como os "cromossomos" que revelam o "DNA da organização" (FULD, 2007).

V. Composição de operadores de busca: os sintagmas podem ser utilizados como "sementes" sintáticas, ou argumentos, de busca em outras fontes por meio de *browsers* (*softwares* buscadores) na *Web*, para complementação da informação necessária para a elaboração de conceitos mais ricos em significados (com uso eventual de Gráficos Conceituais, envolvendo também verbos e outros funtores sintáticos).

### 5.6.4 Geração do reticulado

A atividade seguinte no processo consiste na geração do reticulado *(lattice)* de conceitos com uso do aplicativo *ToscanaJ*, como o exemplo da Figura 5.10, do contexto da organização IBM Research. O gráfico apresenta um círculo na parte mais alta da rede, indicando o início, e círculos que concentram, mas abaixo, os retículos nos pontos em que estes se cruzam, compondo um modelo de informação conceitual sobre os objetos e seus atributos no contexto.

Conforme o método da Análise de Conceito Formal, cada círculo intermediário do reticulado representa um conceito, agrupando objetos com os mesmos atributos, e as linhas do reticulado que unem os círculos representam as relações matemáticas entre esses conceitos. O modelo conceitual da Figura 5.10 apresenta uma alta densidade (ou "poluição") de conceitos e relações porque o número de objetos utilizado, sem um "corte" por faixa de freqüência, é relativamente elevado (perto de mil) para esse tipo de representação da informação.

Observe-se que logo abaixo do primeiro círculo no topo do gráfico da Figura 5.10 aparece o círculo indicando o conceito de "Modelo de Insumo-Produto", que agrupa um conjunto de objetos sintagmáticos com os mesmos atributos no contexto. As caixas de texto, pela restrição de espaço no modelo pictórico, mostram apenas os três primeiros objetos, mas essa interface do aplicativo apresenta um "botão" onde ao se "clicar" no mesmo aparecem os demais objetos do conceito. Os conceitos mais abaixo resultam de diferentes composições de mais de um atributo como, por

exemplo, o conceito resultante da composição dos atributos do “Modelo de Insumo-Produto” e “Modelo de Ambiente Externo”. Os primeiros objetos da lista que compõem esse conceito são: *advances in processing and algorithms* e *advent of technologies*.

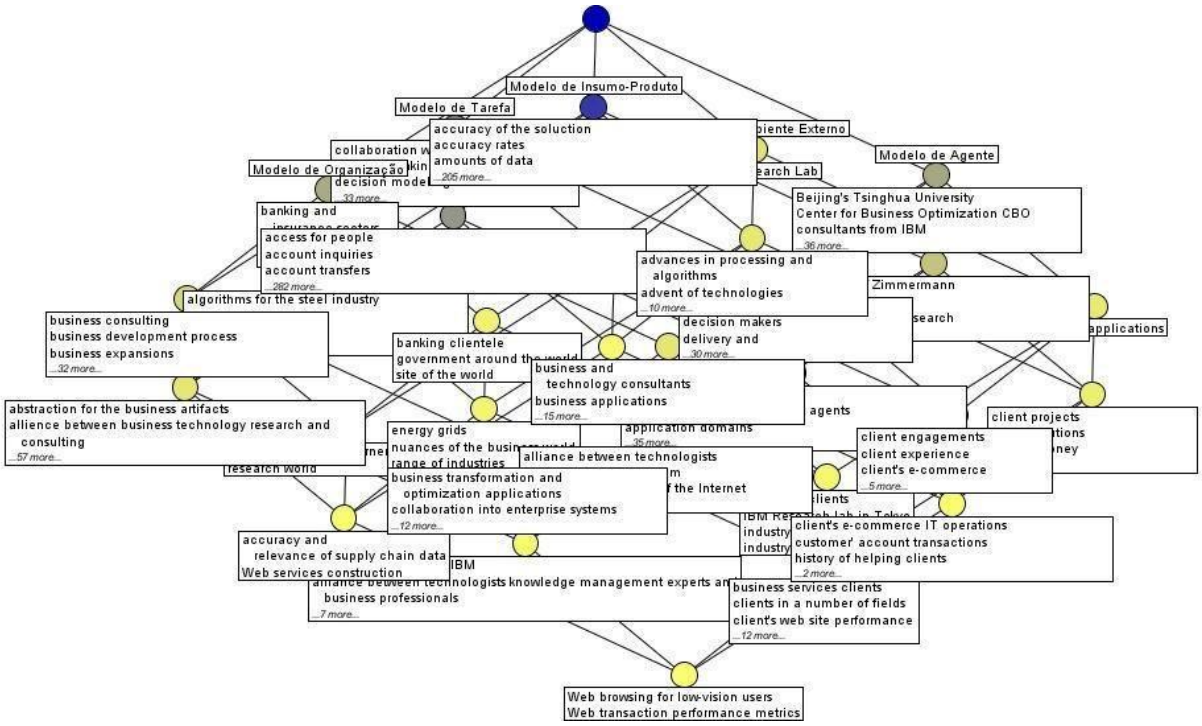



Figura 5.10 Reticulado de Conceitos

(Fonte: do autor da tese)

A noção de *differentiae* de Aristóteles implica que os conceitos situados na parte mais acima no reticulado são mais genéricos, compostos por objetos com menos atributos no contexto, e os conceitos mais abaixo na rede são mais específicos, compostos por objetos com mais atributos no contexto. Essa trama, algo como uma “grelha do ser” de Delbecque (2006), relaciona e mostra ao analista, de modo visual, todas as relações entre conceitos e propriedades ontológicas suportadas pelo método: hiperonímia-homonímia, extensão, intensão, superconceito, subconceito.

Como medida para “despoluição” do reticulado, para clarear o cenário de conceitos, o engenheiro do conhecimento pode optar, num determinado momento, pela supressão das caixas de textos com os objetos classificados e trabalhar apenas com os títulos de atributos e os círculos de conceitos derivados (Figura 5.11). Em particular, isso é recomendado para análises racionais dos conceitos *a priori* dos objetos que o integram, ou seja, análises onde se busca deduzir as implicações racionais dos conceitos a partir de seus atributos e relações hierárquicas na rede. Esse tipo de análise poderia levar o engenheiro do conhecimento a questionar, por exemplo, o significado de um conceito cujos objetos combinassem os atributos de “Modelo de Tarefas” e “Modelo de Insumo-Produto”, como ocorre, de fato, no reticulado da Figura 5.11. Ou, de modo mais complexo ainda, o significado de um conceito cujos objetos em extensão apresentassem os

atributos intensionais: “Modelo de Tarefa”, “Modelo de Insumo-Produto”, “Modelo de Agente” e “Modelo de Ambiente Externo”.

Contudo, considerando-se a extensão do esforço necessário, as possíveis interpretações dos reticulados gerados no experimento (e dos modelos de informação visual conseqüentes), em contextos de Inteligência Competitiva, serão discutidas no próximo capítulo. O item seguinte deste capítulo apresenta uma introdução a esses contextos de inteligência e suas características com base na literatura.

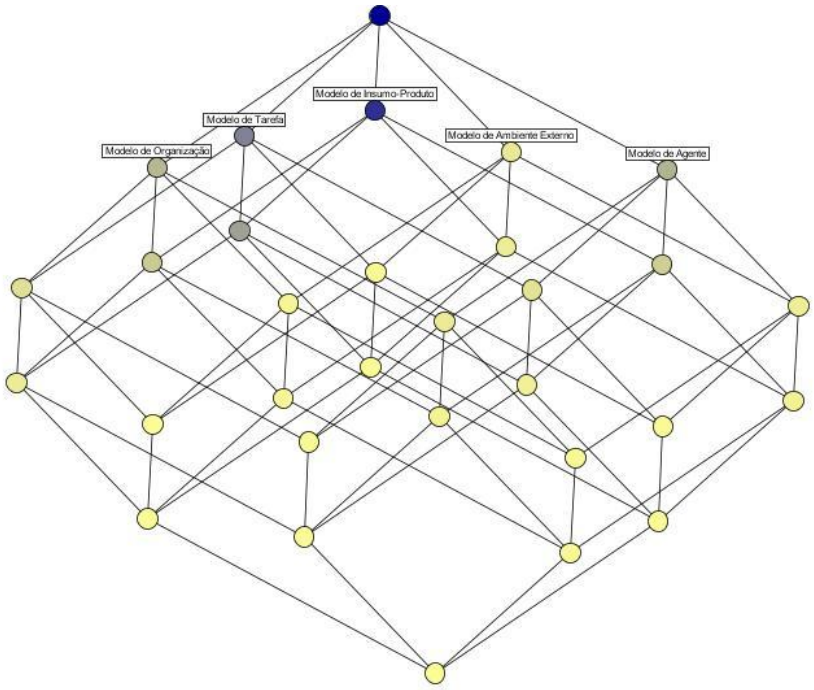



Figura 5.11 Reticulado de Conceitos sem Objetos
(Fonte: do autor da tese)

### 5.6.5 Inteligência analítica e estratégica

Os últimos passos do processo experimental expresso na Figura 5.9 consistem na análise dos conceitos dos reticulados, busca de informações específicas para complementação e ampliação dos modelos anteriores e geração de um modelo mais completo de informação para suporte à Gestão do Conhecimento e Inteligência Competitiva, denominado, nesta tese, “mosaico semântico para Inteligência Competitiva” (utilizado como metáfora, algo parecido com a tela de pintura impressionista inspiradora de Fuld (2007)). Os artefatos, ou modelos gráficos, que comporão esse mosaico se prestam, alguns, a funções analíticas, e outros, a funções estratégicas: por exemplo, os modelos de organização e de processos seriam mais analíticos; os modelos temáticos de produtos e parcerias, ou de modelos envolvendo o ambiente externo, seriam mais estratégicos.

A análise dos processos produtivos das organizações se destina ao estudo dos custos de produção, requisito de inteligência competitiva em mercados de *commodities*, e a análise dos modelos de produtos e parcerias é indispensável para competição em mercados marcados pela diferenciação de produtos ou em mercados de nicho. Fuld (2007, p. 95) coloca a questão dos processos organizacionais, no contexto da Inteligência Competitiva, do seguinte modo:

> *O processo é a mãe dos custos. Compreendendo o processo de um concorrente você pode derivar sua estrutura de custos – e muito mais. Ao juntar as peças de um processo de uma empresa você pode aprender como melhorar a produção de seu próprio produto ou incrementar seu serviço. Conhecer o processo de seu concorrente pode ensiná-lo a respeito do pensamento de gestão e a direção estratégica dele (...). Há mais na análise de custo do que simplesmente colocar peças juntas em um piso plano para representar como a sua empresa produz os produtos e serviços. Seja um banco, uma empresa de software ou uma fabricante de pepperoni, conhecer como, onde e – o mais importante – porque o concorrente gasta dinheiro em uma área ou função em particular pode ensinar a você muito a respeito de como ele pensa estrategicamente, quais mercados deseja conquistar e como pretende chegar lá. Compreenda o processo e você provavelmente entenderá como o seu concorrente se comporta hoje e no futuro. Seus custos ditam o quanto de dinheiro você dispõe para investir em P&D, conduzir campanhas de marketing e publicidade e oferecer incentivos para sua área de vendas e outras.*

Quanto ao nível estratégico da análise de contextos competitivos, Fuld (2007, p. 63-64) adota o modelo analítico de estratégia de Porter, segundo o qual:

> *(...) existem três estratégias genéricas que as empresas podem usar para serem mais lucrativas que a média das empresas no mesmo setor: (1) elas podem diferenciar-se e, assim, obter elevados lucros por meio de seus produtos, (2) elas podem ter baixos custos, ou (3) elas podem focar em um determinado segmento do mercado ou em um nicho geográfico no qual haja menos pressão dos concorrentes e das forças externas.*

O modelo de Porter, comentado por Fuld (2007) como um ecossistema, é baseado na análise de contexto competitivo. Esse autor assim define a importância da análise de contexto:

> *O contexto é tudo em um jogo. É o meio pelo qual você analisa a inteligência e o ímpeto estratégico de seus adversários. O contexto que consideramos ideal para um jogo de guerra são as cinco forças e o modelo dos quatro componentes da análise da concorrência de Michael Porter.*

Os jogos de guerra são eventos onde várias equipes, representando organizações num mercado concorrencial, tentam esboçar estratégicas competitivas para superar seus adversários e expô-las para um grupo moderador do jogo. Fuld (2007, p. 53) assim define esse recurso metodológico:

> *Um jogo de guerra não significa apenas ganhar ou perder. Significa retirá-lo de sua zona de conforto e ajudá-lo a obter uma atual e realista visão da paisagem competitiva. Isso permite a você testar a extensão da força de sua própria estratégia e examinar melhor a de seu concorrente. (...) é uma simulação viva que reflete o mundo real, forçando concorrentes, clientes e possíveis entrantes a exporem suas estratégias.* (FULD, 2007, p. 53)

A elaboração das estratégias deve valer-se de toda informação considerada útil no contexto, tais como notícias de fatos e tendências, além de cenários evolutivos do mercado e das próprias organizações. O grupo moderador deve estabelecer as regras do jogo, conduzir as atividades (coordenadas por um facilitador) e avaliar as estratégias apresentadas pelos grupos concorrentes, atribuindo, ao final, uma menção de competência a cada um pela força da estratégia apresentada.

Como exemplo, Fuld (2007) apresenta o contexto e os resultados de um conhecido jogo de guerra, realizado por alunos da Harvard University e do Massachusetts Institute of Technology (MIT), envolvendo possíveis estratégias competitivas das empresas Google, Yahoo!, America On Line (AOL, uma divisão do grupo de mídia Time Warner) e Microsoft. O bilionário mercado em questão era o de conteúdos na *Web*, composto por produtos e serviços como busca inteligente, plataformas tecnológicas (e de redes de usuários, portanto) e disponibilização de conteúdos específicos para os clientes (como os de economia), além da análise de custos e preços, entre outros fatores de negócio.

As cinco forças de Porter, comentadas por autores como Fuld (2007) e Greenwald e Kahn (2005), são:

- competição entre os jogadores atuais;
- ameaça de novos entrantes;
- ameaça de substitutos;
- poder dos compradores; e
- poder dos fornecedores.

O modelo dos quatro componentes passa pela análise de (1) metas futuras, (2) hipóteses, (3) estratégia e (4) capacidade. As metas constituem a motivação objetiva dos executivos em termos de resultados colimados para a organização no tempo, as hipóteses são as crenças que uma organização tem de seu setor de atuação, a estratégia o "mapa da viagem" que uma empresa escolhe para alcançar lucratividade e valor econômico real, e a capacidade o conjunto de recursos corporativos que a empresa traz para o campo de batalha, tais como capital, pessoas, distribuição global, sua linha de produtos e uma infra-estrutura de informação (FULD, 2007).

Essa área do conhecimento é constituída, basicamente, por crenças empíricas (ou "de trincheiras") de gurus da Inteligência Competitiva como Fuld (2007), Greenwald e Kahn (2005) e outros, e algumas divergências de pensamento merecem destaque. Fuld (2007, p. 60-61), por exemplo, não acredita no modelo de análise SWOT *(Strength, Weakness, Opportunity, Threat)* para o esboço de estratégias competitivas:

> *O que é interessante nos modelos de Porter é o fato de estes serem baseados na idéia de que toda empresa é parte de um amplo ecossistema industrial e que deve agir nesse ambiente. Antes de Porter – isto é, antes do final dos anos 70 – um estrategista podia esboçar as forças, fraquezas, oportunidades e ameaças de uma empresa (...) e comparar as informações dessas quatro caixas com informações similares a respeito de seus concorrentes. Mas este era um de muitos cenários, o que tornava difícil enxergar verdadeiramente o efeito que a descrição de uma SWOT tinha sobre outra. Isso não dava garantia de que você incluiria todos os fatores que pressionavam você e seus concorrentes, tal como mudanças regulatórias, novas tecnologias e força crescente dos clientes. (...) SWOT é somente um conjunto nítido de listas, e não mais do que isso. Algumas vezes, interpretar uma matriz SWOT é como ler folhas de chá; você as interpreta da forma que deseja.*

Greenwald e Kahn (2005, p. 5) reconhecem o valor do modelo de cinco forças de Porter, mas adotam uma visão simplificada em relação à importância de cada uma dessas forças:

> *Concordamos com a visão de Porter que cinco forças – Substitutos, Fornecedores, Potenciais Entrantes, Compradores e Competidores na Indústria – podem afetar o ambiente competitivo. Mas, de modo diverso de Porter e muitos de seus seguidores, não pensamos que essas forças são de igual importância. Uma delas é claramente muito mais importante que as outras. É tão dominante que líderes buscando desenvolver e perseguir estratégias vencedoras deveriam começar por ignorar as outras e focar apenas nela. Essa força é constituída pelas barreiras contra a entrada – a força subjacente aos "Potenciais Entrantes" de Porter.*
>
> *Se não há barreiras para entrada, então muitas preocupações estratégicas podem ser ignoradas. A empresa não tem que se preocupar sobre como interagir com competidores identificáveis ou sobre como antecipar e influenciar seus comportamentos. Simplesmente, há muitos deles com que lidar. (...) Sem a proteção de barreiras na entrada, a única opção que uma empresa tem é operar tão eficientemente e efetivamente quanto possível.*
>
> *Efetividade operacional pode ser pensada como uma estratégia, na verdade, como a única estratégia apropriada em mercados sem barreiras contra a entrada.*

Outra visão um tanto discordante e cética de Greenwald e Kahn (2005, p. 6-7), em relação às crenças gerais sobre competitividade empresarial, se refere às fontes da própria competitividade:

> *Em um ambiente global crescente, com menos barreiras comerciais, transportes mais baratos, fluxos de informação mais velozes e competição implacável tanto de rivais estabelecidos como de economias recentemente liberalizadas, pode parecer que vantagens competitivas e barreiras contra entrada deverão diminuir. (...) mas esta macro visão perde uma característica essencial das vantagens competitivas – que vantagens competitivas são quase sempre embasadas no que são essencialmente circunstâncias "locais".*

Concluindo este capítulo, com base na resenha apresentada anteriormente, sobre fontes e usos da informação útil para desenvolvimento de *insights* de Inteligência Competitiva, propõe-se então o mapa contextual da Tabela 5.1, com as incidências de atributos de usos das principais fontes genéricas.

O exercício busca estabelecer algumas prováveis correlações entre as atividades de elaboração de Inteligência Competitiva descritas anteriormente, aqui denominadas "alerta antecipado", "jogos de guerra", "descoberta do DNA corporativo" e "conhecimento de fatias da realidade", e as possíveis fontes de informação.

Tabela 5.1 Fontes e Usos de Informação para Inteligência Competitiva

| **Fontes de Informação** | | **Usos em Inteligência Competitiva** | | | |
|---|---|---|---|---|---|
| **Nº** | **Denominação** | **Alerta Antecipado** | **Jogos de Guerra** | **DNA Corporativo** | **Fatias da Realidade** |
| 1 | Arquivos Públicos | | | | ● |
| 2 | Congressos | ● | ● | | ● |
| 3 | Informação de Conhecimento Tácito | | ● | | ● |
| 4 | Intercomunicação Pessoal | ● | | | ● |
| 5 | Livros e Volumes Consolidados | | | ● | ● |
| 6 | Periódicos de Notícias | ● | ● | | ● |
| 7 | Periódicos Especializados | ● | ● | | ● |
| 8 | Publicações Governamentais | ● | ● | | |
| 9 | Programas de Rádio e TV | ● | | ● | ● |
| 10 | *World Wide Web* (aberta) | ● | ● | ● | ● |
| 11 | *World Wide Web* (invisível) | ● | ● | ● | ● |

O reticulado da Figura 5.12 mostra os conceitos formados com base no contexto da Tabela 5.1, sugerindo-se a seguinte interpretação para essa estrutura conceitual: a fonte de informação mais completa sob todos os aspectos é a *Web*, tanto a visível como a invisível, postada no final do reticulado (última caixa de objetos embaixo). Essa fonte pode ser útil tanto para alertas antecipados como para jogos de guerra, compreensão de fatias da realidade de um contexto competitivo e até para descoberta do DNA corporativo das organizações no mercado. Compreendendo-se a *Web* como multimídia, essa interpretação se sustenta na medida em que admitamos que inclusive os mecanismos de busca de pessoas, áudio e vídeo poderão ser utilizados como recursos de fonte, tais como o *Facebook*, as redes sociais e outros (para, assim, incluir a "Informação de Conhecimento Tácito" como uma intensão do conceito "Web" no reticulado).[76]

Então, lendo-se o reticulado de baixo para cima, dos objetos mais específicos para os mais genéricos em termos de informação, tem-se a impressão, de acordo com o método da Análise de Conceito Formal, que congressos, periódicos e programas de rádio e TV estão no mesmo nível de especialização (talvez com base no conceito de "notícia"), situando-se num nível mais genérico ainda as publicações governamentais, a intercomunicação pessoal e a informação do conhecimento tácito. E que, no topo do reticulado, os arquivos públicos, os livros e volumes consolidados seriam as fontes mais genéricas de todas.

Contudo, isso representaria um paradoxo da Análise de Conceito Formal se não considerássemos a leitura correta dessa árvore de conceitos. Como os conceitos, no caso, serão tão mais genéricos quanto as fontes de informação (que são seus objetos instanciados) forem mais úteis, servindo a várias metodologias de elaboração de Inteligência Competitiva, intuitivamente a *Web* deve ser a fonte mais inclusiva em termos dessas funções, portanto a mais genérica, e os arquivos públicos, livros e volumes consolidados as fontes mais específicas. Os livros e volumes consolidados, por exemplo, pela sua característica essencial de ausência de notícia, serão mais úteis apenas em contextos de conhecimento de fatias da realidade e descoberta do DNA corporativo do negócio.

[76] Fuld (2007), por exemplo, ressalta inclusive a importância de se pesquisar na *Web* conteúdos de maledicência das organizações, citando como um endereço com esse tipo de conteúdo o *http://www.fuckedcompany.com*.

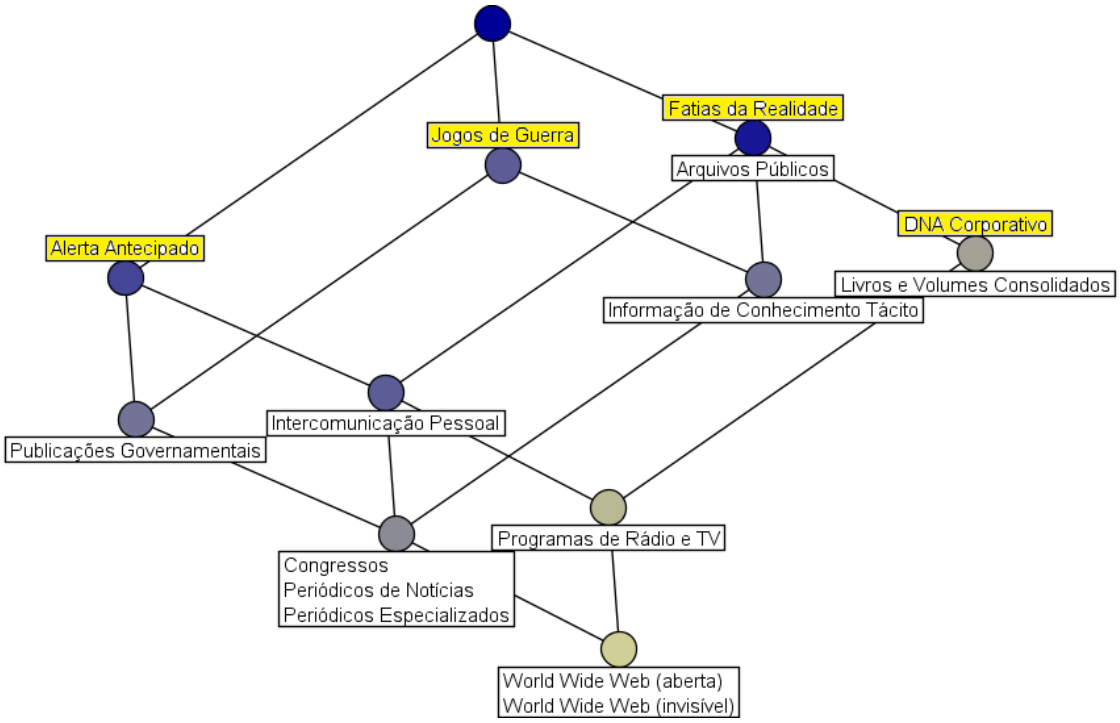



Figura 5.12 Reticulado de Fontes e Usos de Informação para Inteligência Competitiva (Fonte: do autor da tese)

# 6. MOSAICO SEMÂNTICO PARA INTELIGÊNCIA COMPETITIVA

*A inteligência pode ser emergente*.
(FREEDMAN, 1995)

Este capítulo apresenta, ao mesmo tempo, uma análise mais detalhada dos modelos de informação obteníveis a partir da Análise de Conceito Formal, com os resultados da mineração textual na *Web,* e uma consolidação desses modelos no construto gráfico conceitual que se denomina, nesta tese, "mosaico semântico" para suporte ao desenvolvimento de inteligência competitiva nas organizações. É um capítulo destinado à demonstração de como a metodologia proposta pode induzir os engenheiros do conhecimento a pensar sobre os conceitos de negócio minerados, criar significados e desenvolver *insights* para Inteligência Competitiva.

A relevância dos conceitos minerados e modelados nesse mosaico será discutida em cada caso, com base na Teoria da Informação de Shannon (SAYOOD, 2000) e nas teses psicológicas de Bateson (2002), Weick (1995) e Choo (2003) apresentadas no Capítulo 3. Outro aspecto que será desenvolvido neste capítulo se refere aos processos mentais de Engenharia do Conhecimento que podem ser utilizados para solução dos problemas de Inteligência Competitiva nas organizações que exigem raciocínio lógico em contextos informacionais.

Com a análise de modelos de informação específicos, que também podem ser chamados de "modelos conceituais", pretende-se mostrar como esses modelos podem contribuir para o desenvolvimento de uma "inteligência emergente", um conceito de Inteligência Artificial baseado na natureza. Esse conceito sugere, no contexto em estudo, que a partir de fragmentos de modelos de informação conceitual relevante pode-se construir modelos mais completos e complexos, úteis para o desenvolvimento de *insights* em Inteligência Competitiva. Ou seja, propõe-se com esse construto mostrar como se pode obter, com base na Teoria da Informação de Shannon (SAYOOD, 2000) e nas teses de Bateson (2002), Weick (1995) e Choo (2003), "diferenças que fazem diferenças" e "informação sobre informação" útil para Inteligência Competitiva.

Entretanto, surge o seguinte desafio de ordem prática: como integrar os conceitos minerados e com atributos classificados no modelo *CommonKADS* Modificado, conforme os APENSOS III-A, III-B e III-C, a um modelo de uso em Inteligência Competitiva? Como resposta, a Tabela 6.1 e a Figura 6.1 apresentam, respectivamente, uma proposta de classificação dos conceitos obtidos na Análise de Conceito Formal, com base nas cinco forças de Porter (1980) em contextos de Inteligência Competitiva, e o reticulado de conceitos correlato. Os conceitos agrupados em "Estruturas Corporativas", "Processos Produtivos", "Concorrentes", "Forças do Mercado" e "Outros" são então considerados os objetos de análise e as forças de Porter, onde os conceitos se "encaixam" pela sua utilidade, são considerados os respectivos atributos desses objetos, buscando-se correlações entre os conceitos e sua utilidade no modelo analítico de Porter.

Com isso, sugere-se que os conceitos formais compostos por objetos sintagmáticos envolvendo a própria organização que realiza o estudo, assim como outros objetos relativos a formas de organização em geral (como a própria *Web*), com ou sem outros atributos do modelo *CommonKADS* Modificado, se referem a organizações de "jogadores" (ou *players*) no mercado, assim como a seus clientes e parceiros. Como exemplos de objetos sintagmáticos que integram essa classificação (conforme os conteúdos do APENSO III-A, APENSO III-B e APENSO III-C), reunidos num macro-conceito denominado "Estruturas Corporativas" na Tabela 6.1, têm-se: *alliance between technologists knowledge management experts and business professionals, business activity monitoring, bulk of Microsoft researchers, chairman of Microsoft, consultants from IBM, eResearch Infrastructure Council, Fujitsu Group, management issues in partnership, product team with goals* e outros.

Tabela 6.1 Análise de Conceito Formal com as Cinco Forças de Porter

| **Objetos Conceituais** | | **Atributos Competitivos (Forças de Porter)** | | | | |
|---|---|---|---|---|---|---|
| **Macro-conceito Formal** | **Classificação com Atributos do Modelo CommonKADS Modificado** | ***Players Atuais*** | **Substitutos** | **Entrantes** | **Clientes / Consumi-dores** | **Parceiros** |
| **Estruturas Corporativas** | Organização | ● | | | | |
| | Organização, Tarefa | ● | | | | |
| | Organização, Agente | ● | | | | ● |
| | Organização, Insumo&Produto | ● | | | | |
| | Organização, Tarefa, Agente | ● | | | | ● |
| | Organização, Tarefa, Insumo&Produto | ● | | | | |
| | Organização, Agente, Insumo&Produto | ● | | | | ● |
| | Organização, Tarefa, Agente, Insumo&Produto | ● | | | | ● |
| **Processos Produtivos** | Tarefa | ● | | | | |
| | Agente | ● | | | ● | |
| | Insumo e Produto | ● | | | | |
| | Tarefa, Agente | ● | | | ● | |
| | Tarefa, Insumo&Produto | ● | | | | |
| | Agente, Insumo&Produto | ● | | | ● | |
| | Tarefa, Agente, Insumo&Produto | ● | | | ● | |
| **Concorrentes** | Organização, Ambiente Externo | ● | ● | ● | | |
| | Ambiente Externo, Agente | ● | ● | ● | ● | |
| | Ambiente Externo, Tarefa, Agente | ● | ● | ● | ● | |
| | Ambiente Externo, Agente, Insumo&Produto | ● | ● | ● | ● | |
| | Ambiente Externo, Tarefa, Agente, Insumo&Produto | ● | ● | ● | ● | |
| **Forças do Mercado** | Ambiente Externo, Tarefa | | ● | ● | | |
| | Ambiente Externo, Insumo&Produto | | ● | ● | | |
| | Ambiente Externo, Tarefa, Insumo&Produto | | ● | ● | | |
| **Outros** | Organização, ..., Ambiente Externo | ● | ● | ● | ● | |

Objetos como *Fujitsu Group, consultants from IBM* e *bulk of Microsoft researchers* se referem a organizações específicas, mas objetos como *business activity monitoring* e *management issues in partnership* se referem a outras formas de organização genéricas: *business* e *partnership*. Objetos que apresentam atributos de agentes, nesse macro-conceito, são considerados como provavelmente referentes a clientes e parceiros da organização no mercado.

Como exemplo explanatório, um objeto (referente) sintagmático como “Fujitsu Research Institute”, que se refere apenas ao macroconceito (ou ideia) de “Estruturas Corporativas, tem o sentido de uma “Organização” concorrente no contexto de mercado em estudo, por isso devendo ser classificado em “Players Atuais” nas Forças de Porter.

Em outro exemplo, objetos (referentes) que são classificados, em termos de macroconceito e sentido, como “Estruturas Corporativas” e “Organização, Agente” (Tabela 6.1) têm atributos de Porter como “Players Atuais” e, também, como “Parceiros”. Ou seja, esses objetos se referem a organizações que estão no mesmo nicho de mercado que a patrocinadora do esforço de Inteligência Competitiva, mas não são concorrentes, compondo a própria macroestrutura operativa dessa organização.

O macro-conceito seguinte, denominado “Processos Produtivos”, se refere a processos produtivos, atores, insumos e produtos, mas sem referência a qualquer organização, nem mesmo à organização que realiza a análise de inteligência. Considera-se, neste exercício, que os conceitos com tais atributos (tarefa, agente, insumo & produto) são utilizados, nos *Web Portais*, referindo-se ao nicho competitivo onde se encontra a própria organização que patrocinou a mineração de conceitos, assim como seus atuais concorrentes (sendo todas, portanto, *players* nesse mercado). Exemplos extraídos dos contextos estudados são: *after-sales service management, asset monitoring and management, data exploration, expertise in reasoning techniques, information systems, logistics engineering, model solutions for management, production systems, risk assessment, shopping assistant technology, steps toward implementation, system management, warning system for automobiles* e outros.

O bloco de macro-conceito seguinte, denominado “Concorrentes”, classifica os conceitos e objetos que se referem a organizações e pessoas externas à própria organização patrocinadora que, de algum modo, podem ser ou se tornar: substitutas das atuais concorrentes, novas concorrentes (ou “entrantes”), clientes e/ou parceiras. A incidência de atributos correlatos a “ambiente externo” é que determina essa classificação. É importante ressaltar-se, também, que os “parceiros” são considerados como integrantes da própria corporação, motivo pelo qual os objetos de “ambiente externo” não recebem marcação de incidência na respectiva coluna da Tabela 6.1. Como exemplos, pode-se citar: *Hu Jintao Administration, Takeuchi Keiji, automobile manufacturer, brokers with the confidence, customer demands, customer value drivers, end user, end-to-end solution for the client, health care providers, history of helping clients, household products retailer,*

*innovation to clients, oil field operators, privacy of the user, products to costumers, steel manufacturers* e outros.

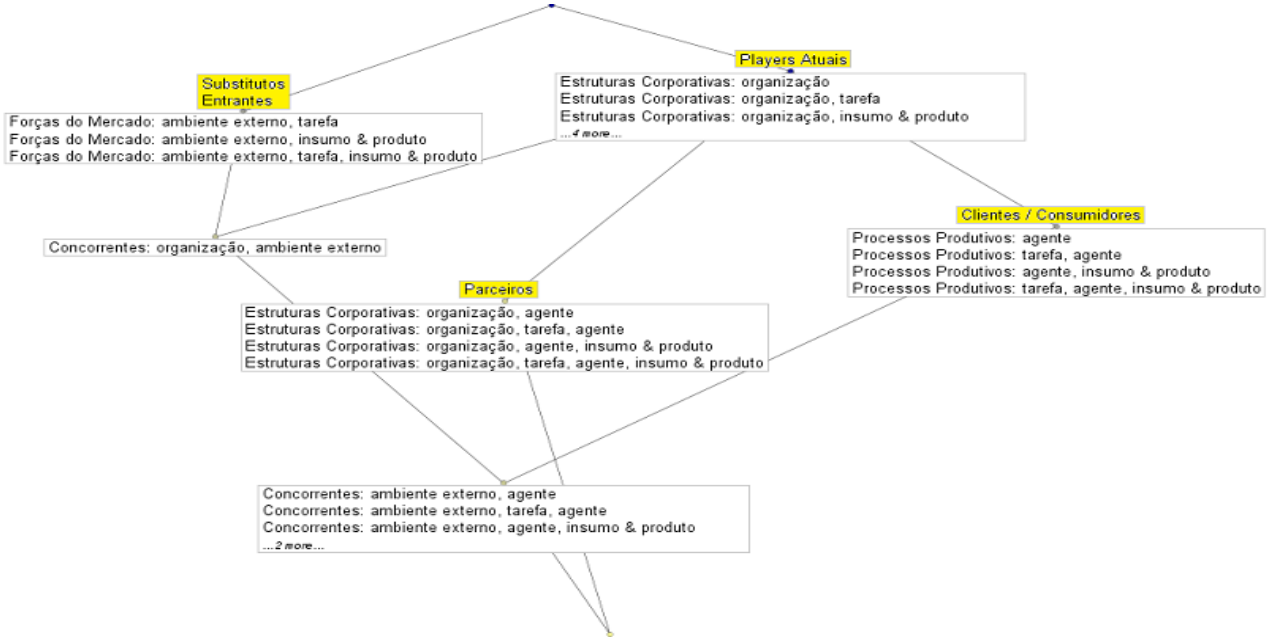



Figura 6.1 Reticulado de Conceitos: *CommonKADS* Modificado & Forças de Porter
(Fonte: do autor da tese)

O penúltimo bloco de macro-conceitos – “Forças de Mercado” – se refere a conceitos gerais mais citados pelas organizações estudadas, mas sem vinculação com organizações, clientes ou parceiros, constituindo-se, geralmente, conceitos corporativamente “neutros”, mas conhecidos como de impacto geral na sociedade. Esses conceitos são assim classificados por não conterem atributos de “agente” nos sintagmas, apesar de se referirem ao “ambiente externo” à organização, tendo-se como exemplos: *air and water quality, body of regulations, chip technology, climate change, enrollment of women, ideas around the world, Immunodeficiency Syndrome, research efforts in nanotechnology, sales revenues, San Francisco, science fairs, silicon chip, spare parts inventory* e outros.

O reticulado da Figura 6.1 é um recurso para orientação preliminar do analista caso queira partir de uma concepção funcional dos objetos de mineração, tal como a função “Substitutos Entrantes”. Com essa função em mente, o analista (Engenheiro do Conhecimento) poderá, talvez, se interessar pelos referentes do macroconceito e sentido “Forças do Mercado: Ambiente Externo, Insumo & Produto”, que no contexto “Fujitsu Research” são (APENSO III-A): “Asia strategy”, “changes in society”, “China-India automobile market”, “China’s future”, “climate change policy”, “energy and environment policy”, “market potential”, “market risk” e “targets for Copenhagen”. Como se pode observar no sentido desses referentes, esta superclasse se refere a temas

(conceitos) do ambiente externo à organização e, também, a insumos e produtos desse meio (que poderão, certamente, impactar profundamente a estratégia do negócio da "Fujitsu Research").[77]

Em suma, esse reticulado é um modelo de representação gráfica da informação na mineração conceitual que permite classificações com vários "cortes" dimensionais, associando-se aspectos funcionais dos referentes (objetos sintagmáticos), com base no construto da Engenharia do Conhecimento, com as Forças de Porter.

O último bloco de macro-conceitos corresponde a conceitos mais completos, contendo todos os atributos do modelo de análise *CommonKADS* modificado, mas composto por sintagmas raros nos textos, tendo sido observados relativamente poucas vezes entre os sintagmas que contêm os 2,5% de substantivos mais freqüentes estudados no experimento. Exemplos: *alliance between technologists knowledge management experts and business professionals, expansion of the Internet, Web browsing technology, Turing Award winner's research, Web and mining logs of Internet queries in search.*

O reticulado de conceitos formais da Figura 6.1 deverá servir de guia para análise dos fragmentos de modelos de informação apresentados a seguir, representando uma plataforma meta-ontológica para Inteligência Competitiva. Observa-se (com uma leitura de cima para baixo), no reticulado, que os objetos conceituais se dividem, originalmente, em um bloco (caixa mais escura) relativo a organizações substitutas ou entrantes (novas) no mercado e a organizações consideradas concorrentes *(players)* atuais. Em seguida, os atributos (ou "intensões") desses objetos são compartilhados, também, por organizações externas (bloco de "Clientes / Consumidores") e internas (bloco de "Parceiros"), derivando também um tipo de organização concorrente mais genérico (sem nome de bloco), que herda atributos tanto das organizações do bloco "substitutos e entrantes" como do bloco "*Players* Atuais".

Os conceitos de "parceiros" e "clientes e consumidores" se originam do conceito de "*players* atuais" e os conceitos de "substitutos", "entrantes", "*players* atuais" e "clientes / consumidores" originam, de modo combinado, um segundo bloco conceitual de concorrentes identificados por objetos do ambiente externo que se referem a agentes no modelo *CommonKADS* Modificado.

A idéia de guia metodológico se refere à necessidade de mapeamento ou encaixe dos conceitos minerados nas cinco forças de Porter e, com isso, compreender-se a utilidade desses conceitos e suas interrelações para o desenvolvimento de *insights* competitivos.

Os objetos, atributos e relações de incidência desses modelos fragmentados podem ser obtidos das planilhas de contextos (APENSO III-A, APENSO III-B e APENSO III-C) e dos respectivos reticulados de conceitos das organizações amostradas, com possíveis interpretações à luz de algum formalismo léxico-conceitual (EVENS, 1988).

[77] Essas informações conceituais ressaltam a posição da organização Fusitsu no mercado asiático e mostram como pode funcionar a ideia de "reforço de aprendizagem" de contexto mencionada por Weick (1995) para o Engenheiro do Conhecimento.

## 6.1 Fragmentos de um mosaico informacional

Conforme Davydov (2001, p. 26), o sucesso na utilização da informação, de um ponto de vista da ampliação do ecossistema de negócios de um modo geral, como em qualquer sistema que inclua pessoas e estruturas organizacionais, depende de uma efetiva resolução de questões introduzidas pela estreita colaboração entre os três maiores componentes (ou aspectos) desse "sistema corporativo":

- Financeiros: investimentos, custos e lucros, por exemplo;
- Tecnológicos: por exemplo, a inteira infra-estrutura de domínio do sistema;
- Organizacionais: personagens e seus papéis específicos e zonas de influência e responsabilidade, por exemplo.

Esses componentes sistêmicos são, essencialmente, os mesmos anteriormente mencionados por Greenwald e Judd (2005). Como se pode observar, por exemplo, nos contextos das três organizações estudadas em detalhe – Fujitsu Research, IBM Research e Microsoft Research, esses três componentes de um sistema corporativo clássico são representados com informações conceituais numa série de sintagmas plurinominais:

- componente financeiro: *account transfers, acquisition costs, banking institutions, business value, cost of producing, customer value, delivery costs, discounting strategies, economics overview, expertise in development finance, group funds grants, increase sales revenues, marketing costs, Microsoft funds research, portions of stock, productions costs, reduction in safety stock requirements, research and development budgets, research grants, revenue optimization, rise in inflation, sale trends, study of economics, team of economists, value pricing, warranty costs*;
- componente tecnológico: *3-D scene, advances into software, analysis of the Web graph, anti-piracy cryptography, area of research, automation of production schedule, bar code readers, business continuity, business modeling/visualization, business process, CAD tools, cell phones, cloud computing, data center architectures, fault-tolerance, hardware systems, information retrieval, information systems, logistics engineering, media player, Microsoft Windows, model solutions for management, natural language processing, petabytes of information, pricing algorithms, research efforts, risk management, social-network mining, support for decision-making, support in collaboration, text mining, Web service, workflow management* e outros;
- componente organizacional: *areas of expertise, area of finance, areas of interest, Asia strategy, banking and insurance sectors, branch office, business category, business networks, business structures, capability areas, computing group, consulting company, enterprise borders, field personnel, FRI's China researchers, Fujitsu Group, hardware division, HiPODS China Lab, IBM Research Lab in Tokyo, institute site, Life Science*

*group, management issues in partnership, middleware in terms of business needs, operations in some corner of the world, partnership with IBM, research facilities, social network, solution team, T. J. Watson Research Center, Venue Tokyo* e outros.

### 6.1.1 Conceitos e estruturas estáticas de informação

Como se mencionou no Capítulo 3, existem, além dos reticulados conceituais, várias estruturas cognitivas com visualização gráfica que podem ser utilizadas para modelagem da informação, tais como a Linguagem de Modelagem Unificada (*Unified Modelling Language* – UML) e os Gráficos Conceituais. Os artefatos gráficos da UML são bastante úteis para aquilo que podemos chamar de modelagem estática da informação, relativa às estruturas da arquitetura da informação corporativa, concentrando-se na modelagem de classes de objetos e das relações entre as mesmas. Os Gráficos Conceituais se prestam melhor à apresentação de predicados inerentes aos objetos e agentes, ou seja, dos aspectos mais dinâmicos da informação corporativa, como fluxos de atividades e tarefas dos processos.

A propriedade de apresentação polissêmica dos substantivos/nomes, nos quadros de contextos com incidência de atributos em objetos utilizados na Análise de Conceito Formal, permite que se construam, com a UML, fragmentos de modelos de classes de objetos. Os modelos fragmentados apresentados nas figuras a seguir são construídos dentro de conceitos reticulados, para mostrar como subconceitos se relacionam internamente nesses conceitos.

Os fragmentos de modelos estáticos de informação constituídos por sintagmas plurinominais, que representam subconceitos dos conceitos formais reticulados, geralmente podem também ser representados, em linguagem unificada (*Unified Modeling Language* – UML), na forma dos diagramas canônicos de classes da Figura 6.2 (sem as “gavetas” de atributos e métodos, nos retângulos de classes, por serem desnecessários no caso).

Esses modelos canônicos, comuns na modelagem de sistemas orientados a objetos (WEILKIENS; OESTEREICH, 2007), sugerem que um objeto representado, linguisticamente, por um sintagma plurinominal pertence a uma classe conceitual e tem uma classe de atributos associada ao mesmo, como uma característica ou propriedade. Essa classe-atributo, por sua vez, também pode pertencer a uma classe ainda mais agregada de conceitos (no caso de um modelo de objetos com três substantivos constituintes), ou superclasse, e ter uma superclasse de atributos mais consolidada que a diferencie como um membro ou objeto na superclasse de conceito.

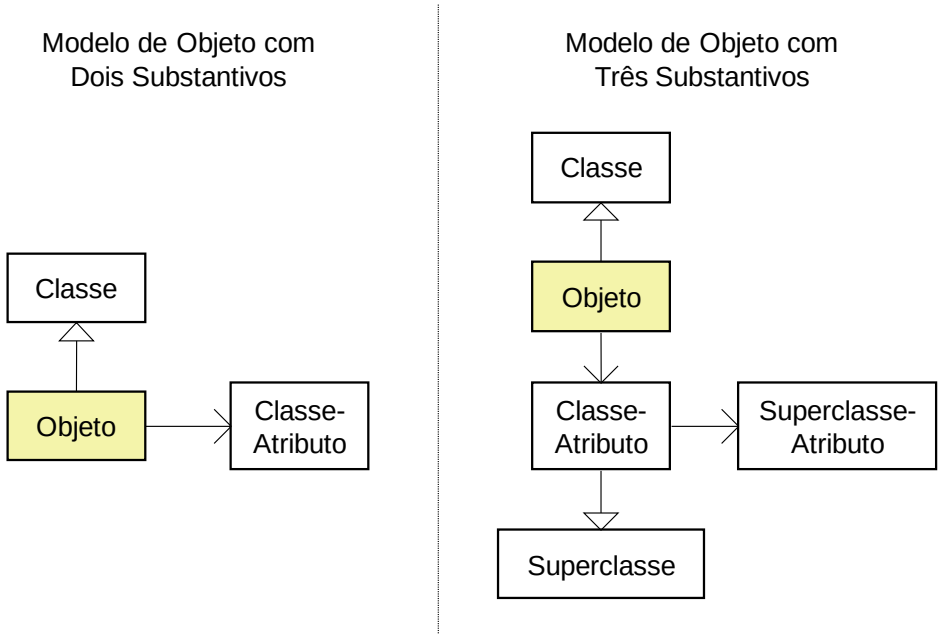



Figura 6.2 Modelos Canônicos de Classes

(Fonte: do autor da tese)

A Figura 6.3, baseada no modelo canônico de Weilkiens e Oestereich (2007, p. 46), esclarece esse tipo de estrutura de conceitos representada com UML.

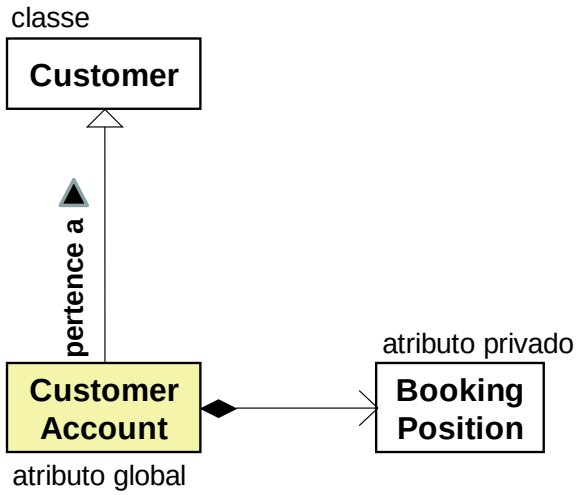



Figura 6.3 Associação/Composição de Classes

(Fonte: WEILKIENS e OESTEREICH, 2007)

Weilkiens e Oestereich (2007, p. 46) explicam essa estrutura canônica da seguinte forma:

> *(...) a classe Customer obteria um atributo, account, mediante a referência a um objeto da classe CustomerAccount, e a classe CustomerAccount obteria um atributo privado, bpos, com um objeto de coleção (ou uma subclasse), que se refere aos objetos de BookingPosition.*

O losango preenchido, na linha de associação entre as classes *Customer Account* e *Booking Position*, significa que os objetos da classe *Booking Position* pertencem unicamente à

classe *Customer Account* e, portanto, não podem ser compartilhados, conceitualmetne, com nenhuma outra classe. E isso se explica por uma razão lógica: o saldo da conta bancária *(booking position)* de um cliente é informação útil apenas para se avaliar a posição financeira *(custommer account)* desse único cliente e não outro.

Como exemplo desses modelos canônicos de classes de conceitos, observe-se a Figura 6.4 representando, graficamente, os sintagmas nominais “sistemas de informação” e “empresas de controle de pragas”.

A utilidade desse tipo de fragmento de modelo de informação, ou de “fatia da realidade” de Fuld (2007), pode ser compreendida a partir de um exemplo bastante pragmático e atual. O mercado de cartões de crédito no Brasil, conforme recente edição do periódico Exame (2010), se encontra num estágio de importante mudança estrutural, com a perspectiva de entrada de novas empresas após a implementação de medidas governamentais de regulação no sentido de ampliação da concorrência – a reportagem de 17 páginas classifica esse evento como “a maior reviravolta da histórica do mercado brasileiro” nesse setor (EXAME, 2010, p. 5).

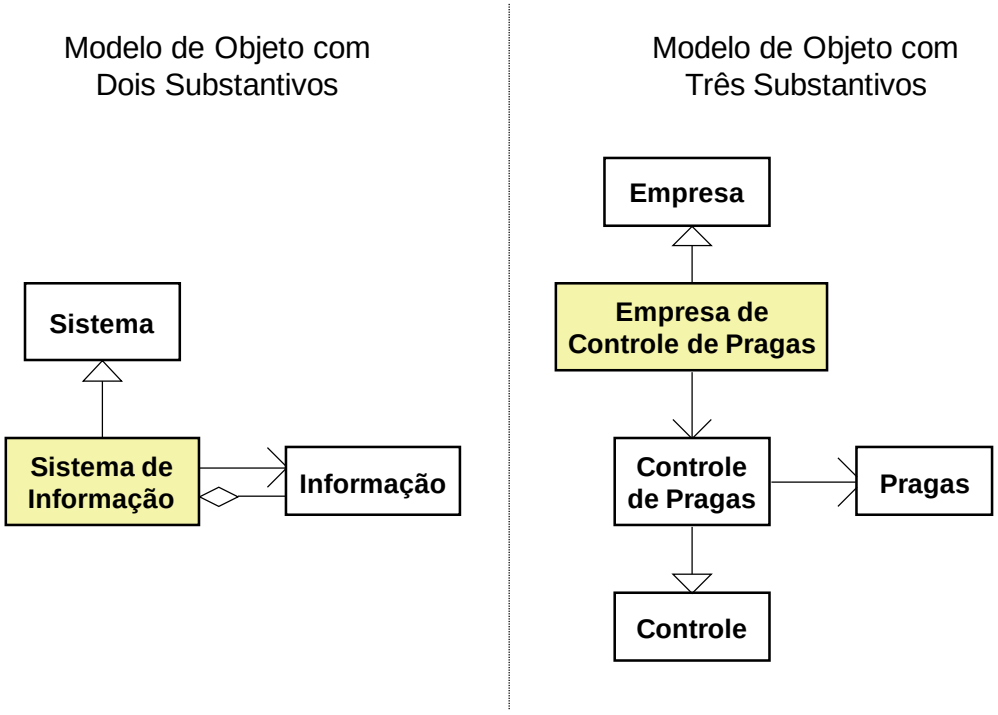



Figura 6.4 Exemplos de Classes Sintagmáticas Canônicas
(Fonte: do autor da tese)

Caso uma organização estivesse interessada em acompanhar, por quaisquer razões, esse mercado e utilizasse, para tanto, o método de mineração conceitual com base em textos de publicação aberta aplicada nos experimentos apresentados nesta tese, tendo como fontes

publicações digitais com os conteúdos do periódico “Exame”, os resultados apresentariam os seguintes sintagmas compostos com os substantivos mais freqüentes:

1) “cartões de crédito” (aparece 17 vezes);
2) “setor de cartões” (aparece seis vezes);
3) “tipos de cartão” (duas vezes);
4) “número de cartões” (duas vezes);
5) “maquininhas de cartões” (duas vezes);
6) “número de transações” (duas vezes).

Outros substantivos que aparecem nos textos dessa reportagem, que poderiam representar, num “mosaico semântico”, mais que um “alerta antecipado”, uma motivação para exercícios de “jogos de guerra” nas empresas envolvidas com o mercado de cartões de crédito e débito no país, são também bastante densos em significados, tais como: *entrada de novas empresas, fim da exclusividade entre Cielo e Visa, embate entre as credenciadoras, funcionário do Banco Central, responsável pela maior transformação do setor, cartões para as massas, classes D e E, fronteira dos cartões, Centro-Oeste, fim dos mercados, aumento da competição entre as credenciadoras, expectativa de mais eficiência e preços, cartões da bandeira Visa, ano zero da competição, empresa de cartões Cielo, credenciadora de cartões no país, bandeira de cartões Visa, cartões de débito, cartões Visa e Mastercard, diretor executivo de cartões do Panamericano, marca de cartão, bandeira de cartões Elo, uso de cartões, cartão Visa, pai da mudança, repercussão do fim do duopólio no setor de cartões, executivos da área de cartões, empresas de cartões, cartão de débito com bandeira Visa, participação do cartão, terminal de cartões.*

Com esse conjunto de objetos conceituais extraídos numa mineração de textos dirigida, um modelo de classes tendo como núcleo o conceito de “cartão de crédito” poderia ser representado como na Figura 6.5. Os engenheiros do conhecimento observariam, nesse modelo de informação, um conceito bastante explícito que poderia alertá-los sobre o que está ocorrendo, como “Repercussão do Fim do Duopólio no Setor de Cartões”, e conceitos mais específicos como “Maquininha de Cartão”. O conceito de “Fronteira de Cartões” é um tanto raro na medida em que o termo “fronteira” é apresentado, no caso, como uma metáfora, despertando curiosidade natural em torno de seu significado, o que poderá levar o engenheiro do conhecimento a buscar mais informações a respeito, podendo chegar ao seguinte excerto (com grifo nosso) da reportagem (EXAME, p. 9):

> *Com o fim da exclusividade, é provável que uma parte dos lojistas que hoje têm duas “maquininhas” decida ficar apenas com uma. Por isso, estima-se que o número de máquinas em operação das empresas líderes saia dos atuais 2,3 milhões para 1,7 milhão em 2014.*
>
> *Para tentar reduzir possíveis estragos, os executivos de Cielo e Redecard estão pensando em alternativas. A Cielo quer que as maquininhas gerem receita para seus clientes com a oferta de serviços financeiros. “Ao usar os aparelhos para pagamento de contas e recarga de celular, o estabelecimento comercial estará atraindo um fluxo maior de clientes para o local – e ainda ganha uma taxa por isso”, diz Rômulo Dias, presidente da Cielo. Esse projeto já começou a funcionar em grandes redes. Desde o*

*começo do ano, os clientes dos supermercados Pão de Açúcar podem sacar até 100 reais nos caixas com o cartão Visa. A Redecard, por sua vez, vai intensificar sua estratégia de disputar mercados com credenciadoras regionais, com a do Banrisul, que atua principalmente no Rio Grande do Sul e em Santa Catarina. Além disso, tanto a Cielo quanto a Redecard pretendem conquistar novos clientes no Centro-Oeste, considerado pelas duas empresas como a nova fronteira dos cartões no país. "O brasileiro ainda usa muito cheque e dinheiro para pagar suas contas. Temos muito espaço para crescer."*

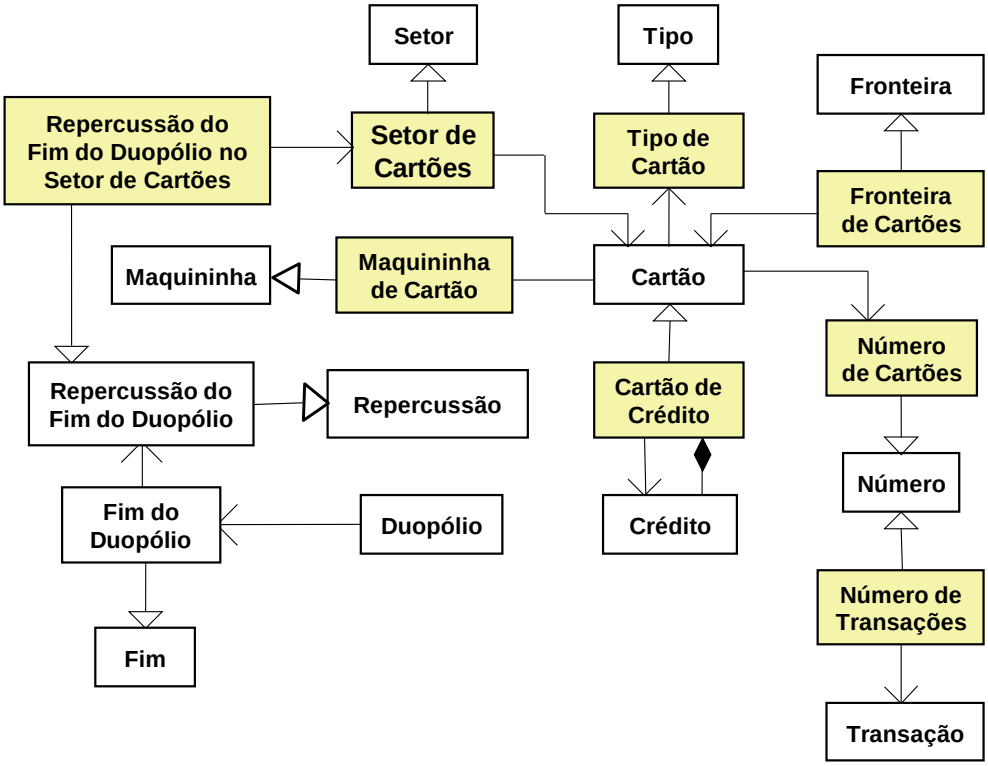



Figura 6.5 Modelo Conceitual com "Cartão de Crédito"

(Fonte: do autor da tese)

O que esse simples exercício com o método de mineração de conceitos proposto sugere? A resposta é: como os conceitos utilizados pelos seres humanos são, em sua maioria, relacionados entre si no discurso, pode-se partir de "conceitos-semente" para buscas de conceitos correlatos, descobrindo-se assim conceitos mais complexos, por composição "vertical" de conceitos (construindo-se conceitos "em pilha", como no caso de "Repercussão do Fim do Duopólio no Setor de Cartões"), e conceitos mais ricos em significados, por associação "horizontal" com outros conceitos, como em "Fronteira de Cartões".

O segundo sintagma mais freqüente com uso do substantivo "cartão" (ou "cartões") – "setor de cartões" – está associado, no texto da reportagem, ao conceito mais complexo descoberto nessa mineração, que exprime, por si só, a idéia central sobre o fenômeno de mercado em estudo – a dos possíveis impactos do fim do duopólio de cartões de crédito e débito sobre o mercado.

Com fragmentos de textos representando informação conceitual relativamente simples sobre um contexto de negócio, mineradas de modo semi-automático, o engenheiro do conhecimento

poderá elaborar composições conceituais mais complexas e ricas em significados. É como se aplica, no caso, o princípio da inteligência emergente da Inteligência Artificial baseada na natureza.

Entre os conceitos modelados na Figura 6.5, pode-se separar os que são mais previsíveis (ou mais prováveis), num contexto de negócio de administração de cartões de crédito e temas correlatos, e os que são menos previsíveis (menos prováveis). Obviamente, os conceitos que se relacionam a “Duopólio” e “Fronteira” são menos prováveis, pois não são típicos desse nicho de mercado, como “Maquininha de Cartão”, “Cartão de Crédito” e “Número de Transações”. Com isso, tem-se mais “surpresa” e, portanto, informação mais relevante para Inteligência Competitiva com “Repercussão do Fim do Duopólio no Setor de Cartões” e “Fronteira de Cartões” que com os demais conceitos minerados nesse contexto.

A avaliação das probabilidades relativas de ocorrência de conceitos, ou de valores de “auto-informação” (SAYOOD, 2000), em publicações de contextos de negócios é atribuição precípua dos engenheiros do conhecimento, que devem conhecer, por dever de ofício, os conceitos mais comuns que estruturam e caracterizam o nicho de negócio da organização onde atuam. Evidentemente, com isso assume-se que o engenheiro do conhecimento não deve conhecer apenas métodos e tecnologias de mineração de conteúdos, mas também, em razoável proporção, também os próprios conteúdos (objetos) do negócio. Essa abordagem é a mesma dos antigos Analistas de Sistemas, que precisavam conhecer os processos de negócio da organização onde atuavam para modelagem de sistemas de informação computacional para automação desses processos.

Então, no exemplo, o engenheiro do conhecimento atuando numa organização no mercado de administração de cartões de crédito, tendo seu sistema cognitivo perturbado com esses conceitos diferentes (pouco prováveis) que aparecem na mineração, inicia um processo de criação de significado que, ao final, poderá lhe proporcionar *insights* para construção de conhecimento e tomada de decisão (CHOO, 2003).

Os exemplos a seguir, destacados do experimento com informações conceituais mineradas dos *Web* Portais de três empresas do mercado mundial de TIC, serão os próximos contextos de análise conceitual discutidos.

### 6.1.2 Contextos e conceitos comparados

#### 6.1.2.1 Diamantes de conceitos

Examinando-se, num primeiro momento, as relações entre os conceitos minerados em cada contexto organizacional – Fujitsu Research, IBM Research e Microsoft Research – amostrado no experimento e, depois, comparando-os, pode-se avaliar a utilidade da informação recuperada e

filtrada, com o método proposto, num ambiente de Gestão da Informação e do Conhecimento para suporte à prática da Inteligência Competitiva. Os reticulados de conceitos desses três contextos, que lembram a estrutura molecular de um diamante, são apresentados nas Figuras 6.6, 6.7 e 6.8, mostrando-se apenas as quantidades de objetos que compõem cada conceito formal (nos cruzamentos de algumas linhas de relacionamento entre os conceitos), tomando-se apenas os sintagmas relativos aos 2,5% de substantivos mais freqüentes em cada contexto.

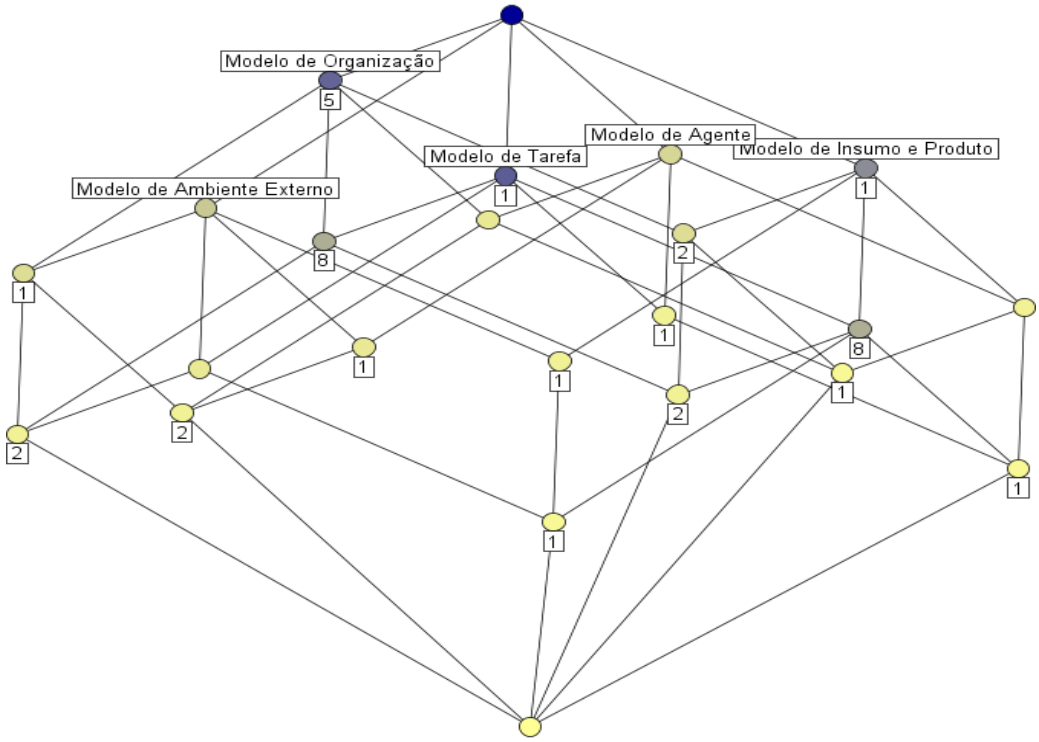



Figura 6.6 Reticulado de Conceitos do Contexto “Fujitsu Research”
(Fonte: do autor da tese)

Observa-se, na Figura 6.6, que no cruzamento intensional dos atributos “Modelo de Tarefa”, “Modelo de Agente” e “Modelo de Insumo e Produto” aparece apenas um objeto sintagmático (círculo mais à direita e mais embaixo no reticulado). Conforme a teoria da Análise de Conceito Formal, esse conceito herda atributos gerais que se referem a tarefas, agentes e insumo e produto, compondo um conceito que é mais específico que seus conceitos antecessores, na “linhagem genética” do reticulado, como os conceitos envolvendo apenas “tarefa” e “agente” (com apenas um objeto instanciando a classe) e “tarefa” e “insumo e produto” (com oito objetos). Esse tipo de conceito com três atributos intensionais é semanticamente rico, com informações que podem ser modeladas de modo a evidenciar estruturas de processos, papéis e insumos e produtos envolvidos nos contextos organizacionais.

Com a estrutura conceitual do reticulado e os sintagmas com substantivos mais freqüentes obtem-se um construto informacional com capacidade para amenizar, senão superar, os

problemas dos sistemas conceituais apontados por Frawley (1988, p. 335-372) na representação do conhecimento:

> *Os problemas que sistemas conceituais apresentam para a análise de textos em sublinguagem científica são claros. Em primeiro lugar, sistemas conceituais perdem inteiramente o rastro do léxico porque eles são sistemas decomposicionais e, portanto, baseiam-se em estruturas primitivas elaboradas, não no léxico em si. Isso elimina, efetivamente, a especificidade requerida para uma sublinguagem científica (...). Segundo, o conjunto de conexões entre as primitivas não é suficientemente grande para capturar toda a informação semântica subjacente no texto, ou mesmo uma parte interessante dela. (...) Isto significa que um conjunto de relações sensitivas-de-superfície é necessário e, por isso, o abstracionismo dos sistemas conceituais aparenta ser uma deficiência no caso da sublinguagem científica.*

Embora o argumento de Frawley (1988) se reporte a sublinguagens de disciplinas científicas, a comparação com os textos e contextos do experimento desta tese é válida para compreensão dos problemas estruturais da mineração conceitual baseada na linguagem natural para os quais se propõe solução. Outros excertos de Frawley (1988, p. 344-345) revelam que mesmo numa era *pré-Internet* tinha-se uma visão integrada da mineração de textos, com a necessidade de abordagens mistas, como a léxico-conceitual, para gestão do conhecimento:

> *Quando se olha para o que a análise de textos em sublinguagem científica na prática tem apresentado [como exemplos, Grishman e Kittredge, 1986; Kittredge e Lehrberger, 1982], parece que a análise de textos científicos de sucesso deve proceder a partir de uma base léxica. Existem duas razões para isso: (1) um modelo léxico de sublinguagem científica é focado em conteúdo; (2) um modelo léxico é colocacional-de-superfície.*
>
> *(...) Existe toda uma gama de argumentos defendendo que o conteúdo supera a forma em textos de sublinguagens científicas e, consequentemente, que o conteúdo é grosseiramente comparável com o léxico. (...) a forma retórica em textos científicos é de menor importância e a coesão em textos científicos é efetivada lexicalmente.*

O reticulado conceitual, no caso do experimento, serve de base ontológica para se aprender as estruturas primitivas da informação recuperada – no caso do modelo testado, com atributos intensionais derivados da metodologia *CommonKADS*, tem-se uma ontologia de objetos de negócio para classificação dos sintagmas extraídos do léxico (ver Tabela 6.1 e Figura 6.1). Os objetos do léxico com atributos de “tarefa”, “agente” e “insumo e produto”, por exemplo, se vinculam, na ontologia gerada no reticulado da Figura 6.1, a processos de negócio dos atuais *players* no modelo analítico de Porter para Inteligência Competitiva, observando-se que tanto parceiros como clientes ou consumidores herdam esses atributos em conjunto.

Quanto aos conteúdos léxicos “superficiais” de Frawley (1988), eles sempre estarão disponíveis para consulta, no construto elaborado com base no método proposto nesta tese, e ainda poderão ser filtrados com base na freqüência dos substantivos que eles representam nos textos. Caso se queira descobrir apenas as estruturas lexicais essenciais, pode-se restringir a Análise de Conceito Formal apenas aos sintagmas representando os substantivos mais freqüentes, definindo-se os parâmetros dessa amostra, que no experimento “calibrou-se” o sistema para operar com 2,5%, 5,0%, 7,5% e 10,0%.

Com isso, conclui-se que a metodologia proposta permite a superação do dualismo dos sistemas de conceitos quanto à ênfase nas informações primitivas ou nas informações de superfície, ao mesmo tempo propiciando a representação gráfica de conceitos superficiais extraídos diretamente dos textos e o aprendizado de ontologias mais profundas (com primitivas) a partir de *insights* indutivos e dedutivos.

Com essa explanação inicial, pode-se prosseguir na exploração conceitual dos demais contextos – IBM Research e Microsoft Research. A Figura 6.7 e a Figura 6.8 apresentam os respectivos "diamantes conceituais" dessas organizações, observando-se, visualmente, as distribuições quantitativas de objetos entre os conceitos.

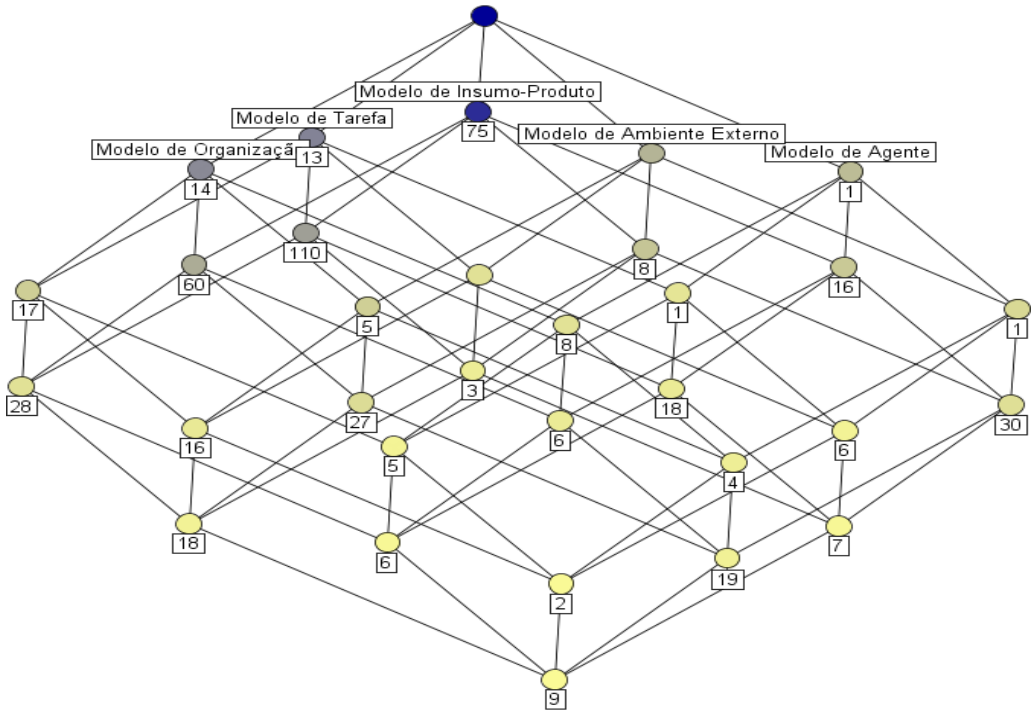



Figura 6.7 Reticulado de Conceitos do Contexto "IBM Research"
(Fonte: do autor da tese)

Observa-se, na Figura 6.7, que existem 13 objetos sintagmáticos classificados como tendo atributos apenas de "tarefa", 75 objetos com atributos de "insumo e produto" e um objeto com atributos apenas de "agente", mas existem 18 objetos com esses três atributos herdados. A estrutura do reticulado permite várias outras análises de relacionamentos entre conceitos, como análises envolvendo *clusters* de objetos com atributos de ambiente externo, que são especialmente interessantes, por razões óbvias, do ponto de vista da Gestão da Informação e do Conhecimento para Inteligência Competitiva.

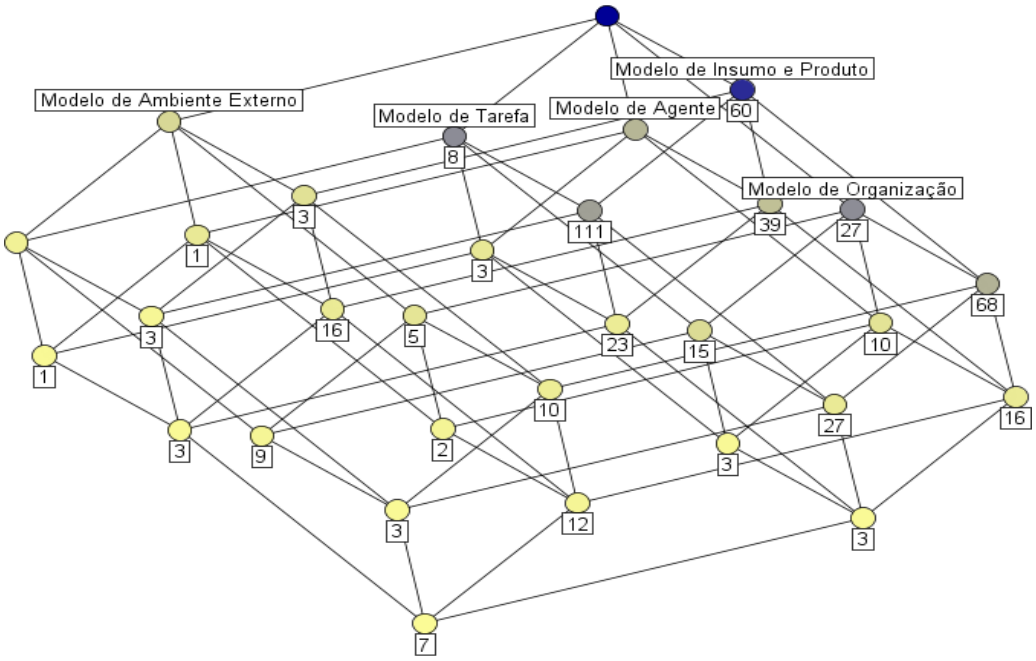



Figura 6.8 Reticulado de Conceitos do Contexto “Microsoft Research”
(Fonte: do autor da tese)

Em relação ao conceito, estudado com mais pormenores, herdando os atributos “tarefa”, “agente” e “insumo e produto”, o contexto “Microsoft Research” apresenta 23 objetos instanciadores. Observa-se, também, 73 objetos compondo conceitos com inserção, inclusive, nos atributos “Modelo de Organização” e “Modelo de Ambiente Externo” no contexto “IBM Research”, e 48 objetos no contexto “Microsoft Research”.

Os 18 objetos sintagmáticos do conceito de processos de negócio do contexto “IBM Research” são apresentados na Figura 6.9 (caixa mais escura). Como se pode observar, uma visualização mais confortável de reticulados de conceitos em geral exige telas maiores que as encontradas na maioria dos microcomputadores pessoais.

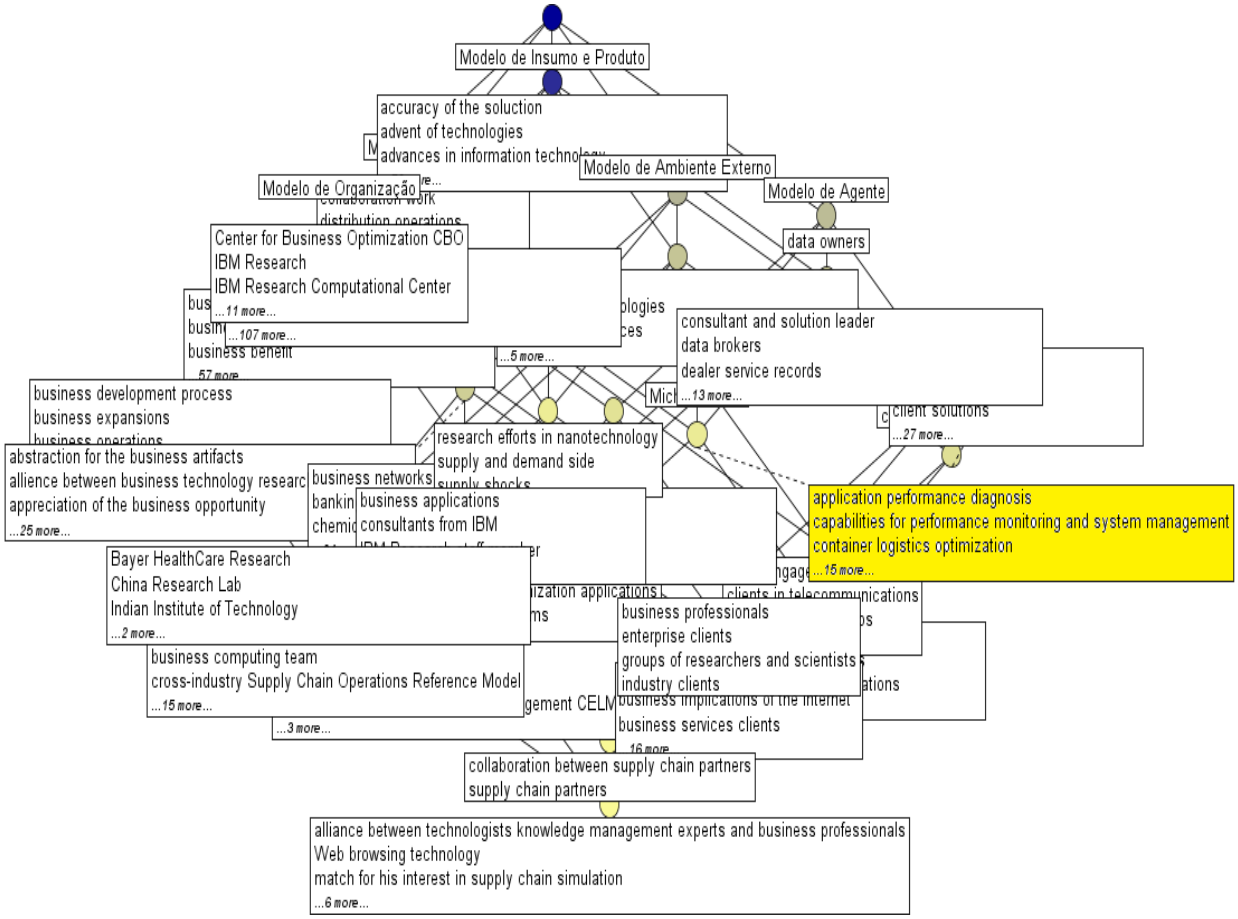



Figura 6.9 Reticulado de Conceitos Formais com Lista de Objetos

(Contexto: “IBM Research”)

(Fonte: do autor da tese)

### 6.1.2.2 Modelos de informação do contexto “Fujitsu Research”

A Figura 6.10 mostra uma composição de fragmentos de modelos conceituais de informação a partir dos seguintes objetos sintagmáticos com substantivos mais freqüentes no contexto “Fujitsu Research”: *delivery costs, delivery routes, value in service innovations, value pricing, visualization of value in service*. A parte superior da figura evidencia a centralidade do subconceito “delivery”, integrando, com relacionamentos de associação, as classes de objetos “Delivery Costs” (associação), “Transportation and Delivery Plans” e “Service”. É importante observar-se que a associação entre “Delivery” e “Service” não é obtida a partir da análise da formação dos sintagmas, mas de uma análise semântica onde se conclui que “Delivery” (Entrega) é um tipo de “Service” (Serviço); ou seja, “Delivery” integra, como instância, uma classe conceitual denominada “Service”.

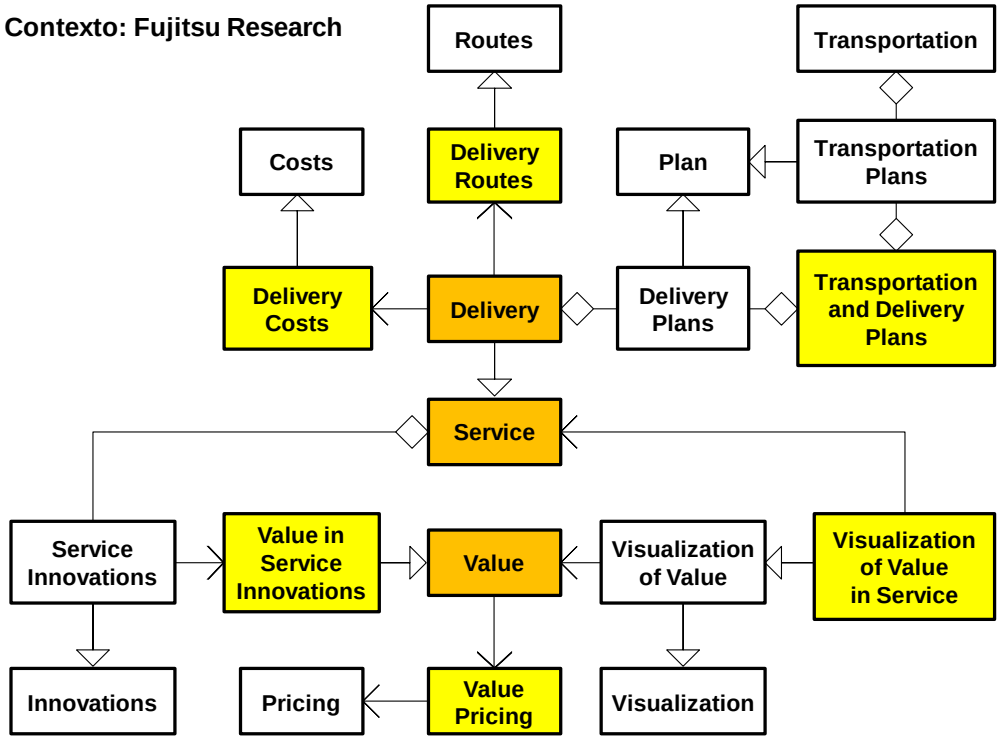



Figura 6.10 Conceitos Integrados em “Delivery”, “Service” e “Value”
(Fonte: do autor da tese)

O resultado experimental da Figura 6.10 também mostra outro aspecto importante: enquanto os objetos sintagmáticos centrados em “Delivery” são encontrados entre os sintagmas que representam os 2,5% de substantivos mais freqüentes no contexto, os objetos centrados em “Service” e “Value” não se encontram nesse grupo de sintagmas, mas entre um grupo de sintagmas que representam substantivos não tão freqüentes, mas entre os 10% mais freqüentes no contexto. Com isso, evidencia-se a vantagem da integração de modelos conceituais contendo substantivos mais freqüentes e menos freqüentes, ainda que dentro de um grupo maior de substantivos mais freqüentes – os 10% mais freqüentes, no caso.

Considerando-se que, historicamente, o nicho de mercado da organização Fijitsu Research é o de soluções tecnológicas com uso de computação eletrônica, pode-se destacar como conceitos de negócios relativamente menos prováveis os da Figura 6.10 (e não os da Figura 6.11): “Delivery Costs”, “Value in Service Innovations”, “Value Pricing”, “Transportation and Delivery Plans” e “Visualization of Value in Service”. E isso se torna evidente na medida em que o engenheiro do conhecimento entenda que os conceitos centrais “Service”, “Delivery” e “Value” são menos prováveis que “Management” no contexto.

Conforme as teses de Bateson (2002), Weick (1995) e Choo (2003), o fragmento de modelo conceitual da Figura 6.10, pela relativa novidade das informações que o compõem, terá maior

poder de atração da atenção do analista, desencadeando assim um processo psicológico de criação de significado.

Em termos de Inteligência Competitiva, resta saber da utilidade desse modelo conceitual para conduzir o raciocínio dos engenheiros do conhecimento. Como recurso para aprofundamento da análise, sugere-se então uma pesquisa específica (recursiva) no Google[78] utilizando-se a expressão-chave *"'Fujitsu Research' and 'service delivery'"*, referindo-se o primeiro termo ao contexto organizacional e o segundo ao alvo da pesquisa, unidos pelo conector lógico *AND*[79]. O resultado da busca aparece com vários endereços URL *(Uniform Resource Locator)* possíveis, sendo o primeiro no *ranking* um endereço que leva a uma apresentação corporativa do laboratório da Fujitsu (revelando conteúdos em formato *.pdf) mostrando o caminho da estratégia de desenvolvimento (*roadmap*, no idioma original) da organização até 2020. O segundo endereço mostra uma página *Web* com notícias da Fujitsu Research publicadas em 2009 e o terceiro leva a páginas da Fujitsu Research no seguinte endereço na *Web*: *http://jp.fujitsu.com/group/fri/en/economic/publications/report/2008/report-310.html.* Esse endereço apresenta o resumo de um relatório intitulado: *Important Factors for Innovation in Physical Distribution and Commercial Distribution (1) - Verification of Factors Derived from Case Studies,* cujo resumo é apresentado a seguir (com grifos nossos):

> *1. Innovation in services is comprised of service product innovation, service delivery innovation and service environment innovation.*
>
> *2. The following important factors for innovation in physical distribution and commercial distribution were derived by Kimura (2008) (Research Report No. 309): (1) that it is possible to adequately acquire the necessary information for business at the business locations from the information systems, (2) a company-wide prevalence of a strong customer-oriented philosophy that begins with top management, (3) the activity of small groups at the business locations ((1) and (2) are factors for both service product innovation and service delivery innovation in physical distribution and commercial distribution, (3) is a factor of service delivery innovation in physical distribution, and of service product innovation and service delivery innovation in commercial distribution). These factors were verified based on the results of a questionnaire survey.*
>
> *3. The results of this verification reject hypothesis (1) regarding service delivery innovation in physical distribution and service product innovation in commercial distribution, but support the other hypotheses. For companies to gain an edge in service innovation among other companies in the same industry, it is necessary that they pay attention to the factors verified as significant.*

Embora o texto na íntegra desse relatório esteja na *Web* invisível, provavelmente disponível apenas para os usuários na rede interna *(Intranet)* do Fujitsu Research Institute (FRI), trata-se, claramente, de uma linha de desenvolvimento competitivo da organização. O termo *innovation* (inovação) permite essa conclusão, pois inovação em serviços, com agregação de valor ao cliente, é uma necessidade competitiva ditada pelas necessidades do mercado em constante mutação. A simples menção do interesse da FRI no nicho de serviços inovadores para clientes

[78] As buscas de aprofundamento relatadas neste capítulo se deram de 3 a 13 de junho de 2010.
[79] Observe-se que nesse tipo de busca o algoritmo do Google precisa encontrar endereços com conteúdos que contenham os dois sintagmas adotados como argumentos, restringindo assim os resultados.

que operam com distribuição física de produtos, apoiando, com sua pesquisa e desenvolvimento, as operações comerciais da *holding* – Fujitsu Corporation, deve servir para "acender uma luz amarela no painel de observação" de uma organização que compete com a FRI no mercado, algo como um "alerta antecipado".

A construção de um conhecimento de primeira ordem (ou preliminar), com mais esse pequeno texto recuperado recursivamente da *Web*, torna-se um processo de resultado imediato, pois o texto evidencia informação relevante (como notícia) no contexto. O último estágio do modelo analítico holístico de Choo (2003), relativo ao processo de tomada de decisão, dependerá muito do modelo de gestão da organização patrocinadora do esforço de Inteligência Competitiva.

Outro agrupamento de objetos sintagmáticos do contexto Fujitsu Research que podem gerar um modelo conceitual integrado corresponde aos objetos centrados no subconceito "management": *management issues, management reform, management science, model solutions for management, risk management, solutions for management issues* (a Figura 6.11 mostra os relacionamentos semânticos entre esses objetos).

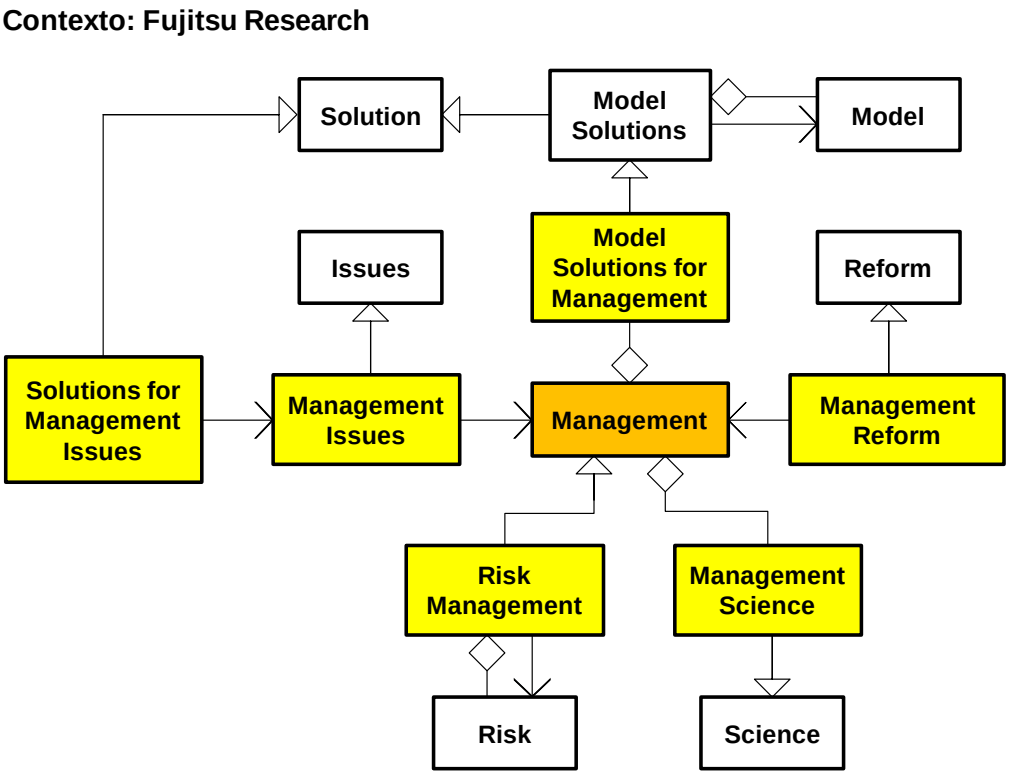



Figura 6.11 Conceitos Integrados em "Management"

(Fonte: do autor da tese)

O mesmo procedimento recursivo de busca focada no Google, com os argumentos "Fujitsu Research" e "solutions for management", resulta em apenas dois endereços, sendo o primeiro

relativo a páginas *Web* contendo informações sobre as atividades de pesquisa e desenvolvimento do FRI *(http://jp.fujitsu.com/group/fri/en/develop/)*, tais como novos campos de pesquisa. O conteúdo dessa página representa, claramente, uma mensagem de publicidade sobre os produtos tecnológicos ofertados pela organização, com o seguinte título: *We apply cutting-edge science and IT to "visualize" management issues, and provide strong support for our customers in optimizing their operations.* Entre outros pontos, a FRI indica, nessa página, a oferta de serviços de engenharia de logística, confirmando assim a centralidade do conceito de "Delivery Services" nos negócios da empresa evidenciado na Figura 6.10:

> *We formulate transportation and delivery plan as well as hub and location plans in consideration of all the issues involved in the form of mathematical models, and investigate and apply the requisite Technologies for their optimization.*

Quanto ao segundo endereço da resposta do Google *(http://jp.fujitsu.com/group/fri/en/economic/publications/report/2007/report-289.html)*, o conteúdo é o resumo de um relatório com um título bastante revelador, que mostra as fraquezas gerenciais das empresas japonesas operando na China: *Problems and Solutions for Management of Japanese Companies in China*. Esse resumo fala por si, numa extensão alarmante, de um ponto de vista competitivo de abertura e conquista de mercado:

> ***ABSTRACT***
>
> *1. Following the rapid growth of the Chinese economy, Japanese companies are aggressively expanding operations in China. These companies, however, are saddled with various management problems. These problems are related to Japanese company systems and Japanese-style management.*
>
> *2. Concerning the management of personnel and labor, when Japanese systems are introduced directly they do not fit with the circumstances in China, and incentives for employees become deficient. Delays in the localization of personnel cast a shadow on various areas such as management costs, procurement, sales and etc.*
>
> *3. High costs are also hindering Japanese companies' development of operations in China. The large number of dispatched employees requires a substantial payroll, and because many business transactions are done among Japanese companies the procurement costs are also high. In addition, excessive attention to quality increases costs. This is also connected with Japanese-style management.*
>
> *4. In terms of organization, the transfer of authority from Japanese companies to subsidiaries in China is insufficient, and these subsidiaries have little independent discretionary power. As a result, strategic development of operations in China is difficult. Profits in China are low, and consequently many companies receive compensation from the home company to stay in the black. Furthermore, the supervising companies in China are not functioning properly.*
>
> *5. The weakening presence of Japanese companies in China, their inadequate commitment to social contribution, their sluggish response to unforeseen incidents or scandals and etc. have hurt the image of these companies among Chinese society, and are having a negative impact on their operations development.*
>
> *6. Regarding market strategy, insufficient entry into the Chinese market and weakness in organizational control represent problems. Strategic procurement and the restructuring of operations in response to market shifts are slow.*
>
> *7. To rectify these problems and improve business, it is necessary that Japanese companies organize business strategies for Chinese operations, encourage the autonomy and independence of their local subsidiaries, and promote the localization*

*of personnel. It is also critical that these companies be aware of the compatibility of Japanese-style business with China.*

6.1.2.3 Modelos de informação do contexto “IBM Research”

Os 18 objetos sintagmáticos minerados no contexto “IBM Research”, representando os 2,5% de substantivos mais freqüentes, são (ver APENSO III-B): *application performance diagnosis, capabilities for performance monitoring and system management, container logistics optimization, disclosure management system, delivery and tracking systems, decision support systems, device management systems, healthcare delivery and tracking systems, healthcare systems, inventory optimization with SAP NetWeaver, Jayant’s research interests, Mary’s work in software reliability engineering, production systems, shopping assistant technology, system management, tool in decision support systems, tracking systems, warning systems for automobiles.*

É evidente, nesse conjunto léxico, uma concentração expressiva de sintagmas compostos, inclusive, com o substantivo “system”. Assim, pode-se representar os possíveis relacionamentos conceituais desse contexto com o modelo de classes da Figura 6.12.

Os fragmentos de modelos de informação integrados na Figura 6.12 mostram coisas evidentes *a priori* mas, também, aspectos que podem se tornar interessantes em contextos de Gestão do Conhecimento para Inteligência Competitiva. Como o contexto “IBM Research” se refere a uma organização do mercado de tecnologias da informação e comunicação (TIC), a centralidade da informação no conceito de “System” não apresenta novidade. Contudo, deve-se observar que a classe “System” aparece com duas funções estruturais distintas: a primeira, como em “System Management”, na condição de atributo de classificação de outro conceito, onde “System Management” é apenas uma instância da classe “Management”; e a segunda, como em “Healthcare Systems”, indicando que “Systems” é uma classe essencial – neste caso, “Healthcare Systems” é um tipo (ou uma instância) de “System” (“Healthcare System” é uma espécie do gênero “System”).

Ou seja, numa interpretação no ambiente de negócios desse contexto corporativo, “Systems” se refere, em alguns casos, a instrumentos de gestão de sistemas de informação (que podem ser outros sistemas de informação, como os aplicativos de gerenciamento de computadores servidores ou de redes de computadores), e, noutros casos, a tipos de sistemas de informação produzidos pela organização. Os conceitos de “Device Management Systems” e “Disclosure Management Systems” são sistemas computacionais destinados ao gerenciamento de outros sistemas computacionais, pois “device”, no caso, se refere a dispositivos de *hardware* e “disclosure” à abertura do acesso a informações de sistemas.

**Contexto: IBM Research**

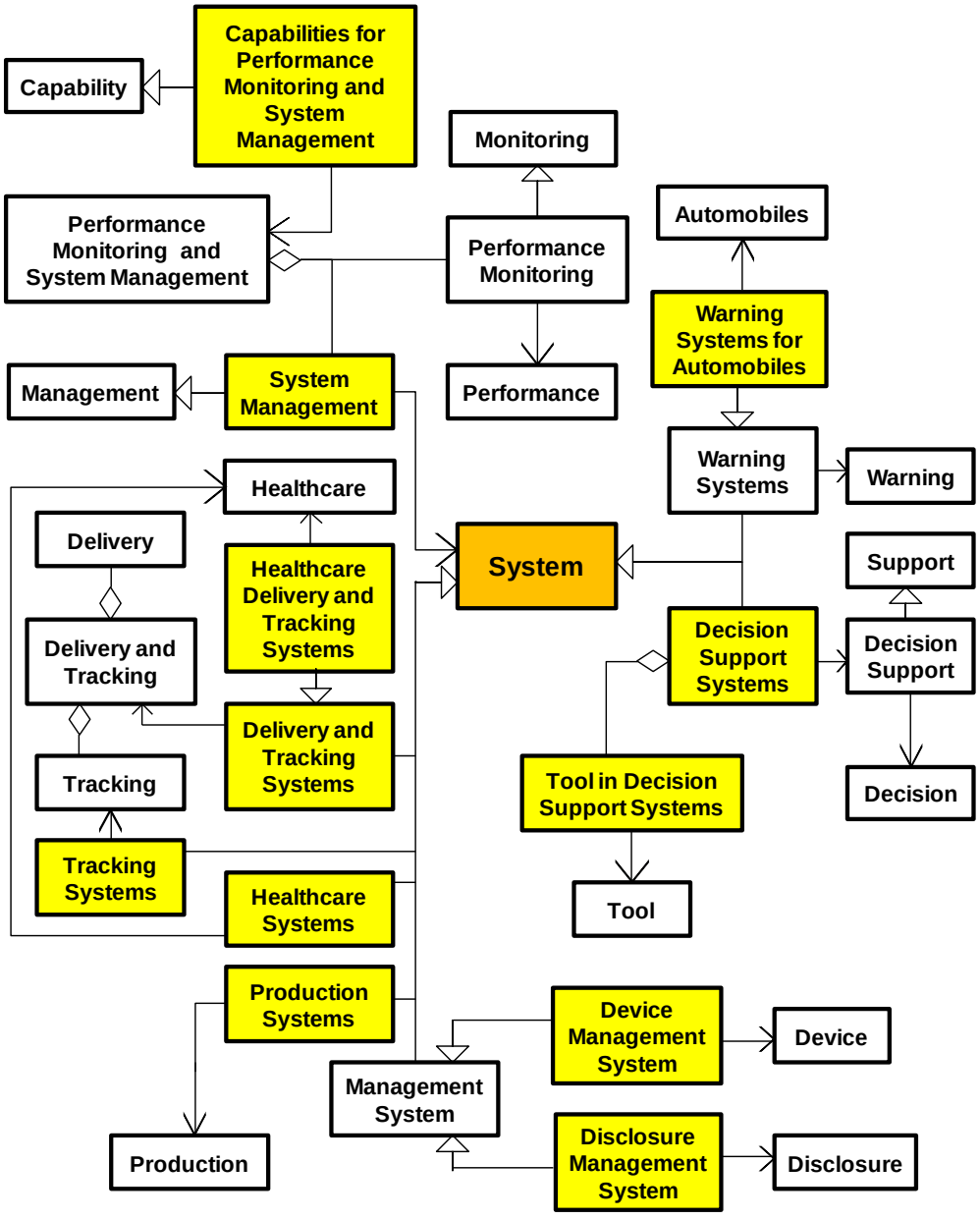



Figura 6.12 Conceitos Integrados em “System”

(Fonte: do autor da tese)

Embora esses conceitos mencionados possam representar temas triviais nesse mercado, outros conceitos presentes no modelo do contexto “IBM Research” apresentam algo diferente, como “Healthcare Delivery and Tracking Systems” e “Warning Systems for Automobiles”. Com uma busca recursiva no Google utilizando-se “’IBM Research’ and ‘Healthcare Delivery and Tracking Systems’” como argumentos, encontra-se um único endereço *(http://domino.watson.ibm.com/odis/odis.nsf/pages/profile.jasinksi.html)*, que apresenta o conteúdo textual a seguir (com grifos nossos):

**Joe Jasinski**
*Distinguished engineer and program director,*
*healthcare and life sciences.*

*In an era when healthcare systems are under pressure around the world, IBM Healthcare and Life Sciences Program Director Joe Jasinski is helping the medical community transform the practice and delivery of healthcare through the help of IT solutions. Today's healthcare environment is facing a global crisis, plagued by rapidly rising costs, poor or inconsistent care quality, and inadequate access for people in many countries. At the same time, advances in medical technology, including genomics, regenerative medicine and information-based medicine, show promise for helping develop new treatment strategies, while advances in information technology are paving the way for improved healthcare delivery and tracking systems. IBM's Healthcare and Life Sciences group has been instrumental in developing and integrating advanced computational hardware and software to provide the infrastructure that enables advances in medical and biological research and healthcare delivery.*

*At IBM's Watson Research Center, Joe Jasinski, IBM Distinguished Engineer and Program Director for Healthcare and Life Sciences Research, is working to develop strategies and coordinate efforts across IBM's Research Division in areas ranging from the use of information technology in payer/provider healthcare to computational studies in molecular biology. "We bring deep technical expertise and unique capabilities, such as the Blue Gene supercomputer, to bear on problems that most other organizations cannot meaningfully address," says Joe. In addition to his Research responsibilities, Joe has contributed his expertise to IBM Research Services, where he has coordinated with Global Business Services to bring his and other researchers' skills directly to the clients.*

*Prior to his current position, Joe served as worldwide operations manager for IBM Life Sciences, where he was responsible for day-to-day operations and strategy for one of IBM's fastest growing new businesses. Joe began his work in Healthcare and Life Sciences in 2000, when as Senior Manager of the IBM Research Computational Center, he oversaw research efforts in nanotechnology, materials chemistry and chemical kinetics. His academic credentials include a Ph.D. from Stanford University and undergraduate degrees in mathematics and chemistry from Dartmouth College. Following post-doctoral work at the University of California, Berkeley, Joe joined the Watson Research Center as a research staff member in 1982.*

*"Healthcare is one of the fastest growing IT opportunities in the developed and developing world," says Joe. "The opportunity to transform the way biomedical research is done and the way healthcare is delivered around the world is pretty energizing. What gets me out of bed in the morning is literally the chance to change the world for the better by helping people discover new treatments for disease and helping people get access to safe, affordable and effective medical care."*

Essa página também apresenta conectores na *Web* para outros conteúdos textuais correlatos, onde a organização apresenta seus casos de sucesso nessa linha de negócios. Com uma simples pesquisa recursiva a partir de um conceito minerado anteriormente, pode-se, portanto, obter informação de qualidade para aprofundamento de temas em ambientes competitivos. Observa-se, no caso, que o conteúdo resultante da busca no Google contém dicas importantes sobre linhas de negócios em expansão na organização e, mais importante ainda, uma revelação sobre o que essa organização pensa acerca desse mercado: “Serviços de saúde representam uma das oportunidades de negócios com TIC com crescimento mais rápido no mundo desenvolvido e em desenvolvimento.” Este trecho representa, de modo auto-evidente para

os conhecedores do mercado, informação relevante para desencadear o estágio de construção do conhecimento no contexto.

O modelo conceitual da Figura 6.12, portanto, apresenta dois conceitos relativamente pouco prováveis em textos sobre organizações do negócio de TIC no mercado: “Healthcare Delivery and Tracking Systems” e “Healthcare Systems”. Estes são os conceitos mais perturbadores nesse tipo de ambiente de informações porque para se compreender o sentido (significado) desses conceitos e associar a eles alguma idéia de negócio será necessário um processo de construção de conhecimento, que poderá ser apoiado na mineração recursiva.

Os demais conceitos da Figura 6.12, referindo-se a objetos mais conhecidos no contexto de mercado de TIC, provavelmente produzirão menos impacto psicológico no sistema cognitivo do engenheiro do conhecimento.

Outro aspecto interessante no texto sobre o pesquisador da “IBM Research” se refere às suas credenciais acadêmicas, revelando sua natureza multidisciplinar em nível de graduação (matemática e química) e um pós-doutorado. A tendência do concurso de profissionais com perfil acadêmico de Ciências Exatas e Engenharias na área médica está se realizando numa disciplina denominada Engenharia Biomédica, multidisciplinar por excelência.

Quanto ao conceito de “Warning System for Automobiles”, uma busca no Google combinando essa expressão com o contexto “IBM Research” resulta em apenas quatro endereços, todos relativos a organizações de pesquisa e desenvolvimento tecnológico da *holding* “IBM Corporation”. O texto dessa página é apresentado a seguir (grifos nossos):

> ***Telematics provides early warnings about quality issues to automobile manufacturers and dealers.***
>
> *The challenge*
>
> *Originally 'telematics' meant the combination of telecommunications and computers. Today, the term covers the many wireless functions originating from an automobile. OnStar™, which combines a global positioning system and wireless communications to provide safety, information and concierge services, is probably the best known. Many of these systems are purchased because of the safety and security benefits they offer.*
>
> *But what if your car could use telematics to continually send information about its performance and tell you the specifics about a problem before you had a breakdown? Unlike the ubiquitous "Check Engine" light, this sort of early warning could deliver real information to auto owners. Add a little bit of intelligence and the early warning could give auto manufacturers "headlights" into quality issues before they become customer satisfaction problems. More importantly, this technology could help control the increase of warranty costs - which add up to more than $8 billion a year in the United States alone.*
>
> *The approach*
>
> *Quality information about cars currently relies on 'lagging' indicators, such as warranty costs, calls to customer service numbers and dealer service records. The trouble is that this information is only available after something has happened. Worse, there is little automatic communication between original equipment manufacturers (OEMs) and suppliers that would allow them to work through the causes of current problems and prevent future ones. IBM is working on a solution that will combine data from multiple sources and provide manufacturers with an Early Warning System to identify potential problems before they become major liabilities.*

> *IBM's Early Warning System uses telematics to capture performance and diagnostic trouble code data from cars as they happen. These leading indicators are combined with the lagging indicators, creating a comprehensive set of information. Complex analysis with clustering and data mining is done to identify trends and vehicles that behave abnormally. Once a problem is identified, OEMs, dealers and suppliers will be able to work collaboratively to track and investigate warranty problems and take corrective action.*
>
> *A portal will provide dashboard views and functionality, failure notification, service reminders for customers and support for collaborative decision-making to track and investigate warranty problems*
>
> *While telematics provides critical functionality for the Early Warning System, one of the more important capabilities is the ability to correlate information across multiple sources and use analytical tools to provide useful information for decision-making and action. IBM Research's strengths in these areas, combined with an understanding of automotive technology and the industry, result in a solution that has the potential to help manage rising warranty costs, contribute to consumer safety and, eventually, help develop a better brand.*
>
> *Next steps*
>
> *A prototype Early Warning System has been built and has monitored 25 parameters on 100 vehicles for three months. The next steps are to integrate additional sources of data and build the collaborative and decision-making support that will make these solutions most valuable for the auto industry.*

Essa simples busca resultou, portanto, nas seguintes informações relevantes no contexto, que podem ser úteis para desencadear um processo de construção do conhecimento para Inteligência Competitiva:

a) a organização IBM Research vislumbra um mercado nascente (“The Challenge”), para um produto inovador – sistemas de alerta em redes telemáticas para automóveis (“Warning System for Automobiles”), valorado em US$ 8 bilhões anuais (que parece bastante atraente para competidores “entrantes”, portanto);
b) a estratégia de pesquisa e desenvolvimento do produto (“The Approach”), que envolve sistemas de informação em rede e serviços aos clientes em potencial, é apresentada em linhas gerais;
c) o atual estágio da pesquisa e desenvolvimento do produto também é informado (em “Next Steps”).

Quando se amplia um pouco a faixa de freqüência de substantivos que compõem os objetos sintagmáticos dos conceitos com os mesmos atributos – Modelo de Tarefa, Modelo de Agente, Modelo de Insumo e Produto, observa-se, no contexto, outros *clusters* de conceitos lexicais úteis, tais como: *application domains, application performance diagnosis, application wiki*. O significado de “application”, neste contexto, geralmente é o de um (aplicativo de) *software*, mas ocorre, também, uma expressão composta com “application” com um significado mais tradicional em linguagem corrente, sem utilidade no contexto: *application for his knowledge* (aplicação para o conhecimento dele).

O modelo de informação integrando os três conceitos específicos mencionados com “application”, no contexto “IBM Research”, é apresentado na Figura 6.13. Observe-se que o

fragmento de modelo resultante pode ser integrado com outro fragmento informacional com "analysis between on-platform agents", partindo-se da premissa que "diagnosis" (diagnóstico) é um tipo específico do gênero "analysis" (análise).

**Contexto: IBM Research**

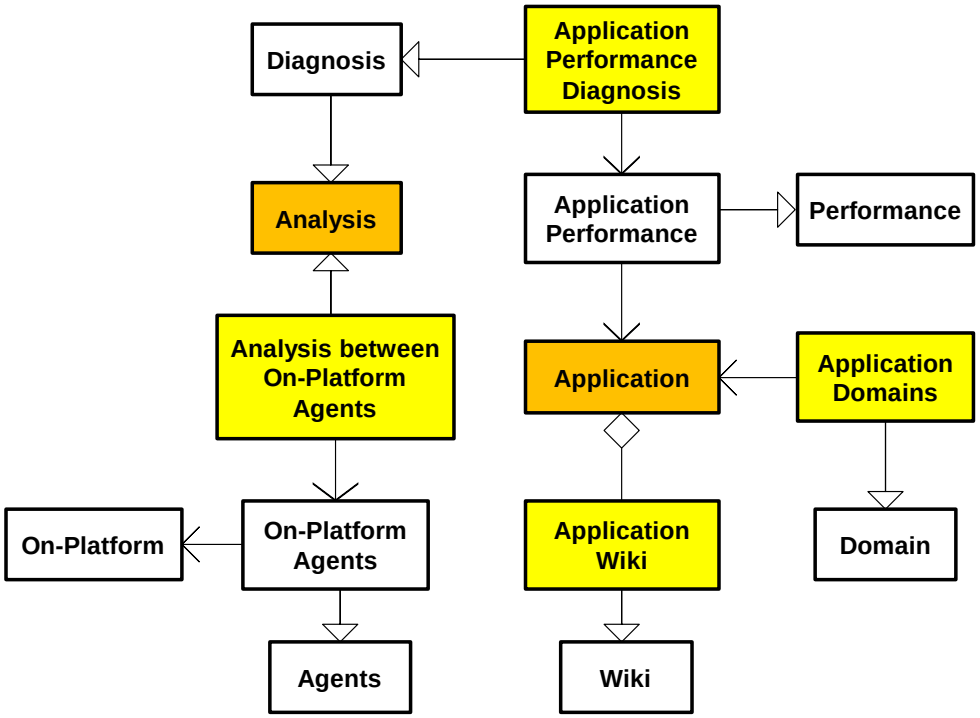



Figura 6.13 Conceitos Integrados em "Application" e "Analysis"
(Fonte: do autor da tese)

O modelo integrado da Figura 6.13 sugere, para conhecedores do mercado de TIC (atributo que se aplica, obrigatoriamente, aos engenheiros do conhecimento envolvidos com Inteligência Competitiva nesse nicho), que a organização do contexto tem produtos para análise do desempenho de aplicações de *software*, inclusive em aplicações processadas em ambientes computacionais compostos por diferentes plataformas tecnológicas, como é comum em sistemas de informação em rede telemática. E deixa evidente, também, que para tanto utiliza agentes[80] de *software* específicos para cada plataforma tecnológica computacional, de modo a capturar informações de controle de desempenho em pontos determinados da rede.[81] Estima-se que esses

[80] O conceito de "agente inteligente", "agente de *software*", ou, simplesmente, "agente", em contextos de TIC, remonta a pesquisas da década de 1970 onde se propunha um *software* autocontido, autônomo e de processamento concorrente com os recursos de um sistema, com missões específicas no ambiente. Esse tipo de *software* atuaria de modo automático respondendo a entradas *(inputs)* do ambiente, com base em sensores, sem comando dos usuários. Os *softwares* de segurança e de monitoramento de desempenho de dispositivos de redes de computadores operam, geralmente, com base nesse conceito.

[81] Em geral, cada plataforma computacional "proprietária" (de fabricante específico, geralmente com características fora dos padrões normativos da indústria) exige que os agentes de *software* de

conceitos não causam tanto impacto no sistema cognitivo do analista como os da Figura 6.12, apresentando, portanto, menores valores relativos de auto-informação. Os conceitos correlatos a "Tracking Systems" e "Healthcare Systems" da Figura 6.12, por envolverem conceitos algo estranhos para o mundo da TIC (conceitos da área de saúde), terão valor relativo maior de auto-informação que os conceitos correlatos a "Application", "Analysis" e "Agents" da Figura 6.13 (que se referem a temas tradicionais da TIC).

Outro objeto sintagmático interessante, no conceito analisado (processos de negócio), é "shopping assistant technology" (objeto nº 828 do APENSO III-B). Embora sua composição lexical apresente, por si, uma idéia do que se trata, uma busca no Google com os argumentos de pesquisa "'IBM Research' and 'shopping assistant technology'" resultam em apenas três endereços de conteúdos publicados pelas organizações de pesquisa da *holding* IBM Corporation. O primeiro endereço, *http://domino.watson.ibm.com/odis/odis.nsf/pages/solution.10.html,* apresenta o seguinte conteúdo temático:

> **Tough Problems Solved**
>
> **Personal Shopping Assistant**
>
> *Technology guides customers through store aisles, suggesting items to purchase, pointing out sales and easing (or avoiding) the check-out line*
>
> *It's not uncommon for busy shoppers to lose their place – and valuable time – amid the dizzying array of products and retail store layouts. And retailers – in the midst of the business of shopping – often let marketing and customer-service opportunities slip through their fingers. But this is an opportunity they cannot afford to squander: The competition to attract and retain customers is a growing challenge, so retailers must find ways to establish stronger customer ties and improve the shopping experience.*
>
> *Many companies now use loyalty card programs that offer discounts to repeat customers and gather data about their buying habits. The success of such programs relies on how companies use and manage that data. Many have spent years gathering information with few practical applications, hoping that technology would one day catch up with their needs and offer a way to turn data into dollars.*
>
> *That day is here. IBM Global Business Services, IBM Research and IBM Retail Store Solutions teamed with Cuesol, an IBM Business Partner, to find a way to help turn customer information into useful, profitable action for a large retail grocery chain. With more than 345 stores in the northeastern United States, the company was seeking to solidify its customer base and trigger incremental sales.*
>
> *Although the grocer already had an existing loyalty card program, it wanted a more precise, cost-effective way to reach customers – not after they made their purchases, but during the decision-making process. The company sought a solution that would allow customers to better manage their time in the store, help them navigate the aisles quickly and reduce wait times. The grocer also wanted to hone its marketing efforts to reach shoppers with relevant, targeted promotions as they traversed the aisles.*
>
> *The solution was the IBM Personal Shopping Assistant called Shopping Buddy, which consists of a data management system, Wi-Fi network, infrared technology and Bluetooth transmissions centered around touch-screen computers mounted on shopping cart handles. It uses IBM Store Integration Framework, IBM WebSphere® Application Server, IBM WebSphere MQ messaging software, IBM Tivoli® management software, IBM Mobile Tablet for Retail, as well as Cuesol's Cart Companion® browser software.*

---

monitoramento de desempenho se adaptem às suas características tecnológicas, sob pena de não se viabilizarem na missão.

> *IBM Research teamed with IBM's Retail Store Solution division to assist in prototyping various aspects of the IBM Personal Shopping Assistant for the Store Integration Framework. An in-store retail commerce server prototype was created, which integrated the Shopping Buddies, the location service and the point-of-sale terminal controller. The service-oriented architecture adopted by the in-store server provided the basis for the prototyping of various store services, such as data replications, offer presentations and gift registry. The in-store server also served as the test environment for the development of the Shopping Buddy prototype.*
>
> *In practice, customers scan in their loyalty cards to activate a familiar, Web-style screen with a variety of display options, such as sale items or a list of products shoppers buy most frequently. A location-tracking system monitored through ceiling-mounted beacons enables the retailer to pinpoint shoppers' locations and deliver relevant real-time information as they move through the store.*
>
> *The system integrates with the company's backend systems, so buying histories and favorite items can be displayed on the screen as a constant reminder of products to buy. The device also lets shoppers order cold cuts from the deli, sending an alert when the order is ready. An attached imaging scanner invites consumers to scan items as they place them in the cart, keeping a running total (of expenses and savings) along the way and completing their transaction using IBM self-checkout systems.*
>
> *Based on the success of the initial engagement, the company plans to extend the technology to as many as 150 locations.*
>
> *Meanwhile, IBM Research is exploring ways to refine Personal Shopping Assistant technology with customer-sensitive features, such as suggesting a wine to go with a meal or providing dietary guidance on specific items. Researchers see Shopping Buddy as a prime candidate for moving Web technology closer to the physical world, with, for example, ads targeted to individual consumers, based on prior purchases, in the precise environment where they are primed to make a purchase – just as camera ads show up on your computer screen when you query your search engine for camera. IBM researchers also say the technology could be extended to other retailers and retail sectors.*
>
> *IBM Personal Shopping Assistant puts technology at shoppers' fingertips, giving them better control over their shopping experience, and provides retailers with a practical on demand solution.*
>
> *For more information on IBM Personal Shopping Assistant and to explore other ways to enhance retail commerce, contact IBM Research Services today.*

Obviamente, trata-se de um tipo de serviço inovador, baseado em sistemas de informação com bastante aporte computacional, destinado, em princípio, para orientação de clientes de grandes lojas de departamentos em suas expedições de compras – daí o conceito de "assistente de compras" *(shopping assistant)*. Deve-se ressaltar, a propósito, que esse tipo de serviço ao cliente está previsto na rota evolutiva prevista para a *Web*, no estágio de *Web* 3.0 e além (conforme a Figura 5.3). O processo de criação de significado, neste caso, se encontra estruturado previamente pelos pesquisadores da evolução provável da *Web*, representando uma opção importante a ser considerada pelo engenheiro do conhecimento no estágio seguinte do construto de Choo (2003): a construção do conhecimento.

Com essas informações relevantes sobre a evolução da *Web*, a questão padrão que se apresenta é evidente: "Quais os impactos mais prováveis dos serviços eletrônicos de assistência a processos de compras sobre a organização que patrocina o esforço de Inteligência Competitiva?". O conhecimento de contexto deverá, pois, ser construído em torno das respostas possíveis a essa

questão. As organizações do mercado de TIC terão, certamente, seus modelos de negócio afetados por esse tipo de tecnologia porque elas operam, na maioria, no conceito de “monte-você-mesmo-seu-produto”, com auxílio de *softwares* para precificação automática de seus produtos nos *Web* Portais corporativos.

Quanto aos demais endereços sugeridos pelo Google, o segundo deles *(http://domino.watson.ibm.com/odis/odis.nsf/pages/focus.01.html)* leva o internauta ao seguinte conteúdo (apresentando-se apenas o cabeçalho):

> ***FOCUS on…***
>
> ***Innovation in Customer Relationship Management***
>
> *IBM Research Services offers clients an assortment of powerful CRM tools designed to assist in sifting through massive amounts of data to learn what motivates and influences consumers' choices and buying patterns, help identify high-value customers, and help transform that information into the knowledge and insight necessary to optimize their most lucrative customer relationships.*

O conteúdo dessa página representa o outro lado da moeda em relação produto inovador da classe “Shopping Assistant Technollogy”: as vantagens para a organização que o emprega em suas lojas. Com esse tipo de tecnologia, o lojista poderá conhecer os padrões de compras e consumo dos clientes e atuar, daí por diante, em duas linhas de ação concorrentes em termos de eficácia e eficiência: (1) reter seus atuais clientes com a sedução da comodidade tecnológica à disposição, pretensamente sem similar nas organizações concorrentes, e (2) competir em posição vantajosa na busca de novos clientes.

O respectivo conteúdo apresentado a seguir, na íntegra, sugere como seriam os processos de negócio para a disponibilização desse tipo de produto aos clientes dos clientes da IBM Corporation (com aporte tecnológico desenvolvido pela IBM Research):

> ***Finding the person hidden in the data***
>
> *Management theories come and go, but experience shows that it can cost many times more to acquire a new customer than it does to retain an existing one. Savvy businesses understand that their ability to compete successfully in a global marketplace depends on attracting new clientele while continuing to nurture and grow their core customer base. At the same time, common wisdom indicates that 20 percent of customers account for 80 percent of sales, so it's imperative that successful firms identify their most valuable customers and target their efforts toward boosting their satisfaction. With those axioms in mind, companies are turning to IBM Research Services - a collaboration between IBM Research and IBM Global Business Services - for assistance in sifting through massive amounts of data to learn what motivates and influences consumers' choices and buying patterns, and which customers are likely to prove most lucrative. IBM Global Business Services then assists clients in transforming that information into the knowledge and insight necessary to optimize their customer relationships. In concert with IBM Research, IBM Global Business Services offers clients a wide range of expertise, techniques and tools geared toward devising customer relationship management (CRM) strategies that will help cement customer loyalty at the highest levels and provide a competitive advantage over rival firms.*
>
> ***Customer experience modeling***
>
> *One key to motivating customer loyalty lies in understanding consumer purchasing behavior and identifying the critical success factors that drive sales. In one engagement with a large U.S. retailer, IBM took a multidisciplinary approach to*

*developing a state-of-the-art model of consumers' decision-making processes using an innovative, scientific methodology that integrated techniques from ethnography, psychology, psychometrics and advanced statistics.*

*Unlike traditional "predictive" marketing models, the new IBM model targets cause-and-effect relationships rather than correlations. Examining more than 1,500 variables reflecting every aspect of the retailer's customer experience - including location, advertising, salespeople and pricing - the model calculated the ability of each one to boost sales and zeroed in on ten top success factors. By applying the results to a series of new marketing strategies, the retailer boosted sales and was even able to double sales in one product category. In addition to removing much of the guesswork in designing marketing campaigns, the model also provided side benefits, including insight into consumers' mental models that helped the client understand its target audience, product-positioning, and whether up-selling strategies were likely to be successful.*

***Customer lifetime value management***

*Cross-channel marketing via lifetime value modeling enables a retailer to leverage the full potential value of its existing customer base to drive long term profits across all of its contact points. When Saks Fifth Avenue sought to optimize its cross-channel marketing, it turned to IBM Research/IBM Global Business Services, which developed a novel methodology to help track a customer's response on one channel as a result of contact on another, and expand the overall customer value. Our approach is unique in the way it provides marketing rules that are optimized with respect to long term profits throughout the lifetime of a customer, and across different channels that the retailer may use. This capability was realized by an intricate blending of advanced predictive modeling technology and mathematical modeling (using the so-called "Markov Decision Processes") of the dynamics of retailer customer interaction. The system was benchmarked against Saks' existing CRM methodology for managing its direct mail and store channels, and initial results suggest that using the IBM approach could help boost store revenues.*

***Customer equity and lifetime management***

*IBM's customer equity and lifetime management (CELM) tool offers clients further insight into customers' long-term value and helps them devise strategies to manage, support and nourish customer relationships. CELM uses specialized algorithms and analytical techniques to capture and analyze customer dynamics in order to model and improve long-term customer relationships in a way that can help boost the value/risk ratio of the overall customer portfolio. Companies may reallocate marketing budgets dynamically to help maximize their return on investment and increase customer lifetime value and loyalty. When Finnair wanted to redesign and improve its customer loyalty management process, it asked IBM for assistance and ended up being an important partner in the practical development of CELM. Using CELM's analytical capabilities, Finnair was able to establish profiles of its high-value customers and focus its marketing strategy on passengers fitting those profiles. As a result, the airline reduced market costs by more than 20 percent and improved response rates by up to 10 percent.*

*In another engagement, IBM Research, IBM Global Business Services and IBM Retail Store Solutions teamed with Cuesol, an IBM Business Partner, to help turn customer information into useful, profitable action for a large retail grocery chain. The company sought a solution that would promote customer loyalty by enabling them to better manage their time in the store. The grocer also wanted to hone its marketing efforts to reach shoppers with relevant, targeted promotions as they traversed the aisles. The solution was the IBM Personal Shopping Assistant, which consists of a data management system, wi-fi network, infrared technology and Bluetooth transmissions centered around touch-screen computers mounted on shopping cart handles. Customers scan in their loyalty cards to activate a familiar, Web-style screen with a variety of display options, such as sale items or a list of products shoppers buy most frequently. The Personal Shopping Assistant creates a map of the store and displays a suggested route. A location-tracking system monitored through ceiling-mounted beacons enables the retailer to pinpoint shoppers' locations and deliver relevant real-*

*time information as they move through the store. Based on a successful trial, the retailer is planning to extend the technology to as many as 150 locations.*

***Text mining***

*Businesses typically produce massive amounts of unstructured data related to customer interactions, and sophisticated tools are needed to analyze the information and put it to use in improving the customer experience. The IBM Business Insights Workbench (BIW) can help clients mine these enormous collections of unstructured data to help companies meet competitive challenges. Unlike more conventional search and retrieve tools, BIW creates natural classifications (taxonomies) of the data objects and analyzes them in conjunction with structured information to find trends, co-occurrences and other insights. Clients use the BIW tool to help improve customer retention by identifying and remedying causes of dissatisfaction. The tool can help pinpoint underlying problems through analysis of common customer complaints and phone calls and may lead to improved product design as a result.*

*In an engagement with a large insurance and financial services company, the BIW tool was used to assist in analyzing e-mail messages sent by customers regarding their credit card and online banking accounts. As a result, the company was able to separate out the e-mails from disatisfied customers and identify common problem areas for resolution. Similarly, a major life insurance firm used BIW to help analyze interactions between its call center help desk and customers. Transcripts of each call were mined to look for patterns, such as repetitive calls from the same person, that indicated an issue was not being resolved. Both companies are developing strategies to use the BIW tool on an ongoing basis to help them target inefficiencies and problems before they develop into major crises.*

***Meeting competitive challenges with CRM***

*In today's speed-of-light, information-saturated business climate, a company's success often depends on turning mountains of data into actionable knowledge that can be used to identify and profile its high-value customers and their buying patterns. IBM Research Services offers clients an assortment of powerful CRM tools designed to assist in analyzing data and generating the information and insight needed to focus business strategies and marketing efforts on retaining and expanding this core customer base, while gaining a competitive edge in the global marketplace.*

É oportuno destacar-se, dos conceitos tratados nesse texto da IBM Research, um conceito mais fundamental utilizado pelas organizações mais competitivas nos mercados: o da identidade comportamental de cada cliente individualmente. Embora a sociedade ocidental contemporânea tenha se tornado uma sociedade de massas, com consumo de massas, o mercado está se voltando para as transações onde o cliente é tratado como se fosse único, buscando-se atender a suas necessidades específicas. O que está implícito nesse conceito de identidade é algo como um DNA do cliente, um padrão comportamental de consumo que o identifique entre outros clientes – obviamente, muitos clientes poderão ser agrupados por similaridade de padrões, sem prejuízo do conceito.

Como vimos ao longo dos modelos apresentados nesse contexto, a organização IBM Research apresenta em seu *website* uma “mensagem para o mundo” bastante competitiva, anunciando projetos de desenvolvimento de produtos tecnológicos tanto para desbravar mercados emergentes (como “entrante”, no modelo de inteligência baseado nas forças de Porter) como para se consolidar em mercados atuais (na condição de *player* atual). Como entrante, a organização promete produtos inovadores para os mercados de serviços de saúde, de proprietários de automóveis e redes lojistas de varejo; e, como *player* atual, mostra como pretende satisfazer

atuais clientes com seus aplicativos de monitoramento e diagnóstico de problemas de desempenho em sistemas computacionais.

6.1.2.4 Modelos de informação do contexto “Microsoft Research”

Os 23 objetos sintagmáticos do contexto “Microsoft Research”, agrupados no conceito de processos de negócio, representando 2,5% dos substantivos mais freqüentes, são: *CAD tools, development of systems, communications systems, computer graphics achievement, computer hacking, computer interaction, computer recognition, computer science instruction, computer science research, computer simulations, computing systems, computing systems of the future, control of a computer, conversion systems, goal for all of Heckerman’s research, human-computer interaction, intersection of human computer interaction, life-support systems, research platform for sensor networks, systems management, technology transfer agents, technology transfer program manager, work in computer security.*

Com a Figura 6.14, têm-se fragmentos de modelos de informação integrados no conceito central de “System”. Como era de se esperar, esse é um dos conceitos essenciais do negócio no contexto “Microsoft Research”, ao lado de “Computer” (Figura 6.15).

**Contexto: Microsoft Research**

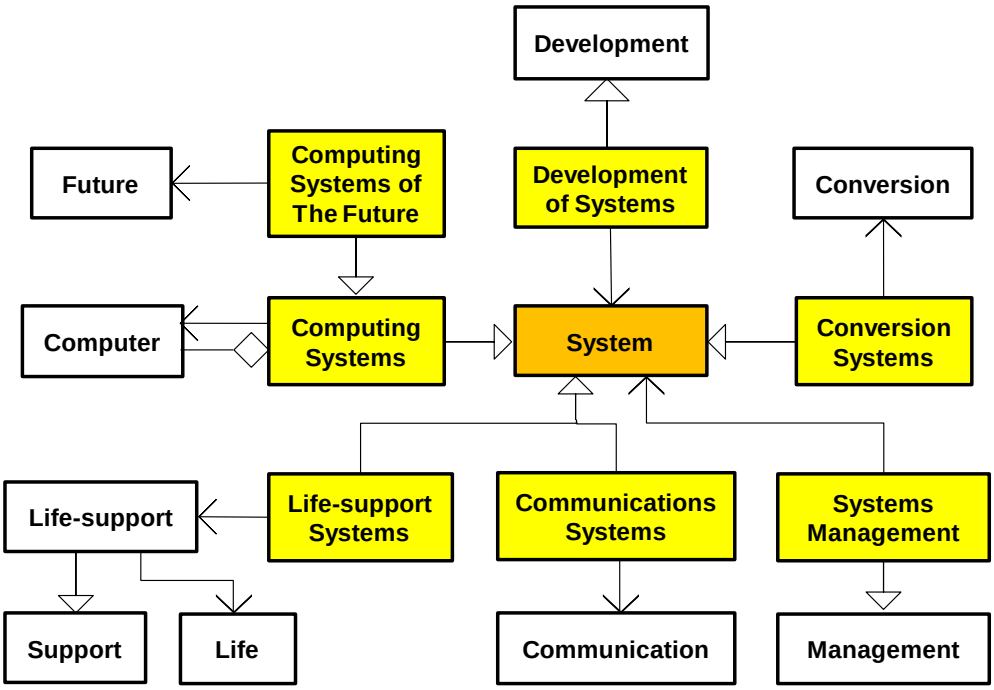



Figura 6.14 Conceitos Integrados em “System”

(Fonte: do autor da tese)

O modelo informacional da Figura 6.14 apresenta alguns conceitos triviais nesse ramo de negócio, como “Development of Systems”, “Communications Systems” e “Systems Management”, mas também conceitos que merecem, até para quem conhece esse nicho de mercado, um aprofundamento das investigações, como “Computing Systems of the Future”, “Life-support Systems” e “Conversion Systems” – esses conceitos apresentam, portanto, maior valor relativo de auto-informação.

Os primeiros 20 endereços URL de uma série, recuperados com o Google utilizando-se os argumentos de busca “‘Microsoft’ and ‘life-support systems’”, se referem todos ao conceito de “Life-support Systems” da organização Microsoft Research. O primeiro deles *(http://research.microsoft.com/en-us/groups/ecology/default.aspx)* revela, de plano, a força semântica desse conceito com o seguinte texto (grifos nossos):

> **Computational Ecology and Environmental Science**
>
> *Developing novel computational tools and methods to predict and mitigate the rapid changes occurring in the earth’s life support systems.*
>
> *Understanding the earth’s life support systems, and predicting and mitigating the rapid changes that are occurring in these systems because of human activities is one of the great global scientific challenges humanity is currently facing. The programme in ecological and environmental sciences aims to contribute to meeting this challenge by identifying critical scientific problems and developing novel computational methods and tools to address these problems. Thus our research is necessarily broad, ranging from fundamental theoretical work; through data gathering and manipulation; to developing advanced predictive models of biotic and coupled biotic and physical systems at scales from local to global.*
>
> *In tackling scientific problems in these areas, we face a broad range of technological challenges, including computational modeling of complex systems, integrating models and data, data acquisition and management, visualization techniques, and tool usability, maintainability and extensibility.*
>
> *Our scientific research is grouped into four Research Units:*
>
> *Plant ecology*
> *Ecological networks*
> *Spatial ecology and biogeography*
> *Behavioural ecology*
>
> *To facilitate the above research, and computational ecology more widely, we develop:*
>
> *Tools for Ecoinformatics and Biodiversity Informatics*
> *Tools for Environmental Management and Education*

Com base nos conteúdos das páginas desse primeiro endereço na *Web*, o programa de pesquisas da Microsoft Research na linha de suporte à vida é bastante amplo e apresenta, curiosamente, duas vertentes em direções opostas, mas partindo do mesmo conceito central baseado nos sistemas naturais: a primeira vertente se refere ao desenvolvimento de sistemas computacionais de apoio às pesquisas sobre a vida no planeta Terra, com temas atuais como a ecologia; a segunda se refere à emulação de sistemas naturais em computadores, ou o que essa organização denomina “computação natural” *(natural computation)*, algo que se parece com a Inteligência Artificial baseada na natureza.

Obviamente, esse programa de pesquisas é multidisciplinar e o autor revela tratar-se de uma iniciativa de longo prazo que deverá ser desenvolvida *na interface da biologia, química, engenharia e tecnologia da informação explorando a computação inspirada na natureza e a computação biológica*.

**Contexto: Microsoft Research**

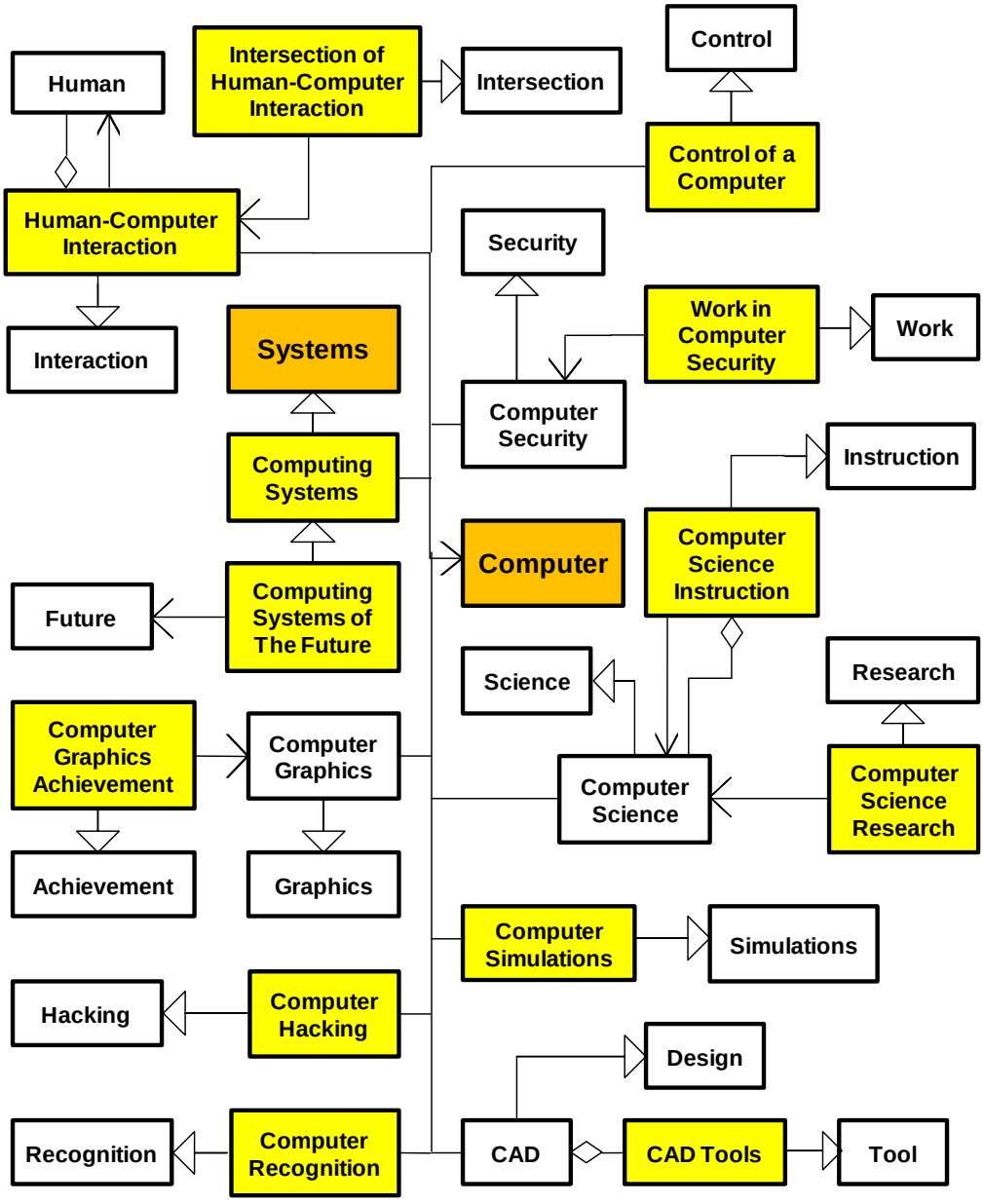



Figura 6.15 Conceitos Integrados em “Systems” e “Computer”

(Fonte: do autor da tese)

Quanto ao conceito de “conversion systems”, de acordo com os conteúdos das páginas recuperadas utilizando-se o mesmo algoritmo no Google, parece tratar-se de uma linha de pesquisas mais antiga da Microsoft Research, que desde o final dos anos 1990 investe no

desenvolvimento de sistemas para conversão de sinais e fonemas e para conversão de linguagens de textos. O documento recuperado como o primeiro endereço URL é um artigo publicado por cientistas da Microsoft Research sobre a tecnologia conhecida como OCR *(Optical Character Recognition)*, indispensável para digitalização de documentos produzidos com suporte em mídia papel. Contudo, resta saber se esses artigos de cientistas publicados na *Web* aberta não significam que a Microsoft está abandonando essa linha de pesquisas, como nos relatos e na lógica de inteligência apresentada por Fuld (2007).

Com uma busca no Google para "'Microsoft Research' and 'computing systems of the future'", apenas três endereços são recuperados, sendo dois deles de páginas do *Web* Portal da conhecida organização AAAI *(Association for Advancement of the Artificial Intelligence)*. O conceito é bastante genérico, referindo-se a vários objetos considerados como de sistemas computacionais do futuro; essa percepção é reforçada com uma busca no próprio *Web Portal* da Microsoft Research, que retorna mais de setenta mil endereços de documentos versando sobre uma ampla gama de sistemas computacionais considerados futuristas em várias épocas.

"Computing Systems of the Future", da Figura 6.14, pode ser integrado no modelo centrado em "Computer", como na Figura 6.15, assumindo-se que "Computing Systems" tem um atributo-componente denominado "Computer" (apenas substituindo-se, no caso, uma necessária classe-atributo "Computing" por "Computer").

Outro conceito que desperta alguma curiosidade, no contexto, é "Computer Simulation", devido ao potencial de uso da simulação baseada em computador como método de pesquisa em todas as ciências. A busca no Google com "'Microsoft Research' and 'computer simulation'" retorna nada menos que 11 mil endereços de documentos, aproximadamente, confirmando o uso em larga escala desse tipo de recurso metodológico e tecnológico em várias áreas do conhecimento e o envolvimento dessa organização com o tema. Conclui-se, portanto, que a maioria dos conceitos correlatos a "System" e "Computer" apenas reforçam o aprendizado preliminar sobre o contexto Microsoft Research, mostrando conceitos mais óbvios e, também, como os vários referentes instanciam os sentidos dos objetos sintagmáticos minerados – uma visão mais periférica do ambiente informacional irradiando-se a partir dos conceitos centrais.

Com os fragmentos integrados no modelo da Figura 6.16 encontram-se três conceitos interessantes, por não serem triviais no mercado de TIC: "Collaboration with AIDS Researchers", "Research Platform for Sensor Networks" e "Goal for Heckerman's Research". O primeiro conceito mencionado, que de fato é um conceito mais específico, pois contem atributos do "Modelo de Ambiente Externo" (partindo-se da premissa que AIDS[82] é um tema de interesse geral da sociedade mundial), é integrado aos outros dois, no modelo de informação, assumindo-se que a classe "Research" tem um atributo denominado "Researchers".

[82] AIDS: *Acquired Immunodeficiency Syndrome*.

**Contexto: Microsoft Research**

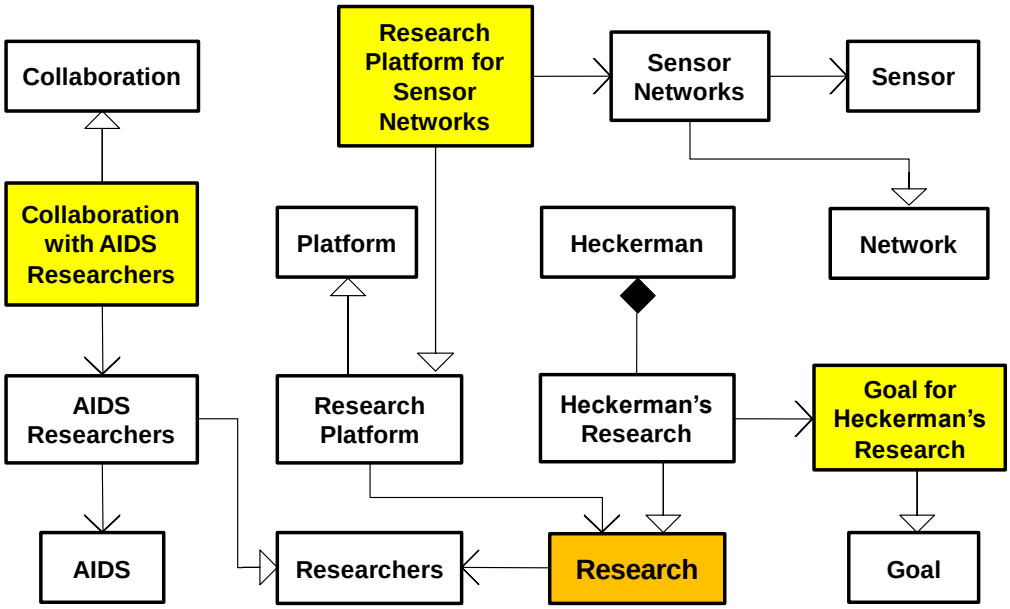



Figura 6.16 Conceitos Integrados em "Research"

(Fonte: do autor da tese)

De novo, como os conceitos básicos dos ambientes de TIC e de saúde não têm conexões epistemológicas auto-evidentes, causa certa surpresa a ocorrência de conceitos com objetos sintagmáticos interligando essas duas áreas do conhecimento. Assim, pode-se estabelecer maior valor relativo de auto-informação para esses conceitos multidisciplinares que para os conceitos mais fundamentais de cada um dos dois nichos de mercado.

Quando se utiliza os recursos do Google para recuperar conteúdos com "Microsoft Research" e "collaboration with AIDS researchers", obtém apenas um endereço: *http://research.microsoft.com/en-us/about/brochure-5.aspx.* O conteúdo da primeira página desse endereço, bastante esclarecedor sobre o envolvimento da Microsoft Reserach com o tema, é apresentado a seguir (grifos nossos):

> **Supporting Advances Throughout the Sciences**
>
> *Advances in computer science don't just benefit computing. They play a growing role in helping scientists in all fields make new contributions, facilitating discovery and advancing the state of the art in everything from health care to education to the environment, while also addressing some of society's most pressing challenges. Since 1998, Microsoft Research's External Research Division has led the organization's collaborative projects with leading scientists and researchers in academia, government, and industry to apply the latest innovations in software to scientific challenges in other fields. The division also creates open tools, technologies, and services to support every stage of the research process. It also supports the aspirations and potential of early-career researchers and promising young scientists around the world through internships, scholarships, and awards. Below is a sampling of how Microsoft researchers and groups are engaging with the broader scientific community:*

*Worldwide Collaboration*

*Microsoft Research collaborates with medical researchers at the Ragon Institute of MGH, MIT, and Harvard; the Fred Hutchinson Cancer Research Center; the University of Washington; and the University of Oxford to design an effective vaccine for HIV/AIDS, in which techniques for machine learning are used to identify the most vulnerable parts within the immune system. Researchers from the Pontificia Universidad Católica de Chile and the Instituto Politécnico Nacional in Mexico collaborate with Microsoft Research on a project to use computer-vision technology to improve quality control in the food industry — a key part of the Latin American economy.*

*In the Swiss Experiment, an international collaboration, scientists are using Web-based tools developed by Microsoft researchers to view huge quantities of environmental-sensor data more efficiently and in richer detail than ever.*

*Microsoft Research Asia's eHeritage program works with academic partners throughout the region to apply the latest technologies to aid the preservation, interpretation, and dissemination of cultural and natural heritage for research, education and protection. The program has funded more than 10 research proposals over the past several years, from universities such as the University of Queensland in Australia and the University of Tokyo.*

*In India, Microsoft teams with the Molecular Biophysics Unit at the Indian Institute of Science on a project called From Genomics to Function. Identifying a number of target organisms, such as pathogens, that play a key role in the lives of humans, researchers are using computers to learn more about how the organisms' genes relate to their functions.*

*Researchers at the University of Queensland and a consortium of government, academic, community, and environmental groups collaborate with Microsoft Research on the Health-e-Waterways project, which is developing data-management tools to help water-resource managers to react to changes in water quality, quantity, and aquatic species.*

*Microsoft has also established more than 30 joint research institutes to support innovative research projects. For example, Microsoft Research opened three in Europe: the University of Trento Centre for Computational and Systems Biology, which uses programming language theory to design new computational tools for biology; the Microsoft Research-INRIA Joint Centre, which focuses on longterm research into formal methods and computing security; and the BSC-Microsoft Research Centre in Barcelona, which focuses on the design and interaction of next-generation processors.*

*Other global joint research institutes include the Games for Learning Institute at New York University, which seeks ways to use computer gaming to improve instruction in science, math, literacy, and other academic disciplines. We also spearhead Microsoft's academic multi-core relations through the Universal Parallel Computing Research Centers, a joint activity between Microsoft and Intel that supports client multi-core research at the University of Illinois at Urbana-Champaign and the University of California at Berkeley.*

*The Microsoft Institute for Japanese Academic Research Collaboration is enabling faculty and students in one of the world's leading economies to combine their tradition of technological development and product innovation with Microsoft's computer-science expertise.*

*Investing in People*

*Microsoft operates the largest Ph.D. internship program in the information technology industry. Each year, nearly 1,000 top computer-science students have the opportunity to work at one of Microsoft Research's locations around the world. A variety of Ph.D. fellowship and scholarship programs have benefited hundreds of students worldwide over the past decade. In 2008, Microsoft awarded scholarships to nearly 80 Ph.D. students, including 10 through our Graduate Women's Scholarship program. We support young computer scientists through an assortment of summer schools, student*

> *clubs, visiting professorships and awards. Microsoft Research runs numerous academic conferences and workshops where students and university faculty members can interact and exchange ideas with top computer-science researchers in academia, government, and industry. In 2000, the first Microsoft Research Faculty Summit drew 150 participants; since then, more than 25,000 scientists, academic researchers, faculty members, and students have attended Microsoft Research-sponsored summits, conferences, and workshops held in Asia, India, the United Kingdom, and Latin America.*

E repetindo-se o mesmo procedimento de busca no Google para "'Microsoft Research' and 'Research Platform for Sensor Networks'", obtêm-se apenas dois endereços: *http://www.groklaw.net/articlebasic.php?story=20070614231022599* e *http://research.microsoft.com/en-us/about/brochure-8.aspx.* O primeiro apresenta uma análise crítica sobre o comportamento da *holding* Microsoft em pesquisa e desenvolvimento, que para o autor trata-se de uma "compra de cérebros" das universidades a preços competitivos (para a Microsoft). O segundo endereço não apresenta o conteúdo sugerido, presumindo-se que tenha sido removido anteriormente, apesar do algoritmo do Google ainda apontar para o mesmo. Todavia, uma pesquisa mais aberta no Google revela que a Microsoft tem patrocinado pesquisas de redes de sensores, consideradas "plataformas de pesquisas" em alguns projetos, como o apresentado no documento do endereço *http://www.csiro.au/files/files/porx.pdf.* Outro documento da própria Microsoft Research, disponível no endereço *http://research.microsoft.com/en-us/labs/asia/msrabrochure_english.pdf*, confirma esse fato.

O conceito de "redes de sensores" trata de dispositivos sensíveis (sensores) para captação de dados relevantes do meio para armazenamento, organização e análise em contextos de gestão do conhecimento (pesquisas científicas de mudanças climáticas, por exemplo). Essa rede é constituída de uma série de nós dispersos, estrategicamente, numa determinada região, relativa ao fenômeno em estudo, onde se encontram os sensores que captam os dados, e de um nó central onde se concentra o poder computacional de processamento dos dados, para posterior análise por especialistas temáticos. Os dados capturados nos nós são transmitidos automaticamente, geralmente por redes sem fio, ao nó central da rede.

O conceito "goal for all of Heckerman's research", com a busca no Google, leva a seis endereços na Web, todos referentes ao mesmo conteúdo. O primeiro endereço, da própria *holding* Microsoft *(http://www.microsoft.com/presspass/features/1999/03-22heckerman.mspx),* apresenta o trecho abaixo, esclarecendo do que se trata (com grifos nossos):

> ***Microsoft Research: Making Computers More Intelligent and Responsive***
> *Microsoft researcher David Heckerman combines probability theory with artificial intelligence to make computers "aware" of users' needs*
>
> *REDMOND, Wash., March 22, 1999 — David Heckerman entered Stanford University medical school in 1980 to learn about the human brain. At the time, he was asking questions like, "What is the nature of human awareness?" and, "Are humans simply fancy computers that can understand and direct their own existence?"*
>
> *Two decades later, Heckerman still contemplates these questions, but from a different perspective. Rather than questioning whether the brain works like a computer, he now asks whether it's possible for a computer to emulate the human brain. Can computers*

> *be "aware?" Can they offer a level of intelligence that resembles the sophisticated processes of the human brain?*
>
> *A senior researcher in the Decision Theory and Adaptive Systems (DTAS) Group at Microsoft Research, Heckerman is approaching these questions armed with a background in statistics, medicine and artificial intelligence. His work centers on using data in sophisticated ways to make computers "anticipate" the desires of users so they can more efficiently serve people's needs.*
>
> *"People use several phrases to describe what I do, including statistics, machine learning, and data mining, but it's really all the same," Heckerman says. "My work centers around learning from data. There's a sea of data on the Web, in computer databases, everywhere. I want to take that data and gain some insight and knowledge out of it, so we can make smarter decisions."*
>
> *A boyish looking man in his early 40s, Heckerman demonstrates energy and passion when discussing his work. Despite the ambitiousness of his research, he has the unusual gift of explaining his work in the simplest terms. A couple of hours talking to him, and it's clear that what most people regard as intensely challenging, Heckerman sees as logical and straightforward, even simple.*
>
> *Heckerman combines data with expert knowledge to make predictions about complex problems. What differentiates his work from that of traditional statisticians is that the predictive models he builds-called Bayesian networks-capture cause and effect relationships about the world.*
>
> *The implications of Heckerman's research are enormous. Already, his work is helping people eliminate junk mail from their e-mail in-boxes and easily obtain a sophisticated level of computer technical support without placing a phone call. It is also enabling businesses to better target customers by predicting the habits of computer users who browse or shop online. While his research has far-reaching implications for how computers will be used in the future, the underlying goal for all of Heckerman's research is to build "intelligence" into the computer to make it a far more useful tool than it is today.*
>
> *"The idea is that when you use your machine, it will form guesses about what you're trying to do and help you," Heckerman says. "It will be like having a butler."*

Essa informação confirma, por outro lado, o que se observou nos conceitos das Figuras 6.14 e 6.15 e na experiência dos jogos de guerra relatados por Fuld (2008), sobre as mais prováveis estratégias a serem desenvolvidas, futuramente, pela Microsoft para manter seu poder competitivo no mercado. Ou seja, pode-se concluir que a Microsoft deverá continuar a concentrar a essência de sua estratégia nos microcomputadores pessoais de milhões de usuários de seu Sistema Operacional *Windows* pelo mundo afora, agregando inovações como as vislumbradas por Heckerman, a cada versão, para retenção do interesse de seus clientes.

Os sensores inteligentes mencionados na introdução a esta tese podem ser implementados, portanto, mediante uso da metodologia de mineração e modelagem conceitual proposta, integrando conceitos “de superfície” com “primitivas” conceituais desenvolvidas com população e aprendizado de ontologias.

#### 6.1.2.5 Fujitsu Research *vs*. IBM Research *vs*. Microsoft Research

A Figura 6.17 representa um modelo mental com os conceitos minerados no contexto competitivo das três organizações: Fujitsu Research, IBM Research e Microsoft Research. Os

conceitos centrais do discurso de cada organização apresentados na *Web* se encontram no círculo interior, tais como “Delivery”, “Service”, “Analysis” e “Computer”; os conceitos derivados se encontram nas elipses dispostas na periferia do círculo central em cada contexto organizacional, como “Value in Service Innovations”, “Warning System for Automobiles” e “Life-support Systems”.

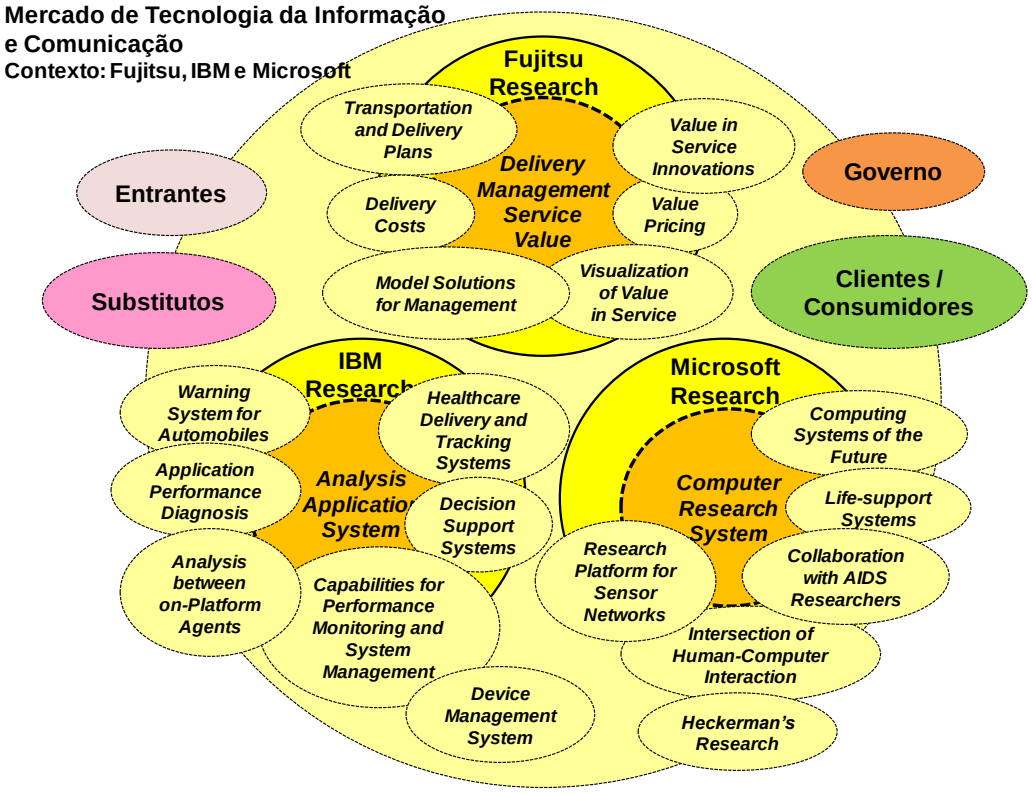



Figura 6.17 Modelo Mental Comparativo dos Conceitos de Negócio Minerados
(Fonte: do autor da tese)

Com esse modelo, pretende-se apoiar a análise comparativa dos conceitos de negócio utilizados pelas três organizações em sua “mensagem para o mundo” numa abordagem mista indutiva e dedutiva, com inspiração nas noções de ontologia superficial *(shallow ontology)* e ontologia profunda *(deep ontology).* A propósito, Staab (2006, p. 99) argumenta que:

> *Nem todas as ontologias têm as mesmas características e, em geral, podemos distinguir ontologias profundas de ontologias superficiais. Ontologias profundas são, frequentemente, aquelas encontradas na ciência e na engenharia, onde consideráveis esforços são empregados na construção e no desenvolvimento da conceptualização. Em domínios tais como proteômica e medicina, a ontologia está num sentido muito real dos dados de interesse. Isso se torna aparente quando usamos uma ontologia para classificar conjuntos complexos de propriedades como constituintes de certos tipos de objeto.*
>
> *Ontologias superficiais contêm poucos termos relativamente imutáveis que organizam volumes de dados muito grandes – por exemplo, termos como cliente, número de conta, e saque a descoberto, utilizados em contextos bancários e financeiros, ou as relações básicas que definem informações geoespaciais.*

O que se propõe para análise dos modelos conceituais minerados no método em tela é considerar-se os conceitos obtidos na Análise de Conceito Formal como uma ontologia superficial, mas buscando-se interpretá-los para se chegar à compreensão de uma ontologia profunda mínima referente ao contexto. Essa ontologia superficial seria baseada no léxico, em particular nos sintagmas nominais minerados, e especialmente nos que representam os substantivos mais freqüentes. Embora não intencional, esse tipo de ontologia seria mais atomista (no conceito de Fodor), pois os segmentos da linguagem natural minerados com o método dispensam, geralmente, desambiguação, pela sua estrutura sintagmática complexa, além de sua utilidade limitar-se ao contexto.

Intuitivamente, pode-se argumentar que provavelmente a ontologia superficial apresentará mais pistas para desencadear o estágio de criação de significado porque ela terá mais objetos com os maiores valores relativos de auto-informação que a ontologia profunda. Em geral, é de se esperar que objetos mais primitivos num contexto, que compõem a ontologia profunda, sejam mais conhecidos dos analistas que conceitos derivados com as composições sintagmáticas, que geralmente representam as inovações.

Os conceitos mais profundos dessa ontologia, como “Computer” e “System” (definidos, originalmente, por engenheiros e biólogos), se destinam a ancorar, epistemologicamente, o raciocínio do engenheiro do conhecimento em tarefas de Inteligência Competitiva, tanto no sentido indutivo como dedutivo, nos contextos em análise. Como exemplo, suponhamos que se queira avaliar que tipo de estratégia uma organização como a Microsoft implementaria para um novo produto no mercado: observando-se a Figura 6.17, tem-se dicas importantes: os conceitos de “computador” e “sistema”. Como se sabe, esses dois conceitos mais primitivos”[83], no contexto da Microsoft, significam sua imensa base instalada do Sistema Operacional *Windows* pelo mundo, da qual a empresa absorve a experiência de milhões de usuários (no APENSO III-C, apoiando esta premissa, o objeto sintagmático “computing experience of millions of people” aparece representando o substantivo muito freqüente “computing”).

Observe-se que esses conceitos – “computer” e “system” são bastante genéricos e, ao mesmo tempo, bastante específicos no contexto do negócio, de modo que podem ser utilizados para se alcançar, efetivamente, uma ontologia profunda. O conceito de “computador” pode ser instanciado como qualquer dispositivo eletrônico processador de dados digitais, o que pode representar, num determinado contexto, um microprocessador instalado num telefone celular ou num *iPod* – idem para o conceito de “sistema” (os objetos “Heckerman’s research” e “intersection of human-computer interaction” sugerem exatamente esse tipo de interesse). Então, pode-se conduzir o raciocínio para se tentar compreender qual seria a linha estratégica natural para a implementação de soluções tecnológicas em áreas como a medicina, como sugerem os objetos

[83] Conceitos mais básicos da linguagem utilizados no contexto de negócio.

da ontologia superficial nesse contexto: “collaboration with AIDS researchers” e “life-support systems”.

Em relação ao contexto da IBM, do mesmo modo pode-se tentar compreender as estratégicas da organização com o desenvolvimento de sistemas nas áreas sugeridas pelos objetos da ontologia mais superficial: “warning system for automobiles” e “health delivery and tracking systems”. Como o mercado de automóveis, no mundo, é imenso, pode-se inferir que a IBM está buscando também, como a Microsoft, um mercado de escala mundial para seus novos produtos tecnológicos. E que, do mesmo modo que a Microsoft, está de olho no mercado de serviços de saúde, com uma escala natural em nível global.

Essas conjecturas, ainda que preliminares, mostram que pode haver competição entre IBM e Microsoft pelo mercado de tecnologias de suporte aos serviços de saúde em nível global.

Quanto à Fujitsu Research, seu alvo de pesquisa e desenvolvimento parece concentrar-se na agregação de valor na gestão de serviços de transporte de bens físicos, declarado nos objetos de ontologia superficial “transportation and delivery plans” e “value in service innovations”. Esse nicho de mercado de TIC não coincide, aparentemente, com os nichos onde atuam, ou pretendem atuar, IBM e Microsoft, e isso pode ser uma estratégia da Fujitsu Corporation.

Com essa análise preliminar dos três contextos e seus conceitos minerados, conclui-se que a organização Fujitsu Research está voltada para um mercado de nicho visado por sua *holding*, enquanto IBM Research e Microsoft Research estão atuando em dois tipos de frentes, sendo uma aparentemente competitiva com sua rival – na área de serviços de saúde – e outra de diversificação de produtos em ambientes não competitivos uma com a outra, como os mercados de sistemas para microcomputadores (no contexto da Microsoft) e de sistemas inteligentes para alarmes em automóveis (no contexto da IBM).

Como uma proposta de ontologia ainda mais profunda para esses três casos, numa visão avançada de Gestão do Conhecimento, pode-se induzir um raciocínio com base: (1) no conhecimento que a Fujitsu Research possa ter acumulado em tecnologia digital de suporte a serviços de transporte, (2) no conhecimento que a IBM Research possa ter acumulado em sistemas digitais inteligentes, e (3) no conhecimento que a Microsoft Research possa ter acumulado em sistemas operacionais para microprocessadores digitais de dispositivos pessoais. A essência dessa ontologia profunda, então, se deslocaria do produto em si para o conhecimento nele embutido, resultado de dezenas de anos de investimentos em pesquisa e desenvolvimento e de experiências de cada organização analisada.

O que se pretende estabelecer com este tipo de tese? Que o óbvio geralmente é informação importante em Inteligência Competitiva, pois uma organização de porte, competitiva no mercado, com uma longa história de sucesso, tem sempre um “núcleo duro” de conhecimento acumulado ao longo de muito tempo, e sobre o qual toda sua missão, visão e estratégias são elaboradas. Os conceitos centrais dos modelos apresentados anteriormente, neste capítulo, compõem a ontologia

mais profunda com objetos que representam esse núcleo de conhecimento, enquanto os objetos da ontologia mais superficial mostram as variações de estratégias e a diversificação de produtos em torno desses conceitos centrais.

Os conceitos mais primitivos dos negócios das organizações competitivas, elicitáveis a partir dos modelos conceituais propostos nesta tese, seriam aquilo que Roberts (1973, p. 8), numa homenagem ao seu colaborador Max Fisch, menciona como "bens do intelecto": "Parsons escreveu: '*Ben dell'Intelletto* no Inferno de Dante significa Bem Intelectual; e Benjamin Peirce, pai de CSP, nós chamamos *Ben dell'Intelletto*.'"

## 6.2 *Insights* para inteligência competitiva

O experimento mostra que pode-se implementar a mineração de conceitos para a construção de modelos de informação em ambientes de Inteligência Competitiva de dois modos, basicamente: (1) com acesso direto aos conceitos minerados e (2) com acesso recursivo a novos textos e conceitos complementares. Esses dois modos de acesso aos objetos de conhecimento (ou "bens de conhecimento", na terminologia *CommonKADS*) e suas implicações epistemológicas são discutidos a seguir.

### 6.2.1 Modelo cognitivo com acesso direto à informação

O uso de modelos mentais para acesso à informação e desenvolvimento do conhecimento é abordado em várias disciplinas científicas, algumas mais diretamente associadas aos processos de aprendizagem, como educação e neurosciência. Ontoria, De Luque e Gómez (2006, p. 22-23), por exemplo, apresentam o tema definindo os mapas mentais como grandes teias de idéias e estabelecendo uma metáfora correlacionando o aprendizado com mapas mentais e a própria estrutura do cérebro humano:

> *O funcionamento neuronal do cérebro, com sua estrutura radial, representa um bom referencial para explicar o significado do pensamento irradiante com sua estrutura radial. Da mesma maneira que esta dinâmica neuronal consiste no estabelecimento de múltiplas relações ou associações ramificadas, quando se usa a expressão "pensamento irradiante", faz-se referência àqueles "processos associativos de pensamento que procedem de um ponto central ou se conectam a ele" (Buzan, 1996, p. 67). Quando uma unidade de informação (sentimento, pensamento, imagem externa, situação ...) chega ao cérebro, são geradas muitas conexões com outros dados disponíveis. Essas relações ou entroncamentos ("irradiações") produzidos são indicadores da emergência do pensamento irradiante. Esse núcleo de conexões pode ser equiparado a uma esfera central acesa que se irradia para diferentes direções. O cérebro humano constitui, assim, um sistema superestruturado de processamento e armazenamento da informação, de tal forma que é considerado uma "gigantesca máquina de associações ramificadas", representada pelas estruturas neuronais do cérebro, que potencializam o pensamento. A grande capacidade de processamento da informação e de aprendizagem possibilita o pensamento irradiante, por viabilizar o estabelecimento de múltiplas conexões com a informação disponível.*

> *(...) O que se busca, portanto, é a criação de estruturas que formem uma "totalidade unificada", o que coincide com o processo do cérebro cuja tendência é a da globalidade ou de formas holísticas de pensamento. (...) A simples utilização da palavra e da imagem como fontes de estímulos e de conhecimento é uma forma de multiplicar o potencial do cérebro.*

O tema também é objeto de pesquisa na Inteligência Artificial desde a origem desta disciplina nos anos 1950, como apresentado por Capuano (2009), com o mesmo enfoque das conexões neuronais no processo cerebral de aprendizado. A abordagem do processo de pensamento "irradiante", de Ontoria, De Luque e Gómez, converge com as idéias de Brooks (1980, p. 131) sobre as relações entre informação e conhecimento:

> *Qual a relação entre informação e conhecimento? Eu percebo conhecimento como uma estrutura de conceitos conectados por suas relações e informação como uma pequena parte dessa estrutura. A estrutura do conhecimento pode ser subjetiva ou objetiva.*
>
> *(...) Eu expressei esse relacionamento por aquela que eu denominei a "equação fundamental": K[S] + ∆I = K[S+∆S], que estabelece, de um modo geral, que a estrutura do conhecimento K[S] é alterada para a nova estrutura modificada K[S+∆S] pela informação ∆I, e ∆S indicando o efeito dessa modificação.*

O modelo de gestão do conhecimento proposto nesta tese, a partir de uma base de informação conceitual minerada em contextos organizacionais competitivos, parte dessas noções fundamentais. Com os sintagmas representando substantivos mais freqüentes nos contextos da mineração textual, como demonstrado no experimento de laboratório, obtêm-se o núcleo conceitual, praticamente livre de problemas semânticos, a partir do qual se poderá irradiar o pensamento, de modo indutivo ou dedutivo, em Inteligência Competitiva. Esse núcleo conceitual, que pode ser transformado, no nível individual e coletivo, no fator K[S] de conhecimento de Brooks (1980), poderá então ser ampliado, com as informações conceituais complementares (fator ∆I) contidas nas ramificações a partir desse núcleo (como apresentado nos modelos conceituais integrados deste capítulo), de modo a se aumentar o estoque de conhecimento do engenheiro do conhecimento e das equipes de negócio em ∆S.[84]

Esse modelo conceitual constituirá o que denominamos, nesta tese, de meio de acesso direto à informação para desenvolvimento de inteligência na organização competitiva. Como vantagem do uso de modelos conceituais para visualização da informação conceitual, Ontoria, De Luque e Gómez (2006, p. 26) argumentam:

> *Com o agrupamento de idéias surgem categorias, combinações, hierarquias, novas associações, etc., que permitirão decidir as idéias ordenadoras básicas (IOB) que se identificam com os ramos principais.*

[84] Com a mesma justificativa de Brooks (1980), representou-se o raciocínio em termos pseudo-matemáticos por entender-se que este é o modo mais compacto de expressar as idéias de transformação da informação em conhecimento no nível individual, como no processo de criação do conhecimento operacional tácito idealizado por Nonaka e Takeuchi (1997).

### 6.2.2 Modelo cognitivo recursivo

O modelo básico de acesso direto poderá, no entanto, ser ampliado com buscas e modelagem de informação complementar, de modo a atender necessidades específicas do contexto, ditadas pelos operadores do negócio, tais como pormenores temáticos conceituais não suficientemente aprofundados no modelo básico. Como exemplo, mostrou-se no experimento a utilidade de buscas complementares de informação no Google para esclarecimento de determinados conceitos algo inovadores apresentados nos modelos conceituais integrados, como “Warning System for Automobiles” e “Life-support Systems”.

Esse modelo ampliado a partir das ramificações do pensamento nucleado é denominado, nesta tese, “modelo recursivo”, pois ele percorre um caminho no sentido inverso do modelo de acesso direto (Figura 6.18), partindo da análise de conceitos minerados específicos, dispostos nos modelos, para as fontes originais de informação minerada, de modo a obter detalhes ampliados desses conceitos. Ou, ainda, buscando informações complementares em outras fontes digitais fora do *Web* Portal da organização analisada.

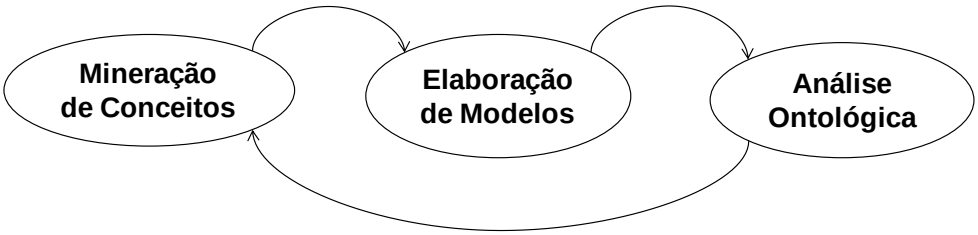



Figura 6.18 Ciclo de Modelagem Conceitual Recursiva
(Fonte: do autor da tese)

Os objetos sintagmáticos resultantes da nova mineração de conceitos poderão, também, serem dispostos em contextos para aplicação do método da Análise de Conceito Formal. Os atributos também poderão ser outros, ao invés dos atributos baseados no modelo *CommonKADS*, de modo a mostrar agrupamentos de objetos com outra finalidade. Como exemplo, pode-se empreender nova mineração conceitual, tendo como conceito central “Warning System for Automobiles”, de modo a se obter informações mais detalhadas acerca dos produtos existentes nesse mercado, seus fabricantes e respectivas aplicações específicas em veículos (alarme contra furto, alarme de temperatura do motor, alarme de problemas mecânicos, alarme de condições de tráfego, etc). Os objetos, no exemplo, poderão ser os produtos (com indicação dos fabricantes) e os atributos as aplicações, obtendo-se, nos *clusters* conceituais, as subdivisões desse nicho de mercado com os respectivos produtos tecnológicos disponíveis (ver modelo genérico ilustrativo na Tabela 6.2 e na Figura 6.19).

Tabela 6.2 Simulação Genérica de Mineração Conceitual Recursiva

(Contexto: “Warning System for Automobiles”)

| Produto e Fabricante | | Aplicação | | | |
|---|---|---|---|---|---|
| Nº | Nomes Comerciais | AW | AX | ... | AZ |
| 1 | Produto I (Fabricante A) | ● | | | |
| 2 | Produto II (Fabricante B) | | ● | ● | |
| 3 | Produto III (Fabricante C) | ● | ● | | |
| 4 | Produto IV (Fabricante D) | | ● | | ● |
| | ... | | | ● | |
| N | Produto M (Fabricante Z) | | | | ● |

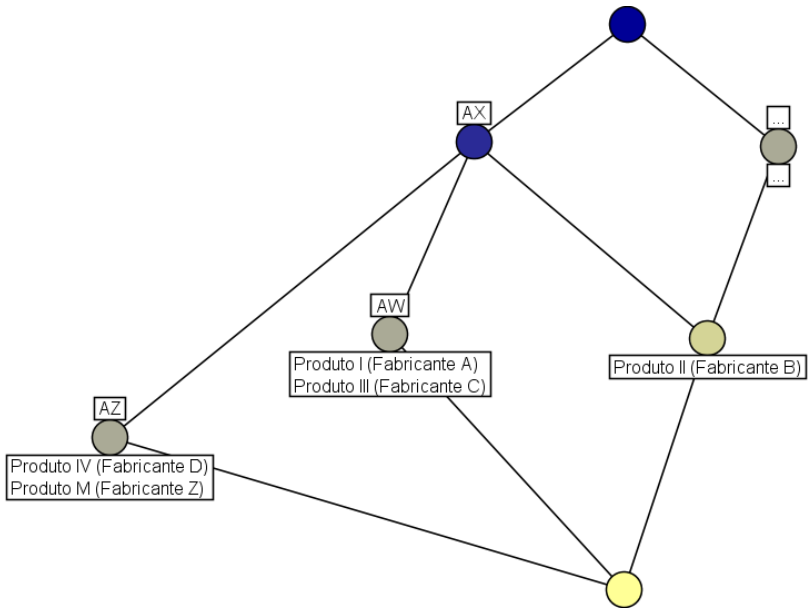



Figura 6.19 Reticulado de Conceitos do Contexto “Warning System for Automobiles”

(Fonte: do autor da tese)

A organização patrocinadora do estudo de inteligência, neste caso, poderá optar por competir com os atuais *players* fabricando outro produto com aplicações similares ou se tornar um competidor “entrante” fabricando um produto com funcionalidades diversas dos produtos existentes nesse mesmo nicho de mercado.

O processo recursivo com mineração, modelagem e análise conceitual poderá ser repetido em vários níveis semânticos, partindo de conceitos mais abstratos para conceitos mais substantivos (Figura 6.20). Quanto aos modelos de classes conceituais, estes também poderão ser ampliados e integrados com classes mais substantivas, podendo-se elaborar modelos mentais tão extensos quanto se tornarem confortáveis para a visualização e compreensão humana.

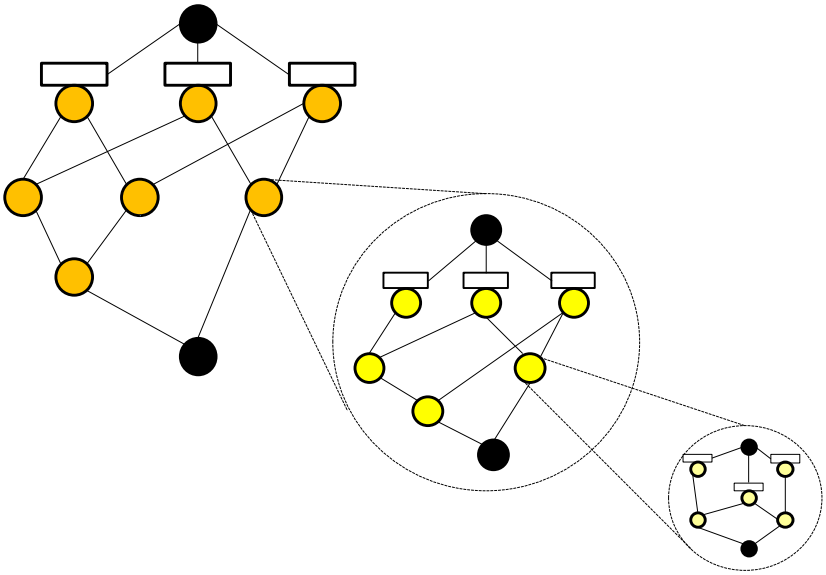


Figura 6.20 Reticulados de Conceitos Recursivos

(Fonte: do autor da tese)

O detalhamento dos componentes de um objeto conceitual poderá ser executado com Gráficos Conceituais (GCs), como no caso de “warning system for automobile” ilustrado na Figura 6.21. Esse modelo se baseia num fragmento do texto “On the Drawing Board”, pesquisado com o Google utilizando-se os argumentos de busca “IBM” e “warning system” com conjunto, sobre os sistemas de alarme para veículos desenvolvidos pela organização “IBM Research”, disponível no endereço eletrônico: *http://domino.research.ibm.com/odis/odis.nsf/pages/board.02.html.*

Os GCs são interessantes quando se busca conhecer melhor detalhes constitutivos de conceitos isolados, ou conceitos-tipo, também chamados de “esquemas” por Sowa (1984), de coisas mais próximas do mundo material, pois seu emprego em conjuntos de conceitos abstratos com muitos relacionamentos e sub-tipagens poderá resultar em gráficos muito “poluídos”, algo desconfortável para consultas visuais (o modelo da Figura 6.21 resultou dos conteúdos de apenas cinco linhas de texto).

Como exemplo de aplicação prática em ambientes de Inteligência Competitiva, executivos do negócio com pouco tempo para leitura de textos técnicos sobre produtos poderão se beneficiar com a “leitura” de Gráficos Conceituais. O poder dos GCs é ressaltado na obra de Sowa (1984) com base em pesquisas psicológicas no tema “cognição”, cuja verificação mais importante é a de que o ser humano não aprende e utiliza a linguagem de modo “atômico”, mas com base em estruturas que conectam a linguagem a conceitos. E, de um ponto de vista filosófico, Sowa (1984, p. 42) argumenta que:

> *Conceitos e percepts são blocos de construção para a elaboração de modelos mentais. Mas regras e padrões são necessários para se organizar os blocos de construção em estruturas maiores. Immanuel Kant (1781) introduziu o termo “esquema” para designar uma regra que organiza as percepções em um todo unitário.*

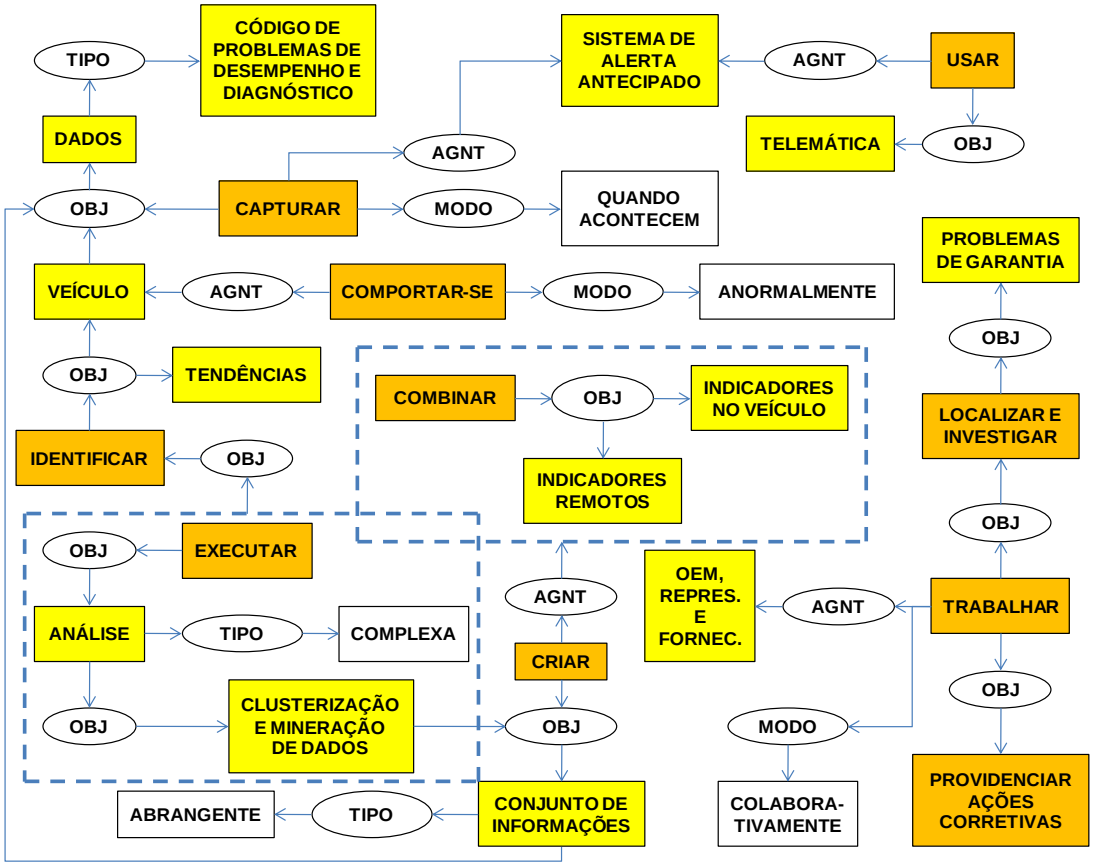



Figura 6.21 Gráfico Conceitual de “Warning System” no Contexto “IBM Research”
(Fonte: do autor da tese)

Em particular, os GCs poderão ser utilizados com certo conforto para modelagem de conceitos em pesquisas recursivas de novidades em termos de produtos, que podem aparecer na mineração de conceitos, como no caso de “warning system for automobile”. A presença de verbos nos objetos sintagmáticos minerados para modelagem com GCs supre a lacuna deixada pelos modelos conceituais com UML, mostrando a ação dos agentes, os objetos manipulados por eles e os relacionamentos entre os objetos e as próprias ações (como verbos transitivos exigem complemento, este tipo de estrutura mostra um detalhamento conceitual maior, baseado nos sintagmas, que os modelos estáticos somente com classes e objetos).

### 6.2.3 Sensibilidade temporal do modelo conceitual

#### 6.2.3.1 Definição do problema temporal

O modelo apresentado até este ponto tem como objetivo propiciar uma visão estática do mosaico semântico do contexto de negócio analisado. Entretanto, o mundo e o mercado são dinâmicos e mudam a todo momento, necessitando-se atualizar, com a maior freqüência possível,

os dados coletados das fontes digitais disponíveis, especialmente dos *Web* Portais, que são o meio pelo qual as organizações atualizam, diariamente, sua “mensagem para o mundo”.

A metodologia proposta poderá atender a esse requisito com as abordagens apresentadas a seguir, com base nos dados experimentais do APENSO IV, relativos aos substantivos/nomes mais freqüentes (nas faixas de 2,5%, 5,0%, 7,5% e 10,0%) encontrados na coletânea inicialmente testada e num texto adicional pesquisado na *Web* posteriormente, ambos do contexto “Microsoft Research”. O texto se refere a uma entrevista concedida pelo executivo-chefe da Microsoft sobre o interesse dessa empresa na aquisição do *Web* Portal de buscas Yahoo! (CNET, 2008).

Observa-se, na planilha do APENSO IV, uma coluna de enumeração de substantivos/nomes mais à esquerda e dois blocos de dados, o primeiro denominado “Coletânea Inicial” e o segundo “Acréscimo à Coletânea”. O primeiro bloco de dados se refere aos substantivos/nomes minerados anteriormente no contexto “Microsoft Research” e o segundo ao novo texto mencionado, cujo impacto na base de informação e nos modelos anteriores será analisado.

Como a metodologia proposta nesta tese é embasada, inclusive, no estudo de freqüências de substantivos/nomes encontrados nas coletâneas digitais de negócios, e os modelos conceituais produzidos se concentram nos substantivos/nomes com freqüências de até 10,0%, deve-se cuidar, inicialmente, para que as novidades eventualmente presentes nos textos acrescentados à base apareçam com substantivos/nomes indicativos nessa faixa de freqüência. Ou seja, considerando-se que os substantivos/nomes mais freqüentes muito provavelmente comporão sintagmas complexos úteis para modelagem conceitual, deve-se providenciar uma técnica para que seus respectivos substantivos/nomes apareçam entre os mais freqüentes substantivos/nomes preexistentes na base, de modo a perturbar (modificar) sua estrutura inicial e, com isso, chamar a atenção dos engenheiros do conhecimento.

Caso se decida apenas adicionar os novos textos à base preexistente e calcular os parâmetros de freqüência novamente, provavelmente as novidades não aparecerão perturbando a base porque, como é lógico se esperar, os novos textos serão bem menores que a base existente e, conseqüente, os substantivos/nomes mais freqüentes terão freqüências muito menores que as freqüências dos substantivos/nomes preexistentes. O cenário, portanto, exige uma solução que torne os substantivos/nomes relativos às novidades tão destacados, em termos de freqüência, quanto os substantivos/nomes da base anterior.

Os dados do APENSO IV mostram o problema do desencontro de magnitudes de freqüência entre coletâneas de tamanhos muito diversos. Enquanto a coletânea do contexto “Microsoft Research” soma 1.159 substantivos/nomes distintos (não repetidos), como se pode observar no APENSO I, o novo texto adicionado à base soma apenas 289 substantivos/nomes distintos. E, por isso, as freqüências dos substantivos/nomes mais “carnudos” também são muito díspares: “research”, o substantivo mais freqüente na base preexistente, aparece 278 vezes, enquanto “thing”, o substantivo mais freqüente no texto novo, aparece apenas 34 vezes. E assim por diante:

“computer”, por exemplo, aparece 80 vezes na base, enquanto “Yahoo” aparece apenas 14 vezes no texto acrescentado à coletânea-base.

6.2.3.2 Solução proposta

Como solução, propõe-se uma abordagem de “normatização” das freqüências dos substantivos/nomes mais freqüentes encontrados nos textos acrescidos à coletânea preexistente, de modo a torná-las relativamente destacadas na base como um todo a partir de uma atualização. O que se propõe, assim, é que o engenheiro do conhecimento possa atualizar a base de informação textual preexistente e, para verificar possíveis perturbações em relação à situação anterior, possa consultar apenas as freqüências relativas de substantivos/nomes. Caso observe alguma alteração nas posições relativas de freqüências, deverá então aprofundar sua análise e buscar, na base indexada, os respectivos novos sintagmas plurinominais que causaram essa alteração e providenciar os novos modelos atualizados de informação conceitual, conforme a metodologia apresentada nesta tese.

Como normatizar, então, as freqüências dos substantivos/nomes acrescentados à base? Ou, numa visão técnica de PLN, como tornar os novos textos tão informativos quanto a base preexistente considerando-se as tendências estatísticas de freqüências de *tokens* em textos de diferentes tamanhos? Como discutido anteriormente, textos maiores tendem a repetir mais palavras e, por isso, apresentam palavras com freqüências maiores.

Então, propõe-se uma normatização baseada em parâmetros lingüísticos mais estáveis de textos, de um ponto de vista semântico, elegendo-se a variável “quantidade absoluta de substantivos/nomes não-repetidos” para tanto. Considerando-se que os conceitos elicitados dos textos residem nos substantivos/nomes e nas combinações entre eles (que são, no caso, os sintagmas complexos), conclui-se que os objetos lexicais que interessam aos sensores conceituais temporais são esses.

O modelo proposto de normatização de freqüências dos substantivos/nomes do novo texto adicionado à base pode ser resumido na seguinte fórmula matemática:

$$F_n(i) = F_o(i) \,.\, (Qs_e \,/\, Qs_n),$$

onde:

$F_n(i)$: frequência normatizada do substantivo/nome novo indexado como “i”;

$F_o(i)$: freqüência não normatizada do substantivo/nome novo indexado como “i”;

$Qs_e$: quantidade de substantivos/nomes não repetidos na base preexistente;

$Qs_n$: quantidade de substantivos/nomes não repetidos no texto novo.

Este modelo de sensor semântico, testado com os dados do APENSO IV, resulta em valores de $F_n(i)$ conforme a Tabela 6.3 a seguir (calculados somente para a faixa dos 2,5% de substantivos/nomes mais freqüentes). O valor do fator de normalização $Q_{Se}/Q_{Sn}$ calculado é 4,0, razão entre 1.159 e 289.

Tabela 6.3 Alterações de Posições Relativas de Frequências com Atualização da Base

| Nº | Coletânea de Base | | Novo Texto Adicionado | | Coletânea Atualizada | |
|---|---|---|---|---|---|---|
| | Substantivo / Nome | Freq.(i) | Substantivo / Nome | $F_o(i)$ | Substantivo / Nome | $F_n(i)$ |
| 1 | Research | 278 | thing | 34 | Research | 278 |
| 2 | Microsoft | 256 | people | 17 | Microsoft | 256 |
| 3 | Computer | 80 | Yahoo | 14 | thing | 136 |
| 4 | Computing | 68 | Gates | 13 | people | 125 |
| 5 | People | 57 | software | 13 | software | 98 |
| 6 | Technology | 57 | lot | 12 | computer | 80 |
| 7 | Systems | 55 | today | 12 | computing | 68 |
| 8 | Areas | 50 | Microsoft | 11 | Technology | 57 |
| 9 | Information | 50 | | | Yahoo | 56 |
| 10 | Science | 49 | | | Systems | 55 |
| 11 | Software | 46 | | | Gates | 52 |
| 12 | Group | 42 | | | Areas | 50 |
| 13 | Data | 37 | | | Information | 50 |
| 14 | Work | 35 | | | Science | 49 |
| 15 | Redmond | 34 | | | Group | 42 |
| ... | ... | ... | | | Data | 37 |
| 29 | Overview | 20 | | | ... | ... |

Embora se deva considerar que alguns substantivos/nomes do novo texto já se encontravam na base preexistente com freqüência elevada, como é o caso de “people”, “software” e “Microsoft”, a normatização das freqüências dos conteúdos novos produzem um efeito que interessa no contexto: a elevação virtual das freqüências de alguns objetos novos como “Yahoo” e “Gates” – “Yahoo”, em particular, deverá se situar em nono lugar no novo *ranking* de freqüências após a normalização. Esses objetos constituem novidades que são destacadas no novo *ranking* de substantivos/nomes mais freqüentes no modelo conceitual e que, portanto, deverão receber mais atenções do engenheiro do conhecimento após a atualização da base.

A descoberta da novidade relacionada com o nome “Yahoo”, no caso, será de importância vital para o analista de inteligência operando numa organização concorrente da Microsoft no mercado, como é o caso da Google, por exemplo. A pergunta que esse analista deveria fazer, de plano, seria a mais simples possível: “o que a Microsoft está querendo com o pessoal do Yahoo!?” Com esse simples experimento resta evidente, mais uma vez, a importância da associação de termos léxicos de conteúdos e de contextos na mineração, filtrando-se melhor os resultados de modo a tornar a busca mais eficiente.

Enfim, o cálculo e a nova ordenação de freqüências da Tabela 6.3 podem ser realizados mesmo antes de atualização da base, comparando-se os dados da base com os dados do novo

texto minerado. Com isso, evita-se um novo processamento da base toda atualizada, poupando-se recursos tecnológicos e humanos.

Este experimento de atualização de base textual no tempo mostra, portanto, que a metodologia de mineração e modelagem de conceitos proposta pode dar conta do problema temporal de modo a produzir os resultados colimados. Com a técnica proposta, estima-se que qualquer novidade sobre o negócio publicada na *Web* ou em outras fontes digitais, cadastradas pelo engenheiro do conhecimento para mineração textual periódica, poderá ser descoberta tempestivamente e tratada adequadamente sob o prisma da Inteligência Competitiva.

Com a capacidade de se “calibrar” a faixa de freqüências dos substantivos/nomes que indexam os sintagmas, a metodologia permite que se executem buscas em diversos níveis e volumes de conteúdos, com parcelas maiores ou menores da base textual recuperada. E com esse recurso de atualização dos conceitos mais relevantes num contexto, ao longo do tempo, publicados pelas organizações que atuam em um nicho de negócios, pode-se monitorar a concorrência numa base permanente e controlada, destacando-se as informações que aparecem como novidades em cada busca e que devem, portanto, receber atenções especiais do engenheiro do conhecimento.

# 7. CONCLUSÕES E QUESTÕES EM ABERTO

## 7.1 Tese e resultados do experimento

Com base nos resultados do experimento desenvolvido, evidencia-se uma viável metodologia de mineração e análise de informação conceitual relevante de fontes digitais para modelagem semi-automática de sistemas de negócio em ambientes de Inteligência Competitiva, alcançando-se o principal objetivo desta tese. A metodologia demonstra como o arsenal epistemológico da Inteligência Artificial (IA) implementado em *softwares* de etiquetagem e de análise sintática pode contribuir, efetivamente, para o desenvolvimento de mosaicos conceituais representando informações essenciais para indução de processos de criação de significados e construção de conhecimento sobre os ambientes de negócios.

Assim, resta evidente que esse tipo de ambiente informacional poderá ser estruturado, pelo menos em parte, *a priori*, com potencial para atendimento tanto de demandas *ad hoc* como para uma atividade de inteligência mais proativa, no sentido de se antecipar às necessidades dos usuários, construindo-se um mosaico de informação abrangente sobre o mercado, incluindo os principais competidores, potenciais entrantes e substitutos, parceiros e clientes, conforme a abordagem epistemológica de Porter.

Os resultados da experiência de laboratório também permitem um refinamento da tese original, que então passa a ser:

***"*É viável a estruturação de uma *praxis* de Gestão do Conhecimento para Inteligência Competitiva com base em informações conceituais relevantes recuperadas de fontes digitais abertas e modeladas *a priori* das demandas dos usuários, com desempenho na revocação de mais de 90,0% dos 2,5% de substantivos/nomes mais freqüentes."**

**A subtese consequente, que contraria a visão tradicional de subjetividade dos critérios de relevância da informação para o usuário, também é defensável na medida em que se mostrou que a metodologia permite uma separação formal, de modo não subjetivo, do conjunto de conceitos mais relevantes num contexto de negócios.**

Esta metodologia representa um primeiro passo – o da construção de uma base epistemológica – para o desenvolvimento de uma solução computacional de engenharia do conhecimento (um *framework*) para a modelagem de conceitos de negócio em atividades de apoio ao desenvolvimento de estratégias competitivas nas organizações contemporâneas. A proposta metodológica desenvolvida e testada para solução do problema, ainda que em escala experimental reduzida, apresenta características de uma solução de engenharia, como previsto por Bates (1999, p. 1049):

> *O posicionamento metodológico fundamental da ciência da informação pode ser descrito como sociotécnico. As duas mais importantes tradições metodológicas que delineamos são as ciências sociais e as ciências da engenharia. (...) para atuar*

> *efetivamente na ciência da informação, deve-se ao menos sentir-se confortável com ambos os lados dessa tradição dual.*

Conclui-se então, com fundamento no raciocínio de Peirce (2010) combinando *retrodução* com *abdução*, que a hipótese da tese se confirmou no experimento. Esse experimento também serviu de teste para a tese defendida, numa abordagem lógico-dedutiva tipicamente popperiana, que resistiu a uma primeira experiência que poderia refutá-la. O experimento poderá ser reproduzido, formalmente, com informações de outros contextos de negócio, o que mostra, assim, sua efetiva contribuição para o avanço da Ciência da Informação além dos estudos de caso.

Outros argumentos apresentados em defesa desta tese, com base no experimento desenvolvido, são os seguintes:

- a metodologia proposta para mineração de conceitos a partir de conteúdos digitais em linguagem natural é eficaz e eficiente em contextos de informação de negócios, mostrando a utilidade de conhecimentos basilares da disciplina "Recuperação da Informação" em buscas na *World Wide Web* (como reclamado por Bates (1999)) – considerando-se que, em média, mais de 90% do conjunto composto pelos 2,5% de substantivos mais freqüentes em um texto de *Web* Portal corporativo estão presentes em sintagmas compostos com dois ou mais substantivos;
- a metodologia implementa uma abordagem de busca, identificação e indexação textual para recuperação da informação conceitual relevante algo inovadora, baseada em sintagmas nominais, inspirada nas idéias de Gottschalg-Duque (2005), apresentando expressivo poder de resolução semântica no próprio processo de busca, com pouca necessidade de busca recursiva para desambiguação de conceitos (com efeito, prevê-se que apenas conceitos relativos a inovações no contexto de negócio exigem desambiguação[85]);
- os modelos de avaliação da relevância da informação baseados na Teoria da Informação de Shannon e nas teses de Bateson, Weick e Choo são válidos, com ideias convergindo para a equação do conhecimento de Brooks (1980), podendo ser utilizados para o desencadeamento de processos de criação de significado acerca dos conceitos minerados e de construção do conhecimento em contextos de negócio;
- a metodologia tem recursos para simulação de cenários, permitindo avaliar-se o impacto de novos conceitos na base de informação conceitual;
- o processo de implementação da metodologia de mineração e modelagem da informação conceitual proposto se encontra em estágio intermediário de automação computacional, com uso de *softwares* dotados de recursos de Inteligência Artificial em algumas atividades críticas, tais como etiquetagem e análise sintática, demonstrando a utilidade

[85] Deve-se lembrar que os engenheiros do conhecimento atuando em áreas de Inteligência Competitiva nas organizações complexas deverão conhecer os principais conceitos referentes ao contexto do negócio, ainda que não em tanta profundidade como os operadores do negócio.

desse recurso e a possibilidade de ampliação de seu uso no contexto (postando-se, neste ponto, solidariamente ao apelo de Bates (1999) no sentido de se evitar a irrefletida "rejeição tecnológica dos Ludditas" na Ciência da Informação);

- os modelos ontológicos propostos podem ser úteis na composição de um ambiente de aprendizagem a partir de informação textual coletada de fontes digitais abertas, numa linha de atividades de Gestão do Conhecimento para apoio ao desenvolvimento de Inteligência Competitiva nas organizações (como mencionado na introdução a esta tese, adota-se uma percepção de gestão do conhecimento por meios indiretos, com base na informação adequadamente representada para o aprendizado individual);
- os modelos conceituais resultantes são expressos em linguagem natural, sendo, portanto, acessíveis aos vários tipos de públicos, com perfis sociotécnicos bastante diversos, que geralmente compõem uma organização de porte (embora a representação da informação não deva, na Ciência da Informação, ser confundida com o conhecimento do conteúdo da informação (BATES, 1999), os modelos conceituais propostos, como se mostrou no Capítulo 6, permitem o desenvolvimento de *insights* nos usuários a partir do acesso direto a conceitos do negócio);
- embora testada apenas com informações da *Web*, a metodologia proposta é bastante robusta, com informação digital sobre negócios recuperada e apresentada em linguagem natural, não exigindo nenhum tipo de tratamento de dados preliminar à mineração.

Considerando-se a vastidão dos ambientes de informação digital disponíveis atualmente, pode-se avaliar o tamanho do desafio que toda solução tecnológica para mineração textual deve enfrentar. A noção de *praxis*, ou de *praxeologia*, neste caso, deve ser encarada como uma prática de engenharia do conhecimento em constante evolução, buscando-se soluções sociotécnicas bastante pragmáticas para os problemas que se apresentam – esclarecendo-se que soluções de engenharia não são, geralmente, nem exatas nem muito genéricas, mas com aproximações suficientes e aplicações genéricas para determinados contextos similares.

Entretanto, pode-se afirmar que a solução metodológica engendrada também apresenta formalismos que embasam sua fundamentação técnica e generalização, como a Análise de Conceito Formal, embora não sem uma certa dose daquilo que se poderia denominar "avaliação conceitual de contexto". A classificação dos objetos sintagmáticos minerados dos textos (que também chamamos de "instâncias de conceitos") bem representa esse tipo de dificuldade, pois tipos característicos de texto poderão exigir modelos de mineração conceitual específicos. Como exemplo, a mineração e modelagem de conceitos a partir dos decretos governamentais, no Brasil, sugere que se deve adotar também, como padrões sintáticos de busca, sintagmas com verbos (no infinitivo) e substantivos/nomes (estilo "VB_DT_NNS", por exemplo), de modo a se elicitar os comandos e ordenamentos normativos caracteristicamente expressos nesse tipo de documento.

Os cálculos estatísticos também emprestam formalismo à metodologia, fundamentando, experimentalmente, a crença indutiva que suporta as conclusões sobre a eficácia da abordagem de “engenharia lingüística” adotada na mineração.

Em relação ao alcance dos objetivos específicos da tese, conclui-se:

a) que o arsenal de Inteligência Artificial tem nos *softwares* de etiquetagem automática de palavras *(tagging),* de análise sintática de textos *(parsing)* e de Análise de Conceito Formal (com elaboração de reticulados de conceitos) os recursos tecnológicos de automação de processos mais críticos para a viabilização da metodologia proposta, condição *sine qua non* para enfrentamento do verdadeiro “oceano” de textos de negócios disponíveis nos repositórios digitais da atualidade;
b) os modelos conceituais gerados são sensíveis ao uso freqüente de conceitos nos textos minerados, podendo-se inferir que quando um novo conceito é introduzido num *Web* Portal, mediante uma série de textos correlatos, ele poderá ser identificado nos modelos conceituais elaborados posteriormente caso a freqüência de seus substantivos/nomes alcance certo limiar, recurso útil para se monitorar variações ou inovações de conceitos de negócio ao longo do tempo;
c) o nível de interoperabilidade sociotécnica das ontologias produzidas com a metodologia é próximo da interoperabilidade da própria sublinguagem natural utilizada no contexto em estudo, pois os conceitos são expressos nessa sublinguagem, restando como única barreira o conhecimento básico de reticulados de conceitos e dos modelos de classes da UML, ou, alternativamente, o conhecimento de modelos mentais mais simples, porém menos compactos, como os apresentados por Ontoria, De Duque e Gómez (2006), com um nível de requisitos cognitivos para compreensão bastante elementar.

## 7.2 Modelo de mineração conceitual proposto

Em síntese, o proposto modelo de mineração e modelagem de conceitos para gestão do conhecimento de negócio, no suporte a uma *praxis* de Inteligência Competitiva, é representado no processo da Figura 7.1. Didaticamente, as atividades que compõem o processo, representadas pelas caixas com arestas arredondadas, são separadas em duas “raias”, sendo uma referente a atividades que, pela indisponibilidade de tecnologias de suporte, necessitam ser desenvolvidas manualmente, e outra referente a atividades semi-automáticas, que contam, atualmente, com *softwares* de apoio.

Como produto de engenharia, a metodologia também apresenta recursos de configuração de desempenho que pode ser regulada conforma a necessidade da situação de inteligência em estudo. Estabelecendo-se a faixa de freqüência dos substantivos do texto que se pretende sejam representados pelos objetos sintagmáticos, pode-se operar com conceitos mais “centrados” ou

conceitos mais “radiais” – conceitos mais utilizados pela organização em sua “mensagem para o mundo” ou conceitos com menor ênfase, mas talvez tão ou mais importantes no contexto.

Contudo, mesmo as atividades que ainda não contam com tecnologia de automação podem vir a ser semi-automatizadas com algum esforço de desenvolvimento de *softwares*, especialmente para a atividade de geração de modelos de classes conceituais. Existem, atualmente, apenas editores de ontologias no suporte a essa atividade, recurso tecnológico muito primitivo ainda, que não pode ser classificado como ferramenta de automação, mas apenas de visualização e armazenamento de modelos digitais. A crença na viabilidade da automação de expressiva parte dessa atividade se fundamenta na verificação empírica da existência de padrões de sintagmas nominais que originam modelos de classe canônicos, tais como os exemplificados na Figura 7.2.

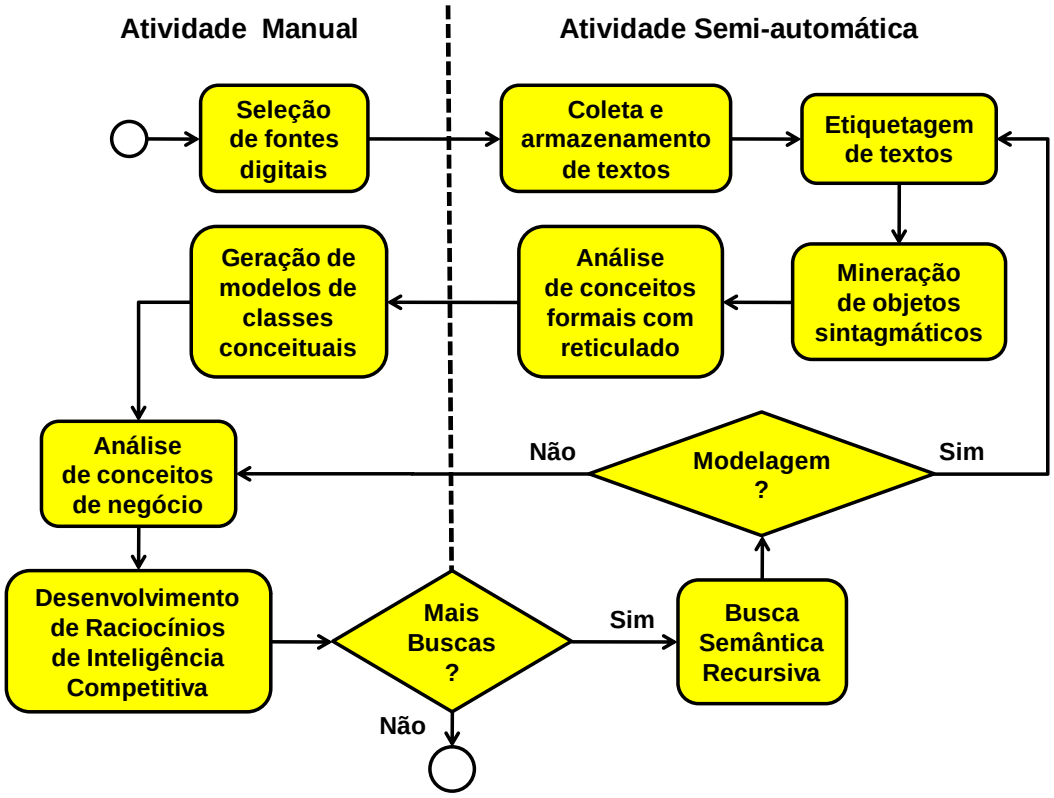



Figura 7.1 Modelo de Mineração de Conceitos de Negócio para Inteligência Competitiva
(Fonte: do autor da tese)

O objeto conceitual “warning system”, por exemplo, apresenta um padrão de composição sintática bastante comum no idioma inglês, por isso passível de representação com um modelo de classe canônico. O *token* “system”, neste caso, é o substantivo de base e o *token* “warning” é o substantivo adjetivador (neste caso, baseado num verbo na forma “ing”), sugerindo um modelo de informação conceitual onde “System” representa a superclasse (ou classe-base), enquanto a classe “Warning” assume o papel de uma classe de atributos de “System”. Ou seja, “Warning

System” é um tipo (instância ou espécie) de “System” com atributo “Warning” (em síntese, “sistemas de alarmes” são apenas objetos ou instâncias da classe “Sistema”).

Quanto ao objeto “systems management”, trata-se de apenas uma instância da classe “Management” com a classe-atributo “System”, pois o *token* “management”, no caso, aparece em segundo lugar na composição léxica, estrutura típica do idioma inglês. Essa estrutura de posicionamentos de *tokens* facilita a padronização de modelos de objetos sintagmáticos.

**Objetos de Contexto:**
***Systems Management, Warning System, Heckerman's Research***

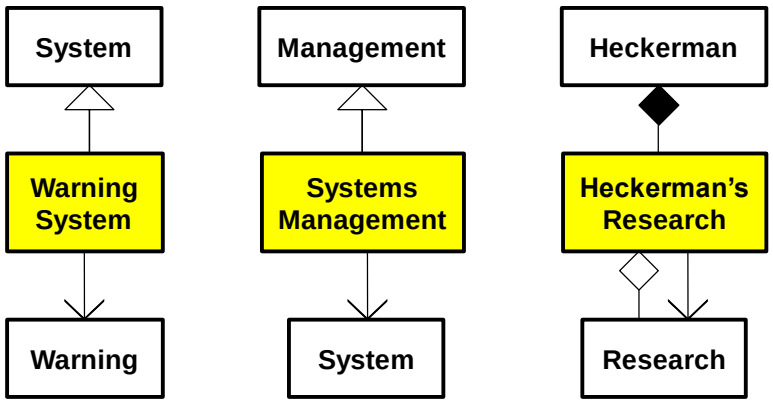



Figura 6.2 Modelos de Classe Canônicos
(Fonte: do autor da tese)

Em outros idiomas, como o português, essas estruturas também existem, ainda que com posicionamentos de *tokens* diversos. Como exemplos em português, pode-se utilizar os mesmos sintagmas: “sistemas de alarme”, “gerenciamento de sistemas”, “pesquisas de Heckerman”, onde “sistemas de alarme” é apenas uma instância da classe “Sistema” (com o atributo “de alarme”); “gerenciamento de sistemas” é uma instância da classe “Gerenciamento” (com o atributo “de sistemas”); e “pesquisas de Heckerman” algo que pertence a uma pessoa de nome “Heckerman” (entendendo-se, também, que “Pesquisa de Heckerman” é um objeto que tem um *token*-atributo denominado “Pesquisa”).

Outra motivação para a crença, nesta tese, que a metodologia proposta possa ser utilizada em outros idiomas é a argumentação de Werner (1988, p. 145):

> *Os trabalhos de Hamill (1978), Hutchins (1980) e outros estão acumulando evidências no sentido que a lógica humana é a mesma em qualquer lugar. E tudo isso parece dar crédito para a idéia que o nível conceitual é mais fundamental, mais similar de linguagem para linguagem (ou de cultura para cultura), e talvez por isso também mais simples.*

Werner (1988, p. 147), comentando sobre sistemas especialistas, também apresenta uma pista importante para se avaliar uma metodologia de automação dessa natureza em questões de escalabilidade de processamento de volumes de informação e conceitos[86]:

> *(...) a máquina deve prover toda informação que um usuário humano requisitar. Ao mesmo tempo, o usuário deve ser alertado pelo sistema especialista sobre outros tipos de informação pertinentes ao tópico de indagação que se pode também possuir.*
>
> *O principal problema que eu prevejo é que uma máquina com uma insaciável "sede de conhecimento" e uma igualmente insaciável "necessidade de informar" é uma "chateação" conversacional que não tem fim – o proverbial cão com um osso.*

## 7.3 Possíveis linhas de pesquisa decorrentes

Considerando que a pesquisa implementada nesta tese se apoiou em três pilares epistemológicos correlatos à Ciência da Informação – *Recuperação da Informação, Engenharia do Conhecimento* (onde se insere a Análise de Conceito Formal e a modelagem de conceitos) e *Inteligência Competitiva*, vislumbram-se, igualmente, três linhas de pesquisa decorrentes. A primeira delas poderia se ocupar de um aprofundamento em questões de Processamento da Linguagem Natural – PLN em contextos de informações de negócios, como o desenvolvimento de métodos de mineração de conceitos para contextos mais específicos. O desenvolvimento de *softwares* com tecnologia de Inteligência Artificial para etiquetagem de textos também representa, naturalmente, um desafio formidável, além da relevância de semelhante iniciativa considerando-se a carência de recursos tecnológicos de PLN para o idioma português.

Outra linha de pesquisa poderia se concentrar em questões de Engenharia do Conhecimento, tais como a elicitação semi-automática de conceitos relevantes, num contexto, para desenvolvimento de ontologias profundas a partir de textos, com modelos canônicos para aplicações específicas – a própria modelagem de arquitetura de sistemas de informação computacionais muito poderia se beneficiar desse tipo de pesquisa, considerando-se o custo (principalmente em termos de tempo) da modelagem de sistemas corporativos atualmente. Esta área parece particularmente promissora para os próximos anos, quando a "Geração Y" (ou geração "pós-*Internet*") estiver no comando das organizações em geral e o conceito de "leitura mosaica" se tornar um imperativo metodológico para superação dos problemas da gestão da informação da "era-Gutenberg".

Cunha, a propósito, na entrevista apresentada em Maia (2010), conclui que:

> *O jovem de hoje está lendo diferente. Com essa nova era digital, nós temos um leitor imersivo, aquele que navega em outra materialidade, que é a da tela do computador. (...) Assim, o jovem está lendo, mas essa leitura é de outra forma. Uma leitura mosaica. E essa leitura não contínua começou com o jornal, ele que preparou o nosso olhar para a Internet. O problema é que a gente (professores) continua tentando trabalhar o nosso leitor como se ele fosse o leitor renascentista que tinha*

[86] Considerando-se o contexto pré-*Internet* da crítica de Werner (1988), ela pode ser aplicada à própria *Web*, na atualidade, devido à "tediosa" recursividade de busca com hipertexto.

> *todo o tempo de decifrar a escrita. Esse leitor contemplativo desapareceu já na Revolução Industrial. No lugar dele, veio um leitor mais distraído que tem à disposição uma série de escritos e ruídos, ou seja, uma série de movimentos que convivem com a sua leitura, como o rádio, a TV e a Internet. Assim, ele mudou o olhar, mudou a percepção e a forma de cognição.*

Quanto à terceira linha de pesquisa sugerida, concentrada em temas específicos da Inteligência Competitiva, pode-se pensar no desenvolvimento de modelos lógicos de raciocínio para missões e situações específicas, tais como as apresentadas por Fuld (2007) e a idéia de "diferença" em Shannon (*apud* SAYOOD, 2000), Bateson (2002) e Weick (1995). Esse tipo de modelagem poderia simplificar o trabalho dos engenheiros do conhecimento em "situações-padrão" e "situações-surpresa" (quem sabe, com estruturas de conceitos canônicas) nas áreas de inteligência das organizações.

É interessante como o desafio original de recuperação da informação continua presente, mas agora despertando o interesse de organizações que desenvolvem tecnologias de ponta, como se percebe no discurso de organizações como a Microsoft Research.[87]

> ***Improving How Systems Store, Retrieve, and Present Information***
>
> *Our lives are more data-driven than ever, yet computing hasn't achieved its full potential to help us manage, protect, visualize, and understand that data. For businesses, it is important to identify quickly the information that matters and make the right decisions in real time.*

[87] Informação disponível em: *http://research.microsoft.com/en-us/about/brochure-4.aspx.*

# 8. REFERÊNCIAS


ACM – ASSOCIATION FOR COMPUTING MACHINERY. Recent AI-related dissertations. *SIGART Bulletin*, v. 1, n. 1, 2007.

ADOMAVICIUS, Gediminas; TUZHILIN, Alexander. Personalization technologies: a process-oriented perspective. *Communications of the ACM*, v. 48, n. 10, out. 2005.

AL NEIMAT, Taimour. *Why IT projects fail*. Disponível em: <http://www.projectperfect.com.au>. Acesso em: 18 jun. 2007.

ARAÚJO JR., Rogério Henrique de. *Precisão no processo de busca e recuperação da informação*. Brasília: Thesaurus, 2007.

ARISTÓTELES. *Metafísica*. Tradução de Leonel Vallandro. Porto Alegre: Globo, 1969.

ARISTÓTELES; SANTOS, Ricardo (Org.). *Categorias*. Porto (Portugal): Porto Editora, 1995.

ARMSTRONG, Deborah J. The quarks of object-oriented development. *Communications of the ACM*, v. 49, n. 2, p. 123-128, fev. 2006.

AUER, Sören. *Powl – A Web-based platform for collaborative ontology Management*. Disponível em: <http://www.semanticscripting.org/SFSW2005/papers/Auer-Powl.pdf>. Acesso em: 30 nov. 2006.

BALLARD, Bruce W.; TINKHAM, Nancy L. A phrase-structured grammatical framework for transportable Natural Language Processing. *Computational Linguistics*, v. 10, n. 2, p. 81-96, abr.-jun. 1984.

BARBUT, M.; MONJARDET, B. Ordre et classification. *Algebre et Combinatoire, Tome II*. Paris: Hachette, 1970.

BARNETT, Jim; KNIGHT, Kevin; MANI, Inderjeet; RICH, Elaine. Knowledge and Natural Language Processing. *Communications of the ACM*, v. 33, n. 8, p. 50-71, ago. 1990.

BATES, Marcia J. The invisible substrate of Information Science. *Journal of the American Society for Information Science*, v. 50, n. 12, p. 1.043-1.050, out. 1999. Disponível em: <http://www.gseis.ucla.edu/research/bates1.html>. Acesso em: 15 set. 2007.

BATESON, Gregory. *Mind and nature: a necessary unity*. Cresskill (New Jersey, EUA): Hampton, 2002.

BELEW, Richard K. A connectionist approach to conceptual Information Retrieval. *Association for Computing Machinery*, 1987.

BELL, Michael. *Service-oriented modeling: service annalysis, design and architecture*. Hoboken, New Jersey (EUA): John Wiley & Sons, 2008.

BENABDELLATIF, M. Process modelling analysis: comparison between activity-oriented models, product-oriented models and decision-oriented models. *In*: ZANASI, A.; BREBBIA, C. A.; EBECKEN, N. F. F. E.; MELLI, P. *Data Mining III*. Southampton, Boston: WIT Press, 2002, p. 251-258.

BERRY, Michael W. (Ed.). *Survey of Text Mining: clustering, classification, and retrieval*. Springer Science, 2004.

BISSON, G.; NEDELLEC, C.; CANAMERO, L. Designing clustering methods for ontology building – the Mo'K Workbench. *In*: *Proceedings of the ECAI Ontology Learning Workshop*, 2000, p. 13-19.

BITNER, Mary Jo; BROWN, Stephen W. The evolution and discovery of services science in business schools. *Communications of the ACM*, v. 49, n. 7, jul. 2006.

BITTELSTONE, Robert. XPL: an expert systems framework in APL. *Association for Computing Machinery*, 1985.

BOOCH, Grady; RUMBAUGH, James; JACOBSON, Ivar. *The Unified Modeling Language user guide*. 2. ed., Addison-Wesley, 2005.

BORKO, Harold. Automatic indexing: a tutorial. *Association for Computing Machinery*, SIGIR Automatic Information Retrieval, 9-11 nov., 1981.

BRACHMAN, Ronald J.; MCGUINNESS, Deborah L. Knowledge representation, connectionism, and conceptual retrieval. *In*: Association for Computing Machinery*, SIGIR '88: Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, maio 1988, p. 161-174.

BREUKER, J. A.; WIELINGA, B. J. Interpretation of verbal data for knowledge acquisition. *In*: O'SHEA, T. (Ed.). *Advances in Artificial Intelligence*. Amsterdam: North Holland, 1987.

BROOKS, Bertram C. The foundations of Information Science. Part I. Philosophical Aspects. *Journal of Information Science*, n. 2, p. 125-133, 1980.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. A Protégé plug-in for ontology extraction from text based on linguistic analysis. *In*: *Proceedings of the 1st. European Semantic Web Symposium (ESWS)*, 2004, p. 31-44.

BUITELAAR, P.; SACALEANU, B. Extending synsets with medical terms. *In*: *Proceedings of the First International WordNet Conference*, 2002.

BUSH, Vannevar. As we may think. *Atlantic Monthly*, maio 1945. Disponível em: <http://www2.theAtlantic.com/atlantic/.../flashbks/computer/tech.htm>. Acesso em: 13 out. 2007.

CAMPOS, Maria Luiza de Almeida. Modelização de domínios de conhecimento: uma investigação de princípios fundamentais. *Ciência da Informação*, Brasília, v. 33, n. 1, p. 22-32, jan./abr. 2004.

CANTELE, Regina C.; ADAMATTI, Diana F.; FERREIRA, Maria A. G. V.; SICHMAN, Jaime S. *Reengenharia e ontologias: análise e aplicação*. Disponível em: <http://www.lti.pcs.usp.br/publicacoes/Cantele-etal04-WWS.pdf>. Acesso em: 22 dez. 2006.

CAO, Tru H.; CREASY, Peter N. Universal marker and functional relation: semantics and operations. *In*: LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual Structures: Fulfilling Peirce's Dream*. Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997. Proceedings, p. 416-430.

CAPUANO, E. A.; NEHME, C. Chauke. Exploring the parameter space of unsupervised ART neural networks for data mining. *In*: ZANASI, A.; BREBBIA, C. A.; EBECKEN, N. F. F. E.; MELLI, P. *Data Mining III*. Southampton, Boston: WIT Press, 2002, p. 461-472.

CAPUANO, Ethel Airton. O poder cognitivo das redes neurais artificiais modelo ART1 na recuperação da informação. *Ciência da Informação*, v. 38, n. 1, p. 9-30, jan./abr. 2009.

CAPUANO, Ethel Airton; CASAES, Julio; DA COSTA, Julio Reis; DE JESUS, Magda Sifuentes; MACHADO, Marco Antonio. Inteligência Competitiva e suas conexões epistemológicas com Gestão da Informação e do Conhecimento. *Ciência da Informação*, v. 38, n. 2, p. 19-34, maio/ago. 2009.

CAPUANO, Ethel Airton. Redesenho de processos e estruturas nas organizações da Administração Pública. *In*: KNIGHT, Peter Titcomb; FERNANDES, Ciro Campos Christo; CUNHA, Maria Alexandra (Org.). *e-Desenvolvimento no Brasil e no Mundo: subsídios e Programa e-Brasil*. São Caetano do Sul (SP): Yendis, 2007, p. 585-625.

CAPURRO, Rafael; HJØRLAND, Birger. O conceito de informação. Tradução de Ana Maria Pereira Cardoso, Maria da Glória Achtschin Ferreira e Marco Antonio de Azevedo. *Perspectivas em Ciência da Informação*, v. 12, n. 1, p. 148-207, jan./abr. 2007.

CHAUKE-NEHME, Cláudio. Foreword. *In*: DO PRADO, Hércules Antonio; FERNEDA, Edilson (Orgs.). *Emerging technologies of text mining: techniques and applications*. Hershey, New York: Information Science Reference, 2008, p. xi.

CHEN, Peter Pin-Shan. The entity-relationship model – toward a unified view of data. *ACM Transactions on Database Systems*, v. 1, n. 1, p. 9-36, mar. 1976.

CHOMSKY, Noam. Three models for the description of language. *I.R.E. Transactions on Information Theory*, v. 2, n. 3, p. 113-124, 1956.

CHOO, Chun Wei. Information seeking in organizations: epistemic contexts and contests. *Information Research*, v. 12, n. 2, jan. 2007. Disponível em: <http://informationr.net/ir/12-2/paper298.html>. Acesso em: 25 jan. 2007.

CHOO, Chun Wei. *A organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões*. Tradução de Eliana Rocha. São Paulo: SENAC, 2003.

CHOO, Chun Wei. Towards an information model of organizations. *The Canadian Journal of Information Science*, v. 16, n. 3, p. 32-62, 1991.

CHURCH, Kenneth W.; RAU, Lisa F. Commercial applications of Natural Language Processing. *Communications of the ACM*, v. 38, n. 11, p. 71-79, nov. 1995.

CIARAMITA, M.; GANGEMI, A.; RATSCH, E.; SARIC, J; ROJAS, I. Unsupervised learning of semantic relations between concepts of a molecular biology ontology. *In*: *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, p. 659-664.

CIMIANO, P.; HANDSCHUH, S.; STAAB, S. Towards the self-annotating Web. *In*: *Proceedings of the 13th World Wide Web Conference (WWW)*, 2004, p. 462-471.

CIMIANO, Philipp. *Ontology learning and population from text: algorithms, evaluation and applications*. Springer Science, 2006.

CNET. *Newsmaker: Gates explains why Microsoft needs Yahoo*. 20 fev. 2008. Disponível em: <http://news.cnet.com/Gates-explains-why-Microsoft-needs-Yahoo/2008-1014_3-6231341.html>. Acesso em: 17 jun. 2010.

COHEN, Michael D.; MARCH, James G.; OLSEN, Johan P. A garbage can model of organizational choice. *Administrative Science Quarterly*, v. 17, n. 1, p. 1-25, mar. 1972.

CONDON, E. U. Statistics of vocabulary. *Science*, n. 1.733, p. 300, 1928.

COOPER, W. S. Bridging the gap between AI and IR. Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. *Association for Computing Machinery*, 1984, p. 259-265.

CZARNEWSKI, Paulo Cezar. SOA: Arquitetura Orientada a Serviços. *In*: *Mostra TIC 2007* – Brasília (DF). Disponível em: <http://wwwsolucoestipublica.gov.br/palestras/Desd_3.4_ paulo_c_czarnewisk.pdf>. Acesso em: 17 jun. 2007.

DAELEMANS, Walter; GAZDAR, Gerald; DE SMEDT, Koenraad. Inheritance in Natural Language Processing. *Association for Computing Linguistics*, 1992.

DALE, Robert; ALIOD, Diego Mollá; SCHWITTER, Rolf. Evangelising language technology: a practically-focussed undergraduate program. *In*: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the Workshop of Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Philadelphia, jul. 2002, p. 27-32.

DAVENPORT, Thomas H. *Ecologia da informação*. 3. ed. São Paulo: Futura, 2000.

DAVENPORT, Thomas H. *Process innovation: reengineering work through information technology*. Boston, Massachusetts: Ernst & Young, Harvard Business School Press, 1993.

DAVIS, Mills. *Project10X's Semantic Wave 2008 Report: industry roadmap to Web 3.0 & multibillion dollar market opportunities – executive summary*, out. 2008. Disponível em: *http://www.isoco.org/pdf/Semantic_Wave_2008-Executive_summary.pdf.* Acesso em: 11 abr. 2010.

DAVYDOV, Mark M. *Corporate portals and e-business integration*. McGraw-Hill, 2001.

DE VILLE, Barry. *Microsoft data mining: integrated business intelligence for e-commerce and knowledge management*. Digital Press, 2001.

DELBECQUE, Nicole. *A linguística cognitiva: compreender como funciona a linguagem*. Tradução de Fernanda Oliveira. Lisboa: Instituto Piaget, 2008.

DIETZ, Jan L. G. The deep structure of business process. *Communications of the ACM*, v. 49, n. 5, p. 59-64, maio 2006.

DING, Yihong; XU, Li. *Evolution of the World Wide Web – part 2: Web evolution theory and the next stage*. Disponível em: *<http://www.deg.byu.edu/ding/WebEvolution/evolution-dream.html>.* Acesso em: 11 abr. 2010.

DOSZKOCS, Tamas E. Natural Language Processing in Information Retrieval. *Journal of the American Society for Information Science*, v. 37, n. 4, p. 191-196, jul. 1986.

DOYLE, L. B. Is automatic classification a reasonable application of statistical analysis of text? *Journal of the ACM*, n. 12, p. 473-489, 1965.

EBECKEN, Nelson Francisco Favilla; LOPES, Maria Celia Santos; COSTA, Myrian Christina de Aragão. Mineração de textos. *In*: REZENDE, Solange Oliveira (Org.). *Sistemas Inteligentes: fundamentos e aplicações*. Barueri, São Paulo: Manole, 2005, p. 337-375.

EKLUND, Peter W.; ELLIS, Gerard; MANN, Graham (Eds.). Conceptual structures: knowledge representation as interlingua. *Proceedings of the 4th International Conference on Conceptual Structures, ICCS '96*, Sidney, Australia: 1996.

ETZIONI, O; CAFARELLA, M.; DOWNEY, D.; POPESCU, A. M.; SHAKED, T.; SODERLAND, S.; WELD, D.; YATES, A. Methods for domain-independent information extraction from the Web: an experimental comparison. *In*: *Proceedings of the 19th National Conference on Artificial Intelligence (AAAI)*, 2004, p. 391-398.

EVANS, R. A framework for named entity recognition in the open domain. *In*: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 2003, p. 137-144.

EVENS, Martha Walton (Ed.). *Relational models of the lexicon: representing knowledge in semantic networks*. Cambridge University, 1988.

EXAME. *Estudo exame: cartões*. Edição 968, n. 9, ano 44, 19 maio 2010.

FAURE, D.; NEDELLEC, C. A corpus-based conceptual clustering method for verb frames and ontology. *In*: VELARDI, P. (Ed.). *In*: *Proceedings of the LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications*, 1998, p. 5-12.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMITH, Padhraic; RAMASAMY, Uthurusamy. *Advances in knowledge discovery and data mining*. Menlo Park (CA): AAAI; MIT Press, 1996.

FERNEDA, Edberto. Redes neurais em sua aplicação em sistemas de recuperação de informação. *Ciência da Informação*, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006.

FERRER, Florencia; LIMA, Christian. Introdução de mudanças tecnológicas no setor público: por onde começar? *In*: KNIGHT, Peter Titcomb; FERNANDES, Ciro Campos Christo; CUNHA, Maria Alexandra (Org.). *e-Desenvolvimento no Brasil e no Mundo: subsídios e Programa e-Brasil*. São Caetano do Sul (SP): Yendis, 2007, p. 626-638.

FRAWLEY, William. Relational models and metascience. *In*: EVENS, Martha Walton (Ed.). *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University, 1988, p. 335-368.

FREEDMAN, David H. *Los hacedores de cerebros – cómo los científicos están perfeccionando las computadoras, creando un rival del cerebro humano*. Tradução para o espanhol de Paulina Matta. Andres Bello, 1995.

FREGE, Gottlob. On sense and reference. *In:* RICHARD, Mark (Ed.). *Meaning*. Blackwell, 2003, p. 36-56.

FULD, Leonard M. *Inteligência Competitiva: como se manter à frente dos movimentos da concorrência e do mercado*. Tradução de Janaína Ruffoni. Rio de Janeiro: Elsevier, 2007.

GAMALLO, P.; AGUSTINI, A.; LOPES, G. Clustering syntatic positions with similar semantic requirements. *Computational Linguistics,* v. 21, n. 1, 2005, p. 107-145.

GAMALLO, P.; GONZALEZ, M.; AGUSTINI, A.; LOPES, G.; DE LIMA, V. S. Mapping syntatic dependencies onto semantic relations. *In*: *Proceedings of the ECAI Workshop on Machine Learning and Natural Language Processing for Ontology Engineering*, 2002, p. 15-22.

GANTER, Bernhard; STUMME, Gerd; WILLE, Rudolf (Eds.). *Formal Concept Analysis: foundations and applications*. Springer, 2005.

GERBÉ, Olivier. Conceptual graphs for corporate knowledge repositories. *In*: LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual Structures: Fulfilling Peirce's Dream*. Proceedings of the Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997, p. 474-488.

GIGUETTE, Ray. Building objects out of Plato: applying philosophy, symbolism, and analogy to software design. *Communications of the ACM*, v. 49, n. 10, p. 66-71, out. 2006.

GILBERT, Nigel. *Computer simulation of social processes*. Disponível em: <http://www.http://sru.soc.surrey.ac.uk/SRU6.html>. Acesso em: 11 dez. 2007.

GODIN, R.; GECSEI, J.; PICHET, C. Design of browsing interface for information retrieval. *In*: BELKIN, N. J.; VAN RIJSBERGEN, C. J. (Eds.). *Proceedings of the SIGIR '89*, p. 32-39, 1989.

GOMEZ-PÉREZ, A.; CORCHO, O. Ontology languages for the Semantic Web. *IEEE Intelligent Systems*, 2002.

GONÇALVES, José Ernesto Lima. As empresas são grandes coleções de processos. *Revista de Administração de Empresas*, v. 40, n. 1, jan./mar. 2000, p. 6-19.

GOOD, I. J. *Speculations concerning Information Retrieval*. Research Report PC-78, IBM Research Centre, Yorktown Heights, New York, 1958.

GOTTSCHALG-DUQUE, Cláudio. *SiRILiCO – Uma proposta para um sistema de recuperação de informação baseado em teorias da Lingüística Computacional e ontologia*. Tese de Doutorado, Belo Horizonte: Escola de Ciência da Informação da Universidade Federal de Minas Gerais (UFMG), 2005.

GREENWALD, Bruce; KAHN, Judd. *Competition demystified: a radically simplified approach to business strategy*. Portfolio, 2005.

GREFENSTETTE, G. *Explorations in automatic thesaurus construction*. Kluwer, 1994.

GREGG, Dawn G.; WALCZAK, Steven. Adaptive Web information extraction. *Communications of the ACM*, v. 49, n. 5, maio 2006.

GUTIÉRREZ, Patricio. Balanço do processo de desenvolvimento do governo eletrônico no Chile. *In*: KNIGHT, Peter Titcomb; FERNANDES, Ciro Campos Christo; CUNHA, Maria Alexandra (Org.). *e-Desenvolvimento no Brasil e no Mundo: subsídios e Programa e-Brasil*. São Caetano do Sul (SP): Yendis, 2007, p. 152-p. 171.

HAMMER, Michael; CHAMPY, James. *Reengineering the corporation: a manifesto for business revolution*. New York: HarperCollins, 2001.

HARRIES-JONES, Peter. A r*ecursive vision: ecological understanding and Gregory Bateson*. Toronto (Canada), Buffalo (New York, EUA) e Londres (Reino Unido): University of Toronto, 2002.

HARRIS, Z. *Mathematical structures of language*. Wiley, 1968.

HAVEY, Michael. *Essential business process modeling*. Sebastopol (CA): O'Reilly, 2005.

HAYS, David G. Dependency theory: a formalism and some observations. *Language*, v. 40, n. 4, p. 511-525, 1964.

HEARST, M. Automatic acquisition of hyponyms from large text corpora. *In*: *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, 1992, p. 539-545.

HEHN, Herman F. *Peopleware: como trabalhar o fator humano nas implementações de sistemas integrados de informação (ERP).* São Paulo: Gente, 1999.

HESSEN, Johannes. *Teoria do conhecimento*. Tradução de António Correia. 7. ed., Coimbra (Portugal): Arménio Amado, Brasil: Martins Fontes, 1976.

HINDLE, D. Noun classification from predicate-argument structures. *In*: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 1990, p. 268-275.

HOBBS, Jerry R.; APPELT, Douglas E.; BEAR, John; TYSON, Mabry. *Robust processing of real-world natural-language texts*. *In*: JACOBS, Paul S. (Ed.). *Text-based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1991.

HOFMANN, T. Probabilistic latent semantic indexing. *In*: *Proceedings of the 22nd ACM SIGIR International Conference on Research and Development in Information Retrieval*, 1999, p. 50-57.

HOSKINS, D. D.; DOBBERNACK, D.; KUPTSCH, C. (Eds.). *Social security at the dawn of the 21st century*. International Social Security Association, 2001.

IBIBLIO. *Internet pioneers: Vannevar Bush*. Disponível em: <http://www.ibiblio.org/pioneers/bush.html>. Acesso em: 13 out. 2007.

KAO, Anne; POTEET, Stephen R. (Eds.). *Natural Language Processing and Text Mining*. Springer, 2007.

KEIL, Frank C.; WILSON, Robert A. The concept concept: the wayward path of cognitive science. *Mind & Language*, v. 15, n. 2 e 3, p. 308-318, abr./jun. 2000.

KING, Margaret. Evaluating Natural Language Processing systems. *Communications of the ACM*, v. 39, n. 1, p. 73-79, jan. 1996.

KNELLER, George F. *A ciência como atividade humana*. Tradução de Antonio José de Souza. Rio de Janeiro: Zahar; São Paulo: EDUSP, 1980.

KOCH, Ingedore Villaça. *O texto e a construção dos sentidos*. São Paulo: Contexto, 1997.

KOCH, I. V.; SILVA, M. C. P. S. *Lingüística aplicada ao português: sintaxe*. São Paulo: Cortez, 1985.

KOLL, Matthew B. WEIRD: an approach to concept-based Information Retrieval. SIGIR Forum, Spring 1979. *Association for Computing Machinery*, v. 13, n. 4, 1979.

KONCHADY, Manu. *Text mining application programming*. Boston: Charles River Media, 2006.

KUZNETSOV, S. Machine Learning and Formal Concept Analysis. *In*: EKLUND, P. (Ed.). *Concept Lattices, 2nd International Conference on Formal Concept Analysis, LNCS 2961*, Berlin: Springer, 2004, p. 287-312.

LANDAUER, T.; DUMAIS, S. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, n. 104, p. 211-240, 1997.

LEASE, Matthew. Natural Language Processing for Information Retrieval: the time is ripe (again). *ACM PIKM' 07*, 9 nov. 2007, Lisboa, Portugal.

LEBOWITZ, Michael. Intelligent information systems. *Association for Computing Machinery*, 1983.

LEE, Juhnyoung. Model-driven business transformation and the Semantic Web. *Communications of the ACM*, v. 48, n. 12, p. 75-78, dez. 2005.

LEIDNER, Jochen L. Current issues in Software Engineering for Natural Language Processing. *In*: *Association for Computational Linguistics*. SEALTS '03: Proceedings of the HLT-NAACL 2003 Workshop on Software Engineering and Architecture of Language Technology Systems, v. 8, maio 2003, p. 45-50.

LIEBOWITZ, Jay. *Strategic Intelligence: Business Intelligence, Competitive Intelligence, and Knowledge Management*. Auerbach, 2006.

LIMA-MARQUES, Mamede; MACEDO, Flávia Lacerda Oliveira. Arquitetura da Informação: base para a gestão do conhecimento. *In*: TARAPANOFF, Kira (Org.). *Inteligência, Informação e Conhecimento*. Brasília: IBICT e UNESCO, 2006b.

LIN, D. Automatic retrieval and clustering of similar words. *In*: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL)*, 1998, p. 768-774.

LIN, D.; PANTEL, P. Induction of semantic classes from natural language text. *In*: *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2001, p. 317-322.

LINDIG, Christian; SNELTING, Gregor. Assessing modular structure of legacy code based on mathematical concept analysis. *ACM ICSE 97*, Boston, MA (USA), 1997.

LISPECTOR, Clarice. *A hora da estrela*. Rio de Janeiro: Rocco, 1998.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research*, n. 2, v. 4, p. 159-165, 1958.

LUHN, H. P. A statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development*, n. 1, p. 309-317, 1957.

LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual structures: fulfilling Peirce's Dream*. Proceedings of the Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997.

LUNA-REYES, Luis F.; ANDERSEN, David F.; RICHARDSON, George P.; PARDO, Theresa A.; CRESSWELL, Anthony M. Emergence of the governance structure for information integration across governmental agencies: a system dynamics approach. *In: The Proceedings of the 8th Annual International Digital Governmental Research Conference*, 2005.

MÄDCHE, A.; STAAB, S. Discovering conceptual relations from text. *In*: *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI)*, 2000, p. 321-325.

MÄDCHE, A.; STAAB, S. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, v. 16, n. 2, 2001, p. 72-79.

MAGLIO, Paul P. *et al*. Service systems, service scientists, SSME, and innovation. *Communications of the ACM*, v. 49, n. 7, jul. 2006.

MAIA, Flávia. Leitor contemporâneo. Entrevista concedida por Maria Zilda da Cunha, especialista em literatura infantil e juvenil. *In: Correio Braziliense*, Brasília, 30 maio 2010.

MANDELBROT, B. An information theory of the statistical structure of language. *In*: *Communication Theory*, p. 486-502, London: Butterworths Scientific Publications, 1953.

MARCH, J. G.; SIMON, H. A. *Organizations.* 2. ed., Blackwell, 1993.

MATHEUS, Renato Fabiano. Rafael Capurro e a Filosofia da Informação: abordagens, conceitos e metodologias de pesquisa para a Ciência da Informação. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 10, n. 2, p. 140-165, jul./dez. 2005.

MCGEE, James; PRUSAK, Laurence. *Gerenciamento estratégico da informação*. Tradução de Astrid Beatriz de Figueiredo. 11. ed. Rio de Janeiro: Campus, 1994.

MEADOW, Charles T.; BOYCE, Bert R.; KRAFT, Donald H.; BARRY, Carol. *Text information retrieval systems*. 3. ed., Elsevier, 2007.

MICHAUD, Claude. Modelos e Conhecimento. *In*: TARAPANOFF, Kira (Org.). *Inteligência, informação e conhecimento*. Brasília: IBICT e UNESCO, 2006b, p. 211-239.

MINEAU, Guy; STUMME, Gerd; WILLE, Rudolf. Conceptual structures represented by Conceptual Graphs and Formal Concept Analysis. *In*: TEPFENHART, W.; CYRE, W. (Eds.). Conceptual Structures: Standards and Practices. *Proceedings of the 7th International Conference on Conceptual Structures*, LNAI 1640, Berlin: Springer, 1999, p. 423-441.

MINSKY, M. *Semantic information processing*. Cambridge, Massachusetts: MIT Press, 1968.

NATIS, Yefim V.; SCHULTE, Roy W. *Business component architecture unites services and events*. Gartner Group, Research, 8 out. 2004.

NAVIGLI, R.; VELARDI, P. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, v. 30, n. 2, p. 151-179, 2004.

NEVO, Dorit; WADE, Michael R. How to avoid disappointment by design: avoid market failure by aligning system performance with stakeholder expectations. *Communications of the ACM*, v. 50, n. 4, p. 43-48, abr. 2007.

NONAKA, Ikujiro; TAKEUCHI, Hirotaka. *Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação*. Tradução de Ana Beatriz Rodrigues e Priscilla Martins Celeste. 16 ed., Rio de Janeiro: Elsevier, 1997.

OLIVEIRA, Saulo Barbará de (Org.). *Gestão por processos: fundamentos, técnicas e modelos de implementação*. Rio de Janeiro: Qualitymark, 2004.

ONTORIA, A.; DE LUQUE, A.; GÓMEZ, J. P. R. *Aprender com mapas mentais: uma estratégia para pensar e estudar*. Tradução de Silvia Mariângela Spada, 2. ed., São Paulo: Madras, 2006.

ORTEGÓN, Edgar; PACHECO, Juan Francisco; PRIETO, Adriana. *Metodologia del Marco Lógico para la planificación, el seguimiento y la evaluación de proyectos y programas*. Instituto Latinoamericano y del Caribe de Planificación Económica y Social (ILPES), Chile: Comissión Económica para el Desarollo de la America Latina y Caribe, Naciones Unidas, jul. 2005.

PAULA FILHO, Wilson de Pádua. *Engenharia de Software: fundamentos, métodos e padrões*. 2. ed. Rio de Janeiro: LTC, 2005.

PEIRCE, Charles S. *Semiótica*. Tradução de José Teixeira Coelho Neto, 4. ed., São Paulo: Perspectiva, 2010.

PEIRCE, Charles Sanders. *Existential Graphs*. Comentários de John F. Sowa, baseado no Manuscrito MS 514, depositado na Houghton Library, Harvard University, 1909. Disponível em: http://www.jfsowa.com/peirce/ms514.htm. Acesso em: 12 jul. 2009.

PENTEADO, Roberto; BOUTIN, Eric. Creating strategic information for organizations with structured text. *In:* PRADO, Hércules Antonio do; FERNEDA, Edilson (Orgs.). *Emerging Technologies of Text Mining: techniques and applications*. Hershey, New York: Information Science Reference, 2008, p. 34-53.

PERINI, M. A. *Gramática descritiva do Português*. 4. ed., São Paulo: Ática, 2003.

PESONEN, Juha Petteri. *Concepts and object-oriented knowledge representation*. MA Thesis, University of Helsinki, Department of Cognitive Science, fev. 2002.

PIDD, Michael. *Computer simulation in management Science*. 2. ed., John Wiley & Sons, 1988.

PRADO, Hércules Antonio do; FERNEDA, Edilson (Orgs.). *Emerging technologies of text mining: techniques and applications*. Hershey, New York: Information Science Reference, 2008.

PRAHALAD, C. K.; KRISHNAM, M. S. The new meaning of quality in the information age. *Harvard Business Review*, v. 77, n. 5, set. 1999.

PRESSMAN, Roger S. *Engenharia de Software*. Tradução de José Carlos Barbosa dos Santos. São Paulo: Makron Books, 1995.

PRISS, Uta. Formal Concept Analysis in Information Science. *In*: CRONIN, Blaise (Ed.). *Annual Review of Information Science and Tecnology*, ASIST, v. 40, 2005.

PULIER, Eric; TAYLOR, Hugh. *Compreendendo SOA corporativa*. Tradução de Marcelo Trannin Machado, Rio de Janeiro: Ciência Moderna, 2008.

QUILLIAN, M. Ross. *Semantic memory*. PhD Dissertation, Carnegie-Mellon University, Pittsburgh, 1968.

QUONIAM, Luc. *Competitive Intelligence 2.0*. Brasília: Apresentação na Faculdade de Ciência da Informação da Universidade de Brasília – FCI/UnB, 5 mar. 2010.

REINBERGER, M. L.; SPYNS, P. Discovering knowledge in texts for the learning of dogma-inspired ontologies. *In*: BUITELAAR, P.; HANDSCHUH, S.; MAGNINI, B. (Eds.). *Proceedings of the ECAI Workshop on Ontology Learning and Population*, 2004, p. 19-24.

RICHARD, Mark (Ed.). *Meaning*. Blackwell, 2003.

RITTO, Antonio Carlos. *Organizações caórdicas: modelagem de organizações inovadoras*. Rio de Janeiro: Ciência Moderna, 2005.

ROBERTS, D. D. *The Existential Graphs of Charles Sanders Peirce*. Mouton, The Hague, Netherlands: 1973.

ROBREDO, Jaime. Planejamento e gerência de sistemas deiInformação sob o ângulo da gestão por processos. *Revista de Biblioteconomia de Brasília*, v. 23/24, n. 4, 1999/2000. Edição Especial.

ROSE, Daniel E.; BELEW, Richard K. Legal information retrieval: a hybrid approach. *Association for Computing Machinery*, 1989.

SALTON, G. Automatic text analysis. Science, n. 168, p. 335-343, 1970.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, v. 24, n. 5, p. 515-523, 1988.

SANDERSON, M.; CROFT, B. Deriving concept hierarchies from text. *In*: *Proceedings of the SIGIR Conference on Research and Development in Information Retrieval*, 1999, p. 206-213.

SANTOS, Cícero Nogueira dos. *Aprendizado de máquina na identificação de sintagmas nominais: o caso do Português brasileiro*. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro, 2005.

SARACEVIC, Tefko. Information Science. *Journal of the American Society for Information Science*, v. 50, n. 12, p. 1051-1063, 1999.

SAUSSURE, Ferdinand de. *Escritos de Lingüística geral*. Organizados e editados por Simon Bouquet e Rudolf Engler. Tradução de Carlos Augusto Leuba Salum e Ana Lucia Franco. São Paulo: Cultrix, 2002.

SAYOOD, Khalid. *Introduction to data compression*. Morgan Kaufmann, 2000.

SCHREIBER, Guus; AKKERMANS, Hans; ANJEWIERDEN, Anjo; DE HOOG, Robert; SHADBOLT, Nigel; VAN DE VELDE, Walter; WIELINGA, Bob. *Knowledge Engineering and Management: The CommonKADS Methodology*. Cambridge, MA (EUA), e London (Reino Unido): MIT, 2000.

SCHREIBER, Guus; WIELINGA, Bob; BREUKER, Joost. *KADS: A principled approach to knowledge-based system development*. Academic Press, 1993.

SCHUTZ, A.; BUITELAAR, P. RelExt: A tool for relation extraction from text in ontology extension. *In*: *Proceedings of the International Semantic Web Conference*, 2005, p. 593-606.

SCHÜTZE, H. Word space. *In*: *Advances in Neural Information Processing Systems 5*, 1993, p. 895-902.

SCOTT, Mike. *WordSmith tools: version 5.0*. Liverpool (UK): Lexical Analysis Software, 2008.

SEARLE, Leroy; KEELER, Mary; SOWA, John; DELUGACH, Harry; LUKOSE, Dickson. Fulfilling Peirce's dream: conceptual structures and communities of inquiry. *In*: LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual Structures: Fulfilling Peirce's Dream*. Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997. Proceedings, p. 1-11.

SHANNON, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, p. 379-423, jul.-out. 1948.

SILVA, E. M.; PRADO, H. A. do; FERNEDA, E. Text Mining: crossing the chasm between the academy and the industry. *In*: ZANASI, A.; BREBBIA, C. A.; EBECKEN, N. F. F. E.; MELLI, P. *Data Mining III*. Southampton, Boston: WIT Press, 2002, p. 351-361.

SILVA, Ivanosca A.; BEDREGAL, Benjamin R. C.; LUCENA, Márcia J. N. R.; GOTTGTROY, Márcia P. B. *Um framework para desenvolvimento de sistemas complexos*. Disponível em: <http://www.dimap.ufrn.br/~bedregal/publications/pub2001/clei2001-Ivanosca.PDF>. Acesso em: 30 abr. 2005.

SIMPSON, Rosemary; RENEAR, Allen; MYLONAS, Elli; VAN DAM, Andries. 50 Years after "As We May Think". The Brown/MIT Vannevar Bush Symposium. *Interactions*, mar. 1996. Disponível em: <http://delivery.acm.org/10.1145/230000/227187/p47-simpson.pdf?key1=227187&key2=2100462911&coll=&dl=acm&CFID=15151515&CFTOKEN=6184618>. Acesso em: 13 out. 2007.

SOWA, John F. *Conceptual structures: information processing in mind and machine*. Reading, MA: Addison-Wesley, 1984.

SOWA, John. *Knowledge representation: logical, philosophical and computational foundations*. Brooks Cole, 2000.

SPOHRER, Jim; RIECKEN, Doug (Ed.). Services science. *Communications of the ACM*, v. 49, n. 7, p. 31-33, jul. 2006.

STAAB, Steffen. The Semantic Web revisited. *IEEE Computer Society*, p. 96-101, maio-jun. 2006.

STONE, John A. *Developing software applications in a changing IT environment*. Mc Graw-Hill, 1997.

STUMME, G. Formal Concept Analysis on its way from Mathematics to Computer Science. *In*: PRISS, U.; CORBETT, D.; ANGELOVA, G. (Eds.). *Conceptual Structures: Integration and Interfaces, 10th International Conference on Conceptual Structures, LNCS 2393*, Berlin: Springer, 2002, p. 2-19.

TALEB, Nassim Nicholas. *The black swan: the impact of the highly improbable*. Penguin Books, 2007.

TARAPANOFF, Kira. Informação, conhecimento e inteligência em corporações: relações e complementaridade. *In*: TARAPANOFF, Kira (Org.). *Inteligência, Informação e Conhecimento*. Brasília: IBICT e UNESCO, 2006a.

TARAPANOFF, Kira (Org). *Inteligência, informação e conhecimento*. Brasília: IBICT e UNESCO, 2006b.

TARAPANOFF, Kira (Org.). *Inteligência organizacional e competitiva*. Brasília: UnB, 2001.

TESNIERE, Lucien. *Elements de syntaxe structurale*. Paris: Librairie C. Klincksieck, 1959.

TONG, Richard M.; ASKMAN, Victor N.; CUNNINGHAM, James F; TOLLANDER, Carl J. RUBRIC: an environment for full text information retrieval. *Association for Computing Machinery*, 1985.

TURCATO, D.; POPOWICH, F.; TOOLE, J.; FASS, D.; NICHOLSON, D.; TISHER, G. Adapting a synonym database to specific domains. *In*: *Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval*, 2000.

VALTCHEV, P.; MISSAOUI, R.; GODIN, R. Formal Concept Analysis for knowledge discovery and data mining: the new challenges. *In*: EKLUND, P. (Ed.). *Concept Lattices, 2nd International Conference on Formal Concept Analysis, LNCS 2961*, Berlin: Springer, 2004, p. 352-371.

VAZ, Maria Elizabete. Integração de dados além da tecnologia. *Mostra TIC 2007* – Brasília (DF). Disponível em: <http://www.solucoestipublica.gov.br/grade_palestras>. Acesso em: 11 maio 2007.

VELARDI, P.; NAVIGLI, R.; CUCHIARELLI, R.; NERI, F. Evaluation of OntoLearn, a methodology for automatic population of domain ontologies. *In*: BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. (Eds). *Ontology Learning from Texts: Methods, Applications and Evaluation*. Frontiers in Artificial Intelligence and Applications, IOS Press, n. 123, 2005, p. 92-106.

WEICK, Karl E. *Sensemaking in organizations*. SAGE, 1995.

WEILKIENS, Tim; OESTEREICH, Bernd. *UML 2 certification guide: fundamental and intermediate exams*. San Francisco (CA): Morgan Kaufmann, 2007.

WERNER, Oswald. How to teach a network: minimal design features for a cultural knowledge acquisition device or C-KAD. *In*: EVENS, Martha Walton (Ed.). *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge University, 1988, p. 141-166.

WILKS, Yorick. Methodology in AI and natural language understanding. *In*: *Proceedings of the 1975 workshop on Theoretical issues in natural language processing,* p. 130-133, Cambridge, Massachusetts: ACM Special Interest Group on Artificial Intelligence, 1975.

WILLE, Rudolf. Conceptual Graphs and Formal Concept Analysis. *In*: LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual Structures: Fulfilling Peirce's Dream*. Proceedings of the Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997, p. 290-303.

WILLE, Rudolf. Formal Concept Analysis as mathematical theory of concepts and concept hierarchies. *In*: GANTER, Bernhard; STUMME, Gerd; WILLE, Rudolf (Eds.). *Formal Concept Analysis: Foundations and Applications*. Springer, 2005, p. 1-33.

WILLE, Rudolf. Restructuring lattice theory: an approach based on hierarchies of concepts. *In:* RIVAL, I. (Ed.). Ordered Sets. Dordrecht-Boston: Reidel, 1982, p. 445-470.

WINOGRAD, T. *Understanding natural language*. Edinburgh, 1972.

WINOGRAD, Terry. *Language as a cognitive process – volume I: syntax*. Addison-Wesley, 1983.

WITTGENSTEIN, Ludwig. *Philosophical investigations*. Traduzido para o inglês por G. E. M. Anscombe. 3. ed., Malden (Massachusetts, EUA): Blackwell, 2001.

YOON, Jeonguon. Realizando a u-Coréia. *In*: KNIGHT, Peter Titcomb; FERNANDES, Ciro Campos Christo; CUNHA, Maria Alexandra (Org.). *e-Desenvolvimento no Brasil e no Mundo: subsídios e Programa e-Brasil*. São Caetano do Sul (SP): Yendis, 2007, p. 120-129.

ZANASI, A.; BREBBIA, C. A.; EBECKEN, N. F. F. E.; MELLI, P. *Data mining III*. Southampton, Boston: WIT Press, 2002.

ZARRI, Gian Piero. RESEDA, An information retrieval system using Artificial Intelligence and Knowledge Representation Techniques. *ACM SIGIR Forum*, v. 17, n. 4, p. 189-195, 1983.

ZEMAN, Jay. Peirce's Graphs. *In*: LUKOSE, Dickson; DELUGACH, Harry; KEELER, Mary; SEARLE, Leroy; SOWA, John (Eds.). *Conceptual structures: fulfilling Peirce's dream*. Fifth International Conference on Conceptual Structures, ICCS'97, Seattle, Washington, USA, August 1997. Proceedings, p. 12-24.

ZEMAN, Jay. *The graphical logic of C. S. Peirce*. PhD Dissertation, University of Chicago, 1964.

ZIPF, G. K. *Human behavior and the principle of least effort*. Addison-Wesley, 1949.

## APENSOS