RENATO ROCHA SOUZA

Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Doutor em Ciência da Informação

Área de concentração: Organização e Tratamento da Informação.

Orientadora: Profa. Dr. Lídia Alvarenga

Belo Horizonte

Escola de Ciência da Informação

2005

S729p Souza, Renato Rocha
Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais / Renato Rocha Souza. – Belo Horinzonte: ECI/UFMG, 2005.

197f.; 29,7 cm.

Tese (Doutorado) – Escola de Ciência da Informação, UFMG, 2005.

1. Indexação automática – Sintagmas nominais. I. Título

CDU 025.4.034

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"UMA PROPOSTA DE METODOLOGIA PARA ESCOLHA AUTOMÁTICA DE DESCRITORES UTILIZANDO SINTAGMAS NOMINAIS".

Renato Rocha Souza

Tese submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de **"Doutor em Ciência da Informação"**, linha de pesquisa **"Organização e Utilização da Informação (OUI)"**.

Tese aprovada em: 04 de maio de 2005.

Por:

Profa. Dra. Lídia Alvarenga –ECI/UFMG (Orientadora)

Profa. Dra. Beatriz Valadares Cendon –ECI/UFMG

Profa. Dra. Maria Eugênia Albino Andrade –ECI/UFMG

Prof. Dr. Hélio Kuramoto –IBICT/Brasília

Profa. Dra. Renata Vieira -UNISINOS

Aprovada pelo Colegiado do PPGCI

Profa. Maria Eugênia Albino Andrade
Coordenadora

Versão final Aprovada por

Profa. Lídia Alvarenga
Orientadora

Dedicatória

À minha esposa, Karina, com quem o convívio se traduz na consciência diária de todos os valores pelos quais se deseja viver junto a alguém;

Ao meu filho Theo porque, como o próprio Deus, me faz conhecer através do amor o sentido da vida;

A meus pais, Roberto e Ana Maria; meus irmãos, Leonardo e Pedro; e minha inteira família; por me mostrarem na vida o valor maior de “ser” e “conhecer”, antes de “ter”;

E a “mon chien” Guguinho, porque – parafraseando Drummond – “tem quatro patas e o sentimento do mundo”.

Agradecimento especial

À minha estimada orientadora, Lídia Alvarenga, por ter acreditado em meu trabalho quando ainda era um devir, e por todo o prazer da convivência neste processo de orientação;

Agradecimentos

À Universidade Federal de Minas Gerais;

Aos professores Hélio Kuramoto, Renata Vieira e parceiros – na UNISINOS e na Universidade de Évora; membros do colegiado do NITEG; Eckhard Bick e Kothi Raghavan; pelo apoio, a cessão de computadores, ferramentas, e inestimáveis contribuições;

Aos inumeráveis colegas e amigos; professores, funcionários e alunos da Escola de Ciência da Informação da UFMG e da PUC-MG, pelas acolhidas, atenção, carinho, idéias, sugestões, apoio, críticas, e o privilégio de conhecê-los e trabalhar convosco;

A cada um dos colegas e amigos de doutorado, em especial ao Carlos Alberto Ávila Araújo e ao Rivadávia C. D. Alvarenga Neto, pelas intensas trocas de idéias.

Epígrafe

“Quem lê tanta notícia?”
(Caetano Veloso)

“Onde não há texto, também não há objeto
de estudo e de pensamento”
(Mikhail Bakhtin)

Resumo

Desde que se tornaram inviáveis em alguns contextos os processos manuais de indexação de documentos, buscam-se alternativas eficazes que possibilitem a representação automática dos assuntos principais desses documentos. Os processos mais comuns de indexação automática descrevem os documentos através de uma lógica simplista advinda da análise de freqüência das palavras que neles ocorrem. Buscando propor processo de indexação mais eficaz, que analise as palavras e expressões no âmbito de seus contextos lingüísticos, três pressupostos são definidos: (1) a utilização de sintagmas nominais como descritores apresenta vantagens em relação ao uso de palavras-chave; (2) a extração de sintagmas nominais de textos de documentos digitalizados é possível e viável com ferramentas tecnológicas atualmente disponíveis e (3) é possível estabelecer processo automatizado e eficaz para escolha de descritores significativos para documentos digitalizados, utilizando sintagmas nominais. O objetivo da presente pesquisa é apresentar uma metodologia para viabilizar o processo de atribuição de descritores a textos digitalizados – indexação – através da extração de sintagmas nominais e da análise de fatores como a freqüência de ocorrência desses sintagmas nominais nos textos dos documentos, no conjunto dos documentos; a estrutura dos sintagmas nominais; o nível dos sintagmas nominais e a ocorrência desses em tesauro de um campo de conhecimento específico. Para atingir esse objetivo são analisados (a) um *corpus* de 15 documentos dos quais foram extraídos os sintagmas nominais manualmente, para testar o processo de extração automática e (b) um *corpus* de 60 documentos provenientes de publicações eletrônicas da área de ciência da informação. A metodologia proposta foi aplicada inicialmente a parte do *corpus* para validação e parametrização das variáveis do algoritmo, e então novamente aplicada, com alterações, à totalidade do *corpus*. Os resultados apresentados demonstraram grande pertinência dos descritores atribuídos aos documentos e permitiram concluir que a metodologia obtém sucesso inequívoco nas condições estudadas.

Palavras-chave: sintagmas nominais, sistemas de recuperação de informações, indexação automática.

SOUZA, R. R. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais. 2005. 197 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte.

Abstract

Since manual indexing was found impossible for some document processing contexts, researchers seek alternatives to represent documents' subjects automatically. The most common processes try to determine documents' subjects through the analysis of words' frequencies. Searching for a better indexing process which analyses words and expressions within their linguistics contexts, three assumptions are made: (1) using noun phrases as descriptors is better than using keywords; (2) the extraction of the noun phrases from digitalized textual documents is possible and viable with the software tools available and (3) it is possible to establish an automated and functional process to choose good descriptors for documents using noun phrases. The aim of this research was to develop a methodology that would enable the indexation of digitalized documents through the extraction of the noun phrases and analysis of characteristics such as: (1) the frequency of occurrence of the noun phrases in the text of the document; (2) The frequency of occurrence in the whole set of documents; (3) the structure of the noun phrase; (4) the level of the noun phrase and (5) the occurrence of the noun phrase in a thesaurus of the subject's field. In order to reach this goal, the following pieces were analyzed (a) a corpus made of 15 documents from winch the noun phrases were extracted manually, to test the automatic extraction and (b) a corpus made of 60 documents coming from the field of information science. The methodology proposed was applied initially to part of the corpus for validation and calibration purposes, and then it was again applied, with some changes, to the whole corpus. The results presented showed a great deal of adequateness of the descriptors associated to the documents and this led to the conclusion that the methodology is unequivocally successful in the studied conditions.

Keywords: noun phrases, information retrieval systems, automatic indexing.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

CGI	**C**ommon **G**ateway **I**nterface
D	**D**eterminante
HTML	**H**yper **T**ext **M**arkup **L**anguage
MB	**M**egabytes
N	**N**ome
PDF	**P**ortable **D**ocument **F**ormat
PERL	**P**ractical **E**xtraction and **R**eporting **L**anguage
PLN	**P**rocessamento de **L**inguagem **N**atural
RAM	**R**andom **A**ccess **M**emory
RDF	**R**esource **D**escription **F**ramework
SGBD	**S**istema **G**erenciador de **B**anco de **D**ados
SGML	**S**tandard **G**eneralized **M**arkup **L**anguage
SN	**S**intagma **N**ominal
SRI	**S**istema de **R**ecuperação de **I**nformações
SV	**S**intagma **V**erbal
TXT	**T**e**xt**o Simples
VISL	**V**irtual **I**nteractive **S**yntax **L**earning
WWW, WEB	**W**orld **W**ide **W**eb
XML	E**x**tensible **M**arkup **L**anguage
XSL	E**x**tensible **S**tylesheet **L**anguage

Sumário

1 INTRODUÇÃO

Uma das características dos trabalhos de pesquisa no campo das ciências sociais aplicadas é a possibilidade de se adotar posicionamento que as distanciam tanto do positivismo solipsista, comum às ciências exatas – que não raro apresentam objetos de pesquisa como *fins em si*, totalmente dissociados dos contextos que os motivam e das conseqüências de seus resultados – mas também distante de certo subjetivismo, que por vezes acometem as pesquisas oriundas de ciências humanas, o que faz também com que seus resultados não se traduzam em benefícios generalizáveis. A tese que ora se apresenta nasceu de um contexto social amplo que o justifica, e se materializa em uma metodologia para resolver um problema bem definido, apresentando soluções viáveis.

É propósito desta introdução oferecer os subsídios necessários ao correto entendimento da passagem desses amplos contextos aos problemas que neles se originam, em especial aquele para o qual se pretende apresentar proposta de solução. Nesse sentido, são apresentados nesta introdução: o contexto social da pesquisa; a gênese do objeto de pesquisa; a delimitação do problema; os objetivos e pressupostos. Ao final, apresenta-se a estrutura do trabalho.

Diversos teóricos procuram abarcar, em suas análises, o fenômeno da concretização de previsões sobre uma “sociedade da informação”, ou “do conhecimento”, em que a maior força motriz para geração de bens comuns está baseada na informação e nos diversos sistemas especialistas e mediáticos que a manipulam ou dela dependem (TOFFLER, 1980; SCHAFF, 1990; GIDDENS, 1991; LEVY, 1993 e 1999; CASTELLS, 1999; TAKAHASHI, 2000; MATTELART, 2002). Alguns desses teóricos apontaram por vezes as facetas mais insidiosas desse processo (SANTOS, 2000; POSTMAN, 1984; BECK, 1992); é inegável, porém, a importância que os sistemas de informação[1], seus subprodutos e suas tecnologias associadas, assumiram na constituição das estruturas sociais (GIDDENS, 1991; CASTELLS, 1999), e a confiança quase atávica dos usuários nesses sistemas (GIDDENS, 1991). Se olharmos à nossa volta, podemos perceber as inúmeras dependências entre essas tecnologias e a sociedade, o que pode ser ilustrado

[1] Entende-se, no escopo deste trabalho, que sistemas de informação são sistemas que desempenham atividades de comunicação de informações, integrando tecnologias e grupos humanos, nas diversas configurações políticas e sociais.

por uma miríade de exemplos, como os sistemas de comunicação do mercado financeiro, os sistemas de controle de tráfego terrestre, marítimo e aeroviário, de telecomunicações e telefonia, sistemas de folha de pagamento, sistemas de controle comercial, a Internet e a *word wide web*, entre outros.

Lado a lado aos problemas sociais de exclusão digital que impedem que grande parcela da população possua os meios tecnológicos e as ferramentas cognitivas para compreensão, acesso e utilização dos acervos disponíveis nas redes eletrônicas, convivem problemas não menos importantes, relativos à gestão das informações que são produzidas continuamente pelas atividades humanas, e necessárias a todo instante para preencher nossas lacunas de conhecimento, nos vários âmbitos sociais. Esses problemas devem ser atacados de forma concomitante, porque a ignorância de qualquer desses aspectos pode gerar atrasos onerosos no desenvolvimento da sociedade.

Os sistemas de informação e de comunicação permeiam e viabilizam virtualmente todas as atividades sociais, e não mais podemos conceber a sociedade sem sua acentuada imbricação com as tecnologias de informação que nela surgem e a modificam. Acompanhando o desenvolvimento dessas tecnologias, os repositórios de informações que são produzidos durante o desempenho das inúmeras atividades humanas vêm migrando para o ambiente *on-line*, de forma que, parafraseando SHERA & CLEVELAND (1977), "*os registros da aventura intelectual humana*" estejam cada vez mais em formatos digitais, acessíveis através de redes e sistemas de computadores. Nas palavras de FOSKETT (1997, p. 3), "*as necessidades humanas de informação estão crescendo, na medida em que crescem as dependências de informação da sociedade, para sobreviver e florescer*".

Para suprir a necessidade de registrar as informações, criadas continuamente em ritmos vertiginosos, e a demanda por essas informações, são necessárias mudanças estruturais nas entidades que atuam como "centros de cálculo" (LATOUR *in* BARATIN e JACOB, 2000, p. 21), como as bibliotecas, repensando seus processos e instrumentos à luz das novas configurações sócio-técnicas. Esses centros de cálculo há muito vêm se beneficiando da existência de sistemas de recuperação de informações[2] (SRIs), que

[2] Entende-se, no escopo da presente tese, que os sistemas de recuperação de informações são sistemas, usualmente baseados em tecnologias digitais, que lidam com a organização e o acesso aos itens

utilizam diversas tecnologias mecânicas e digitais de computação, para gerenciar grandes acervos de documentos. São exemplos os sistemas de controle de acervo de bibliotecas tradicionais e também, em fenômenos mais recentes, a Internet, as intranets empresariais com seus portais corporativos, e as bibliotecas digitais.

Nesse contexto, o objeto de pesquisa em questão nasceu como contribuição para se enfrentarem alguns dos muitos desafios que surgem, quando lidamos com massivas quantidades de dados textuais, como nos grandes acervos de documentos digitais, notadamente quando estes precisam ser regularmente organizados e pesquisados, visando recuperar em tempo hábil informações relevantes para algum objetivo específico.

Com o aparente esgotamento[3] das estratégias tradicionais de busca de informação em SRIs, entendemos que a melhoria da eficácia do serviço aos usuários dos sistemas depende de esforços em diversas linhas de pesquisa, em todo o espectro da cadeia de processos de organização da informação. Algumas das opções de trabalho são as seguintes:

1) a exploração das informações semânticas intrínsecas aos documentos, de forma a expandir a compreensão das unidades e padrões de significado em textos, imagens e outras mídias;

2) o desenvolvimento de novas possibilidades de marcação semântica dos dados utilizando-se metalinguagens, criando registros de metadados acoplados aos próprios documentos com termos amplamente consensuais e não ambíguos, para que esses possam ser mais facilmente manipulados e identificados por computadores e outros dispositivos e, como conseqüência, pelos usuários;

3) o desenvolvimento de estratégias de apresentação da informação recuperada nas buscas, de forma altamente significativa e contextual[4] – como em algumas interfaces gráficas – de forma que as relações entre os conceitos, e em conseqüência, os contextos, sejam evidentes; e também de estratégias que

---

de informação, desempenhando as atividades de representação, armazenamento e recuperação desses itens.

[3] As estratégias tradicionais de busca e recuperação de informações em SRIs baseiam-se na modelagem do assunto dos documentos a partir da distribuição de suas palavras-chave. Embora existam inúmeras propostas de avanços, parece haver um limite para a eficácia de muitas dessas estratégias.

[4] Informação apresentada sem desprezo do contexto que lhe confere sentido.

busquem estimular os vários órgãos sensoriais ao mesmo tempo – como nas ferramentas multimídias – para que a absorção das informações pelos usuários seja maior. Através dessas interfaces e estratégias, as informações podem ser apresentadas de forma a possuírem conexões visuais com os contextos de origem, permitindo ao usuário refinar os resultados através da definição das conexões pertinentes e a exclusão das conexões geradas pelo ruído informacional;

4) A construção e a manutenção de perfis personalizados de utilização, de forma que o SRI "aprenda" com a forma de trabalho do usuário e possa utilizar essas informações específicas para melhorar a estratégia de busca do SRI.

Vejamos a representação das estratégias, seus desdobramentos e possíveis vantagens no mapa conceitual da FIG. 1 a seguir:

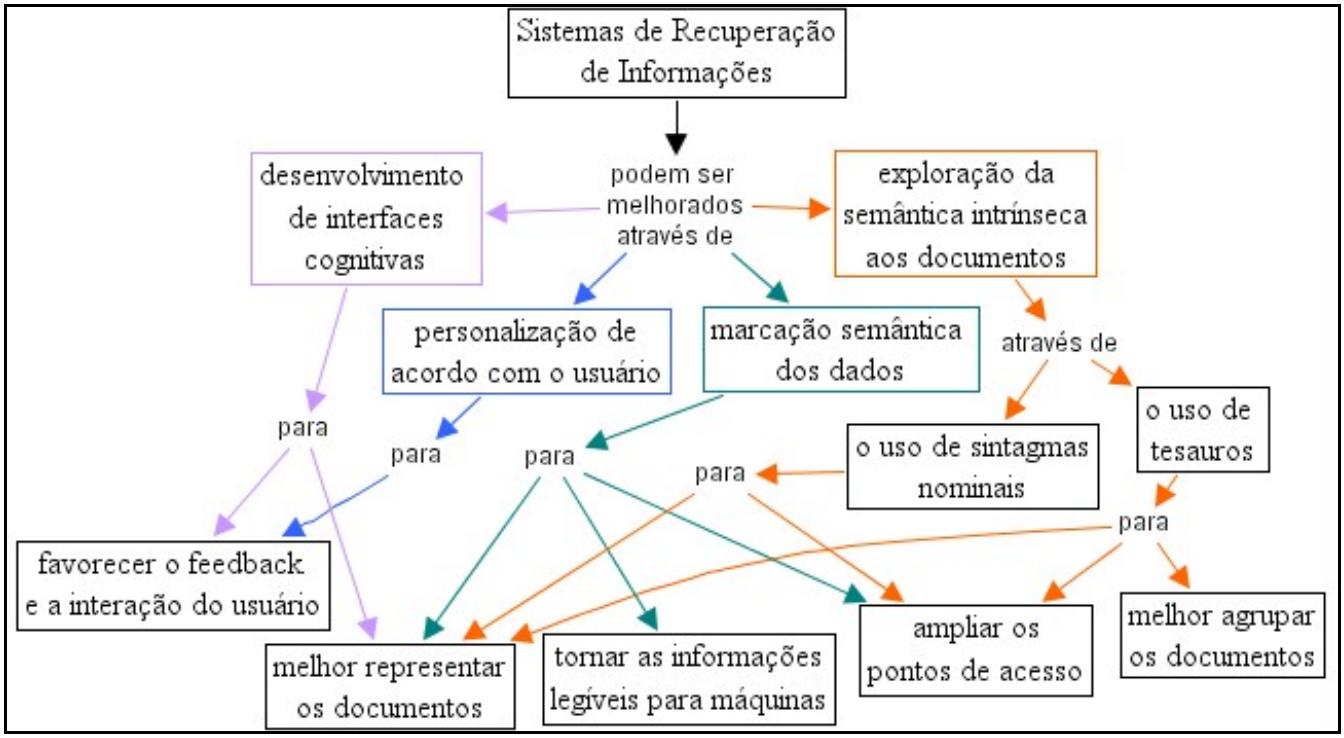


Figura 1 – Mapa Conceitual representando estratégias alternativas para melhoria dos SRIs.

Uma abordagem completa para a organização e a recuperação de informações, visando à melhoria dos SRIs atuais, poderia reunir essas estratégias e soluções, buscando:

- possibilitar a indexação dos documentos utilizando representações mais significativas, de modo a aumentar e melhorar os pontos de acesso e a relevância das informações recuperadas;
- prover forma adequada de apresentar as informações recuperadas aos usuários, de maneira que sejam intuitivas e facilmente compreensíveis;
- utilizar no processo de indexação padrões universais de registros de metadados para que os vários sistemas de informação sejam interoperáveis;
- adaptar-se continuamente aos usuários, sendo preferível que possam aprender com a forma com que trabalham, de modo que as buscas sejam continuamente refinadas através de trabalho de personalização.

Existem hoje diversas tentativas, mais ou menos coordenadas, de se abordarem estas ações fundamentais, mas a real integração demandaria estudos concomitantes em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a lingüística, a ciência da computação, a psicologia cognitiva, a comunicação, a sociologia, a antropologia, entre outras.

De maneira isolada, há pesquisas que incluem desde o estudo de interfaces gráficas, que procuram estabelecer uma plataforma de utilização mais intuitiva por parte dos usuários de SRIs (LAMPING et al, 1995; CAÑAS et al, 1999), até as tecnologias que vêm sendo exploradas no contexto da *web* semântica[5] (BERNERS-LEE et al, 2001; SEMANTICWEB.ORG, 2003), com vistas ao projeto e à implementação de padrões de metadados, que adicionem aos dados informações significativas sobre seus contextos, marcando-os semanticamente; e mecanismos de busca que levem em conta estes dados marcados. Ainda no âmbito da *web* semântica, há pesquisas e desenvolvimento de programas de computador comumente chamados *agentes inteligentes*, que têm a possibilidade de fazer a colheita (ou *harvesting*) de informações em outros computadores, agentes e dispositivos eletrônicos, para então tomar decisões baseadas em heurísticas embutidas. Esses agentes também executariam tarefas comuns de usuários de forma

[5] "*Web* semântica" é o nome genérico do projeto capitaneado pelo World Wide Web Consortium que pretende embutir inteligência e contexto nos códigos XML utilizados para confecção de páginas web, de modo a melhorar a forma com que programas podem interagir com estas páginas e também possibilitar o uso mais intuitivo pelos usuários. Esse tópico será explorado adiante na presente tese.

automática e personalizada, por meio da construção de perfis personalizados (HERMANS, 1996; NWANA, 1996; WOOLDRIDGE & JENNINGS, 1995 e 1998).

Pouco explorada, entretanto, é a utilização da semântica embutida nos próprios documentos, ou seja, as potencialidades intratextuais da linguagem natural, para automatizar e melhorar as tarefas de indexação, organização e recuperação de informações. Os SRIs usualmente utilizam como descritores[6] e unidades de recuperação as palavras isoladas que, embora sirvam de forma bastante razoável aos propósitos de recuperação de informações, falham em grande parte justamente por não considerarem o contexto informacional implícito em toda a consulta (LAWRENCE, 2000; RAGHAVAN et al, 1999), porque não estão preparados para lidar com a forma com que estas palavras ou conceitos estão relacionados. Esses relacionamentos, na prática, determinam as minúcias e especificidades dos assuntos pesquisados. Dessa forma, perdem-se informações fundamentais sobre o escopo em que as palavras estejam sendo utilizadas e, em conseqüência, a pertinência da pesquisa diminui. Tais problemas estão relacionados a questões lingüísticas como polissemia[7] e sinonímia[8] que são constantes em bases de dados textuais, e que tornam inviáveis as abordagens clássicas de recuperação de informação (RAGHAVAN et al, 1999). O problema é agravado por uma acentuada dificuldade dos usuários médios em traduzir suas necessidades de informação em termos significativos isolados – usualmente palavras-chave – utilizados para as buscas em sistemas de recuperação de informações.

Pesquisas nessa área incluem o uso de estruturas profundas da linguagem natural, como os sintagmas verbais e nominais, para indexação e recuperação (KURAMOTO, 1996 e 1999; MOREIRO et al, 2003); e de ferramentas de representação de relacionamentos semânticos e conceituais, como os tesauros, para ampliar a gama de informações recuperadas e aferição de contextos (SPARCK JONES & WILLETT, 1997, p. 15-20), além de outras estratégias derivadas da lingüística e da ciência da informação. Todas essas estratégias são fortemente atreladas ao idioma, o que faz com que os

[6] Descritores são considerados, no escopo deste trabalho, como termos de indexação relativos a um documento, usualmente palavras ou conjuntos de palavras que representem conceitos relacionados aos assuntos principais desses documentos.

[7] Qualidade de uma única palavra ter diferentes significações.

[8] Figura pela qual se exprime a mesma coisa ou se repete a mesma idéia por palavras sinônimas.

possíveis resultados da pesquisa tenham aplicação circunscrita ao contexto lingüístico da comunidade em questão. As metodologias, entretanto, são generalizáveis e sua aplicabilidade a outras linguagens é perfeitamente possível.

1.1 – Delimitação do problema

Nesta tese, embasado na epistemologia da ciência da informação (na sub-área de organização de informação), com aportes da lingüística e das tecnologias oriundas da ciência da computação, foi investigado o potencial de uso dos sintagmas nominais em processos de indexação automática. Partiu-se do pressuposto de que os sintagmas nominais, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave[9] usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos, tais como os observados nos mecanismos de busca da Internet, ou em sistemas de leitura das palavras-chave fornecidas pelo autor dos documentos.

A problemática da representação dos documentos e sua posterior recuperação é percebida diariamente através da vivência do autor da presente tese em ambientes onde se experimentam as dificuldades e se observam os problemas relacionados à recuperação de informações em grandes bases eletrônicas, e inspira-se também no embasamento teórico e metodológico advindo do campo da ciência da informação.

Alguns trabalhos, entretanto, se apresentam como marcos a partir dos quais se pretendeu avançar. Dentre eles, a pesquisa sobre a viabilidade do uso dos sintagmas nominais para sistemas de recuperação de informações de KURAMOTO (1996 e 1999) e as ferramentas para marcação sintática de frases da língua portuguesa e automatização da extração de unidades sintáticas, como os sintagmas nominais, desenvolvidas no âmbito dos projetos da Southern Denmark University (BICK, 2000) e da Unisinos (VIEIRA, 2000; VIEIRA e QUARESMA, 2001).

A partir dessas pesquisas e das ferramentas produzidas, pretendeu-se apresentar e validar uma metodologia para a indexação de documentos digitalizados de texto completo através da extração e seleção dos sintagmas nominais representativos. Como

[9] Costuma-se denominar “palavra-chave” qualquer palavra ou conjunto de palavras utilizado como termos de indexação.

subproduto, também foi sugerida uma metodologia para atualização semi-automática de tesauros monolíngües.

1.2 – Objetivos e pressupostos

De forma explícita, os objetivos desta pesquisa são os seguintes:

Objetivo geral

- **Desenvolver uma metodologia para a escolha automática de descritores para documentos textuais digitalizados, em língua portuguesa, utilizando as estruturas lingüísticas conhecidas como sintagmas nominais.**

Objetivos específicos

- **Testar a eficácia relativa de um conjunto de ferramentas para a extração automática de sintagmas nominais, comparando a extração automática com a extração manual;**
- **Analisar a possibilidade de a metodologia proposta ser utilizada para o auxílio na atualização de tesauros de língua portuguesa;**

Os principais pressupostos desta pesquisa foram:

1. a utilização de sintagmas nominais como descritores em processo de indexação automática apresenta vantagens em relação ao uso de palavras-chave, devido ao fato de esses possuírem, em comparação, maior densidade informacional, e serem mais bem relacionados ao contexto semântico do documento (como exposto na seção 2.3.2). Esse pressuposto é posto à prova na subseção 6.3.1;
2. a extração automática de sintagmas nominais é possível através do uso de ferramentas de *software* (apresentadas na seção 4.3), com desempenho qualitativamente comparável ao processo de extração manual. Este pressuposto é posto à prova na seção 5.1, e extensivamente ao longo do capítulo 5;
3. é possível estabelecer processo automatizado e eficaz para a escolha de descritores significativos para textos digitalizados, utilizando sintagmas nominais. Esse pressuposto, central em relação ao trabalho, foi discutido ao longo do capítulo 5.

A estrutura da tese é a seguinte:

Nesta ***Introdução*** foram apresentados o contexto social da pesquisa, a gênese do problema, os objetivos, os pressupostos e a forma em que foram encadeadas as temáticas a serem tratadas ao longo do trabalho.

No segundo capítulo, ***Fundamentos conceituais***, são discutidos os conceitos que compreenderam o construto teórico deste trabalho, como os fundamentos lingüísticos necessários e a teoria advinda dos sistemas de recuperação de informações.

No terceiro capítulo, ***Contextos de aplicabilidade***, são apresentados os ambientes tecnológicos e informacionais que justificaram – e por vezes possibilitaram – o desenvolvimento das metodologias, como as tecnologias da *web* semântica e as bibliotecas digitais.

No quarto capítulo, ***Metodologia e ferramentas***, são apresentados em detalhes a metodologia prospectiva e os recursos computacionais utilizados em todos os testes empíricos. Também são tecidas considerações sobre os *corpora* utilizados.

No quinto capítulo, ***Resultados da aplicação da metodologia prospectiva***, são analisadas comparativamente a extração manual e automática de sintagmas nominais, com o uso de *corpus* do qual os sintagmas nominais foram extraídos, tanto manualmente quanto automaticamente. Em seguida, são apresentados os resultados da aplicação da metodologia prospectiva proposta para a extração automática de descritores num *corpus* inicial reduzido – com o objetivo de testar e refinar a metodologia.

No sexto capítulo, ***A metodologia consolidada***, o conhecimento apreendido com a aplicação da metodologia prospectiva foi utilizado para o desenho da metodologia consolidada. Em seguida, essa metodologia consolidada foi explicitada e aplicada à totalidade do *corpus*. Por fim, a metodologia foi analisada à luz dos resultados atingidos.

No sétimo e último capítulo, ***Considerações finais e possibilidades em aberto,*** são tecidas considerações sobre os resultados empíricos obtidos a partir da metodologia consolidada à luz dos fundamentos teóricos e conceituais; e são comentadas as possíveis dificuldades que podem ser encontradas para a viabilização da metodologia em ferramentas computacionais. Finalmente, são apresentadas especulações teóricas sobre

os campos que podem ser ainda explorados a partir das metodologias e dos resultados apresentados.

2 FUNDAMENTOS CONCEITUAIS

Neste capítulo, são apresentados os marcos teóricos necessários para o completo entendimento da proposta desta tese.

Na primeira seção, apresentam-se os fundamentos lingüísticos necessários para o entendimento da estrutura do sintagma nominal, com algumas considerações sobre a limitação do modelo sintagmático e a adequação desse para metodologias automatizadas.

Na segunda seção, são apresentados os princípios de funcionamento dos sistemas de recuperação de informações, que fundamentaram teleologicamente a presente pesquisa, com ênfase nos processos de representação, armazenamento e recuperação de informações. Cabe ressaltar que a metodologia desenvolvida nesta tese não compreende projeto completo de sistema de recuperação de informação, mas sim um de seus subprocessos, a saber, a indexação, que é uma forma de representação.

Nas duas seções subseqüentes são apresentados os relacionamentos entre os sistemas de recuperação de informações e os sintagmas nominais; e entre os sistemas de recuperação de informações e os tesauros, complementando o ferramental epistemológico necessário para a compreensão da metodologia utilizada.

2.1 – Fundamentos lingüísticos

As digressões e definições a seguir são concernentes ao crescente campo, de certa forma relacionado à ciência da informação, que é o processamento automatizado da linguagem natural; e foram preciosos auxiliares para que tivéssemos a necessária contextualização do objeto de pesquisa e completo entendimento de alguns caminhos evolutivos dos sistemas de recuperação de informações, quando estes adotam estratégias baseadas na sintaxe e na semântica dos documentos textuais armazenados. Tendo sido os sintagmas nominais objetos de estudo da presente tese, foi necessário debruçar-se nas especificidades dos modelos de gramáticas gerativas de forma geral e de suas especificidades para a língua portuguesa.

2.1.1 – Algumas palavras sobre a linguagem

Dentre os vários conceitos expressos pela palavra **linguagem**, adotamos as definições que aproximavam seu significado do termo “língua”, ou seja “o conjunto das palavras e expressões usadas por um povo, por uma nação, e o conjunto de regras da sua gramática; idioma” (FERREIRA, 1999). A linguagem é, segundo PERINI (1985, p. 15), “o mais importante, o mais onipresente dos fenômenos sociais, e um pré-requisito para a existência das sociedades humanas”. Poderíamos acrescentar o fato de que a linguagem é a grande mediadora das relações humanas; o instrumento mais evidente em nossa interação social. VYGOTSKY defende o papel do aprendizado da linguagem para o desenvolvimento da cognição, quando postula que, inicialmente, a linguagem teria como objetivo permitir a comunicação interpessoal, mas também possibilitaria o aflorar do diálogo interno, que se torna a base da abstração reflexiva, que é etapa fundamental no desenvolvimento da inteligência (1987, p. 38-44 e 127-132). Defende, portanto, que pensamento e linguagem estão intimamente entrelaçados na constituição da inteligência de cada indivíduo. De maneira semelhante, CHOMSKY, em seus trabalhos mais recentes, postula que os princípios subjacentes às estruturas das linguagens são de tal modo específicos e tão altamente articulados que deveriam ser vistos como biologicamente determinados e geneticamente transmitidos (apud LYONS, 1983). Ainda é de interesse notar a importância que atribui WITTGENSTEIN à linguagem em suas investigações filosóficas (1967), e a forma com que DAHLBERG explicita o papel da linguagem na formação dos conceitos (1978).

Sendo a forma natural de mediação das relações humanas e o veículo mais evidente de suas idéias, é de se esperar que as parcelas majoritárias dos registros de informação concernentes às atividades humanas estejam codificadas em forma de textos, em alguma linguagem natural específica. O fato de as linguagens naturais serem instrumentos tão evidentes para o intercâmbio cognitivo entre seres humanos, e devido à escassez de abordagens na área da ciência da informação que tratem do assunto, nos faz acreditar que o desenvolvimento e intensificação das pesquisas visando à recuperação de informações através da análise e do processamento dos aspectos profundos e semânticos da linguagem natural possa proporcionar grandes saltos qualitativos na concepção de sistemas de recuperação de informações.

2.1.2 – A lingüística e as gramáticas

Dentre as várias possíveis definições para **gramática**, adotamos o “estudo da morfologia e da sintaxe de uma língua”, e a **gramática gerativa** como a “teoria lingüística que procura estabelecer, com base em princípios universais, um modelo geral de gramática, do qual derivariam as gramáticas de cada língua em particular” (FERREIRA, 1999). Ou mesmo a definição de HOUAISS “descrição de uma língua que usa regras formalizadas, constituindo um conjunto de instruções inteiramente explícitas e de aplicação mecânica, e que são capazes de gerar todas as frases gramaticais de uma língua e nenhuma agramatical” (2001). O estudo da linguagem pertence ao campo da lingüística, e o produto do trabalho do lingüista é a gramática; que nunca deve ser prescritiva ou normativa, mas antes deve almejar explicitar os mecanismos de uma linguagem específica. Usando a terminologia de CHOMSKY, podemos dizer que a gramática “gera” – definindo como gramaticalmente válidas – todas as possíveis sentenças no escopo de determinada língua (idioma) em particular (1969).

A descrição da linguagem compõe-se essencialmente de três elementos: a descrição **formal**, a descrição **semântica** e o sistema que relaciona o plano semântico com o formal. A descrição formal compreende os elementos **fonológicos** (relativos à pronúncia), **morfológicos** (relativos à forma, a composição em morfemas e possibilidades de variação) e os **sintáticos** (a forma como os elementos se inserem nas orações, e suas funções sintáticas); enquanto a descrição **semântica** (relativa ao sentido, ao significado) se relaciona a todos os elementos anteriores através das regras de interpretação semântica. As regras fonológicas, morfológicas e sintáticas definem as construções possíveis na língua, enquanto as construções semânticas relacionam as construções e seus significados (PERINI, 1985 e 1995).

Somente com o correto entendimento desses âmbitos da gramática, pudemos analisar o objeto da presente pesquisa em profundidade, com exceção feita à análise fonológica, uma vez que a pesquisa e a metodologia adotada tiveram como objeto empírico informação registrada em textos. Leva-se em conta que grande parte dos sistemas automatizados de recuperação de informações englobam processos de indexação automática, e estes processos se valem amiúde de reduções morfológicas para as operações de indexação. A utilização de sintagmas nominais nesta tese

demandou processamento que abarcasse principalmente os aspectos sintáticos, mas, ainda assim, um completo entendimento das estruturas sintagmáticas não prescindiu da atenção aos aspectos semânticos da linguagem.

Os modelos e definições apresentados nesta seção são aplicáveis à grande maioria das linguagens naturais, mas a metodologia e as ferramentas apresentadas neste trabalho foram desenvolvidas especificamente para o uso com a língua portuguesa, ainda que possam ser adaptadas para outros idiomas.

2.1.3 – Aspectos morfológicos

Embora as menores unidades da sentença sejam as palavras, a menor unidade sintática é o **morfema**. Como exemplo, temos a palavra *redistribuição.* Sabemos que existem aí diversos elementos gramaticais identificáveis pela forma e sentido: o prefixo *re* (que designa uma ação ou fenômeno repetido), a raiz "*distribui*" e o sufixo "*ção*" (que forma substantivos abstratos). Esse mesmo prefixo aparece em diversas palavras, como "*refazer*" e "*reaparecer*"; a raiz está no verbo "*distribuir*" e em seus compostos e derivados, e o mesmo sufixo aparece em substantivos como "*introdução*", "*interpretação*", "*suplementação*", e muitos outros. Podemos então dizer que a palavra se divide em (pelo menos) três morfemas, "*re-*", "*-distribui-*" e "*-ção*" (PERINI, 1985, p. 51-52). Perini aponta a dificuldade de se definir precisamente o que seja morfema e a arbitrariedade em delimitar quais são os morfemas constituintes de uma palavra de forma categórica. Mesmo o morfema "*distribui*" pode ser pensado como dois morfemas, a saber, "*dis-*" (que aparece em "*dispersar*", "*distrair*", etc.) e "*-tribui*" (que aparece em "*atribuir*", "*contribuir*", etc.) fazendo com que a análise sintática tenha componente subjetivo.

Cada uma das sentenças de uma língua em particular é formada por uma cadeia de elementos léxicos (palavras e morfemas) em seqüência, sendo que estes elementos também formam unidades intermediárias hierarquicamente dispostas (PERINI, 1985, p. 16). O **léxico** é o conjunto de palavras e morfemas que fazem parte de uma linguagem. Costuma-se usar o termo **palavra** para designar formas individuais como *pedra* e também um conjunto de formas relacionadas, como "*pedra*" e "*pedras*". Apesar de "*pedra*" e "*pedras*" serem formas (palavras) diferentes, são agrupadas lexicamente de modo sistemático. Chamamos a essa unidade de agrupamento **lexema**. Lexema é o conjunto de palavras que diferem apenas quanto a morfemas flexionais. Como exemplo, o lexema

“*pedra*” agrupa as palavras “*pedra*” e “*pedras*” (PERINI, 1995, p. 345). Um dicionário, por exemplo, pode ser considerado como uma lista dos lexemas da língua, sendo menos abrangente do que um léxico.

As palavras são formadas por morfemas simples ou por processos de **flexão** (variação da forma) ou **derivação** de morfemas, sendo que os dois processos se diferem pela sistematicidade em que ocorrem. A flexão ocorre em plurais (*pedra*, *pedras*) e nas várias formas de um verbo, como *pegar*, *pego*, *pegando*. Todas as possíveis palavras geradas por flexões do mesmo morfema constituem um único lexema. A derivação ocorre em relações não generalizáveis, como em “*livro*” e “*livresco*”, ou como em “*fazer*” e ”*desfazer*”. Cada uma das palavras geradas por derivação é um item léxico, ou lexema, diferente (PERINI, 1995, p. 345).

Muitos sistemas de recuperação de informações efetuam operações de redução de palavras a morfemas, no processo denominado *steeming*[10], para eliminar diferenças morfológicas que não correspondam a diferenças semânticas significativas. Essa operação permite a geração de índices mais concisos e aumenta a revocação[11], pois multiplica os pontos de acesso a determinados documentos, na medida em que um único morfema pode estar associado a muitas palavras diferentes.

2.1.4 – Aspectos sintáticos

O termo **frase** é utilizado para designar uma unidade do discurso bastante difícil de definir. PERINI (1986, p. 61-62) adota a perspectiva simplista, que diz “a frase é delimitada por uma maiúscula no início e por certos sinais de pontuação (./?/!/...) no final”. Uma frase pode conter nenhuma, uma ou mais de uma oração, sendo que um conjunto de orações de uma frase é também uma oração. **Oração** é uma frase (ou parte de uma) que apresenta determinado tipo de estrutura interna, incluindo sempre um predicado e freqüentemente um sujeito. Tradicionalmente emprega-se também a designação **período** para o conjunto das orações que constituem uma frase, sendo que um período é sempre uma oração. Por outro lado, nem toda oração é um período, já que muitas orações não são coextensivas com a frase de que fazem parte. O estudo das orações é a análise

[10] O s*teeming*, ou “redução à raiz” é o processo através do qual se identificam raízes gramaticais comuns em palavras distintas.

[11] A revocação é definida em detalhes na seção 2.2 deste trabalho.

sintática, e o estudo das funções sintáticas é a análise do papel dos constituintes imediatos da oração.

Dentre as funções sintáticas, podemos destacar o sujeito e o predicado. O **predicado** possui um **núcleo do predicado**, que tem sua função desempenhada sempre por um verbo. Em alguns casos, o núcleo do predicado equivale ao predicado, e em outros, há um **complemento do predicado**. O **sujeito** é o termo da oração que está em relação de concordância com o núcleo do predicado[12] (PERINI, 1995, p. 71-90). Também são elementos opcionais da oração o **objeto direto**, o **predicativo**, o **atributo**, a **negação verbal**, o **adjunto adverbial**, o **adjunto oracional**, o **adjunto circunstancial** e o **vocativo**, cujo estudo mais profundo extrapolaria o objetivo desta explanação.

2.1.5 – Alguns modelos sintáticos da gramática gerativa

Após as considerações anteriores sobre as estruturas morfológicas e as funções sintáticas da linguagem, cabe apresentar um pouco mais dos vários modelos de sistemas de gramáticas existentes, através do diálogo com trabalhos de renomados pesquisadores. Como alguns assuntos apenas tangenciam o objeto nuclear desta pesquisa, esses serão apresentados a partir de compilações realizadas por estudiosos de lingüística.

Como foi dito, uma das tarefas dos lingüistas na construção de modelos de gramáticas é o estudo das possíveis frases – frases bem construídas – que compõem uma linguagem. Os resultados do estudo são modelos de gramáticas gerativas. CHOMSKY propõe que as gramáticas sejam avaliadas pela sua capacidade gerativa fraca – conjunto das linguagens, como conjunto de frases, que a gramática consegue engendrar – e pela capacidade gerativa forte, ou seja, o conjunto de descrições estruturais que podem ser enumeradas pelo mesmo tipo de gramática. A capacidade gerativa forte, desejável para uma gramática “robusta”, engloba a fraca, que é utilizada como condição necessária, porém não suficiente (apud RUWET, 1975, p. 123-29).

Ao buscar um modelo sintático para o estudo das orações, os lingüistas assumem que estas sejam compostas de seqüências finitas de morfemas, sendo esses compostos por diferentes fonemas. A partir desses pressupostos, procuraram estabelecer modelos

[12] A exposição está bastante simplificada, contendo somente o que se considera necessário para o correto entendimento do que sejam os sintagmas nominais. Um estudo aprofundado pode ser encontrado em LIBERATO (1997)

gramaticais simples capazes de engendrar as frases. Os três modelos mais conhecidos são o **modelo dos estados finitos**, o **modelo sintagmático** (ou de estrutura de frase) e os **modelos transformacionais** (CHOMSKY, 1969; RUWET, 1975, p. 83-86; LYONS, 1983, p. 46-81).

No modelo dos estados finitos, defendido por Martinet, Jacobson e Hjelmslev, que teve eco na lingüística estrutural de Sausurre, considera-se que certos morfemas podem assumir lugares e posições específicos numa oração, e estes sejam relacionados entre si apenas do ponto de vista da sucessividade, da ordem linear, dando origem à noção de *relações sintagmáticas* (RUWET, 1975, p. 86-90). Para cada posição em uma frase é possível a escolha de um número finito de morfemas, o que permite a definição de *classes de morfemas* que podem pertencer à mesma posição. Pode-se representar esse modelo como uma máquina de calcular bastante banal, que passa por um número finito de estados, com um estado inicial e um estado final, e a cada estado é gerado um morfema. Uma máquina desse tipo define uma linguagem, a saber, como o conjunto de seqüências de morfemas que podem ser emitidos, e as linguagens produzidas são *linguagens a estados finitos* (RUWET, 1975, p. 86-90). Entretanto, CHOMSKY demonstra que esse modelo – o primeiro que estudou – não atende nem mesmo à capacidade gerativa fraca, ou seja, não possibilita a construção de gramáticas de estados finitos para representar todas as possíveis orações de linguagens naturais (CHOMSKY, 1956, p. 115, 1957a, p. 21-22 *apud* RUWET, 1975, p. 89-92; LYONS, 1983, p. 52-53).

Um modelo mais tradicional e, no entanto, mais poderoso que o modelo de estados finitos é o modelo sintagmático, que procura representar as frases através de uma estrutura hierarquizada de constituintes imediatos. Esse modelo, o segundo estudado por CHOMSKY, é explorado em detalhes mais adiante, e embasou a metodologia desta tese.

Ressalta-se que apesar do modelo da gramática sintagmática ser elegante, relativamente simples, e válido para o estudo da grande maioria das frases bem formadas em determinada linguagem natural, ele apresenta deficiências para algumas linguagens, em casos específicos, não possuindo nem mesmo a capacidade gerativa fraca. Além disso, em grande variedade de casos de ambigüidade sintática, as gramáticas sintagmáticas apresentam problemas para descrever corretamente as estruturas sintagmáticas das orações (RUWET, 1975, p. 120-147; LYONS, 1983, p. 60-63). Essas

limitações, entretanto, não são suficientes para descartarmos seu uso na concepção de sistemas de recuperação de informações, uma vez que os analisadores sintáticos automatizados (*parsers*) que se baseiam em modelos de gramáticas sintagmáticas podem ser altamente robustos (BICK, 1996). Além disso, as frases para as quais o modelo sintagmático apresenta falhas, por serem construções mais rebuscadas e conseqüentemente infreqüentes, apresentam incidência bastante baixa em textos científicos, sendo muito mais importante, para fins de automatização, a robustez do *parser* na identificação dos sintagmas. A estrutura do *parser* será comentada no capítulo 4 e sua eficácia foi posta a prova durante a manipulação dos dados empíricos.

Na busca por modelos de gramáticas mais abrangentes, e dados os problemas da análise sintagmática na sua incapacidade de explicar frases com constituintes descontínuos (separados por morfemas) em uma oração, CHOMSKY propôs novos modelos de gramáticas que pudessem lidar com estes aspectos das linguagens: os chamados modelos transformacionais, ou mesmo gerativo-transformacionais. A gramática transformacional é uma "gramática gerativa que inclui também o conceito de transformação, ou seja, a aplicação de um conjunto de regras que convertem uma *estrutura profunda* de uma língua em estrutura superficial" (HOUAISS, 2001). A **estrutura profunda** é a "representação da frase em nível abstrato, na qual se estabelecem as relações semânticas básicas entre os itens lexicais, cuja ordem linear pode ser modificada com a aplicação das transformações que forem necessárias para derivar a estrutura superficial, mantendo as relações semânticas iniciais na estrutura subjacente" e a **estrutura superficial** é a "organização sintática da frase tal como esta efetivamente se apresenta, e resulta das *transformações* realizadas a partir da estrutura profunda" (HOUAISS, 2001).

Esses modelos pressupõem concepção mais abstrata das estruturas das frases, e utilizam o modelo sintagmático para realizar uma espécie de pré-processamento das orações, cujo resultado tem relação apenas indireta com a ordem com que ocorrem os elementos na forma final das frases. Essa forma final é obtida através de regras conhecidas como *transformações*, e as transformações também tratam, no âmbito da análise sintática, da questão das conjugações que os verbos assumem nas orações. A grande vantagem dos modelos transformacionais é a capacidade de associar como semanticamente equivalentes frases com sintaxe distinta, evidenciando as possíveis

transformações que as associam. Também possibilita a identificação de ambigüidades semânticas em algumas frases (CHOMSKY, 1968; RUWET, 1975, p. 155-212 e 223-279; LYONS, 1983, p. 64-81).

O estudo detalhado dos modelos transformacionais foge ao escopo desta tese, mas pode ser necessário para o desenho de *parsers* mais robustos, no futuro, permitindo a construção de sistemas de recuperação de informações mais poderosos.

2.1.6 – Os sintagmas nominais

Como vimos, o estudo das orações é a análise sintática, e na concepção do modelo sintagmático baseamos-nos na noção de constituintes de uma oração para o estudo das hierarquias de componentes (RUWET, 1975, p. 99-119; LYONS, 1983, p. 54-63). Entendemos por **sintagmas** certos grupos de unidades que fazem parte de seqüências maiores, mas que mostram certo grau de coesão entre eles (PERINI, 1995). Segue-se o exemplo didático de PERINI (1986, p. 44-45):

- “A casa de Lulu é azul e branca”

Os interlocutores percebem que [A casa de Lulu] forma uma unidade, o que não se verifica com [Lulu é azul]. Dizemos então que o primeiro é um constituinte, enquanto o segundo não. As frases são formadas de constituintes, muitas vezes aninhados. A frase acima poderia então ser dividida nos seguintes constituintes:

- “A casa de Lulu é azul e branca”
- “A casa de Lulu”
- “casa de Lulu”
- “azul e branca”
- “é azul e branca”

Esta estruturação é freqüentemente representada através de um diagrama em árvore, também chamado de “indicador sintagmático”, exemplificado na FIG. 2:

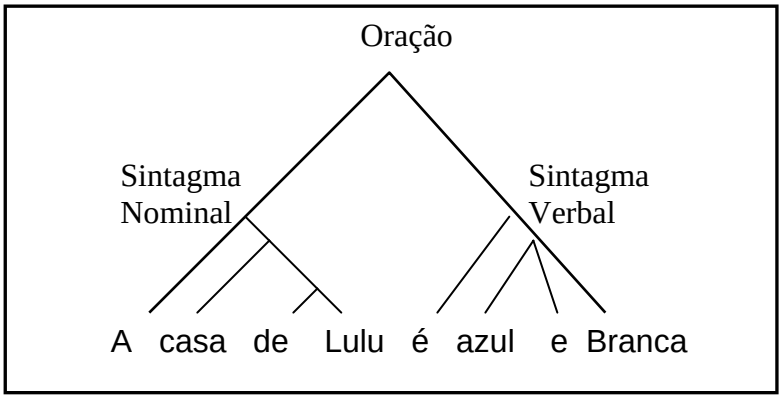


Figura 2 – Exemplo de indicador sintagmático.

Podemos perceber que existem divisões para as orações que são mais satisfatórias do que outras. Na frase acima, podemos concordar que a divisão mais natural seria:

- [“A casa de Lulu] [é azul e branca”]

E não:

- [“A casa de] [Lulu é azul] [e branca”]

Como vimos, os constituintes costumam receber uma “função” na análise tradicional: [*a casa de Lulu*] é sujeito, e [*é azul e branca*] é o predicado, sendo que [*é*] é o núcleo do predicado e [*azul e branca*] é predicativo do sujeito. Já a seqüência [*Lulu é azul*] não recebe função alguma, pois não é um constituinte. As subdivisões “naturais” das orações se denominam sintagmas, e o sintagma é uma unidade do ponto de vista semântico, pois possui significado único e coerente. Eles são classificados segundo as funções que podem ocupar. Se estiverem desempenhando funções típicas de substantivos (sujeito, objeto) são chamados de **sintagmas nominais** (SN), ao passo que se desempenham a função tradicionalmente chamada de “predicado”, são chamados de **sintagmas verbais** (SV) (PERINI, 1985, p. 43-44).

Perini (PERINI et al, 1996) define o SN como a classe gramatical com comportamento sintático de sujeito, de objeto direto e também – se precedido de preposição – de adjunto adnominal ou de objeto indireto. Segundo LIBERATO (1997), o SN é a parte do enunciado que representa conceitos ou referentes. Os referentes podem

ser entidades abstratas ou concretas; podem ser identificados por nomes próprios ou através do sintagma nominal descritivo; podem ter uso referencial, quando representam uma entidade; ou uso atributivo, representando um papel.

De acordo com PERINI (1985, p. 84-86 e 152-161), uma oração típica e bem formada pode seguir as estruturas:

1 - Oração = SN + SV, na qual o sintagma nominal é denominado “sujeito”;

Exemplo: “**O governo** vai mudar”;

2 - Oração = SN1+SV, com SV = (Verbo + SN2), na qual o sintagma nominal é denominado objeto.

Exemplo: “Fulano deixou **o cargo**”;

E ainda temos os sintagmas nominais preposicionados, com estrutura:

3 - Oração = SN^1+SV, com SV = (Verbo + preposição + SN^2).

Exemplo: “Jorginho levou trote n**a faculdade**”;

Para elucidarmos de maneira geral os casos freqüentes de ocorrência dos sintagmas nominais e verbais, podemos indicar a seguinte notação para as funções sintáticas:

O	oração
N	nome
V	verbo
Det	determinante (por exemplo, um artigo)
SN	sintagma nominal

Tabela 1 – Notação para as funções sintáticas

O sintagma verbal ou nominal pode aparecer de acordo com as seguintes estruturas (PERINI, 1985, p. 84-86 e 152-161):

SN = O	O sintagma nominal equivale à oração
SN = N	O sintagma nominal é um nome
SN = Det + N	O sintagma nominal é formado por um determinante mais um nome
SN = SN + O	Um novo sintagma nominal é formado com a junção de um sintagma nominal e uma oração
SV = V	O sintagma verbal é formado pelo verbo
SV = V + SN	O sintagma verbal é formado pelo verbo mais um sintagma nominal

Tabela 2 – Estruturas sintagmáticas possíveis

Analisando as estruturas citadas, pode-se notar que os sintagmas nominais podem aparecer recursivamente na oração, aninhados em outros sintagmas nominais, integrando sintagmas verbais ou mesmo ligados através de preposição – os chamados sintagmas nominais preposicionados. Embora as estruturas sintáticas sempre redundem para as estruturas básicas demonstradas acima, essas estruturas simples escondem uma infinidade de possibilidades. KURAMOTO (1999) apresenta no Anexo C de sua tese de doutorado uma taxonomia de estruturas verificadas para os SNs muito mais detalhada (323 estruturas diferentes), ao analisar seu *corpus* de 15 documentos. Não é objetivo deste projeto reproduzi-las.

Vejamos abaixo o indicador sintagmático do exemplo de RUWET (1976) com a frase “O homem recebe o livro do menino”:

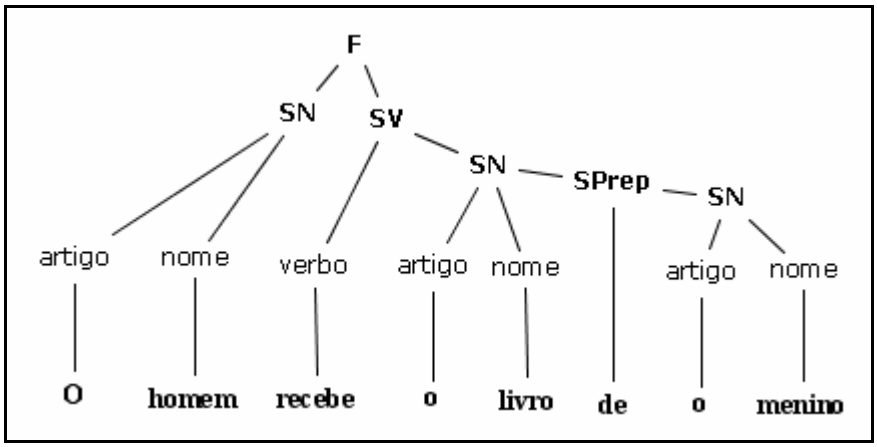


Figura 3 – Exemplo de divisão sintagmática.

Um outro exemplo que mostra o potencial de estruturação dos sintagmas nominais por meio de relações de encadeamento é o seguinte (KURAMOTO, 1995):

SN: “As características do ambiente do mundo dos negócios”

Esse sintagma nominal engloba os seguintes:

SN1: “Os negócios”.

SN2: “O mundo dos negócios”.

SN3: “O ambiente do mundo dos negócios”.

SN4: “As características do ambiente do mundo dos negócios”.

Através dessas relações de encadeamento, podemos classificar o nível dos sintagmas nominais pela quantidade de outros sintagmas que esses englobam, sendo que, no exemplo citado, o sintagma nominal original é de nível 4.

2.1.7 – Funções sintáticas no SN

Segundo PERINI (1995, p. 92-123), a análise da estrutura interna do sintagma nominal é muito mais complexa do que a análise sintática tradicional de orações, que procura dividi-las em seus constituintes imediatos, como sujeito, predicado, adjuntos adnominais, etc. Nos parágrafos a seguir, apresentamos alguns aspectos da análise proposta por Perini, sem, entretanto, esmiuçar suas especificidades e justificativas.

A gramática tradicional distingue, no sintagma, duas funções, a saber, *núcleo*, e os *adjuntos adnominais*. Vejamos o exemplo de Perini:

- “Aqueles seus livros de psicologia”

Nesse sintagma nominal, podemos distinguir o núcleo [*livros*] e os adjuntos adnominais [*aqueles seus*] e [*de psicologia*]. Na análise de Perini, distinguem-se as diferentes funções que podem assumir os adjuntos adnominais. Para permitir a análise, divide-se o sintagma nominal entre área à esquerda e área à direita do núcleo.

A área à esquerda, ou seja, dos elementos que precedem o núcleo, compreende seis posições fixas e quatro posições variáveis, para os elementos opcionais do sintagma nominal. As posições fixas definem seis funções, denominadas (na ordem em que podem ocorrer) **determinante**, **possessivo**, **reforço**, **quantificador**, **pré-núcleo externo** e **pré-núcleo interno**. As posições variáveis ocorrem nos intervalos entre as posições fixas, exceto entre os dois pré-núcleos, nos quais não pode ocorrer nenhum item; e têm sempre a mesma função de **numerador**. PERINI (1995, p. 99) aponta os principais itens lexicais que podem desempenhar cada uma das funções:

Função	**Exemplos de itens que podem desempenhá-la**
Determinante	o, este, esse, aquele, algum, nenhum, um
Possessivo	meu, seu, nosso, etc.
Reforço	mesmo, próprio, certo
Quantificador	poucos, vários, diversos, muitos, muitos, único, primeiro (segundo, terceiro, etc.)
Pré-núcleo externo	mero, pretenso, meio, suposto, reles, inesquecível, ilusório, simples, bom, velho, novo, etc.
Pré-núcleo interno	mau, novo, velho, claro, grande, bom
Numerador	outro, dois (três, quatro, etc.)

Tabela 3 – Funções desempenhadas pelos itens lexicais na estrutura do SN

O sintagma nominal dito **máximo** é de ocorrência muito pouco provável nas construções usuais das linguagens, mas pode ser usado para exemplificar cada um dos elementos, como no exemplo a seguir:

- “Aqueles meus mesmos dois únicos pretensos bons **amigos**”

Perini ainda aponta a existência de itens cuja análise de função é duvidosa, e itens que podem desempenhar mais de uma função.

Da mesma forma, podemos analisar a área à direita do núcleo, também chamada de área dos **modificadores**, embora Perini aponte que a pesquisa neste campo está bem menos avançada que a da área à esquerda. Segundo Perini, distinguem-se, na área à direita, três funções: **núcleo do sintagma nominal**, **modificador interno** e **modificador externo**. Tomemos o exemplo de PERINI (Ibidem):

“Um **ataque** cardíaco fulminante”

No exemplo acima, [Um] é determinante e faz parte da área à esquerda, [ataque] é o núcleo do sintagma nominal, [cardíaco] é o modificador interno e [fulminante] é o modificador externo. Apesar da estrutura à direita ser mais simples, a polivalência funcional dos itens dificulta sua análise, enquanto os itens da área à esquerda são mais especializados e facilmente identificáveis. Para finalizar, Perini ainda sugere a possibilidade da existência de sintagmas nominais sem núcleo, descartando em seguida a hipótese, pois considera que nesses casos alguns elementos da área à direita estariam desempenhando o papel de núcleo do sintagma nominal.

2.1.8 – Identificação e extração dos SNs

Assim como os morfemas, os constituintes ou sintagmas podem ou não ser facilmente identificáveis, sendo que por vezes é necessário recorrer a outros recursos para que seja feita a “demarcação” sintática. Esta característica dos sintagmas dá margem a uma série de posicionamentos, alguns dos quais favoráveis à possibilidade de extração automática dos sintagmas nominais, e outros mais céticos quanto a isso. PERINI acredita que a intuição “subjetiva, mas nem por isso duvidosa” que nos permite separar a oração em seus constituintes imediatos possa ser caracterizada através de critérios puramente formais (1985, p. 42-43), mas há quem defenda que a identificação dos constituintes é somente completa através de uma abordagem cognitiva e amplamente

contextual (LIBERATO, 1997), que só é esperada na análise do discurso[13] e na pragmática[14]; ou através de outros modelos gramaticais, como a análise transformacional (RUWET, 1975, p. 155-212 e 223-279). Para a análise semântica, há também o problema das situações anafóricas, que ocorrem quando a estrutura de uma oração se apresenta reduzida porque ocorre na vizinhança de outra estrutura oracional de certa forma paralela, dependendo dessa para sua total compreensão (PERINI, 1986, p. 57).

Todavia, existem soluções de compromisso para processos automatizados de extração de sintagmas nominais. De acordo com MIORELLI (2001), os sintagmas nominais podem ser entendidos – e tratados – de forma sintática, privilegiando a forma; ou semântica, buscando os significados maiores, cada uma com suas especificidades e implicações. A abordagem semântico-pragmática, utilizada por LIBERATO (1997), não prescinde de um "interpretador de contextos", natural na cognição humana, mas dificilmente implementado em heurísticas de inteligência artificial. Liberato procura discutir em profundidade alguns aspectos isolados da estrutura do sintagma nominal, relacionando os enunciados das sentenças a seus significados, sem a preocupação de estabelecer a estrutura geral (MIORELLI, 2001). A forma sintática dos sintagmas nominais, como analisados por PERINI (1986, 1995 e 1996) está mais relacionada à estrutura das orações em si, e é mais facilmente tratada computacionalmente. Assim como no trabalho de MIORELLI (2001), segue essa abordagem KURAMOTO (1999) que, em sua tese de doutorado, procurou explicitar e analisar as freqüências de ocorrências de cada estrutura possível para os sintagmas nominais, ao projetar um sistema de recuperação de informações baseado nos mesmos. Ainda nesta mesma linha, e nas regras advindas de uma "gramática de restrições", baseiam-se as heurísticas de funcionamento do analisador sintático (*parser*) de BICK (1996), utilizado para extrair os sintagmas nominais dos *corpora* utilizados na presente tese. Talvez essa forma de modelar a estrutura dos sintagmas nominais seja utilizada em quaisquer abordagens, e com quaisquer ferramentas, que busquem a automatização de extração dos sintagmas nominais.

---

[13] Estuda a estrutura e a interpretação dos textos.

[14] Ocupa-se da relação dos enunciados lingüísticos com a situação extralingüística em que se inserem (PERINI, 1995).

No projeto de sistemas de recuperação de informações, em conjunto com a análise puramente sintática das sentenças, podemos agregar soluções adicionais para o tratamento semântico das estruturas lingüísticas, como os tesauros, como no caso desta tese, ou mesmo as ontologias e as bases de conhecimento. Dessa forma, são contempladas as situações que poderiam gerar possíveis ambigüidades semânticas e amplia-se o escopo de aplicabilidade das soluções.

Acredita-se que esta análise, longe de ser exaustiva, apresente os elementos mínimos necessários para o correto entendimento dos aspectos lingüísticos das metodologias utilizadas ao longo desta tese. Alguns aspectos suplementares do uso de sintagmas nominais são apresentados adiante, quando os contextualizarmos como possibilidades na construção de sistemas de recuperação de informações.

2.2 – Sistemas de recuperação de informações

Desde que os grupos humanos abandonaram o nomadismo e se estabeleceram em comunidades em locais geográficos fixos ao longo de grandes períodos, vêm-se apoiando em alguma forma de comunicação supra-oral para registrar e, com isso, decifrar e disseminar as regularidades percebidas no ambiente. As metodologias e tecnologias associadas às ciências da informação surgiram como respostas às necessidades causadas pelo papel cambiante que tomaram esses registros do conhecimento humano através dos tempos (WERSIG, 1993). Com o advento da imprensa de tipos móveis de Gutenberg e, posteriormente, com o aumento das coleções e acervos de livros e documentos, surgiram diversas técnicas e metodologias para o arranjo mecânico destes documentos em disposições que facilitassem a recuperação sistemática de suas informações para uso posterior.

Com o fenômeno contemporâneo da crescente disponibilização de documentos em formato digital, vimos disseminar o uso dos sistemas – mecanizados, ou mais propriamente, informatizados – de recuperação de informações (SRIs), para lidar com os crescentes volumes de documentos, em diferentes formatos, em meios digitais, ou mesmo para administrar e facilitar o acesso aos documentos em formatos tradicionais.

Para podermos discutir as metodologias que foram utilizadas nesta pesquisa, faz-se necessário entendimento aprofundado dos conceitos pertinentes aos sistemas supracitados, o que fazemos em seguida.

2.2.1 – Conceituação de SRI

A dificuldade de conceituação do que seja um sistema de recuperação de informações advem, a princípio, da ambigüidade dos conceitos de sistema e de informação em si (ARAÚJO, 1995). No âmbito dos sistemas de recuperação de informações, costuma-se evidenciar o conceito de *informação como coisa*, ou seja, registros de conhecimentos em documentos (BUCKLAND, 1991), em detrimento de outras definições e contextos. Sem embargo, há, no contexto específico supracitado, extensa literatura especializada das áreas de ciência da informação e ciência da computação, na qual podemos encontrar uma dezena de definições razoavelmente consensuais, das quais pinçamos as apresentadas a seguir.

KORFHAGE (1997) ressalta o caráter pessoal da informação, e aponta o fato de que sistemas de recuperação de informações armazenam dados, distinguindo as informações que foram armazenadas por um usuário das que serão apropriadas por outro. Os SRIs seriam os intermediários nesse processo mediado de troca de informações. Para LANCASTER & WARNER (1993 p. 4-5), os SRIs são a interface entre uma coleção de recursos de informação, em meio impresso ou não, e uma população de usuários; e desempenham as seguintes tarefas: aquisição e armazenamento de documentos; organização e controle desses; e distribuição e disseminação aos usuários. Essa visão é abrangente, e inclui tarefas que são desempenhadas em conjunto com atores humanos. LANCASTER (1968) já havia anteriormente apontado o fato de que os SRIs não informam o usuário – no sentido de mudar seu conhecimento sobre objeto de sua questão –, mas apenas o informam sobre a possível existência de documentos atinentes à questão, além de características desses documentos; e procura, em outro trabalho, analisar os SRIs subdividindo-os em seis subsistemas: de documentos, de indexação, de vocabulário, de busca, de interface com o usuário e de *matching*[15] (LANCASTER, 1979). CHOWDHURY entende que o conceito de recuperação de informações – e como conseqüência, o conceito de sistemas de recuperação de informações – é auto-explanatório, e divide os SRIs em subsistemas de documentos, de usuários, e de busca/recuperação; detalhando cada um desses subsistemas (1999, p. 1-11). Para CHOWDHURY (Ibidem), os SRIs

[15] *Matching* pode ser definido nesse contexto como o casamento das necessidades de informação com os itens que fazem parte do acervo do sistema e que podem satisfazer esta necessidade.

servem de ponte entre o mundo dos criadores de informações e os usuários dessas, e para isso, colecionam-nas e as organizam. SALTON & MCGILL (1983, p. 1), e mais tarde BAEZA-YATES & RIBEIRO-NETO (1999, p. 1), definem SRIs como sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação.

Há que se notar que as definições procuram apreender um fenômeno atemporal – as necessidades de informação – e as várias metodologias e tecnologias que, através dos tempos, foram engendradas para atender a essas necessidades, desde as atividades de organização de coleções de documentos em acervos bibliográficos, até os modernos sistemas informatizados que lidam com documentos em formato digital. Partindo das definições citadas, assumimos que SRIs **organizam** e viabilizam o **acesso** aos itens de informação, desempenhando as atividades de:

- **Representação** das informações contidas nos documentos, usualmente através dos processos de **indexação** e **descrição** dos documentos;
- **Armazenamento** e gestão física e/ou lógica desses documentos e de suas representações;
- **Recuperação** das informações representadas e dos próprios documentos armazenados, de forma a satisfazer as **necessidades de informação** dos usuários. Para isso é necessário que haja uma **interface** na qual os usuários possam descrever suas necessidades e questões, e através da qual possam também examinar os documentos atinentes recuperados e/ou suas representações.

Sem que seja necessário o aprofundamento da discussão conceitual sobre as diferenças entre *dado* e *informação*, há que se distinguirem os **sistemas de recuperação de informações** (SRI) dos **sistemas de gestão de bancos de dados** (SGBD). Dados podem ser definidos como seqüências de símbolos para os quais são atribuídos significados; símbolos estes que podem ser codificados, interpretados e manipulados por programas de computador, e enviados através de redes e dispositivos de comunicação. O conceito de **informação** já carrega um grau maior de abstração. A informação não prescinde do sujeito que a depreenda a partir dos dados, no ato conhecido como interpretação. No sentido estrito do conceito, nenhum programa de computador lida, sob o

ponto de vista da máquina, com informações, a não ser que possua alguma capacidade de arrazoamento, e, assim mesmo, a utilização do termo dá margem a discussões. No uso corrente, porém, ambos os termos são utilizados para sistemas, apesar das diferenças entre os sistemas de recuperação de informações e sistemas de recuperação de dados, como os SGBDs. Essas diferenças, comentadas por KOBASHI (1994) podem ser sumarizadas através da TAB. 4:

	Recuperação de Dados	**Recuperação de Informações**
Modo de Inferência	dedutivo	indutivo
Modelo Lógico	determinístico	probabilístico
Linguagem de especificação das necessidades	formal (SQL e assemelhados)	natural (como um objetivo)
Necessidade especificada	completa	parcial
Casamento de necessidades e resultados	exato	melhor casamento possível
Objeto da busca	registros que satisfaçam à questão (*query*)	itens relevantes para o usuário

Tabela 4 – Diferenças entre a recuperação de dados e a recuperação de informação (adaptado de RIJSBERGEN, 1979).

Em **sistemas gerenciadores de bancos de dados**, os símbolos são armazenados em uma estrutura matricial em campos determinados, com metadados que lhes conferem certo sentido ontológico. Para recuperar dados específicos, basta especificar as restrições necessárias aos campos de pesquisa e codificá-las numa questão ou *query* (argumento de entrada no sistema) para que se tenha a resposta exata, fruto de busca completa e exaustiva.

A recuperação de informações traz dificuldades intrínsecas ao conceito de “informação”, como a dificuldade da determinação da real necessidade do usuário e do seu melhor atendimento com os documentos que fazem parte do acervo do sistema (FOSKETT, 1997, p. 5). A associação entre os registros e seus conteúdos informativos é

vaga, e isso pode acarretar problemas nas respostas a questões específicas, como baixas taxas de **revocação**[16] e **precisão**[17]. Um sistema de recuperação de informações deve buscar boa relação entre os índices de revocação e precisão, para oferecer, em resposta a determinada consulta, referências ao maior número possível de documentos atinentes, ordenados por critérios de relevância, e o menor número possível de documentos pouco ou não atinentes, de acordo com as necessidades de informação dos usuários.

Dentre os diversos diagramas que descrevem o processo de recuperação de informações em sistemas (CESARINO, 1980, p. 33; LANCASTER, 1993, p. 2), escolhemos o proposto por BAEZA-YATES & RIBEIRO-NETO (1999), apresentado na FIG. 4, que enfatiza o processo da forma em que é realizado nos sistemas automatizados:

[16] A Revocação, ou “recall” ou mesmo “abrangência”, é a razão do número de documentos atinentes recuperados sobre o total de documentos atinentes disponíveis na base de dados. A revocação mede o sucesso do SRI em recuperar documentos pertinentes

[17] Razão do número de documentos atinentes recuperados sobre o total de documentos recuperados. A precisão mede o sucesso do SRI em não recuperar documentos que não sejam relevantes de acordo com a necessidade de informação.

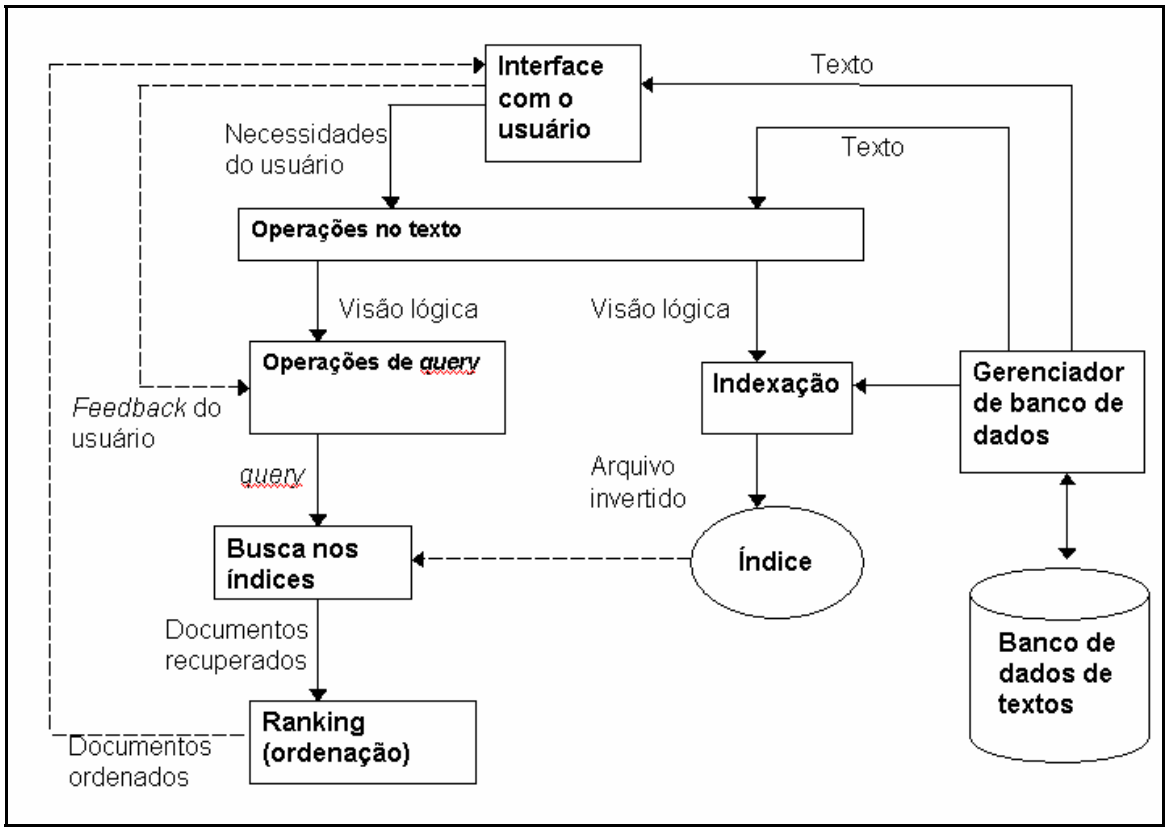


Figura 4 – O processo de recuperação de informações (adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 10)

A FIG. 4 explicita as atividades de representação (operações no texto, indexação e criação do índice); armazenamento e gestão (dos documentos presentes no acervo do banco de dados de textos e do índice), e a recuperação, que se inicia através da análise da necessidade do usuário e redunda na apresentação de um conjunto ordenado de documentos, possivelmente permitindo ao usuário *feedback* sobre os documentos apresentados. No exemplo mostrado acima, o índice é implementado através de um arquivo invertido, que é visto em detalhes adiante.

Vamos examinar em detalhes a seguir as atividades de representação, armazenamento e recuperação de informações em SRIs.

2.2.2 – Representação de documentos em SRIs

Ao procurar descrever os objetos do mundo, os autores de documentos primários (textos, imagens, sons e vídeo) realizam processos *ontológico*s de representação do que é conhecido. No processo de tratamento ou processamento dos registros de

conhecimento para fins de armazenagem nos sistemas de informação, é requerido novo estágio de representação, não mais de cunho ontológico, mas partindo do acervo de conhecimentos sobre essas coisas e seres, objetos da *epistemologia* (ALVARENGA, 2003, p. 5). Acrescenta-se que a coleta de informações descritivas, com vistas ao preenchimento de itens de catálogos, poderia ser considerada nova etapa ontológica, no âmbito da representação. Capturar as informações potencialmente registradas nos documentos e representá-las para permitir acesso posterior é o objetivo e o grande problema dos SRIs.

Segundo ROBREDO & CUNHA (1994, p. 201), o tratamento dos documentos inclui algum tipo de análise de seu conteúdo, o que permite separá-los e ordená-los por grupos ou classes mais ou menos afins, possibilitando sua localização (ou recuperação) posterior. Entretanto, quando o volume de documentos no acervo atinge certo patamar, ou quando os assuntos dos documentos adquirem certo grau de especificidade, não é mais possível ordená-los por meio de grandes classes de assuntos, pois essas classes não são mais suficientemente informativas para representar adequadamente o conteúdo dos documentos e discriminá-lo em relação a outros documentos. Torna-se necessário então utilizar processos de **catalogação** e de **indexação** eficazes, de forma que a recuperação das informações que contêm, de acordo com as necessidades dos usuários, seja a mais eficaz possível.

Há razoável consenso quanto ao fato de a catalogação constituir o processo de coleta de informações bibliográficas dos documentos. A catalogação é também chamada de **análise descritiva**, e enfoca as características objetivas inerentes ao próprio documento, como a autoria, data de publicação, entre outras. Entretanto, observamos várias abordagens, comportando diversas nomenclaturas, para o processo de indexação. Há certa concordância sobre o fato de a indexação ser um processo composto de duas fases razoavelmente independentes: a **análise de assunto**[18] (ou análise conceitual) e a **tradução**. Na análise de assunto, o conteúdo do documento é analisado com o propósito

[18] A análise de assunto também recebe os nomes de análise documentária, análise conceitual, análise temática, entre outros, com algumas pequenas diferenças em suas acepções.

de determinar sua atinência[19], ou seja, de que trata o documento. Na tradução, os assuntos pertinentes identificados são representados por meio de linguagem de indexação, que podem ser códigos de classificação, palavras-chave em um vocabulário controlado, símbolos, etc. A análise de assunto pode ser realizada por um indexador humano ou pode ser automatizada. (CESARINO, 1980; UNISIST, 1981; NAVES, 1996; FOSKETT, 1997, HUTCHINS, 1997).

KOBASHI (1994) aponta o fato de existirem muitas pesquisas voltadas para o processo de tradução, enquanto que poucos estudos procuram sistematizar metodologias para o processo de análise conceitual, que é tratado como puramente intelectual ou dependente do bom senso dos indexadores.

LANCASTER & WARNER (1993) defendem que o propósito principal da indexação é a elaboração de índices e resumos (*surrogate files)* para constituírem representações temáticas de documentos publicados em uma forma que se preste à sua inclusão em algum tipo de base de dados.

Em relação à caracterização do processo de indexação, podemos destacar as seguintes categorias de análise (adaptado de ROBREDO & CUNHA, 1994, p. 203-204):

1) Em relação ao nível de abrangência da análise conceitual:

- **categorização,** que é o reconhecimento dos aspectos dominantes, segundo alguma subdivisão por assuntos preexistente;
- **indexação superficial,** que permite obter os conceitos principais tratados no documento;
- **indexação profunda,** que consiste em obter todos os conceitos considerados fundamentais.

2) Em relação às partes do documento analisadas, a indexação pode-se fazer:

- com base no **título**;
- com base no **resumo**;

[19] Também chamada de concernência, temática, assunto, tema, *aboutness*; sendo todos esses nomes relacionados à determinação daquilo de que trata o documento, seus assuntos ou temáticas principais, sua mensagem no processo de comunicação.

- com base no **título** e no **resumo**;
- com base em partes determinadas (sumário, introdução, conclusão, etc.);
- com base no **documento completo**.

3) Em relação ao procedimento de indexação, é:

- indexação **manual**;
- indexação **automática**;
- indexação **mista**.

4) Em relação à linguagem de indexação utilizada, ocorre:

- indexação em **linguagem natural**
    - Linguagem natural **livre**, que utiliza as palavras extraídas do próprio documento;
    - Linguagem natural **controlada**;
        - **não estruturada**, como as listas de descritores padronizadas com eliminação de sinônimos e as listas de cabeçalhos de assuntos.
        - **Estruturada**, como os tesauros e as classificações facetadas.
- indexação em **linguagem artificial**, controlada e codificada.
    - **não estruturada**, que incluem alguns esquemas de categorização que utilizam símbolos não estruturados;
    - **estruturada**, que incluem os esquemas de classificação bibliográficos como o *Library of Congress Classification* (LCC) e a Classificação Decimal Universal (CDU).

Dois conceitos importantes a serem apresentados para a avaliação do processo de indexação são os de **exaustividade**[20] (em oposição à seletividade) e de

[20] A exaustividade, também chamada de “profundidade” da indexação, cresce à medida que aumenta o número de descritores utilizados na indexação, ou seja, o número de termos atribuídos ao documento de forma a procurar representar o assunto do mesmo.

especificidade[21]. O aumento da exaustividade na indexação costuma aumentar a revocação e diminuir a precisão na recuperação de documentos, enquanto o aumento na especificidade na indexação costuma aumentar a precisão e diminuir a revocação na recuperação de documentos.

O processo de indexação a ser utilizado costuma ser definido no momento em que os sistemas são projetados, e esse processo deve funcionar continuamente à medida que novos documentos são adicionados ao acervo do sistema. O processo de indexação escolhido interfere fortemente no sucesso da posterior recuperação das informações contidas nos documentos.

Embora as primeiras experiências de automação na indexação costumassem levar em conta apenas informações do título dos documentos, os modernos sistemas automatizados de recuperação de informações usualmente procuram realizar a **indexação profunda**, para obter todos os conceitos fundamentais para a representação do documento, e por isso, na maioria das vezes processa-se o **documento completo** para a escolha de descritores. O processo de indexação nesses sistemas é em grande parte ou totalmente **automático**, apesar de alguns sistemas de recuperação de informações utilizarem técnicas **mistas** (também chamadas de **híbridas**) de indexação automática e manual[22].

A linguagem de indexação é quase sempre **natural**, sendo que em muitos SRIs ela é **livre**, e utilizam-se como descritores as palavras do próprio documento após o processo descrito na FIG. 5 a seguir:

---

[21] A busca da maior especificidade é considerada como um princípio da indexação, que postula que um tópico de assunto no documento deve ser representado pelo termo mais específico que o descreva completamente, ao invés de termos genéricos.

[22] Como exemplo, podemos apontar alguns diretórios da *web*, como o mecanismo de busca *Yahoo* (http://www.yahoo.com), no qual há um misto de técnicas automatizadas e não automatizadas para a classificação dos documentos.

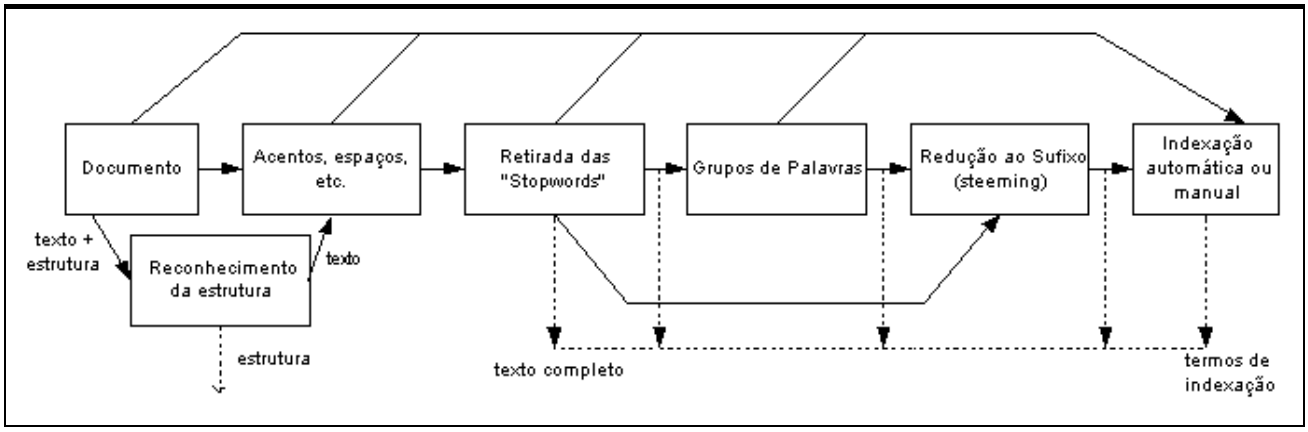


Figura 5 – Visão lógica do documento através das várias fases do processamento do texto (adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 166).

No esquema acima ilustrado, podemos perceber as várias – e opcionais – operações realizadas sobre o documento na sua preparação para a indexação. Após o reconhecimento da estrutura – para o caso de utilização de partes escolhidas do documento na indexação – retiram-se os caracteres indesejáveis, como espaços, acentos, entre outros. Em seguida, são eliminadas as palavras com baixa significação para o processo de indexação, chamadas de *stopwords*. *Stopwords* são palavras que, para um dado idioma, apresentam baixo conteúdo informacional, sendo irrelevantes como descritores, e usualmente eliminadas dos índices. Estas palavras podem ser utilizadas em uma frase numa query, mas nunca são utilizadas individualmente como termos de busca.

No processo de eliminação de *stopwords*, está implícito o fato de que algumas palavras tenham um peso maior do que outras para o propósito de indexação. Uma lista de *stopwords* é também chamada de *stoplist*.

O passo seguinte no processamento do texto é o reconhecimento de agrupamentos de palavras, estruturas sintáticas, gramaticais, frasais, entre outras, como no caso da indexação por sintagmas nominais. Em seguida, as palavras são reduzidas às suas raízes gramaticais no processo de *steeming* e, finalmente, armazenadas em um índice.

É interessante explicitar o processo de eliminação de *stopwords* e de escolha de termos-índice, na medida em que processo semelhante é apresentado na metodologia utilizada nesta pesquisa. Sabe-se que as várias palavras em uma língua apresentam valores diferentes para o propósito de representação do documento, e existem diversas metodologias para a extração das palavras com maior significado, em termos de

representação e discriminação do assunto do documento. Essas metodologias partem da observação de que as palavras com menor densidade informacional ocorrem com maior freqüência do que as de maior densidade, no fenômeno conhecido como **lei de Zipf**. Essa lei postula que a multiplicação do valor da freqüência de ocorrência de determinada palavra num texto pelo seu valor de significância tem como resultado um valor aproximadamente constante.

Utilizando-se a lei de Zipf como ponto de partida, LUHN (apud SALTON & MCGILL, 1983, p. 60-62) criou metodologia que busca as palavras mais significativas dos documentos através da eliminação daquelas com freqüência muito baixa – por considerá-las de pouca valia na representação do documento – e também aquelas que possuem freqüência alta demais – por considerá-las com baixo poder informacional. Algumas outras propostas metodológicas foram desenvolvidas, a partir das considerações de que a metodologia de Luhn seria muito simplificada, e de pouca valia para ser implementada em SRIs. Algumas delas são detalhadas adiante (SALTON & MCGILL, 1983, p. 59-71, MEADOW, 1992, p. 32-47):

- **pesos relacionados à freqüência inversa nos documentos**: a proposta original de Luhn, ao considerar somente a freqüência absoluta dos termos no espaço de um documento, não leva em consideração que os termos escolhidos para indexação também possuam a função de distinguir cada documento dos documentos restantes no acervo. Esse modelo relaciona cada termo à sua freqüência no escopo de um documento, mas também no escopo de todos os documentos do acervo.
- **valor discriminatório dos termos**: este modelo, conceitualmente semelhante ao anterior, procura mensurar matematicamente o poder de discriminação que cada termo possui como descritor de um documento em relação aos outros documentos do acervo, e assume que os termos com alto poder de discriminação sejam bons candidatos a descritores.
- **razão entre sinal e ruído**: baseado na teoria matemática da comunicação de Shannon (1948), este modelo considera a significância a partir da freqüência inversa de ocorrência de uma palavra em cada conjunto de palavras no texto, de forma que as palavras com ocorrência mais “concentrada” sejam mais

significativas. Este modelo não apresentou resultados satisfatórios em ambientes de recuperação de informações.

Podemos perceber que cada uma das etapas do processo de indexação, segundo a ilustração da FIG. 5, é opcional, de acordo com a metodologia de indexação adotada, e algumas das etapas revelam-se mais prováveis de ocorrer na grande maioria das metodologias, enquanto outras ocorrem apenas em metodologias específicas. Como exemplos, podemos citar metodologia de indexação por texto completo que acontece em alguns mecanismos de busca da *web*[23], que nem sempre elimina as chamadas *stopwords* ou perfaz o *steeming*; e metodologia de indexação por sintagmas nominais proposta neste trabalho, que não reduz as palavras às suas raízes gramaticais; e nem elimina a priori as *stopwords* antes que sejam identificados os sintagmas nominais.

Os modelos testados no escopo desta tese para a extração dos sintagmas nominais representativos e significativos são detalhados adiante, quando da apresentação da metodologia de trabalho. Podem-se esperar, entretanto, diversas particularidades metodológicas em relação ao esquema apresentado acima, pelo fato de se estar lidando com sintagmas nominais ao invés de palavras-chave.

Alternativamente à linguagem natural com termos livres, alguns SRIs, utilizam na indexação linguagens naturais **controladas estruturadas**, como tesauros, e **não estruturadas**, como vocabulários controlados, para a escolha de termos preferenciais para a indexação, ao invés de considerar somente os termos presentes no texto dos documentos. Nesses casos, conseguem-se ampliar os pontos de acesso através da utilização de termos preferenciais e sinônimos como descritores na indexação. Os tesauros e sua utilização nos SRIs são explorados com mais detalhe em seção subseqüente deste documento.

Como se pode notar, as atividades de representação e recuperação de informações em SRIs estão intimamente interligadas, e são mutuamente interdependentes. O sucesso da recuperação de informações está condicionado à forma como os documentos constituintes do acervo foram representados.

---

[23] Como exemplo, podemos apontar o mecanismo de busca *Google* (http://www.google.com), que indexa os documentos a partir de seu texto integral.

2.2.3 – Armazenamento em SRIs

Por armazenamento, entendemos a gestão física ou lógica que os sistemas de recuperação de informações realizam dos acervos de documentos e de representações destes (índices, catálogos, etc.). Não fez parte do escopo deste trabalho explorar as tecnologias de hardware e *software* utilizadas para o armazenamento de documentos e seus índices, mas é interessante explorar as várias implementações lógicas de armazenamento para ampliar o entendimento da metodologia utilizada. Cabe ressaltar que o modo como os documentos e seus índices são armazenados em SRIs está intimamente atrelado ao processo utilizado na indexação dos documentos.

Podemos destacar os seguintes modelos de armazenamento de arquivos, dentre os vários existentes (SALTON & MCGILL, 1983, p. 12-21; BAEZA-YATES & RIBEIRO-NETO,1999, p. 191-228; KORFHAGE, 1997, p. 305-311):

- **Arquivos seqüenciais**: em arquivos seqüenciais, como o nome indica, os registros são armazenados seqüencialmente, sem nenhuma espécie de ordenação. É um dos métodos mais simples de armazenar documentos e suas representações (*surrogates*), usualmente compostas por um conjunto de informações descritivas do documento, tanto físicas quanto temáticas. É eficaz no momento do armazenamento, uma vez que não há necessidade de tipo algum de reorganização dos registros existentes quando são adicionados novos registros ao índice. É, porém, um dos menos eficazes no tocante à recuperação das informações, pois as buscas pelos documentos devem ser seqüenciais, o que pode se tornar proibitivamente lento no caso de grandes índices. É um método de armazenamento quase que completamente independente do processo de indexação utilizado, pois índices e documentos são armazenados na mesma estrutura de arquivos.
- **Arquivos seqüenciais ordenados** (*hashed files*): nesses arquivos, um valor usualmente extraído de um dos campos dos *surrogates* (ex: nome do autor) é escolhido como *chave* de ordenação para os documentos, o que possibilita que as buscas sejam mais rápidas. Cada novo documento e seus registros associados devem ser posicionados em locais apropriados na seqüência

existente, no momento de sua inclusão. A busca, porém, para ser eficaz, deve limitar-se ao dado representado no campo de ordenação.

- **Arquivos indexados**: nesses arquivos, é imposta uma estrutura adicional de índices, e os documentos podem ser divididos em seções, para que se possa especificar a seção a ser pesquisada no processo de recuperação. Usualmente, são utilizados como descritores os termos extraídos durante o processo de indexação. São dois os tipos principais de arquivos indexados:
    - **Arquivos diretos**: são aqueles em que documentos e seus índices são armazenados na mesma estrutura e a pesquisa do conteúdo dos documentos é feita através do acesso aos próprios documentos, na busca por termos pertinentes à necessidade de informação.
    - **Arquivos invertidos**: são os utilizados na grande maioria de sistemas de recuperação de informações. Nesses arquivos, existem duas estruturas distintas e inter-relacionadas de armazenamento; a dos documentos em si e a de seus termos índices. A pesquisa do conteúdo dos documentos é feita através do acesso aos índices ordenados, que por sua vez, são divididos em listas de vocabulário (as palavras utilizadas como descritores de todos os documentos presentes no sistema) e listas de ocorrências (ponteiros para os documentos onde cada uma das palavras ocorre).
- **Estruturas arbóreas** (*tree-structured files*): nesses arquivos, os registros são armazenados em uma estrutura de árvore, que modela algum tipo de relação intrínseca dos registros e documentos (ex: documentos associados a estruturas hierárquicas como organogramas e árvores genealógicas). Os nós das árvores podem ser palavras, sufixos, ou outras unidades de significado. Essas estruturas, se utilizadas em conjunto com outras formas de armazenamento, podem facilitar a recuperação de informações em determinadas situações.
- **Arquivos agrupados** (*clustered files*): nestes arquivos, escolhem-se critérios para permitir o agrupamento de documentos que apresentam algum tipo de similaridade, de forma que sejam recuperados em conjunto ou relacionados entre si. Os agrupamentos podem ser fechados ou podem compartilhar documentos

com outros agrupamentos. Essas estruturas, como as arbóreas, podem otimizar a recuperação de informações em alguns contextos específicos.

- **Arquivos ligados em rede** (*netted files*): os arquivos ligados em rede são similares aos arquivos agrupados, exceto pelo fato de que não há critérios explícitos para realizar o agrupamento. Uma rede hipertextual é criada para estabelecer ligações conceituais entre os registros.

Dentre os vários modelos apresentados, podemos destacar aquele baseado em arquivos invertidos, pela elegância e simplicidade, e por serem utilizados na grande maioria dos SRIs. Há, porém, interesse crescente por modelos arbóreos, de agrupamento e em rede, na medida em que permitem melhor representação das associações entre os documentos, possibilitando melhor recuperação. Na possibilidade de utilização da metodologia apresentada nesta tese, os sintagmas nominais escolhidos para descritores devem dar origem a índices com ponteiros para os documentos, na estrutura conhecida como arquivo invertido.

2.2.4 – Recuperação de documentos em SRIs

Um dos problemas centrais da recuperação de informações em SRIs é a predição de quais são os documentos relevantes e quais devem ser descartados, e essa tarefa de "escolha" é executada por algum tipo de algoritmo que, baseado em heurística previamente definida, decide quais são os documentos relevantes a serem recuperados e os ordena a partir dos critérios estabelecidos (BAEZA-YATES & RIBEIRO-NETO,1999, p. 19). Nesta subseção, após relacionar as estratégias de recuperação associadas a cada uma das possibilidades de armazenamento citadas anteriormente, vamos conhecer alguns modelos de algoritmos de recuperação de informações, tomando como referência os arquivos indexados invertidos baseados em palavras-chave.

Para cada estrutura de armazenamento são possíveis algumas estratégias de recuperação, explicitadas a seguir (SALTON & MCGILL, 1983, p. 12-21; KORFHAGE, 1997, p. 305-311; BAEZA-YATES & RIBEIRO-NETO, 1999, p. 191-228):

- **Recuperação em arquivos seqüenciais**: a busca em arquivos seqüenciais é simples, mas pouco eficaz, pois os descritores devem ser buscados seqüencialmente, percorrendo-se cada um dos registros.

- **Recuperação em arquivos seqüenciais ordenados** (*hashed files*): a busca em arquivos seqüenciais ordenados é mais eficaz somente se estivermos usando o valor extraído dos campos dos *surrogates* (utilizado como chave de indexação) para nossa busca. Em outros casos, é tão ineficaz quanto a busca em arquivos seqüenciais.
- **Recuperação em arquivos indexados**: nesses arquivos, em vez de pesquisar o documento em si, a busca é realizada no índice, que pode ser seccionado para que se busque em alguma parte específica do documento. Para realizar a busca são utilizados os termos extraídos durante o processo de indexação e algum algoritmo de busca que realize o casamento de padrões (*pattern matching*), dentre os vários examinados adiante nesta seção. A diferença da busca nos arquivos diretos e nos invertidos se dá pelo fato de que nos primeiros a pesquisa do conteúdo é feita através do acesso aos próprios documentos, na busca por termos pertinentes à necessidade de informação, enquanto nos arquivos invertidos, a pesquisa ao conteúdo dos documentos é feita através do acesso às listas de vocabulário dos índices ordenados. Quando há um casamento de padrões, busca-se a lista de ocorrências para acessar os documentos considerados atinentes. O algoritmo para busca em um arquivo invertido segue três passos gerais:
    - **Busca no vocabulário**: as palavras e padrões presentes na *query* são isolados e é realizada a busca no vocabulário do arquivo invertido. As frases são divididas em suas palavras constituintes;
    - **Recuperação de ocorrências**: a lista das ocorrências das palavras e frases é recuperada;
    - **Manipulação das ocorrências**: as ocorrências são processadas para que sejam resolvidas questões como frases, operadores booleanos e operadores de proximidade.
- **Recuperação em estruturas arbóreas** (*tree-structured files*): a busca em estruturas arbóreas pode ocorrer de várias maneiras, basicamente através de algoritmos que utilizam critérios de decisão para navegação na estrutura.

- **Recuperação em arquivos agrupados** (*clustered files*): o agrupamento de arquivos segundo algum critério facilita a recuperação de arquivos correlatos, mas a estrutura de arquivos agrupados geralmente se constrói sobre estrutura básica com base em arquivos indexados. O agrupamento mais básico acontece com arquivos de documentos que contêm termos em comum, mas este agrupamento pode ser expandido com a utilização de um tesauro, de forma a identificar relacionamentos semânticos entre os termos dos documentos.
- **Recuperação em arquivos ligados em rede** (*netted files*): como não existe indexação específica associada a essa forma de armazenamento, para realizar a recuperação de arquivos ligados de forma hipertextual é necessário percorrer os *links* que os conectam entre si através de alguma estratégia de navegação (*browsing*) como na *web*, exemplo de rede estruturada dessa forma. Para a recuperação de informações na *web*, foram criados mecanismos de busca que constroem índices a partir de palavras-chave dos documentos disponíveis nos servidores.

Os algoritmos de ordenação dos resultados utilizados na recuperação de informações operam segundo premissas de acordo com o conceito de relevância dos documentos, e premissas diferentes levam a conjuntos de respostas diferentes. A FIG. 6 ilustra a taxonomia proposta por BAEZA-YATES & RIBEIRO-NETO (1999, p. 20), ilustrando 15 modelos de recuperação de informações. Esses modelos são detalhados de acordo com a forma como são apresentados pelos autores, na medida em que ajudam a ampliar o entendimento e a contextualização da proposta desta tese.

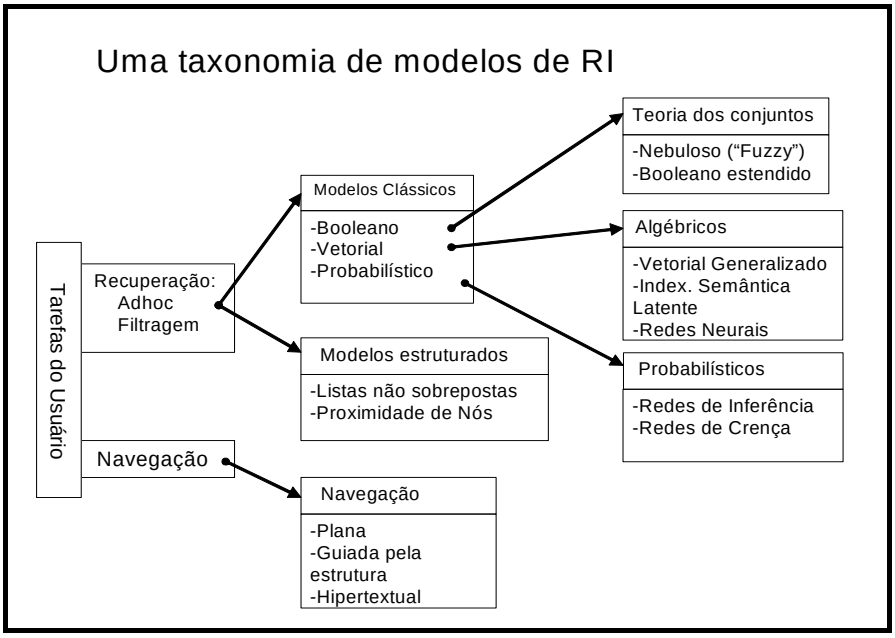


Figura 6 – Uma taxonomia de modelos de RI (adaptado de BAEZA-YATES & RIBEIRO-NETO, 1999, p. 21).

Nos sistemas de recuperação de informações, há usualmente interface através da qual o usuário traduz sua necessidade de informações em forma de questões ou palavras-chave, ou mesmo examina os documentos na busca de informações pertinentes. Essas ações são consideradas como papel do usuário (*user task*). Os dois modos de buscar informações são classificados em modelos de recuperação (***retrieval***) e os modelos de navegação (***browsing***). Nestes últimos, o usuário não propõe uma questão (*query*) ou necessidade de informação ao sistema. Em vez disso, navega através dos documentos – que não foram necessariamente indexados previamente – buscando informações de interesse. A busca em estruturas de arquivos ligados em rede é usualmente executada através de navegação do tipo hipertextual. Nosso interesse, no escopo desta tese, referiu-se aos modelos de recuperação, pois somente nesses modelos faz sentido a metodologia de escolha de descritores.

Quando o acervo de documentos sofre poucas alterações enquanto novas *queries* são submetidas ao sistema, chama-se o modo de operação de “recuperação ***adhoc***”. Quando as *queries* se mantêm relativamente estáticas enquanto novos documentos são adicionados, chama-se a esse modo de operação de filtragem (***filtering***). A filtragem

acontece usualmente em processos de monitoração de fontes de informação, enquanto a recuperação *adhoc* representa as buscas usuais em SRIs.

Os modelos de recuperação se dividem em modelos **clássicos** e modelos **estruturados**. Nos modelos clássicos, cada documento é descrito por um conjunto de palavras-chave representativas – também chamadas de termos de indexação – que busca representar o assunto do documento e sumarizar seu conteúdo de forma significativa. Essas palavras são escolhidas após o processamento do texto, como vimos anteriormente na FIG. 5. Nos modelos estruturados, podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto (como seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras informações.).

Os modelos clássicos de recuperação são três: o modelo **booleano**, o modelo **vetorial** e o modelo **probabilístico**. Para cada um deles, há modelos alternativos que visam estendê-los em funcionalidade e o desempenho. Vamos examinar brevemente esses modelos adiante:

- **Modelo booleano**: esse modelo, baseado na teoria dos conjuntos, é simples e elegante, embora não seja dos mais eficazes. Para cada *query*, são recuperados todos os documentos que possuem os termos nas condições especificadas pelo usuário, que ainda pode utilizar os operadores booleanos *or*, *and* e *not* para estabelecer relações específicas de ocorrência com as palavras-chave, de forma a especificar os documentos a serem recuperados. Sua maior desvantagem é o fato de trabalhar de forma binária, ou seja, os documentos são analisados sob o critério dualista relevante / não relevante, e não é criada nenhuma espécie de ordenação dos resultados que atendam às condições de consulta. Existem alguns modelos alternativos ao booleano, apresentados a seguir:
    - **Lógica difusa ou nebulosa (fuzzy)**: nesses modelos, busca-se estender o conceito da representação dos documentos por palavras-chave, assumindo que cada *query* determina um conjunto difuso e que cada documento possui um grau de pertencimento a esse conjunto, usualmente menor do que 1. O grau de pertencimento pode ser determinado pela ocorrência de palavras expressas na *query*, tal como no modelo booleano, mas pode também utilizar um instrumento – como um tesauro – para

determinar que termos relacionados semanticamente aos termos índice também confiram algum grau de pertencimento ao conjunto difuso determinado pela *query*.

- **Booleano estendido**: nestes modelos, busca-se a superação do problema das decisões binárias do modelo clássico, por meio da aferição de pesos aos termos, aproximando o modelo original do modelo vetorial, a seguir.

- **Modelo vetorial**: nesse modelo, os documentos são modelados como "sacos de palavras" (*bags of words*), e são representados como vetores no espaço n-dimensional, onde n é o total de termos índices (palavras) de todos os documentos no sistema. No modelo, que é não binário, pode-se calcular um grau de similaridade a ser satisfeito pelos documentos para serem considerados relevantes (ex: que as palavras apareçam ao menos duas vezes, etc.) e determinar o grau de similaridade, com vistas a construir um *ranking*. O modelo vetorial é a base da grande maioria de sistemas de recuperação de informações, mais notadamente os que têm como objeto a Internet, embora estes utilizem também outras técnicas[24] para determinar o *ranking* de documentos como resposta à uma consulta. Em seguida, apresentamos alguns modelos que se propõem a estender a funcionalidade do modelo vetorial:

  - **Vetorial generalizado**: nesses modelos, questiona-se a independência dos termos índices, assumida nos modelos clássicos, e abre-se a possibilidade de considerar que certas palavras sejam relacionadas. Uma das formas de determinar relações entre palavras é examinar a co-ocorrência dessas palavras no texto de cada documento, além do exame das relações semânticas estabelecidas por um tesauro, como foi comentado.

[24] Nos mecanismos de busca da Internet de terceira geração, além do modelo vetorial, utilizam-se, para determinar a ordenação dos documentos, técnicas como a análise de *links*, que contabiliza a quantidade de documentos que apontam para um documento específico através de *links* hipertextuais; a análise de autoridade, que investiga a idoneidade e importância da instituição que hospeda o documento em seus servidores; e outras técnicas, como as utilizadas nas redes de inferência e redes de crença.

  - **Indexação semântica latente**: nesses modelos, questiona-se a significância das palavras-chave como candidatos a descritores, e busca-se estabelecer o casamento *conceitual* entre documentos e *queries*. Se nos modelos anteriores buscava-se estabelecer um mapeamento em um espaço booleano ou vetorial de palavras, no modelo em questão busca-se mapear cada documento e cada *query* em um espaço menor, construído a partir dos conceitos relevantes que possuem os documentos no acervo.
  - **Redes neurais**: nesses modelos, utiliza-se o poder das redes neurais para realizar o casamento de padrões entre as *queries* e os documentos do acervo do sistema. Cada *query* "dispara" um sinal que ativa os termos índice, que por sua vez propagam os sinais aos documentos relacionados. Estes, por sua vez, retornam os sinais a novos termos índices, em interações sucessivas. O conjunto resposta é definido através desse processo, e pode conter documentos que não compartilhem nenhum termo-índice com a *query*, mas que tenham sido ativados durante o processo.

- **Modelo probabilístico**: nesse modelo, supõe-se que exista um conjunto ideal de documentos que satisfaz a cada uma das consultas ao sistema, e que este conjunto pode ser recuperado. Através de tentativa inicial com um conjunto de documentos (para a qual podem-se utilizar técnicas de outros modelos, como o vetorial) e do *feedback* do usuário em sucessivas interações, busca-se aproximar cada vez mais deste conjunto ideal, por meio de análise dos documentos considerados pertinentes pelo usuário. O valor desse modelo está em considerar a interação contínua com o usuário como um caminho para refinar o resultado continuamente. Os modelos que procuram ampliar o escopo do modelo probabilístico são os seguintes:
  - **Redes de inferência**: nesses modelos, associam-se variáveis aleatórias ao evento do atendimento de uma *query* específica por um documento específico. Essas variáveis podem ser alteradas de acordo com os eventos futuros, de forma a estabelecer relacionamentos baseados nos eventos observados.

- **Redes de crença (belief networks)**: nesses modelos, similares às redes de inferência, documentos e *queries* são modelados como subconjuntos de um espaço de conceitos. A cada documento, associa-se a probabilidade de que o mesmo cubra os conceitos presentes no espaço de conceitos. Cada *query* é mapeada no espaço de conceitos, que por sua vez, está conectado ao espaço de documentos.

Os modelos apresentados são apenas uma amostra do que vêm sendo pesquisado, em um campo que contém muitas frentes de pesquisa, que não poderiam ser enumeradas neste trabalho. Grandes avanços vêm sendo conseguidos, por exemplo, na recuperação de informações em ambientes de muitas mídias, como áudio e vídeo (SPARCK JONES & WILLETT, 1997, p. 493-502 e 503-512; BAEZA-YATES & RIBEIRO-NETO, 1999, p. 345-363).

A adoção de descritores através da escolha de sintagmas nominais significativos pode permitir a construção de SRIs que utilizem estratégias de busca booleana, vetorial ou probabilística, sendo que os sintagmas nominais, pelo seu maior nível de significado em comparação com as palavras-chave, realizam uma aproximação com o espaço de conceitos que é utilizado na indexação semântica latente e nas redes de crença. A utilização de algoritmos para mensurar e analisar a proximidade das palavras pode aproximar a metodologia utilizada dos modelos estruturados, de forma mais significativa do que os modelos que se baseiam em palavras-chave. Podem-se ainda imaginar melhores técnicas de filtragem, quando se incorpora a camada semântica provida pela transição das palavras-chave para os sintagmas nominais.

A importância do estudo das estruturas e lógicas que embasam o funcionamento dos SRIs, realizado ao longo desta seção, fica evidente a partir da próxima seção, na qual se examina mais detidamente a proposta de uso de sintagmas nominais como termos de indexação, em alternativa às palavras-chave.

2.3 – Sintagmas nominais e sistemas de recuperação de informações

Como foi dito anteriormente, SRIs usualmente adotam termos índices para a indexação de documentos, sendo que esses termos são usualmente palavras-chave. Há a idéia fundamental embutida nesse processo de que a semântica dos documentos e das necessidades de informação do usuário podem ser expressas através desses conjuntos

de palavras, o que é, claramente, uma grande simplificação do problema, porque grande parte da semântica do documento ou da requisição do usuário é perdida quando se substitui o texto completo por um conjunto de palavras (BAEZA-YATES & RIBEIRO-NETO, 1999, p. 19). Com os autores, também concorda LE GUERN (apud KURAMOTO, 1996, p. 3), ao afirmar que:

> *"Não constitui finalidade do descritor a sua visualização mediante a abstração do valor referencial de suas ocorrências no acervo de documentos. As palavras da língua, enquanto palavras da língua, possuem apenas atributos sem qualquer substância, até que façam parte do discurso. Quanto ao descritor, ele representa uma entidade segundo a filosofia de Aristóteles. Assim, o descritor não pode ser considerado, a exemplo das palavras da língua, como um símbolo sem referência*".

Através dessas constatações, muitas pesquisas são realizadas para ampliar o processamento da linguagem natural de modo a identificar o significado expresso em suas estruturas semânticas profundas.

Vamos apresentar algumas das abordagens de processamento de linguagem natural para, em seguida, examinar mais detidamente as que consideram o uso específico dos sintagmas nominais. No capítulo em que a metodologia é apresentada, enumeram-se algumas pesquisas sobre a extração automática de sintagmas nominais.

2.3.1 – SRIs baseados no processamento de linguagem natural

Desde o advento dos sistemas automatizados, possibilitados por computador, são projetados SRIs baseados no processamento de linguagem natural. Na área da ciência da informação, são exemplos os sistemas e metodologias KWIC e KWOC; POPSI, PRECIS, entre outros. (BHATTACHARYYA, 1979; AUSTIN, 1984; LANCASTER, 1993, p. 43-60 e 229-272). Não foi objetivo desta tese enumerar e explicitar o funcionamento dos referidos sistemas. Dentre essas iniciativas, há, na literatura, centenas de registros de tentativas de otimizar a indexação e organização dos documentos em SRIs através de processamento **aprofundado** da linguagem natural.

Embora o tratamento lexical de qualquer texto de documento possa ser considerado como processamento de linguagem natural (PLN), usualmente o termo é utilizado para o caso em que estejam envolvidos aspectos sintáticos, semânticos, pragmáticos ou

dialógicos dos documentos (CHURCH, 1988; JACOBS & RAU, 1988; SMEATON, 1989; BLAIR, 1990; HERMAN & CANDELA, 1990 *apud* KORFHAGE, 1997, p. 238-240; FOSKETT, 1997, p. 371-191). Apesar de ser antigo o interesse nas análises sintática e semântica, e serem inúmeras as propostas metodológicas, não há registros de grandes sucessos destas em relação às análises puramente lexicais (KORFHAGE, 1997, p. 238-240). Acredita-se, no escopo desta tese, que as tentativas anteriores tenham falhado pela dificuldade da análise de estruturas complexas da linguagem – como, por exemplo, os sintagmas nominais – sem ferramentas metodológicas e tecnológicas adequadas para tal.

Uma das técnicas mais difundidas de processamento de linguagem natural – apresentada anteriormente – é a indexação semântica latente, na qual se busca estabelecer um espaço conceitual intermediário que relaciona as *queries* de usuários e os documentos do acervo. O problema dessa abordagem é a forma com que se constitui o espaço de conceitos, que deve ser construído através de algum tipo de análise semântica dos documentos. Outros rumos de pesquisa ligados à inteligência artificial envolvem as análises de diálogos homem-máquina, na busca de melhor interpretação das necessidades dos usuários. Também existem tentativas de implementação de processos de indexação por questões, ou seja, usar as questões que podem ser eventualmente respondidas através da análise do documento a guisa de termos de indexação. Mas essas abordagens esbarram em dificuldades relativas às necessidades de interpretação subjetiva do conteúdo informativo dos documentos.

Devemos considerar ainda os sistemas de recuperação que buscam analisar *queries* em linguagem natural, modelando-as como um documento no espaço vetorial; ou algoritmos de casamento de padrões (*pattern matching*) entre excertos de textos e *queries* (*string matching*), numa busca por similaridade. Essa similaridade deve ser definida por meio de convenções sintáticas que venham a ocorrer em trechos de textos (BAEZA-YATES & RIBEIRO-NETO et al, 1999, p. 103-106 e 286-288; KORFHAGE, 1997, p. 291-300). Outras metodologias similares implantadas em SRIs permitem a busca de expressões regulares, ou mesmo analisam a proximidade da ocorrência de alguns termos, expandindo o conceito de palavra-chave para frases ou outras hierarquias lexicais (LANCASTER, 1993, p. 43-60 e 229-272; NAVARRO; RASMUSSEN in BAEZA-YATES & RIBEIRO-NETO, 1999, p. 219-220 e 406-407; SALTON & LESK in SPARCK JONES &

WILLETT, 1997, p. 60-84; KORFHAGE, 1997, p. 122-123; SMEATON, 1992 ; SALTON & MCGILL, 1983, p. 87-89; FOSKETT, 1997, p. 371-191).

Alguns sistemas de filtragem são projetados para extrair informações conceituais de documentos baseados em heurísticas de inteligência artificial (RAU in SPARCK JONES & WILLETT, 1997, p. 527-533). Outros trabalhos recentes buscam explorar e operar sobre o léxico do sistema de forma a depreender outros significados possíveis para cada item lexical, de forma semelhante ao que se faz com tesauros (ABRAHÃO, 1997; GONZALEZ, 2000-1).

ZIVIANI aponta SRIs que utilizam a técnica de identificação de grupamentos de substantivos (*noun groups*), ao invés de palavras-chave, como estratégia para seleção de termos de indexação, assumindo que os substantivos costumam carregar a maior parte da semântica de um documento, o que não ocorre com artigos, verbos, adjetivos, advérbios e conectivos (BAEZA-YATES & RIBEIRO-NETO, 1999, p. 169-170). Os grupamentos de substantivos, no escopo dessas propostas, são conjuntos de nomes para os quais a "distância sintática" (medida pelo número de palavras entre dois substantivos) não excede um limite predefinido. Devem-se considerar, porém, as características das áreas de conhecimento das quais fazem parte os textos analisados, pois podemos esperar que apareçam sensíveis diferenças nos processos de indexação, dependendo da terminologia e dos estilos textuais característicos de cada área.

Uma metodologia que segue esta linha, mas extrapola a proposta de identificação de grupamentos de substantivos é a identificação dos sintagmas nominais, visando ao seu uso como descritores. Há que se observar, porém, que nem todos os sintagmas nominais podem ser considerados descritores *a priori*. Na proposta desta tese, buscou-se utilizar um tesauro para auxiliar a identificação dos possíveis descritores dentre os sintagmas nominais extraídos e considerados "válidos".

2.3.2 – O uso de SNs como descritores

SALTON & MCGILL (1983, p. 90-94) discutem algumas abordagens teóricas para o uso de métodos lingüísticos na recuperação de informações; dentre elas, a análise da estrutura sintática (*parsing*) dos documentos de forma a identificar as estruturas sintagmáticas. Esses autores, entretanto, apontam as dificuldades intrínsecas ao processo de análise semântica através da análise sintática e exemplificam casos em que

é impossível o reconhecimento não ambíguo de relações semânticas através dos componentes da sentença, sugerindo que um modelo baseado em gramáticas transformacionais poderia trazer melhores resultados. Nesse ponto, parecem então concordar com LIBERATO (1997), que entende que a análise completa das estruturas semânticas só é possível através da análise cognitiva dos contextos. Ao indicar a maior eficácia relativa dos algoritmos de geração de frases baseadas em freqüência de palavras, talvez apontem para o fato de que o algoritmo proposto neste trabalho está na contramão dos resultados até então encontrados. Alternativa apontada é a interferência humana no processo de desambiguação através de uma interface, o que seria pouco desejável no processo que pretende ser, em sua máxima extensão possível, automático.

Um importante caminho de pesquisa que visa auxiliar a resolução dos problemas de desambiguação semântica através da análise dos contextos é a resolução de correferência, ou resolução anafórica (VIEIRA, 1998 e 2000; SANT'ANNA, 2000 ; ROSSI et al, 2001; GASPERIN et al, 2003). A cadeia de correferência é uma seqüência de expressões em um discurso que se referem à mesma entidade, objeto ou evento. Essas cadeias são úteis para a representação semântica do modelo de domínio, e podem melhorar a qualidade dos resultados em diversas aplicações de processamento de linguagem natural, como recuperação e extração de informações, geração automática de resumos, traduções automáticas, entre outros (ROSSI et al, 2001). O processo de resolução de correferências envolve a identificação e a extração dos sintagmas nominais.

LE GUERN e BOUCHÉ (apud KURAMOTO, 1999) apontam o sintagma nominal como a menor unidade de informação contida em um texto, e LE GUERN explicita a transformação que ocorre nas palavras integrantes do universo do discurso, quando analisadas sob a ótica dos sintagmas nominais:

> *"A princípio a palavra, enquanto palavra da língua, enquanto unidade lexical, está no nível N. Antes que faça parte do sintagma nominal, a palavra passa por um nível intermediário (N') onde ela incorpora seus valores dentro do universo do discurso. A distinção entre estes dois níveis é que no nível N, a palavra não é senão um conjunto de propriedades; ela não designa nenhum objeto qualquer que seja. Ela não faz então nenhuma referência a um objeto do mundo real. Ao contrário, quando*

está no nível N', ela designa um objeto ou ao menos faz referência a uma classe de objetos." (1999, p. 27, tradução nossa).

O grupo de pesquisas SYDO, ao qual pertencem esses pesquisadores, tem como fundamento teórico a utilização de sintagmas nominais como descritores (Ibidem, 1996). Ao trabalhar em parceria com esse grupo, KURAMOTO (1999), em sua tese de doutorado, desenvolveu pesquisa fundamental para a consideração de se utilizarem sintagmas nominais como descritores. Já em um trabalho anterior, KURAMOTO (1996) vislumbrou a maquete proposta na tese e já apontava o potencial natural de organização dos sintagmas nominais, que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta ao processo de busca de informação.

Em sua tese, toda a argumentação é fundamentada com o objetivo de demonstrar as vantagens – em termos da semântica apreendida pelos descritores – da utilização de sintagmas nominais ao invés de palavras. Após as considerações sobre a sua viabilidade, apresenta-se um protótipo de interface para sistemas de recuperação de informações baseados em sintagmas nominais, extraídos do próprio acervo de documentos do sistema. A idéia era que, a partir de uma palavra chave introduzida pelo usuário, o sistema pesquisasse todos os sintagmas nominais extraídos do acervo que contivessem a palavra, de forma que o usuário pudesse escolher um sintagma nominal significativo e o sistema possa assim refinar a consulta.

O sistema desenvolvido por Kuramoto pode ser considerado como uma das inspirações para a presente tese, na medida em que, em ambos, busca-se uma alternativa para melhor indexação, utilizando-se sintagmas nominais. Entretanto, em sua maquete, segundo o autor, "a extração dos sintagmas nominais foi realizada de forma manual, simulando extração automática. Esse procedimento foi adotado em função da não-existência ainda de sistema de extração automática de SNs em acervos contendo documentos em língua portuguesa." (1996, p. 6). Alguns sistemas desse tipo, entretanto, se encontram disponíveis atualmente, como o que foi disponibilizado para o propósito do presente trabalho (BICK, 2003; GASPERIN et al, 2003). Outra diferença fundamental refere-se ao objetivo: se no projeto de Kuramoto buscava-se apresentar maquete de um SRI baseado em sintagmas nominais, o objetivo desta tese foi desenvolver uma

metodologia de auxílio à indexação automática utilizando uma técnica aplicada sobre os sintagmas nominais extraídos automaticamente. Diferenças a parte, o fundo filosófico é bastante comum.

Na próxima seção, além da conceituação básica, apresentaremos os tesauros como instrumentos de recuperação de informações. No contexto metodológico desta pesquisa, os tesauros são auxiliares na seleção dos descritores significativos e, além disso, são passíveis de serem atualizados à medida que se aplica a metodologia a diferentes *corpora*.

2.4 – Tesauros e sistemas de recuperação de informações

Os instrumentos para a representação da informação para indexação, armazenamento e recuperação de informações são considerados linguagens documentárias. As linguagens documentárias mais conhecidas são os tesauros e os sistemas de classificação bibliográfica. Alguns autores consideram os tesauros como linguagens artificiais (MEC/MCT, 1990), enquanto outros os consideram linguagens naturais controladas (ROBREDO & CUNHA, 1994). De fato, os tesauros procuram normalizar para uma área do conhecimento as mais propícias formas verbais que denotam os referentes, segundo a teoria do conceito (DAHLBERG, 1978), e essas formas verbais são retiradas das linguagens naturais. Entretanto, a estruturação das relações semânticas e lógico-funcionais pressupõe detalhado estudo para sua construção, o que poderia justificar sua inclusão dentre as linguagens artificiais.

Uma boa definição de tesauro, utilizada na área da ciência da informação, é a da UNESCO (1973, p. 6, apud CAMPOS, 2001, p. 90-91), que o apresenta sob dois aspectos:

a) Segundo a estrutura: “É um vocabulário controlado e dinâmico de termos relacionados semântica e genericamente cobrindo um domínio específico do conhecimento”.

b) Segundo a função: “É um dispositivo de controle terminológico usado na tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem do sistema (linguagem de documentação, linguagem de informação) mais restrita”.

Estas definições vêm sendo usadas na literatura até os dias de hoje (CAMPOS, 2001; FOSKETT in SPARCK JONES & WILLETT, 1997, p. 111-134). Usualmente, um tesauro é uma ferramenta para mapeamento e controle do vocabulário em uma área do conhecimento, através do estabelecimento dos termos preferencialmente utilizáveis (*preferred terms*), em detrimento de outros, que podem ser sinônimos (*non-preferred terms*) ou termos relacionados. De acordo com o Manual para Elaboração de Tesauros Monolíngües (MEC/MCT, 1990), as relações mapeadas pelos tesauros podem ser de três tipos:

- Relações lógicas:
    - relação genérico-específica;
    - relação analítica;
    - relação de oposição;
- Relações ontológicas:
    - relação partitiva;
    - relação de sucessão;
    - relação de material-produto;
- Relações de efeito:
    - relação de causalidade;
    - relação instrumental;
    - relação de descendência;

Ao mapear as relações lógicas, ontológicas e de efeito, o tesauro estrutura os conceitos (CAMPOS, 2001).

Assim, o tesauro típico contém as seguintes estruturas:

- descritores: são palavras ou grupos de palavras que representam conceitos;
- definições: necessárias para a apreensão do significado de um determinado conceito, relacionando-o a outros conceitos;

- relações semânticas: que relacionam os conceitos entre si, através de indicadores *Broader Term*, *Narrower Term*, *Related Term*, e outros.

Os tesauros ainda possuem dois tipos de apresentação, sistemática e alfabética. Na sistemática, os termos aparecem de acordo com suas relações hierárquicas, o que permite a escolha pelo usuário do melhor termo para exprimir uma idéia sem que haja conhecimento prévio desse termo. Na alfabética são apresentadas as relações de ordens lógicas, ontológicas e de equivalência para cada termo, que é listado em ordem alfabética (MEC/MCT, 1990).

De acordo com MEC/MCT (1990), para a constituição de um tesauro é imprescindível que se realize uma pesquisa terminológica prévia, seguindo os procedimentos:

1. formação de uma equipe interdisciplinar constituída de elementos das áreas de classificação, de lingüística, e da área em que estiver sendo construído o tesauro;
2. determinação do campo conceitual básico sob o qual se estruturará o tesauro, sendo este passível de modificação;
3. ter o uso dos termos como parâmetro essencial para elaboração de um tesauro;
4. identificar a literatura relevante;
5. coletar os termos pertinentes a uma área.

O uso de tesauros em sistemas de recuperação de informações é amplamente coberto na literatura (SALTON & MCGILL, 1983, p. 75-89; LANCASTER & WARNER, 1993, p. 89-107 ; FOSKETT, 1997, p. 76-95; KORFHAGE, 1997, p. 138-139; SPARCK JONES & WILLETT, 1997, p. 15-20; BAEZA-YATES & RIBEIRO-NETO, 1999, p. 170-173; CAMPOS, 2001, p. 87-100). FOSKETT (Op. cit) enumera os sete maiores propósitos dos tesauros:

1. prover um mapa de uma dada área de conhecimento, indicando como conceitos ou idéias sobre conceitos são relacionados entre si, o que ajuda a um indexador a entender a estrutura do campo;

2. prover um vocabulário padronizado para uma dada área de conhecimento, que assegure que indexadores sejam consistentes ao escolherem termos de indexação em um SRI;
3. prover um sistema de referências entre termos que garanta que apenas um termo de um conjunto de sinônimos seja usado para indexar um conceito, de forma consensual entre os indexadores, e para prover guias para termos que não sejam relacionados a nenhum outro, seja por meio de estruturas classificatórias ou garantias literárias;
4. prover um guia para usuários dos SRIs de forma que possam escolher corretamente um termo para uma busca por assunto, o que aumenta a importância das referências cruzadas.
5. ajudar a localizar novos conceitos em um esquema de relacionamentos a partir de conceitos existentes, de forma que faça sentido aos usuários do sistema;
6. prover hierarquias classificatórias, de forma que uma busca possa ser ampliada ou restringida sistematicamente, se uma primeira escolha de termos para busca produz, respectivamente, poucos ou muitos resultados;
7. prover maneiras de padronizar os termos em um dado campo do conhecimento (propósito desejável).

Podemos notar que os propósitos 2, 3, 4, 5 e 6 são especificamente aplicáveis aos SRIs. De maneira similar, SALTON & MCGILL (1983, p. 75-89) destacam o uso de tesauros para fornecer termos com maior poder discriminatório do que os que apresentam freqüências muito altas ou muito baixas, por meio do exame das associações (ex. “ciência da informação“, ao invés de “ciência“ e “informação”). Ora, essa abordagem é muito semelhante ao uso de sintagmas nominais, sendo que caberia ao tesauro realizar a verificação dos agrupamentos. Os autores também consideram os tesauros como ferramentas para aumentar a revocação na indexação ou na recuperação, por meio da substituição dos termos extraídos dos textos ou das *queries* por termos preferenciais, ou mesmo da adição de termos mais abrangentes ou mais específicos em uma cadeia de relacionamentos semânticos.

SALTON & MCGILL (Loc. cit.) e BAEZA-YATES & RIBEIRO-NETO (Op. cit., p. 130-137) ainda delineiam um algoritmo para construção automática de tesauros de similaridade para expandir o alcance das *queries* dos usuários. Algoritmos semelhantes são examinados quando apresentarmos as possíveis extensões da metodologia utilizada nesta tese.

JOYCE & NEEDHAM (in SPARCK JONES & WILLETT, 1997, p. 15-20.) e ZIVIANI (apud BAEZA-YATES & RIBEIRO-NETO, Op. cit.) também destacam o papel dos tesauros no campo da recuperação de informações, com aplicações possíveis na reformulação e na ampliação das *queries* dos usuários, ou na ampliação (ou padronização) dos pontos de acesso aos documentos. ZIVIANI, porém, aponta os problemas existentes nesta abordagem, pois os contextos locais dos termos nos textos raramente são captados pelos relacionamentos descritos nos tesauros.

Existem tesauros em diversas áreas do conhecimento, e atividades humanas, como metalurgia, medicina, química, etc.

No escopo desta tese, utilizou-se na metodologia prospectiva, um tesauro específico da área de ciência da informação para verificar a pertinência dos sintagmas nominais extraídos dos documentos do *corpus* utilizado, extraído de publicações na área de ciência da informação. O tesauro utilizado para validação de termos segundo a metodologia de escolha de descritores foi o Tesauro da Ciência da Informação (CNPq/IBICT, 1989) que se encontra bastante defasado, havendo mesmo iniciativas para lançar uma versão mais atual.

Como previsto em objetivo específico, aventou-se a possibilidade de que a metodologia utilizada para a consecução do objetivo geral pudesse ser utilizada para a escolha automática de descritores, o que constituiria subsídio para uma metodologia semi-automática para atualização de tesauros.

3 CONTEXTOS DE APLICABILIDADE

Neste capítulo, são apresentados dois marcos tecnológicos e conceituais que nortearam o panorama de aplicabilidade das novas tecnologias digitais de tratamento da informação. Em primeiro lugar, pela importância da filosofia subjacente e das tecnologias que embasam sua concepção, apresentamos a *web* semântica e suas tecnologias associadas. Em segundo, as bibliotecas digitais, pois se configuram ambientes para onde vão convergir os resultados de todas as pesquisas que hoje são realizadas sobre a melhoria dos SRIs. O objetivo desta seção foi apenas oferecer um contexto onde as metodologias de escolha automática de descritores podem encaixar-se, mas no caso da *web* semântica, alguns conceitos – tais como a estrutura das metalinguagens, como o XML – serão importantes para que se possa ter melhor idéia do funcionamento das ferramentas apresentadas no escopo desta tese. As subseções estão dispostas na seguinte ordem:

Na seção 3.1, apresenta-se o panorama da *web* semântica, com ênfase na metamorfose da *web* tradicional nesse novo repositório, com embasamento filosófico e capacidade tecnológica para comportar e representar os significados inerentes aos documentos e suas ligações. Nessa subseção, apresentam-se as tecnologias das linguagens de marcação, os padrões de metadados e as ontologias, na forma como são apropriadas pela ciência da computação. Ao final, apresenta-se esta *web* modificada como uma entidade muito mais próxima de um sistema de recuperação de informações típico, da forma como o apresentamos anteriormente.

Na seção 3.2 apresenta-se uma breve introdução às bibliotecas digitais, um dos ambientes informacionais característicos de nossa época, que demanda que sejam desenvolvidas técnicas mais eficazes para recuperação de informações.

3.1 – A *web* e a *web* semântica

Surgida no início dos anos 1990 a *word wide web*[25], ou simplesmente *web*, é hoje tão popular e ubíqua que, não raro, no imaginário dos usuários, confunde-se com a própria Internet – a infra-estrutura de redes, servidores e canais de comunicação que lhe

[25] Na tradução literal, “teia de alcance mundial”.

dá sustentação, que foi concebida nos Estados Unidos no final dos anos 1960, tendo começado a funcionar no início dos anos 1970. Se a Internet surgiu como proposta de um sistema distribuído de comunicação entre computadores para possibilitar a troca de informações na época da guerra fria, o projeto da *web*, ao implantar de forma magistral o conceito de hipertexto imaginado por Ted NELSON (1982) e Douglas ENGELBART (1962), buscava oferecer interfaces mais amigáveis e intuitivas para a organização e o acesso ao crescente repositório de documentos que se tornava a Internet. Entretanto, o enorme crescimento – além das expectativas – do alcance e tamanho desta rede, além da ampliação das possibilidades de sua utilização, tornaram necessária nova filosofia de trabalho, com suas tecnologias subjacentes, e a ampliação da infra-estrutura tecnológica de comunicação.

Embora tenha sido projetada para possibilitar o fácil acesso a, intercâmbio e a recuperação de informações, a *web* foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica, e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar a informação de que temos necessidade. Não há estratégia alguma abrangente e satisfatória para a indexação dos documentos nela contidos, e a recuperação das informações, possível através dos “motores de busca” (*search engines*), é baseada primariamente em palavras-chave, contidas no texto dos documentos originais, o que é muito pouco eficaz. A dificuldade de determinar os contextos informacionais tem como conseqüência a impossibilidade de se identificar de forma precisa a atinência dos documentos. Além disso, a ênfase das tecnologias e linguagens atualmente utilizadas nas páginas *web* focaliza os aspectos de exibição e apresentação dos dados, de forma que a informação seja pobremente descrita e pouco passível de ser consumida por máquinas e seres humanos. Nesse contexto que surge a proposta da *web* semântica.

3.1.1 - A *web* semântica

“A *web* semântica não é uma *web* separada, mas uma extensão da atual. Nela a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas”. Com essas palavras, Berners-Lee (BERNERS-LEE et al,

2001) define os planos de seu grupo de trabalho no World Wide Web Consortium[26] (W3C) para operar a transformação que irá modificar a *web* como a conhecemos hoje. "*web* semântica" é o nome genérico desse projeto, capitaneado pelo W3C, que pretende embutir inteligência e contexto nos códigos XML utilizados para confecção de páginas *web*, de modo a melhorar a forma com que programas possam interagir com essas páginas e também possibilitar seu uso mais intuitivo por parte dos usuários (DECKER et al, 2000; BERNERS-LEE et al, 1999). O uso da conotação "semântica" para esta *web* ampliada se justifica se observarmos as aumentadas possibilidades de associações dos documentos a seus significados, através dos metadados descritivos. Além disso, as ontologias construídas em consenso pelas comunidades de usuários e desenvolvedores de aplicações permitem o compartilhamento de significados comuns.

Berners-Lee (BERNERS-LEE et al, 2001) imagina um mundo em que programas e dispositivos especializados e personalizados, chamados agentes, possam interagir através da infra-estrutura de dados da Internet, trocando informações entre si, de forma a automatizar tarefas rotineiras dos usuários. O projeto da *web* semântica, em sua essência, é a criação e a implantação de padrões (*standards*) tecnológicos para permitir tal panorama, que não somente facilite as trocas de informações entre agentes pessoais, mas principalmente estabeleça língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral.

Para atingir tal propósito é necessária a padronização de tecnologias, de linguagens e de metadados descritivos, de forma que todos os usuários da *web* obedeçam a determinadas regras comuns e compartilhadas sobre como armazenar dados e descrever a informação armazenada, de forma que a informação possa ser "consumida" por outros usuários humanos ou não, de maneira automática e não ambígua. Com a existência da infra-estrutura tecnológica comum da Internet, o primeiro passo para este objetivo está sendo a criação de padrões para descrição de dados e de linguagens que permitam a construção e codificação de significados compartilhados. Para melhor entender esses padrões e linguagens, discutiremos a seguir um pouco mais sobre esses conceitos.

---

[26] Consórcio de empresas, profissionais, cientistas e instituições acadêmicas, que é responsável pela criação de padrões tecnológicos que regulam a World Wide *web*.

3.1.2 - SGML, HTML e XML

Um documento na *web* é composto por uma mistura de dados e metadados. “Meta” é o prefixo de auto-referência, de forma que “metadados” sejam “dados sobre dados”. Os metadados em documentos na *web* têm a função de especificar características dos dados que descrevem, a forma como serão utilizados, exibidos, ou mesmo seu significado em um contexto.

A linguagem ainda utilizada atualmente para a construção da maioria das páginas *web* é o HTML, ou *HyperText Markup Language* (linguagem de marcação em hipertexto). A linguagem HTML é derivada do padrão SGML (*Standard Generalized Markup Language*), que é, na verdade, uma *meta-linguagem*, ou seja, uma linguagem para descrever outras linguagens. O padrão SGML é baseado na idéia de que documentos contenham estrutura e outros elementos semânticos que podem ser descritos sem que se faça referência à forma como esses elementos são exibidos. O conjunto de todas as *tags*[27] passíveis de serem utilizadas por qualquer linguagem derivada do SGML é chamado de DTD, ou *Document Type Definition*.

A linguagem HTML é um conjunto definido de *tags*, ou uma DTD específica do SGML, e foi criada tendo em mente a necessidade de construção de documentos para serem exibidos em dispositivos de computador (na *web*), daí sua vocação para tratar do formato que os dados contidos no documento vão assumir ao serem exibidos. O navegador ou *browser*, ao ler um documento HTML, interpreta as *tags* que este documento contém para decidir como serão exibidos os dados também nele contidos. Os navegadores atuais interpretam o HTML porque a DTD para definição do HTML é fixo, e é conhecido a priori pelo interpretador do navegador. Assim mesmo, podem ocorrer navegadores diferentes interpretando definições de exibição de forma particular, com resultados distintos no dispositivo de saída. A estrutura do HTML é rígida, não existindo a possibilidade de adição de novos comandos de marcação (*tags*) sem que haja a redefinição do DTD da linguagem – e conseqüente atualização dos navegadores para que interpretem essas novas *tags*. A última especificação do HTML lançada pelo W3C foi a

[27] Os tags são marcações sintáticas que descrevem os dados e comandos para a manipulação do documento.

versão 4.01 (dezembro de 1997), e desde então a linguagem não tem sofrido mais modificações.

A partir das limitações do HTML, e das necessidades de uma linguagem que pudesse descrever o conteúdo semântico e os significados contextuais, além da estrutura e da forma de exibição de documentos, foi criado o XML (eXtensible Markup Language). O XML é uma recomendação formal do W3C e, em determinados aspectos, se assemelha ao HTML. Ambas são derivadas do SGML e contêm *tags* para descrever o conteúdo de um documento. Mas enquanto o HTML tem como objetivo controlar a forma como os dados são exibidos, o XML se concentra na descrição dos dados que o documento contém. Além disso, o XML é flexível no sentido de que podem ser acrescentadas novas *tags* à medida que forem necessárias, bastando para isso que estejam descritas em um DTD específico; ou seja, qualquer comunidade de desenvolvedores pode criar suas marcações (*tags*) específicas que sirvam aos propósitos de descrição de seus dados. Isso possibilita que os dados sejam descritos com mais significado, abrindo caminho para embutir semântica em documentos da *word wide web* e nas Intranets. O HTML 5.0 ou XHTML é o HTML 4.0 reescrito como se fosse uma DTD específica que segue o padrão XML.

Os dados contidos nos documentos XML podem ser exibidos em uma infinidade de maneiras, dependendo do dispositivo em que são manuseados (telas de computador, celulares, PDAs, e outros). Os documentos XML não contêm, em si, as diretivas para exibição dos dados, e para cada dispositivo-destino específico, pode-se realizar uma transformação do documento originalmente em formato XML para um documento passível de ser exibido ao usuário ou entendido e utilizado por outro dispositivo tecnológico. Esta transformação é realizada, utilizando-se a linguagem XSL (eXtensible Stylesheet Language), e cada arquivo XSL contém as definições necessárias à transformação do arquivo XML original em arquivo HTML específico ou mesmo em outro formato, para manipulação por alguns dispositivos (tela do computador, tela do celular, impressora, coletores de dados, outros sistemas de informação, entre tantos.), no formato que melhor convier (tabelas, gráficos, seqüência de caracteres, e outros.) e extraindo-se os dados que forem necessários. Dessa forma, o trio composto pelos XML, sua DTD específica e o XSL se apresenta como um conjunto de padrões que possibilitam o armazenamento, descrição significativa, intercâmbio e exibição dos dados de forma personalizada.

O padrão XML é aceito como o padrão emergente para troca de dados na *web*. Mas apesar de possibilitar aos autores a criação de suas próprias *tags*, na perspectiva computacional, há muito pouca diferença entre as *tags* <AUTHOR> e <CREATOR>. Para que as marcações semânticas criadas sejam utilizadas de forma não ambígua por comunidades maiores, são necessários alguns padrões de compartilhamento mais universais. O W3C e as comunidades de usuários têm procurado prover esses padrões, como abordamos em seguida.

3.1.3 - Metadados e o padrão Dublin Core

Não basta possuir linguagem flexível como o XML para se construírem metadados. Para compartilhar um significado, é necessário que esse seja consensual e inteligível, de forma não ambígua, dentre todos os participantes da comunidade. Para resolver o problema da explosão de nomenclaturas diferentes e as várias situações, nas quais a interpretação dos dados de maneira unívoca não seja possível, foram criados, no escopo do projeto da *web* semântica, alguns padrões de metadados, a serem utilizados como marcações na linguagem XML, e a nova significação para o termo ontologias, como veremos a seguir.

O padrão Dublin Core é uma iniciativa para criação de um conjunto de metadados para a descrição de documentos eletrônicos, baseada no pressuposto de que a escolha de elementos informacionais para documentos devem ser independente do meio em que estes estejam armazenados. É composto de 15 elementos de metadados (DCMI, 2003) e se baseia no padrão MARC[28]. Seus elementos são *title* (o nome dado ao recurso, ou título), *creator* (a pessoa ou organização responsável pelo conteúdo), *subject* (o assunto, ou tópico coberto pelo documento), *description* (descrição do conteúdo), *publisher* (o responsável por tornar o recurso ou documento disponível), *contributor* (aqueles que contribuíram para o conteúdo), *date* (data em que o recurso foi tornado disponível), *type* (categoria preestabelecida para o conteúdo), *format* (o formato no qual o recurso se apresenta), *identifier* (identificador numérico para o conteúdo, tal como uma URL[29]),

[28] O MARC – MAchine Readable Cataloging é um padrão para comunicação de informações referentes aos diversos tipos de documentos de forma que se possibilite o entendimento por dispositivos eletrônicos. Foi uma iniciativa da biblioteca do Congresso dos EUA.

[29] A URL, ou *Uniform Resource Locator* é um caso particular dos URI (*Uniform Resource Identifier*), que são os endereços que identificam um “ponto de conteúdo” da *World Wide Web*, seja esse uma página de texto,

source (fonte de onde foi originado o conteúdo), *language* (a linguagem em que está escrito), *relation* (como o conteúdo se relaciona com outros recursos, como, por exemplo, se é um capítulo de livro), *coverage* (onde o recurso está fisicamente localizado) e *rights* (ponteiro ou *link* para uma nota de copyright). A DCMI - Dublin Core Metadata Initiative teve seu início em 1995, ganhando o nome da localidade onde se deu o encontro inicial, Dublin, no estado de Ohio, USA. Sua aceitação foi rápida e é hoje padrão internacional, com participantes de mais de 20 países.

Existem duas formas para o padrão Dublin Core, a forma *simples* e a *qualificada*. Enquanto a forma simples apenas especifica os padrões para os 15 possíveis pares de atributo e valor, a qualificada aumenta a especificidade dos metadados com informações adicionais sobre cada *tag* e outras orientações para o processamento dos documentos.

3.1.4 - Ontologias

A palavra “ontologia” deriva do grego *onto* (ser) e *logia* (discurso escrito ou falado). Na filosofia, a ontologia é a teoria sobre a natureza da existência, dos tipos de “coisas” que existem; a ontologia como disciplina filosófica estuda tais teorias. Os projetistas da *web* e os pesquisadores de inteligência artificial adaptaram o termo aos seus próprios jargões. Nesse contexto, ontologia é explicitada em um documento e define formalmente as relações entre termos e conceitos, e também as relações entre os conceitos em si. Nesse sentido, as ontologias mantêm semelhanças com os tesauros, utilizados para definição de vocabulários controlados. Nas palavras do SEMANTICWEB.ORG (2003), “Uma ontologia é uma especificação de uma conceituação. É designada com o propósito de habilitar o compartilhamento e reuso de conhecimentos, de forma a criar ‘compromissos ontológicos’, ou definições necessárias à criação de um vocabulário comum”.

As ontologias se apresentam como modelos de relacionamento de entidades e suas interações, em algum domínio particular do conhecimento ou específico a alguma atividade. O objetivo de sua construção é a possibilidade de troca de informações entre os membros de uma comunidade, sejam eles humanos ou agentes inteligentes. Essa troca

---

vídeo, imagem, som, e outros. O tipo mais comum de URI é a URL, que descreve o endereço da página na *web* (o servidor que a hospeda e o nome do documento nesse servidor) e o mecanismo (protocolo) utilizado para o acesso (HTTP, FTP, e outros).

só acontece quando há uma concordância “ontológica”, ou seja, o uso de terminologias compartilhadas e a definição formal de entidades e seus relacionamentos.

3.1.5 – A *web* e a semântica

A partir dos conceitos de sistemas de recuperação de informações e das tecnologias apresentadas, vamos entender um pouco mais o grande panorama da *web* semântica, e as possíveis convergências com a pesquisa apresentada nesta tese. Observemos a ilustração a seguir:

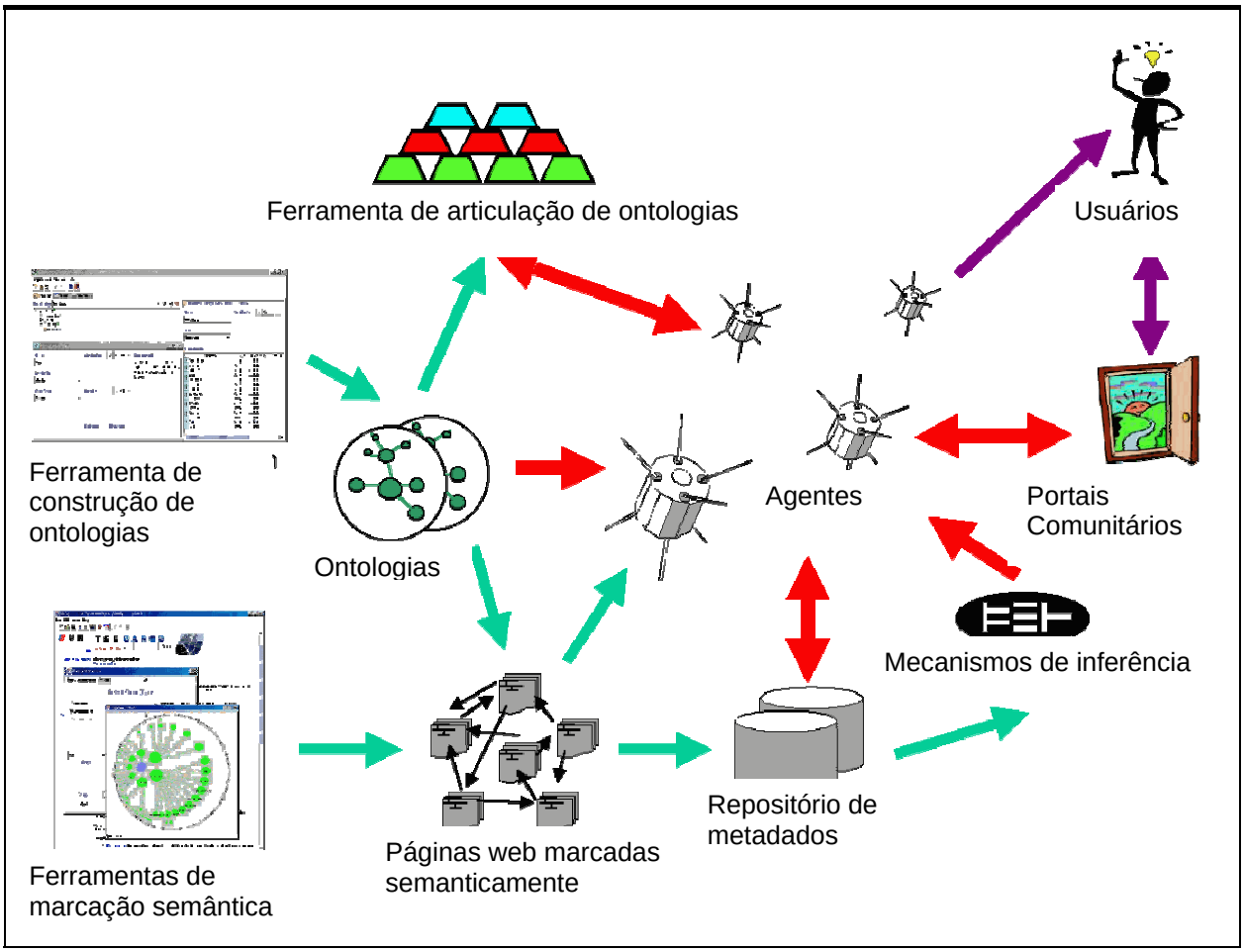


Figura 7 – O roadmap da *web* semântica (adaptado de SemanticWeb.Org, 2001).

Na FIG. 1, que ilustra o *roadmap* da *web* semântica (SEMANTICWEB.ORG, 2001), podemos entender como as tecnologias se articulam entre si, e como a *web* semântica aproxima a *web* da funcionalidade plena do sistema de recuperação de informações.

Vamos discutir as várias entidades representadas e suas funcionalidades discutidas a seguir:

No âmbito da representação dos documentos, temos as **ferramentas de marcação semântica** das páginas *web* e de **construção de ontologias** compartilhadas. Essas ferramentas possibilitam a existência cada vez mais ampla e disseminada de **páginas *web* marcadas semanticamente** por **metadados** descritos em **repositórios** de domínio público, e por conteúdo semântico compartilhado em seu significado pelas comunidades e usuários da *web* através das **ontologias**. As ontologias criadas são **articuladas** entre si através de **ferramentas específicas**. Com estratégia padronizada de indexação, podemos projetar sistemas mais funcionais para recuperação de informações armazenadas.

No âmbito da recuperação e uso dos documentos, os **agentes**, associados aos **mecanismos de inferência** executam o *harvesting* (colheita) de informações nos documentos anotados semanticamente de maneira eficaz, porque são capazes de "compreender" seus conteúdos, de modo que a informação seja mais significativamente utilizada pelos **usuários** (humanos e não humanos) da *web*. Estes podem acessar essas novas tecnologias através dos **portais comunitários** ou mesmo dos portais *corpora*tivos das organizações.

Podemos esperar que a busca de informações na *web* apresente grande melhoria dos índices de revocação e precisão, no atendimento às necessidades de informação, porque a semântica embutida nos documentos permite aos dispositivos de recuperação evitar os problemas comuns de polissemia e sinonímia, além de considerar as informações em seus contextos de significado.

As tecnologias para implementação, assim como os protótipos dessas ferramentas, já se encontram disponíveis, e o processo de atualização da *web* está em pleno curso, e podemos notar que a *web* semântica trata da adoção de padrões de metadados e de compartilhamento desses padrões, de forma que se possa melhor utilizar o vasto repositório de informações disponível da *web* de maneira mais produtiva, ágil e significativa.

Mesmo sendo a proposta da *web* semântica claramente ligada à marcação dos dados na origem, (enquanto nesta tese buscou-se seguir o caminho da exploração da

semântica intrínseca dos textos dos documentos), podemos imaginar algumas convergências, principalmente quando levamos em conta o imenso acervo de documentos já estabelecido, disponível na web atual. A extração de sintagmas nominais podem embasar levantamentos terminológicos para a construção e a validação das ontologias em diversas áreas do conhecimento, que, uma vez construídas, podem auxiliar, como os tesauros, na busca por relacionamentos semânticos expressos em documentos, de forma a favorecer a escolha de descritores.

Finalizando, é importante notar que a linguagem XML há pouco apresentada é a infra-estrutura conceitual que oferece o suporte tecnológico às ferramentas de extração de sintagmas nominais, que são apresentadas no capítulo seguinte, relativo à metodologia desta pesquisa.

3.2 – Bibliotecas digitais

O escopo em que as metodologias e conceitos sugeridos nesta tese devem ser entendidos, fica claro, à luz das novas construções sociotécnicas para registro e utilização da produção intelectual humana, e estas estruturas podem ser entendidas no movimento de construção de grandes repositórios imbricados, multimídia e hipertextuais, de documentos. Segundo FOX e SORNIL (BAEZA-YATES e RIBEIRO-NETO, 1999, p. 414-432) a visão da grande maioria das pessoas em relação às bibliotecas digitais está fundamentada na idéia de bibliotecas tradicionais, em que os documentos são capturados e digitalizados. Mas bibliotecas digitais são muito mais do que coleções digitalizadas. A digitalização de coleções sugere apenas a mudança de arranjo particular espacial e estrutural, com ênfase na preservação baseada na digitalização, enquanto que as bibliotecas digitais pressupõem mais. Os autores colecionam algumas definições para bibliotecas digitais, dentre as quais que se seguem:

> *"Bibliotecas Digitais são construídas – coletadas e organizadas – por comunidades de usuários. Suas funcionalidades dão suporte às necessidades e usos de informação de uma comunidade. São uma extensão, melhoria e integração de uma variedade de instituições de informação enquanto espaços físicos, onde os recursos são selecionados, coletados, organizados, preservados e acessados para dar suporte a uma comunidade de usuários."*

e

> *"O nome genérico para estruturas federativas que provêem a usuários humanos acesso físico e intelectual para as imensas e crescentes redes mundiais de informação, codificada em formatos digitais e multimídia".*

Para FOX e SORNIL, a recuperação de informações é essencial para o sucesso das bibliotecas digitais, e nessa área tem surgido muitos dos esforços de pesquisa. Pulliam, citado por PISTORI (1999), define biblioteca digital como infra-estrutura de informações eletrônicas, na forma padronizada que permite o armazenamento distribuído de dados sobre uma região geograficamente grande, e que procura e acessa informações através de elos (*links* hipertextuais), oferecendo operações transparentes ao usuário final.

A despeito das definições, as bibliotecas digitais têm sido abordadas segundo os pontos de vista da gestão de bancos de dados, interação homem-máquina, ciência da informação, biblioteconomia, sistemas de informação multimídia, redes e comunicação e processamento de linguagem natural, como importantes elementos de transformação das atividades e do direcionamento de esforços dessas áreas. Na área acadêmica, já vimos percebendo um aumento de produtos da atividade científica em taxas exponenciais nas últimas décadas. Esse fenômeno tem relação direta com a disponibilização de acervos em formato eletrônico, em bases de dados e bibliotecas digitais. Some a isso o fato de que as barreiras temporais, geográficas e culturais são derrubadas pela interface onipresente e única dos navegadores da *web*, e pelas possibilidades da produção e consumo assíncronos.

Se observarmos as tendências, podemos supor que estas estruturas são as responsáveis por verdadeira revolução na produção científica, nas formas de disponibilizar, acessar e intercambiar documentos. Esse fato se junta ao fenômeno – retratado por vários autores (MATTELART, 2002; SCHAFF, 1990; TAKAHASHI, 2000; SARACEVIC, 1996) da explosão informacional que tem marcado a atividade científica como um todo.

Ainda no contexto das bibliotecas digitais e na questão da representação do conhecimento, notamos a ruptura nas formas tradicionais de registro de documentos, que não podem mais se basear na imagética trazida pelo livro como ente físico, objeto da biblioteca tradicional. Há o nascimento de uma nova linguagem e de nova noção de documento, com possibilidades ainda não vislumbradas, que nascem da comunhão de

usuários e tecnologias, modificando todo o jeito de produzir conhecimento. As propriedades que surgem nessa nova entidade, “documento digital”, estão ainda por ser descobertas, mas caminho possível é a exploração dos significados inerentes ao arranjo das idéias no texto, representadas pelas seqüências de palavras.

Podemos imaginar transformações na autoria e na utilização dos acervos que ocorrem quando tratamos da construção de hiperdocumentos (CAMPOS, 2001), ou quando exploramos as novas interfaces de acesso ao conhecimento, com o auxílio de agentes, e as possibilidades de *feedback* por parte dos usuários; ou mesmo com as novas metalinguagens e suas marcações semânticas (DECKER et al, 2000; BERNERS-LEE et al, 1999; HEARST in BAEZA-YATES & RIBEIRO-NETO, 1999, p. 257-323).

Embora as bibliotecas digitais que vêm sendo estruturadas apresentem apenas uma ínfima parcela da miríade de possibilidades que o novo meio digital oferece, acreditamos que, no esteio dessas transformações, observaremos grandes mudanças nas possibilidades de busca de informação, de interfaces evoluindo com o usuário e o surgimento de novas estratégias de processamento de linguagem natural, ligadas à exploração das semânticas intrínsecas e contextuais. É nessa frente de pesquisa que a presente tese buscou inserir-se.

Pode-se ainda realizar pequena digressão de cunho estratégico. No caso de nosso país, e devido à necessidade de desenvolvimento de tecnologias e metodologias adequadas a cada linguagem, podemos destacar a importância de pesquisas como a desenvolvida nesta tese na busca de autonomia e possível vanguarda no âmbito das comunidades lusófonas.

4 METODOLOGIA E FERRAMENTAS

Espera-se que neste momento todo o cabedal teórico necessário ao entendimento do contexto no qual se inseriu a presente pesquisa já tenha sido discutido e possa ser corretamente entendido, excetuando-se alguns conceitos específicos que podem ainda vir a ser introduzidos, pois decorreram dos resultados e da manipulação dos dados empíricos.

Neste capítulo, comentar-se-á inicialmente sobre os *corpora* adotados, tanto para a análise de eficácia das ferramentas utilizadas quanto para a análise da metodologia em si. Em seguida, será apresentada a metodologia prospectiva para extração de descritores, com suas etapas e produtos. Na seqüência, são apresentados as ferramentas e processos tecnológicos que dão suporte à metodologia e, ao final, são apresentados os instrumentos para avaliação dos descritores extraídos.

4.1 – Considerações sobre os *corpora* utilizados (material)

Os *corpora* utilizados no escopo desta dissertação foram dois, a saber:

a) O *corpus* escolhido inicialmente para a validação da extração automática de sintagmas nominais é composto pelos 15 textos utilizados pelo professor doutor Hélio KURAMOTO no escopo de sua tese de doutorado (1999). Esse *corpus* é apresentado no Anexo A de sua tese, e reproduzido parcialmente no **Anexo E** desta tese.

b) O *corpus* escolhido para a validação da metodologia utilizada nesta pesquisa para a escolha automática de descritores, consta de 60 documentos textuais de língua portuguesa (escolhidos dentre 75 inicialmente coletados). Os 75 documentos originalmente selecionados constituíram a totalidade dos artigos publicados durante os anos de 2002 e 2003 em duas publicações científicas de meio eletrônico, específicas da área de ciência da informação. Após o descarte dos documentos em línguas estrangeiras, e aqueles que, por sua estrutura ou tamanho, tornaram a análise proibitiva, permaneceram 60 documentos, compondo este *corpus*. As referências necessárias à identificação dos documentos, como a revista onde foram publicados, o título, o resumo e a autoria, se encontram no **Anexo A** desta tese.

As publicações escolhidas na seleção dos documentos foram a revista **DataGramaZero**[30] (29 documentos) e a revista **Ciência da Informação**[31] (31 documentos) do IBICT, porque são reconhecidas pelo programa *Qualis*[32] da CAPES[33] como publicações renomadas na área de ciência da informação. Além disso, estão disponíveis para acesso através da *web*, e formatos de armazenamento conhecidos e facilmente manipuláveis. Os documentos da revista eletrônica **DataGramaZero** estão disponibilizados no formato HTML, e os documentos da revista eletrônica **Ciência da Informação**, do IBICT, são disponibilizados nos formatos PDF e HTML.

O segundo corpus, de 60 documentos, foi disposto da seguinte maneira, para a aplicação da metodologia prospectiva e da metodologia consolidada:

- *Corpus* utilizado no teste inicial da metodologia prospectiva, composto por 6 textos provenientes da revista **DataGramaZero**, constantes no Anexo A deste documento, com os textos numerados de 1 a 6;
- *Corpus* utilizado na validação da metodologia consolidada, composto por dois conjuntos, a saber:
    - O primeiro com 30 textos, sendo que 29 provenientes da revista **DataGramaZero**, e 1 provenientes da revista **Ciência da Informação**, constantes no Anexo A deste documento com numeração de 1 a 30.Este *corpus* engloba aquele utilizado no teste inicial;
    - O segundo com 30 textos, todos provenientes da revista **Ciência da Informação**, constantes no Anexo A deste documento com numeração de 31 a 60.

Como se pode ver, o *corpus* completo de 60 textos, utilizado na metodologia consolidada, divide-se em duas metades com características peculiares – notadamente,

---

[30] Disponível na Internet no endereço: http://www.dgz.org.br.

[31] Disponível na Internet no endereço: http://www.ibict.br/secao.php?cat=Revista%20Ciência%20da%20Informação.

[32] *Qualis* é uma base de dados criada para a classificação dos periódicos e revistas utilizados pelos programas de pós-graduação, na divulgação da produção intelectual de seus docentes e alunos. Acessível na Internet a partir do endereço: http://qualis.capes.gov.br/.

[33] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Acessível na Internet a partir do endereço: http://www.capes.gov.br/ .

porque provêem de publicações diferentes. Ao processar a metodologia consolidada de maneira isolada em cada um dos corpora, pudermos intuir alguma diferença quando da apresentação dos resultados finais.

A necessidade do contexto temático específico se justificou pela própria característica da metodologia de escolha de descritores, que utilizou um tesauro também específico. Essa metodologia, fortemente contextual, deve ser adotada para uma área de conhecimento especificada a priori.

A escolha dos 60 textos do *corpus* atendeu ainda a alguns critérios quantitativos e qualitativos:

- Quantidade compatível com a possibilidade de processamento em curto período, utilizando as ferramentas atualmente disponíveis;
- Quantidade significativa, de forma a ressaltar a ampliação das possibilidades de processamento, em comparação com a extração manual de sintagmas nominais;
- Atualidade dos textos (2002/2003), para que os SNs extraídos refletissem conceitos contemporâneos e salientassem a possibilidade e necessidade de atualização dos tesauros utilizados na metodologia;
- Fidelidade às temáticas mais reconhecidas como pertencentes ao campo da Ciência da Informação.

Foi difícil, porém, estabelecer o que seria uma amostra significativa neste caso. Se considerarmos todo o universo de publicações com temática relacionada ao campo da ciência da informação, este seria virtualmente ilimitado. São dezenas de publicações pertinentes, muitas delas apresentando seus documentos em formato digital. Nesse caso, a escolha da quantidade de textos se pautou pelos critérios anteriormente descritos, tendo sido então a quase totalidade dos textos em português, publicados por duas revistas eletrônicas da área.

Os resultados das análises realizadas nesse conjunto de documentos, porém, se tomados qualitativamente, apontaram para conclusões seguras que permitiriam a avaliação da viabilidade da metodologia.

4.2 – A metodologia prospectiva

É importante ressaltar que a metodologia delineada neste capítulo foi apenas prospectiva, e foi aplicada apenas a um conjunto reduzido de documentos do *corpus* total. O aprendizado adquirido no teste inicial produziu modificações a serem incorporadas em novas versões – hipoteticamente melhoradas – dessas metodologias. A FIG. 8 a seguir exemplifica o processo empírico desta pesquisa:

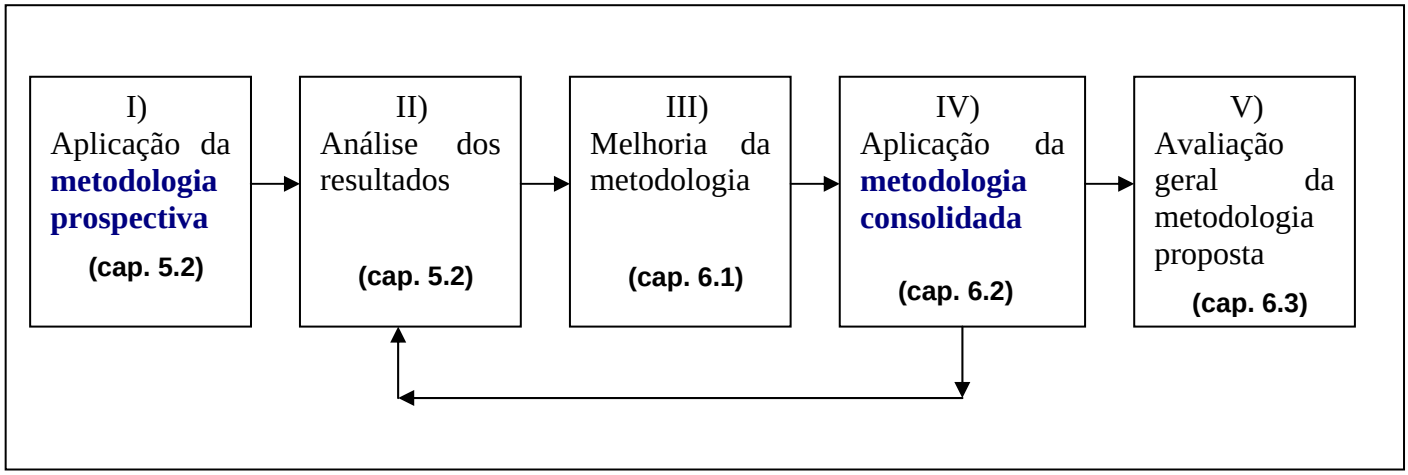


Figura 8 – Seqüência de aplicação e avaliação da metodologia

I) Inicialmente, a metodologia apresentada neste capítulo foi aplicada ao *corpus* inicial. II) Os resultados foram analisados, gerando subsídios para a III) melhoria da metodologia prospectiva e a construção da metodologia consolidada. IV) Esta foi aplicada ao *corpus* completo, e então foi V) avaliada novamente. A pesquisa não se esgota nos resultados obtidos com a metodologia consolidada utilizada nesta tese, e abre margens para melhorias sucessivas em trabalhos futuros.

Cabe ainda ressaltar que, nesse ciclo de aplicações, avaliações e alterações, certas etapas da metodologia puderam ser suprimidas ou novas etapas introduzidas, e algumas delas possivelmente automatizadas. Na metodologia consolidada prescindimos da análise de eficácia que ora é realizada, em contexto avaliativo.

Para maior vinculação dos objetos de estudo à metodologia, cabe associar a cada um dos objetivos apresentados anteriormente os métodos de trabalho utilizados para sua consecução. Seguem adiante os objetivos desta tese, como estabelecidos na introdução, e os passos que foram necessários às suas respectivas consecuções.

É importante ressaltar que, para atingir o objetivo geral, apresentado no item A), dependemos do sucesso da verificação da extração automática, objetivo específico apresentado no item B). O objetivo específico apresentado no item C) é apenas uma possibilidade considerada, como um subproduto do processo de extração e tratamento dos sintagmas nominais. Apesar de delineado nas figuras que representam a metodologia, não será explorado efetivamente no escopo desta pesquisa.

A) Para o objetivo geral: "*Desenvolver uma metodologia para a escolha automática de descritores para documentos textuais digitalizados em língua portuguesa, utilizando as estruturas lingüísticas conhecidas como sintagmas nominais*"; pretendem-se perfazer os seguintes passos, ilustrados na FIG. 9, e em seguida explicitados e comentados:

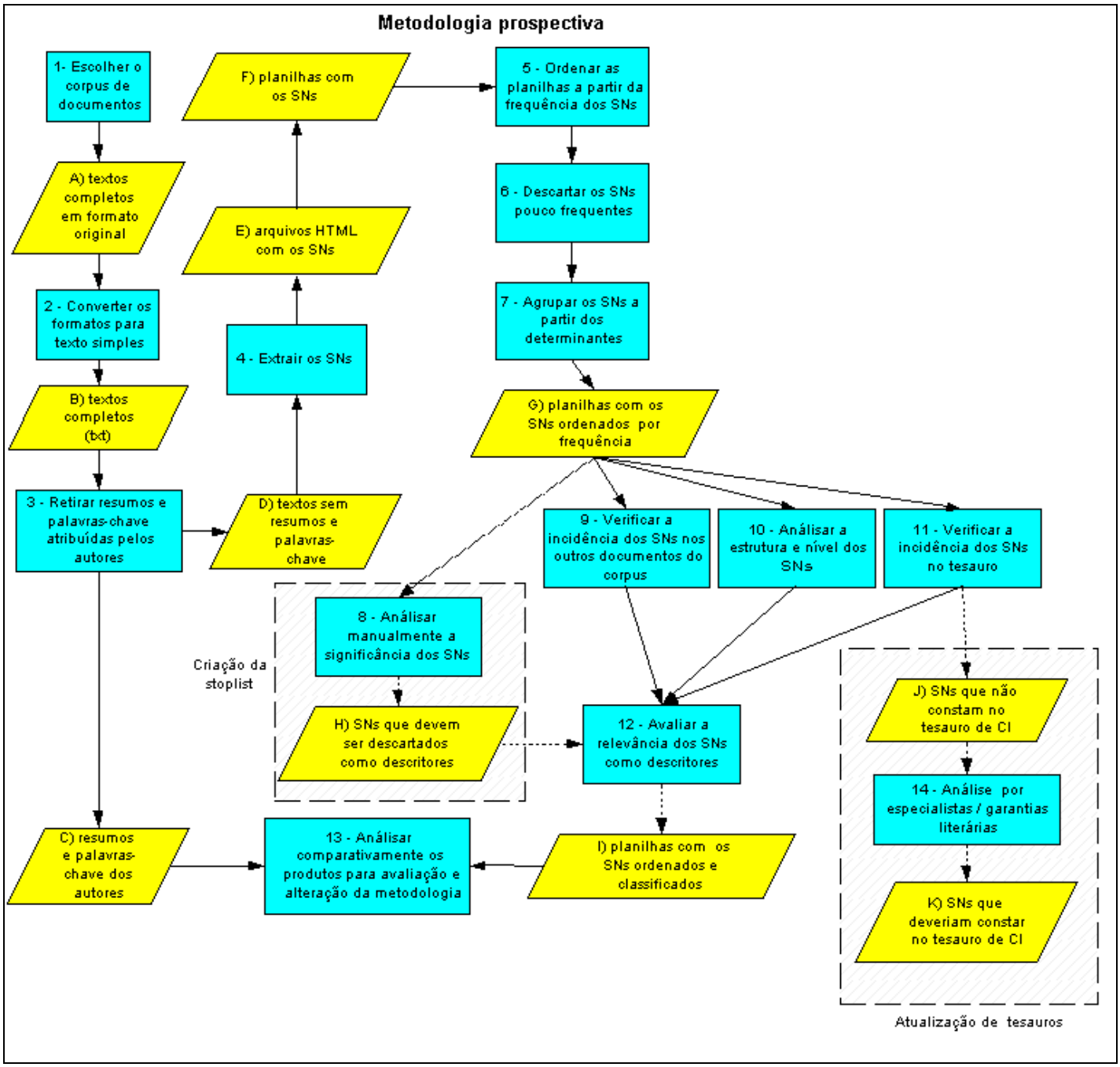


Figura 9 – Fluxograma da metodologia prospectiva

A FIG. 9 apresenta os passos gerais da metodologia, que são detalhados nos itens abaixo. Os processos (em azul-claro) são identificados por seus respectivos números e os produtos (em amarelo) estão identificados por letras.

1. **Escolher um *corpus* significativo de documentos reconhecidamente inseridos dentro de uma área de conhecimento, como universo empírico desta pesquisa;**

A importância de delimitar o assunto dos textos em uma área específica – no caso, a ciência da informação – foi devida à necessidade de escopo e contextualização. Um dos passos metodológicos previstos para a validação dos descritores pré-escolhidos foi realizado com a utilização de tesauro, no caso, da área de ciência da informação (CNPq/IBICT, 1989). Da teoria apresentada nos fundamentos lingüísticos, também pudermos retirar justificativas que destacavam a importância dos contextos para a escolha dos descritores.

O *corpus* de textos escolhido para análise proveio, como já citado, de publicações específicas da área de ciência da informação, como está detalhado em subseção a seguir. Não há, entretanto, restrições de aplicabilidade da metodologia para documentos textuais oriundos de outras áreas do conhecimento, desde que sejam adotadas as ferramentas adequadas – tesauros específicos do respectivo campo de conhecimento – e sejam processados documentos com contextos semelhantes.

O produto desta primeira etapa foram os textos originais em formato digital (produto **A**).

2. Converter os formatos de arquivo para texto simples;

As ferramentas tecnológicas utilizadas nesta metodologia necessitava de documentos submetidos em formato de arquivos de texto simples. Como os documentos digitalizados – notadamente na *web* – se encontravam usualmente em formatos mais complexos, como PDF[34] ou HTML, esses documentos precisam ser convertidos para o formato texto simples (produto **B**). As ferramentas de software utilizadas na conversão são apresentadas adiante.

3. Retirar os resumos e as palavras-chave atribuídas pelos autores

A separação do corpo do texto dos documentos (produto **D**) e das palavras-chave a eles atribuídas e dos resumos preparados pelos autores (produto **C**) foi um artifício metodológico utilizado apenas para possibilitar a análise posterior do sucesso do procedimento automático de extração de descritores, através da comparação simples dos

[34] O PDF, ou Portable Document Format, é um formato proprietário da empresa Adobe (http://www.adobe.com) que, entretanto, disponibiliza gratuitamente o visualizador dos arquivos (Adobe Acrobat Reader). Para transformar os documentos em padrão texto simples, é necessário, entretanto, o software completo (Adobe Acrobat).

SNs atribuídos automaticamente e as palavras-chave atribuídas pelos autores dos artigos. Esse passo não é mais necessário na medida em que a metodologia tenha sido avaliada e considerada bem sucedida.

4. **Extrair os sintagmas nominais do corpo do texto**

Os SNs foram extraídos dos documentos em formato de texto simples através de processo quase que inteiramente automático, utilizando as ferramentas que – conjuntamente com o processo de extração – foram detalhadas adiante. O produto da extração constituem arquivos em formato HTML contendo os SNs na ordem de sua ocorrência nos textos originais (produto **E**). A partir destes arquivos em formato HTML (produto **E**), foram criadas planilhas (produto **F**) utilizando o software MICROSOFT EXCEL (apresentado adiante). Essas planilhas contêm pastas específicas para cada texto, onde serão realizadas todas as operações posteriores.

5. **Ordenar os SNs nas planilhas através da verificação da freqüência de ocorrência dos sintagmas nominais nos documentos;**

Após o agrupamento, os SNs foram inicialmente ordenados nas pastas das planilhas de acordo com a freqüência de ocorrência de cada um no corpo do documento.

6. **Descartar os SNs que apresentavam freqüências de ocorrência inferiores a um patamar preestabelecido;**

Os SNs que apresentavam uma freqüência inferior a certo patamar foram considerados descritores insignificantes e descartados para as operações posteriores. O patamar estabelecido depende de análises que levassem em consideração a relevância dos SNs extraídos em cada faixa de freqüências, além do tamanho dos textos originais.

7. **Agrupar os SNs remanescentes a partir dos determinantes de suas formas “canônicas”, e reordená-los;**

Nesta etapa, ainda realizada manualmente, os SNs que diferiam apenas pelos determinantes iniciais foram agrupados e representados unicamente pela a soma das freqüências, e o representante do agrupamento assumiu a forma canônica, segundo as normas de construção de tesauros. Os determinantes (artigo, pronome ou numeral) foram usualmente composto pelas estruturas apresentadas na TAB. 5:

{a | as | o | os |
dois | três | quatro | cinco | ... | mil |
essa | essas | esse | esses | esta | estas | este | estes | aquela | aquelas | aquele | aqueles |
mesma | mesmas | mesmo | mesmos | tal | semelhante |
meu | meus | teu | teus | tua | tuas | seu | sua | nosso | nossos | vosso | vossa |
vossos | vossas | seus | suas |
um | uma | uns | umas | alguma | algumas | algum | alguns | nenhum | nenhuma |
toda | todas | todo | todos | cada | qualquer |
certa | certas | certo | certos | outra | outras | outro | outros | muita |
muitas | muito | muitos | pouca | poucas | pouco | poucos }

Tabela 5 – Determinantes comuns

Essas planilhas ordenadas com os SNs agrupados, tendo sido descartados os de freqüência abaixo de um patamar preestabelecido (produto **G**), foram utilizadas nos três passos posteriores. Esses passos acrescentaram informações aos SNs de modo a embasar a decisão sobre a relevância de cada um.

8. Analisar manualmente os SNs pré-escolhidos e decidir sobre a sua relevância como descritores, para fins de construção de uma *stoplist*;

Esta etapa opcional pode ser adotada para otimizar o funcionamento posterior da metodologia automática. Os SNs – escolhidos através de julgamento humano – que vierem compuseram a lista de *stopwords* (produto opcional **H**) puderam ser descartados de qualquer conjunto posterior de SNs extraídos. Os passos posteriores foram realizados ainda com as planilhas representadas no produto **G**.

9. Verificar a incidência dos SNs nos outros documentos do *corpus*;

A análise da incidência dos SNs no conjunto de documentos do *corpus* foi um dos critérios considerados na análise da relevância. Pressupôs-se que quanto maior a incidência de um SN no conjunto de documentos, menor a sua relevância como descritor.

10. Analisar a estrutura e o nível dos SNs;

A análise da estrutura sintática e do nível dos SNs, como apresentado na subseção 2.1.6, foi um dos critérios a serem considerados na análise da relevância. Pressupôs-se

que quando a estrutura e o nível do SN estivessem diretamente relacionados à sua relevância como descritor.

11. Verificar a ocorrência destes SNs – de forma total ou parcial – em tesauro específico;

Uma vez que tenham sido escolhidos os SNs pré-candidatos a descritores, houve necessidade de classificá-los segundo suas estruturas sintáticas e segundo seus níveis (como apresentado nas seções 2.1.6 e 2.1.7), para subsidiar o processo de escolha dos SNs mais significativos. Foi também necessário verificar sua ocorrência em um tesauro da área do conhecimento a que pertencem os documentos do *corpus*.

Esta etapa da metodologia foi ainda realizada manualmente, mas pode ser implementada através de processo automatizado no futuro. O resultado dessas etapas foi incorporado às tabelas das planilhas de análise (produto **G**) como informações relativas a cada sintagma nominal.

12. Avaliar a relevância dos SNs como descritores;

Neste ponto talvez resida uma das partes mais importantes da metodologia prospectiva. A lógica para escolha dos sintagmas nominais mais significativos e relevantes como descritores dos documentos foi estabelecida através da avaliação dos dados empíricos, gerando subsídios para o estabelecimento da heurística de escolha, a ser adotada na metodologia consolidada. Para essa avaliação, relacionaram-se a relevância dos SNs como descritores e os fatores: a) freqüência de ocorrência do SNs no texto do documento; b) a incidência dos SNs no conjunto de documentos; c) seus níveis; d) suas estruturas sintáticas e e) sua ocorrência no tesauro da área.

As considerações advindas do cálculo das freqüências foram embasadas na teoria subjacente a alguns dos algoritmos de extração de palavras-chave, baseados na lei de Zipf, que estabelece relação inversa entre a freqüência de ocorrência das palavras-chave e sua significância como descritores. Foram estes os algoritmos: a) análise de freqüência simples com descarte dos picos; b) análise de pesos relacionados à freqüência inversa nos documentos; e c) análise de valor discriminatório dos termos (como apresentados na seção 2.2.2 desta tese).

Houve necessidade de se fazerem adaptações necessárias ao fato de não se manipularem palavras-chave, mas sim sintagmas nominais.

O tesauro foi utilizado para a validação dos sintagmas selecionados no contexto do assunto escolhido para o *corpus*. Como o melhor tesauro de língua portuguesa conhecido e disponível no momento na área de ciência da informação se encontra bastante defasado (CNPQ/IBICT, 1989), esta etapa na escolha dos descritores foi analisada na aplicação inicial da metodologia e sua utilização efetiva avaliada para aplicações posteriores. Aventou-se a possibilidade da consulta em outros tesauros disponíveis, em outras linguagens, como o tesauro da ASIS[35], de acordo com os resultados.

Para verificar a incidência de cada sintagma nominal no tesauro, considerou-se a verificação – para cada sintagma nominal – da ocorrência da estrutura de maior nível (como apresentado na seção 2.1.7), para então se procurar pela ocorrência dos sintagmas nominais aninhados, sucessivamente, e finalmente dos lexemas componentes.

Adotada uma *stoplist* (passo **8**, produto **H**), os SNs presentes puderam ser descartados do conjunto dos candidatos a descritores.

O resultado desta etapa foi a ordenação dos SNs sob os critérios de relevância estabelecidos, nas tabelas das planilhas com os candidatos a descritores (produto **I**). A partir dessa ordenação pôde-se escolher a quantidade desejada de descritores – essa discussão será realizada adiante na seção 4.4.1.

13. Analisar comparativamente os produtos – palavras-chave e resumos dos documentos originais e os SNs escolhidos como descritores – para avaliação da metodologia

Após as etapas de escolha dos sintagmas nominais candidatos a descritores; esses serão comparados às palavras-chave e aos resumos dos documentos originais do *corpus* para o primeiro julgamento de relevância e análise de sucesso da metodologia. Essa análise baseou-se no julgamento do autor desta pesquisa. Essas comparações embasaram o relacionamento entre a significância percebida dos SNs como descritores e suas características intrínsecas, ou relativas à freqüência de ocorrência. As ferramentas matemáticas para permitir a comparação e avaliação serão apresentadas na seção 4.4.2.

14. Análise por especialistas / garantias literárias.

Este item da metodologia, identificado em processo a parte, foi opcional, e não foi levado em consideração na aplicação da metodologia. Esta etapa é parte do objetivo específico **C)** desta tese, exposto adiante.

Os resultados de cada uma destas etapas na aplicação da metodologia prospectiva puderam, eventualmente, determinar mudanças para as próximas aplicações. Como foi apontado, a metodologia prospectiva que ora se apresenta foi apenas uma proposta preliminar a ser testada no *corpus* inicial reduzido.

Os títulos dos artigos do *corpus* selecionado (produto **A**) e suas palavras-chave e resumos (produto **C**) estão indicados no **Anexo A** desta tese.

B) Para o objetivo específico: "*testar a eficácia relativa de um conjunto de ferramentas para a extração automática de sintagmas nominais, comparando a extração automática com a extração manual*"; pretendeu-se perfazer os seguintes passos, em seguida explicitados e comentados:

1. **Submeter ao processo de extração automática de sintagmas nominais os quinze textos utilizados por KURAMOTO (1999) na sua tese de doutorado;**
2. **Validar a extração automática dos sintagmas nominais através da comparação de resultados da performance das ferramentas com os sintagmas manualmente extraídos;**

Nesta etapa, que de fato precedeu as etapas detalhadas no objetivo geral, procurou-se avaliar a performance obtida pelas ferramentas de extração automática em comparação com a extração manual. Os critérios de comparação e categorias de análise foram:

- Tempo gasto na extração dos sintagmas nominais;
- Quantidade de sintagmas nominais identificados;
- Qualidade da identificação dos sintagmas nominais;

---

[35] Disponível na Internet no endereço http://www.asis.org/Publications/Thesaurus/isframe.htm.

Os resultados dessa validação foram levados em conta na avaliação da performance da metodologia estabelecida para o objetivo geral, inteiramente dependente do sucesso na identificação dos SNs.

C) Para o objetivo específico: "*analisar a possibilidade de a metodologia proposta ser utilizada para o auxílio na atualização de tesauros de língua portuguesa*" buscou-se utilizar um subproduto da aplicação da metodologia mencionada no objetivo geral para levantamento terminológico nos *corpora*, levantamento este que poderia embasar, posteriormente, o processo de atualização semi-automática de tesauros através da análise de textos em um domínio do conhecimento, além da construção de *stoplists*.

Este processo paralelo, representado na FIG. **9** em separado, aconteceria da seguinte forma:

1. **O conjunto dos SNs minimamente freqüentes não constantes no tesauro pode ser armazenado em estrutura para posterior validação como descritores em potencial (o conjunto J);**
2. **A análise dos termos reunidos por uma comunidade de pesquisadores da área pode decidir por aqueles que devem ser incorporados (o conjunto K).**

No caso específico desta etapa, houve restrições temporais e conjunturais que permitiram somente o levantamento da coleção de termos, sem que fosse possível a análise quanto à incorporação no tesauro. Esta funcionalidade potencial do instrumento fica apenas como uma indicação de trabalho futuro.

Cabe ressaltar novamente que a proposta metodológica apresentada para a consecução do objetivo geral foi prospectiva, e sofreu alterações à medida que os dados empíricos fossem manipulados e analisados. Esse trabalho, entretanto, não teria sido possível sem as ferramentas de extração automática que, assim como o *corpus* de validação da extração automática, foram gentilmente cedidas pelos proprietários e desenvolvedores. Em seguida passamos à descrição dessas ferramentas e os processos envolvidos em sua utilização.

4.3 – Ferramentas utilizadas

O trabalho de análise necessário à consecução da metodologia acima descrita pressupôs um enorme esforço computacional, ao longo do processo. Para que fosse

possível a análise dos descritores, os SNs tiveram que ser extraídos automaticamente e de forma bastante veloz, mas esse processo foi composto por várias etapas. A FIG. 10 explicita os relacionamentos entre os processos e as ferramentas de *software*:

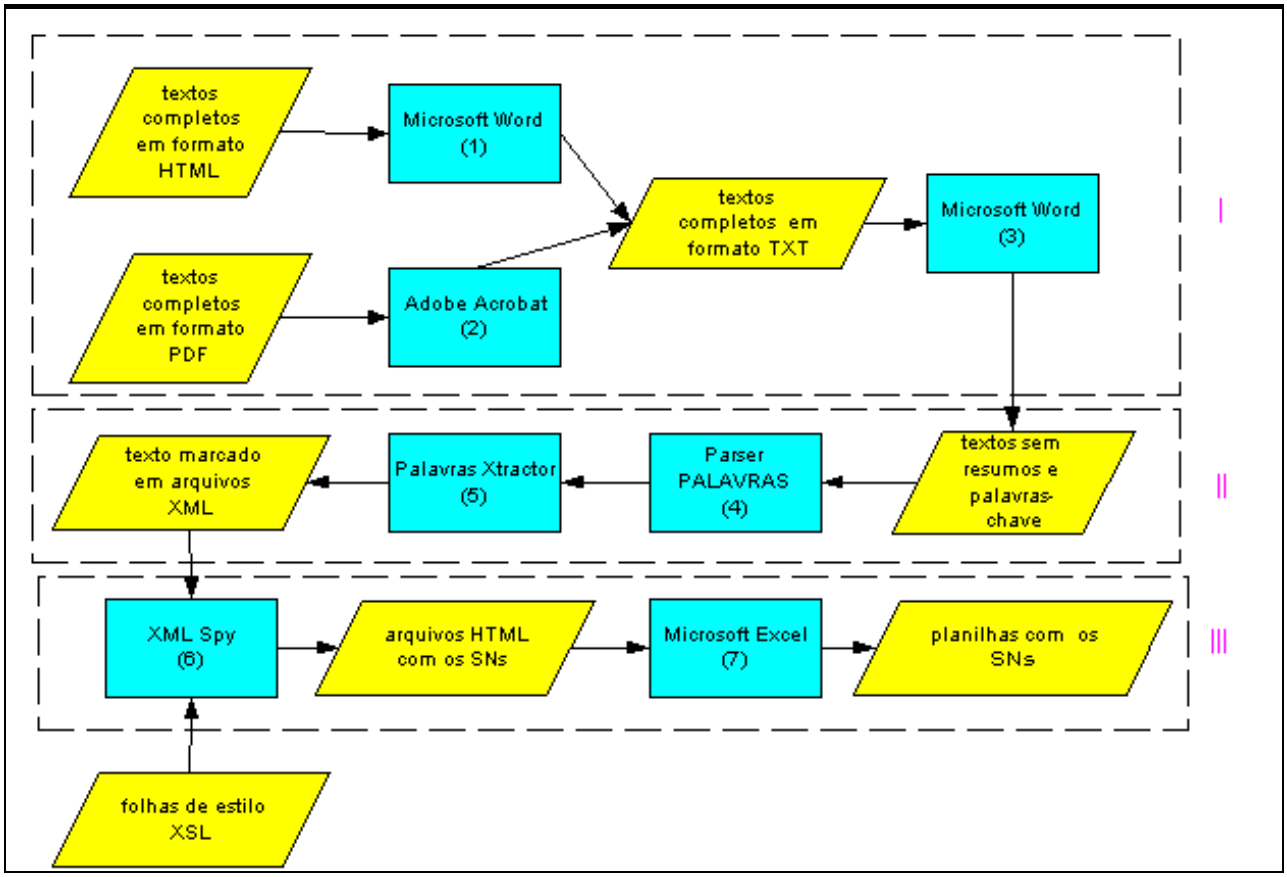


Figura 10 – Ferramentas utilizadas na metodologia

As ferramentas foram utilizadas na seguinte seqüência:

I. Os textos dos *corpora* foram escolhidos pelo autor desta tese e transformados em formato de texto simples, sem caracteres especiais, utilizando as ferramentas ADOBE ACROBAT[36] e MICROSOFT WORD[37];

II. Em seguida, os textos tratados foram submetidos sucessivamente ao processamento do analisador sintático (*parser*) “PALAVRAS”, da Southern University of Denmark, e ao software “Palavras Xtractor”, desenvolvido em conjunto pela Universidade do Vale do Rio dos Sinos (Unisinos) de São

[36] Informações no endereço da Internet http://www.adobe.com/products/acrobat/main.html.

[37] Informações no endereço da Internet http://office.microsoft.com/pt-br/FX010857991046.aspx.

Leopoldo, RS, e a Universidade de Évora, em Portugal, tendo como resultado os documentos sintaticamente marcados em arquivos XML;

III. Após a identificação sintática das palavras dos textos, foi utilizado o software XML SPY[38] para aplicação da transformação XSL nos arquivos XML com uma folha de estilos específica (como explicado na seção 3.1.2), para extração de arquivos HTML com os SNs, e estes SNs foram tratados estatisticamente utilizando o *software* MICROSOFT EXCEL[39].

Os pesquisadores da Unisinos e da Universidade de Évora cederam, para os propósitos desta tese, interface integrada através da qual grande parte do processamento automático envolvido; o desempenhado pelo *parser* do *site* dinamarquês foi realizado, durante os meses de agosto e setembro de 2003. Em seguida vamos descrever em mais detalhes as principais ferramentas, utilizadas na fase **II** descrita na FIG. 10.

4.3.1 – O VISL e o processador "Palavras"

A Southern University of Denmark desenvolveu e tornou público uma ferramenta de processamento morfossintático de textos digitalizados em português chamada "Palavras", que faz parte de um conjunto de ferramentas multilingües chamado **VISL**[40] (Virtual Interactive Syntax Learning).

No VISL, para cada idioma suportado, há ferramentas que operam em modo automático ou semi-automático, nas quais um usuário submete sentenças ou textos completos em uma das linguagens admitidas (dentre as quais o português) e recebe de volta os textos marcados. As análises podem ser feitas em diferentes níveis (morfológico, sintático, semântico) e o site VISL oferece uma interface gráfica que permite aos usuários diversidade de opções de consulta em várias formas de visualização, como textos simples, árvores sintáticas ou marcação com cores (BICK, 1996, 2001 e 2003). O processador Palavras é baseado em uma interface de páginas HTML, *scripts* CGI[41],

---

[38] Informações no endereço da Internet http://www.altova.com

[39] Informações no endereço da Internet http://office.microsoft.com/pt-br/FX010858001046.aspx

[40] Disponível no endereço da Internet: http://visl.sdu.dk/visl/.

[41] Informações no endereço da Internet http://searchdatabase.techtarget.com/sDefinition/0,,sid13_gci213846,00.html

aplicativos Java[42] e *scripts* em PERL[43]; um conjunto de ferramentas chamadas de "*Constraint Grammar*" (gramática de restrições), para a análise automática dos textos submetidos.

A abordagem da gramática de restrições analisa o texto na perspectiva dos lexemas, grupos de palavras e das próprias orações, nos níveis ortográfico, sintático e semântico. Cada oração e seus componentes são marcados, inicialmente, em todas as suas possibilidades sintáticas e semânticas, através do uso do analisador morfológico baseado em léxico. Essa lista provisória e cheia de ambigüidades é então processada através da análise, no contexto da sentença, de que formas sintáticas são impossíveis (que são descartadas), quais são possíveis (que persistem) e quais são mandatárias (que são escolhidas). Através da aplicação sucessiva e repetida dessas regras, resolvem-se paulatinamente as ambigüidades da classificação sintática na sentença e, ao final, resta apenas uma e somente uma possível classificação para cada palavra, o que caracteriza a abordagem como extremamente robusta. Mesmo em textos sintaticamente mal-construídos, há algum resultado ao final, mesmo que incorreto (BICK, 1996, 2001 e 2003; VISL, 2003).

O *parser*, na versão atual, apresenta os seguintes módulos e níveis de análise:

1. Um **analisador morfológico** que trata as categorias morfossintáticas, inflexões, derivações, expressões fixas e os verbos. O analisador utiliza um léxico manualmente construído composto de 70.000 entradas, representando cerca de 50.000 lexemas;

2. Um **desambiguador morfológico** que utiliza 1700 regras da gramática de restrições;

3. Um **"mapeador" sintático** que associa às palavras as possíveis funções sintáticas, utilizando 400 regras de atribuição de funções baseadas em contexto;

4. Um **desambiguador sintático** que utiliza 1500 regras da gramática de restrições;

5. Um **desambiguador de valência** e o **desambiguador de classes semânticas**, ainda não totalmente operacionais, baseados em 2200 regras da gramática de restrições;

---

[42] Informações no endereço da Internet
http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci212415,00.html

A submissão sucessiva a estes módulos permite que se obtenha um resultado único para a classificação morfossintática, como descrito.

A FIG. 11 mostra o resultado do *parsing* do excerto de documento, com o texto "Considerações iniciais na companhia de Edgar Morin", submetido ao processamento morfossintático do processador PALAVRAS, na qual podemos ver a análise realizada:

```
Considerações  [consideração] <*> N F P
iniciais       [inicial] ADJ M/F P      [inicial] N F P    [iniciar] V PR 2P IND VFIN
em             [em] <*> <sam-> PRP
a              [o] <-sam> <artd> DET F S
companhia      [companhia] N F S
de             [de] PRP
Edgar=Morin    [Edgar=Morin] <*> PROP M/F S/P
(...)
```

Figura 11 – Resultado de um texto submetido ao processador PALAVRAS

Observamos na FIG. 11 que em cada linha do arquivo de saída aparecem a forma do lexema, tal qual ocorre no texto submetido, e em seguida a forma canônica do lexema e por fim a classificação morfossintática deste. No exemplo acima, temos para o lexema "considerações" a forma canônica "consideração", e as classificações N (substantivo), F (feminino), P (plural); e na segunda linha, as três classificações possíveis para a palavra "iniciais", a saber, ADJ (adjetivo – "inicial"), N (substantivo – "inicial"), e V (verbo – "iniciar"), com suas inflexões e gêneros respectivos. Para consultar o conjunto de símbolos completo do VISL, pode-se visitar o endereço na Internet: http://visl.sdu.dk/visl/pt/info/symbolset-manual.html.

Além da possibilidade da submissão de textos e sentenças do usuário, o *site* do VISL ainda mantém grandes *corpora* de sentenças previamente assinaladas, disponíveis para estudiosos e pesquisadores. Além disso, os usuários têm acesso a dicionários e ferramentas de tradução de textos.

Uma das possibilidades de marcação oferecidas pelas ferramentas do *site* indica as categorias gramaticais e a função de cada palavra no contexto de uma oração. Através desta marcação e processamento posterior, é possível extrair os sintagmas nominais das sentenças de um texto. Esse pós-processamento pode ser feito manualmente, através da

[43] Informações no endereço da Internet http://searchenterpriselinux.techtarget.com/sDefinition/0,,sid39_gci214291,00.html

análise das funções marcadas, ou pode ser automatizado. Na subseção seguinte será apresentada a abordagem para esse pós-processamento baseada no padrão XML e nas folhas de estilo XSL.

O projeto VISL é altamente orientado a produtos e processos, uma vez que novas ferramentas têm sido constantemente disponibilizadas gratuitamente na Internet na medida em que os protótipos se mostre funcionais. A grande falha do processador PALAVRAS é a fraca interoperabilidade do sistema, causada pela falta de padrões para os arquivos de saída, além de problemas específicos no vocabulário do sistema, que ainda não permitem uma análise sintática próxima do nível de perfeição esperado de um analisador humano. Podemos esperar, entretanto, que essa situação venha a melhorar, haja vista que o processador está sendo continuamente refinado.

4.3.2 – A extração automática de SNs

A partir da ferramenta computacional “Palavras” do VISL, o Laboratório de Engenharia da Linguagem do Programa Interdisciplinar de Pós Graduação de Computação Aplicada da Universidade do Vale do Rio dos Sinos, sob a coordenação da professora doutora Renata Vieira, em parceria com o departamento de Informática da Universidade de Évora, de Portugal, desenvolveu, no escopo do projeto de cooperação DIRPI (PROJETO DIRPI, 2001), um conjunto de programas de interface e de pós-processamento dos resultados, chamados internamente de “Palavras Xtractor”. Os programas estabelecem acesso ao *site* VISL, enviam textos para o analisador sintático PALAVRAS para o português (BICK, 2000 apud GASPERIN et al, 2003). O resultado do processamento dos arquivos de texto submetidos ao analisador é convertida em um conjunto de três arquivos em formato XML: arquivo com o conjunto das palavras, arquivo com as categorias morfossintáticas, e de agrupamentos; exemplificados a seguir:

```
<word id="word_27">Desenvolver</word>
<word id="word_28">capacidades</word>
<word id="word_29">de</word>
<word id="word_30">controle</word>
<word id="word_31">e</word>
<word id="word_32">incremento</word>
<word id="word_33">de</word>
<word id="word_34">o</word>
<word id="word_35">fluxo</word>
<word id="word_36">de</word>
<word id="word_37">o</word>
<word id="word_38">conhecimento</word>
```

Figura 12 – Arquivo de palavras

A FIG. 12 exemplifica um trecho do primeiro dos três arquivos, de terminação “words.xml”. Esse arquivo contém, em cada linha, os lexemas do texto original, etiquetados pelas *tags* <word>, cada uma trazendo a informação do número de ordem da palavra na seqüência do texto. No trecho, exemplificado acima, vemos a análise do excerto de texto “Desenvolver capacidades de controle e incremento do fluxo do conhecimento”.

A FIG. 13 exemplifica um trecho do segundo dos três arquivos, de terminação “pos.xml”, que contém, entre conjuntos de *tags* <word>, informações relativas às categorias morfossintáticas respectivas a cada um dos lexemas do texto original.

```
<word id="word_27">
<v canon="desenvolver">
<inf/>
</v>
</word>
<word id="word_28">
<n canon="capacidade" gender="F" number="P"/>
</word>
<word id="word_29">
<prp canon="de"/>
</word>
<word id="word_30">
<n canon="controle" gender="M" number="S"/>
</word>
```

Figura 13 – Arquivo de Categorias Morfossintáticas

No trecho exemplificado acima, podemos observar a análise das quatro primeiras palavras do excerto apresentado na FIG. 12.

E finalmente a FIG. 14 exemplifica um trecho do terceiro dos três arquivos, de terminação “chunks.xml”, que contém informações sobre as estruturas sintáticas das sentenças do texto original – etiquetados pelas *tags* <sentence> - que, por sua vez, fazem parte de um parágrafo – etiquetado pelas *tags* <paragraph>.

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1" span="word_1..word_26">
<chunk id="chunk_1" ext="sta" form="fcl" span="word_1..word_25">
<chunk id="chunk_2" ext="subj" form="np" span="word_1..word_2">
<chunk id="chunk_3" ext="n" form="adj" span="word_1">
</chunk>
```

Figura 14 – Arquivo de agrupamentos

O excerto acima exemplificado descreve o início do primeiro parágrafo, com uma sentença que contém as palavras 1 a 26 do texto e alguns agrupamentos (*chunks*) que

ocorrem nessa sentença. Nos agrupamentos é que se identificam os lexemas que compõem os sintagmas nominais.

A partir destes três arquivos em formato XML, gerados para cada documento submetido, pode-se trabalhar com desenvoltura, em comparação com o arquivo de saída do site VISL, pois através do uso de folhas de estilo (XSL) específicas é possível então extrair os sintagmas nominais de qualquer texto ou *corpus* da língua portuguesa. Assim como são extraídos os sintagmas nominais, é possível extrair outras instâncias morfossintáticas, como sintagmas verbais, verbos, pronomes, e outros, dependendo do interesse da pesquisa em questão, bastando para tanto o desenho de uma nova folha de estilo.

Os sintagmas nominais utilizados nesta tese foram obtidos, utilizando-se a folha de estilo específica para extração de sintagmas nominais, cedida gentilmente pela pesquisadora da Unisinos Claudia Camerini Correa Perez.

Finalmente, cabe registrar que o equipamento utilizado para todo o processamento local – que exclui aquele realizado pela interface oferecida pela Unisinos – foi um computador AMD Athlon XP 2600+ de 256 MB de memória RAM, gentilmente cedido pelo Núcleo de Informação Tecnológica e Gerencial (NITEG), da Escola de Ciência da Informação - UFMG. Não é o equipamento ideal, entretanto, pois o processamento eficaz de documentos maiores exigiria equipamento mais veloz e com mais recursos de memória.

4.4 – Critérios de corte e avaliação dos descritores extraídos

Para que a metodologia proposta anteriormente fosse corretamente parametrizada e avaliada, foi necessário estabelecer os critérios de corte – para estabelecer a quantidade desejada de descritores – e os instrumentos de avaliação da relevância, determinando a viabilidade do processo. Esses tópicos são apresentados a seguir:

4.4.1 – Considerações gerais sobre a quantidade de descritores extraídos

O primeiro parâmetro a ser estabelecido para a metodologia automática de atribuição de descritores a documentos foi a quantidade desejada desses. Embora a limitação última possa ser considerada a quantidade total de SNs extraídos, isto pode não ser desejável, pelas razões que serão expostas em seguida. Deve-se procurar responder

à questão: qual seria um número razoável de descritores para um determinado documento textual? Ou seja, qual é a exaustividade desejada para o índice?

LANCASTER (1993, p. 20-41), considerando o uso de palavras-chave, aponta para a grande variação nas faixas de termos selecionados, e aconselha que não sejam estabelecidos limites absolutos para as quantidades, e sim parâmetros indicativos, e que o grau de importância do item para os usuários do sistema justificaria uma indexação mais ou menos exaustiva.

Usualmente, observamos quantidades que variam entre 5 a 25 descritores por documento, mas em documentos de algumas áreas do conhecimento – como a química, por exemplo – não é incomum observarmos uma centena ou mais de descritores. LANCASTER ainda aponta, no caso da indexação manual, o fenômeno da diminuição da coerência da indexação, a medida que aumenta a quantidade de termos índices escolhidos (1993, p. 61-74). Entretanto, essa coerência certamente aumentará se o processo for automatizado e seguir determinado algoritmo para a seleção de descritores, em oposição à subjetividade da indexação manual (1993, p. 235-239). Mesmo que em processos automáticos não seja possível a adoção de algum tipo de indexação ponderada nos mesmos moldes em que acontece com a indexação manual – o indexador atribui grau de importância aos descritores escolhidos (LANCASTER, 1993, p. 174-187); é possível adotar um ranking criado automaticamente, de acordo com parâmetros de seleção e corte.

KOBASHI (1994) associa a quantidade de descritores no processo de indexação à completa caracterização de informações fundamentais presentes no texto, num processo que considera a estrutura temática do texto analisado, a seleção de categorias fundamentais para a caracterização da temática e a política de indexação do sistema. Essa análise estrutural é possível de ser implementada em metodologias automáticas, embora não seja o propósito desta pesquisa.

Já se mencionou o fato de que o aumento do número de termos descritores aumenta a revocação dos documentos no processo de recuperação, diminuindo conseqüentemente a precisão. Contudo, se após o processo de análise conceitual automatizada os descritores forem apresentados de forma ordenada, em termos de importância semântica, pode-se realizar uma indexação “modulada”, em que a alteração

de parâmetros – maior precisão ou maior revocação – permita a escolha de quantos descritores sejam desejáveis, segundo a conveniência do usuário, ou as determinações presentes na política de indexação. Observa-se que no caso de metodologias automatizadas baseadas em freqüência, essa parametrização é facilmente implementada no processo de seleção de descritores, desde que estes sejam apresentados em *ranking* relativo de importância semântica.

O pressuposto adotado é o fato de que, idealmente, quanto maior o número de descritores extraídos – número este que está relacionado à estrutura e ao tamanho dos documentos, e à metodologia de identificação e seus parâmetros – maior é a caracterização do assunto do documento. Entretanto, um número excessivo de descritores pode não ser conveniente, por diminuir em demasia a precisão das buscas baseadas nesses índices, o que nos impele a desenvolver uma metodologia flexível e parametrizada, que permita a escolha *a priori* ou *a posteriori* de qualquer quantidade desejada de descritores, dependendo da escolha por maior taxa de precisão ou revocação, quando da recuperação destes documentos.

Na aplicação da metodologia prospectiva não foram excluídos descritores freqüentes *a priori*. Na metodologia consolidada, esse recurso pode ser adotado.

4.4.2 – Critérios de avaliação da metodologia

A metodologia que pretendia extrair descritores para avaliar a relevância semântica dos SNs candidatos a descritores, definimos os conceitos de “Pontuação” e “Taxa de Relevância”. Para efeitos de pontuação, associamos os seguintes valores aos SNs, de acordo com a relevância semântica percebida, segundo o mesmo esquema cromático apresentado no **Anexo B**, dos resultados da aplicação da metodologia prospectiva:

Relevância descritiva do SN	Símbolo	Valor associado
SN extremamente relevante como descritor	SN***	1,0
SN razoavelmente relevante como descritor	SN**	0,5
SN moderadamente relevante como descritor	SN*	0,25
SN não relevante como descritor	SN –	0,0

Tabela 6 – Valor atribuído ao SN de acordo com sua relevância

Computamos valores ponderados (pontuação) relativos à qualidade dos SNs como descritores, segundo a fórmula a seguir:

$$\text{Pontuação}(\text{desc}) = (Núm.SN\ ***) + 0{,}5x(Núm.SN\ **) + 0{,}25x(NúmSN\ *)$$

E definimos também a taxa de relevância dos SNs, para determinada freqüência:

$$Tx\,\text{Re}lev = \left(\frac{Pontuação(desc)}{soma\ das\ ocorrências} \right)$$

A pontuação foi mensurada atribuindo-se valor numérico arbitrário aos SNs de acordo com sua relevância percebida como descritores, e a taxa de relevância apresentou esse valor normalizado. Quanto maior a taxa de relevância, melhor seria a representação do assunto pelos descritores, sendo que o valor máximo é 1 – valor este que seria alcançado se a totalidade dos descritores fosse extremamente relevante, caso bastante incomum mesmo para processos de indexação manual.

Pode-se objetar quanto a certo grau de subjetividade envolvido no processo de julgamento de relevância, uma vez em que foi o próprio autor desta tese que classificou os SNs entre extremamente relevantes, razoavelmente relevantes, moderadamente relevantes e não relevantes como descritores. Entende-se, porém, que a subjetividade está necessariamente presente quando se propõe a escolha de descritores no processo de análise de assunto (CESARINO, 1980; UNISIST, 1981; NAVES, 1996).

Os valores arbitrários de 1,0, 0,5 e 0,25 atribuídos de acordo com a relevância relativa dos descritores não foram considerados absolutamente, mas apenas como parâmetros para a possível avaliação das aplicações da metodologia.

Esses valores e as fórmulas utilizados nesta investigação são discutidos nos capítulos a seguir, quando da análise dos dados.

5 RESULTADOS DA APLICAÇÃO DA METODOLOGIA PROSPECTIVA

Este capítulo descreve a experimentação empírica e conseqüentes análises, necessárias à confirmação dos pressupostos apresentados na introdução e nas afirmações que permeiam este trabalho. Tem como ponto central à validação da metodologia prospectiva como um processo viável para a escolha automática de descritores. A metodologia (prospectiva), na forma inicial, e as ferramentas necessárias à sua consecução, foram apresentadas e delineadas no capítulo anterior.

Este capítulo está dividido da seguinte maneira:

- Na seção 5.1 foram comparadas as extrações manual e automática em um *corpus* anteriormente processado de forma manual. Os resultados apresentados permitiram estabelecer algumas considerações sobre o processamento automático;
- Na seção 5.2 foram apresentados e discutidos os dados provenientes da aplicação da metodologia prospectiva, delineada no capítulo 4, ao *corpus* de testes, gerando subsídios para que essa seja refinada.

5.1 – A validação da extração automática de sintagmas nominais

Nesta seção, pretendem-se apresentar considerações de ordem qualitativa e quantitativa para tecer possível comparação entre os processos manual e automático de extração de sintagmas nominais. Para essa avaliação, tomamos apenas os 15 documentos do primeiro *corpus* apresentado na seção 4.1. Esses documentos foram previamente analisados de forma manual e seus sintagmas nominais foram extraídos e classificados (KURAMOTO,1999).

Embora não tenha sido objetivo desta investigação esmiuçar detalhes da conformação dos sintagmas nominais extraídos automaticamente, como realizado no âmbito da extração manual citada, alguns comentários comparativos são tecidos, a título de avaliação. Uma análise comparativa completa, porém, demandaria tempo demasiado e estaria além dos objetivos propostos, ficando como uma sugestão de pesquisa futura, que poderia ser aplicada ao desenho de melhores *parsers* e à correção de possíveis problemas com os atualmente disponíveis.

As categorias de análise previstas na metodologia para avaliação comparativa das extrações manual e automática dos sintagmas nominais foram:

- Tempo gasto na extração dos sintagmas nominais;
- Quantidade e qualidade dos sintagmas nominais identificados.

5.1.1 – Considerações sobre o tempo gasto no processo

O processo conjunto de extração automática de sintagmas nominais dos 75 textos completos das revistas eletrônicas, inicialmente selecionados, e dos 15 textos analisados manualmente pelo professor Dr. Hélio KURAMOTO (1999, Anexo A) tomou cerca de 130 horas de processamento computacional semi-assistido, em diversos equipamentos, sendo que dessas 130 horas, apenas cerca de 5 horas foram devotadas à extração dos SNs dos 15 textos analisados manualmente. Considerando o conjunto dos *corpora*, objetivemos média aproximada de uma hora e meia de processamento, dedicado a cada documento.

Embora a submissão dos artigos ao processador sintático PALAVRAS e o pós-processamento no programa Palavras Xtractor tenha tomado, para os dois *corpora* selecionados, e ainda os 15 documentos descartados, apenas cerca de doze horas, a aplicação das folhas de estilo utilizando o software XML SPY – necessária para a extração específica dos sintagmas nominais do *corpus* marcado em XML – tomou cerca de três semanas, com a média de dedicação de oito horas diárias, contribuindo para a maior parte do tempo necessário ao processo completo de extração dos SNs dos documentos. Aqui não se considera o tempo gasto na escolha dos SNs significativos, dentre os extraídos.

A característica recursiva do processo de extração dos sintagmas, o tamanho dos documentos originais e dos arquivos gerados pelo processador Palavras Xtractor a partir destes, somados à indisponibilidade de equipamentos PC compatíveis com velocidade de processamento e memória de trabalho suficientes determinaram o tempo tomado pelo processo. Também podemos supor que alguns defeitos no gerenciamento de memória no software XML SPY e na estrutura aninhada dos sintagmas nominais possam ter causado os diversos problemas de insuficiência de memória de trabalho (na memória RAM) do computador, observados durante o processamento dos textos, que adicionaram ao total muitas horas extras de trabalho. É de se esperar que este tempo total de processamento

pudesse ser reduzido consideravelmente com a utilização de equipamentos e *software* mais velozes, e à medida que partes do processo fossem automatizadas, caso a metodologia se mostrasse eficaz para o propósito.

As informações de que dispomos sobre o tempo gasto na identificação manual dos sintagmas nominais, conseguidas por meio de trocas de mensagens e colóquios informais entre o autor e o professor Dr. Hélio KURAMOTO indicaram para o processo manual uma duração muito variada, e pode-se razoavelmente supor que, embora não tenha sido possível mensurar, o tempo gasto na extração automática fosse bastante inferior, em média, ao processo manual.

5.1.2 – Considerações quantitativas e qualitativas sobre os SNs identificados

O Anexo B da tese de doutorado de KURAMOTO (1999) apresenta os sintagmas nominais extraídos manualmente a partir dos 15 textos de seu *corpus*, ordenados alfabeticamente. Não há discriminação de SNs por texto de onde foram extraídos; então as comparações tecidas nesta seção levaram em conta o *corpus* como um todo. A TAB. 7 apresenta alguns dados relevantes:

	Extração Manual	Extração Automática
Total de Sintagmas Nominais identificados	8818	6655 (75%)
Sintagmas Nominais válidos identificados	8818	6462 (73%)
Sintagmas Nominais únicos e válidos	5982	5183 (86%)

Tabela 7 – Comparações quantitativas entre os processos de extração de SNs

Além das diferenças de performance apontadas pelas percentagens relativas (apenas 75% dos SNs totais foram identificados), estimou-se que quase 3% dos SNs identificados pelo analisador automático pudessem ser considerados não válidos, o que diminui o valor dos identificados para cerca de 73% dos SNs originalmente identificados. No caso de um esforço futuro para automatização completa da metodologia desenvolvida nesta pesquisa, sugere-se algum tipo de tratamento desse “refugo”. Quando analisamos os SNs únicos e válidos identificados automaticamente, a percentagem aumentou para 86% dos SNs únicos identificados manualmente.

Podemos identificar alguns problemas específicos do processo, que redundaram na constatação de SNs não válidos, ou na não identificação de SNs válidos:

- Falhas do processador PALAVRAS, na identificação errônea de sinais especiais de formatação (ex. números seguidos por um ponto, números romanos, títulos de seções do texto sem pontuação final, abreviaturas, sinais gráficos como $, &, etc.);
- Falhas e incompletudes no léxico utilizado para a análise sintática do processador PALAVRAS, como apresentado na seção 4.3.1. (ex. nomes próprios, palavras não reconhecidas, etc.);
- Falhas do processador PALAVRAS na identificação correta de palavras em outra língua, como o inglês;
- Falhas oriundas das conversões dos formatos originais dos documentos (PDF, HTML) para textos simples (TXT), onde a estrutura "visual" do documento for perdida;
- Falhas do programa XML SPY na geração dos arquivos de saída, em virtude de problemas de memória e arquivos XML malformados;

Um olhar mais atento e minucioso permitiu verificar que a identificação manual oferece tratamento melhor para a exploração de todos os SNs presentes nas estruturas das orações. Dos números apresentados na TAB. 7, podemos perceber perda aproximada de 27% dos sintagmas nominais totais, no processo automático, para o *corpus* analisado.

Mesmo considerando os problemas apontados e a eficácia qualitativa, se compararmos as performances levando em conta a velocidade relativa dos processos de extração e o grande percentual de SNs extraídos corretamente, consideramos que o primeiro pressuposto apresentado na introdução se verificou – temporariamente – correto. Estivemos, porém, condicionados ao fato de que a metodologia demonstrasse seu valor. Caso contrário, a análise manual do assunto do documento ainda seria a melhor opção para a escolha de descritores adequados.

5.2 – A análise dos dados da aplicação da metodologia prospectiva

A comparação realizada na seção anterior sugeriu um posicionamento levemente cauteloso quanto aos resultados da aplicação da metodologia prospectiva ao *corpus* de

textos. Acreditava-se, porém, que a metodologia fosse capaz de prover resultados satisfatórios, se comparada às metodologias tradicionais de escolha a partir de freqüência de palavras-chave isoladas. A partir deste argumento, deixemos que os resultados, ao final, falem por si.

Nas seguintes subseções, detalhamos os resultados da aplicação da metodologia prospectiva, apresentada no capítulo anterior, ao *corpus* de testes, ou seja, à amostra reduzida, composta de 6 documentos (10% dos documentos totais), escolhidos dentre aqueles pertencentes ao *corpus* de trabalho – composto na íntegra por 60 documentos. As operações realizadas nesse subconjunto e seus resultados permitiram a avaliação do processo e subsidiaram os ajustes e melhorias possíveis. Daí então, no capítulo seguinte, os documentos em sua totalidade são processados a partir da metodologia consolidada, então avaliada.

Como apresentado no capítulo anterior, a metodologia de seleção dos SNs significativos para descritores dos textos, considerada a maior contribuição deste trabalho, levou em consideração os seguintes fatores:

- As freqüências e a relevância semântica dos SNs que ocorriam nos textos dos artigos (fator analisado na subseção 5.2.1);
- A quantidade de ocorrências dos SNs na totalidade do *corpus* (fator também analisado na subseção 5.2.1);
- Os níveis e as estruturas sintáticas dos SNs relevantes como descritores (fator analisado na subseção 5.2.2);
- A ocorrência no tesauro da CI (1989) dos SNs freqüentes e relevantes (fator analisado na subseção 5.2.3).

A partir da análise desses fatores, considerados a partir de suas influências individuais e também correlacionados entre si, foi possível avaliar a metodologia proposta e modificá-la de forma a tornar-se mais eficaz.

Para as análises de **freqüência**, **ocorrência** e **relevância semântica** de SNs como descritores, consideramos *insights* teóricos de algumas das metodologias utilizadas para a seleção de palavras-chave significativas, como apresentado na subseção 2.2.2 deste trabalho. Dentre os algoritmos, destacamos os seguintes:

- Cálculo de freqüências com limites de corte inferior e superior (Lei de Zipf);
- Pesos relacionados à freqüência inversa
- Valor discriminatório dos termos.

O critério adotado para a avaliação da **relevância semântica** dos SNs escolhidos baseou-se em considerações do autor desta tese a partir de análises de semelhança semântica entre esses e as palavras-chave e resumos originais produzidos pelos autores dos documentos do *corpus*. Como apontado anteriormente, esse critério, apresenta componente subjetivo, e numa situação ideal a decisão sobre a relevância dos descritores extraídos deveria ser realizada por um grupo de especialistas.

Para as análises e comparações entre a **relevância semântica** e os **níveis** e **estruturas sintáticas** dos SNs, utilizamos a teoria apresentada nas subseções 2.1.6 e 2.1.7 desta tese, além de aportes teóricos advindos do trabalho de KURAMOTO (1999).

Finalmente, realizaremos comparações entre os SNs relevantes, segundo os critérios anteriores, com aqueles que **ocorrem parcial** ou **exatamente** no **Tesauro da CI** (1989), completando os passos metodológicos a serem validados.

Os seis primeiros artigos do *corpus* apresentado no **Anexo A** desta tese foram processados segundo a metodologia apresentada no capítulo 4, e os dados necessários às primeiras análises, obtidos a partir do processamento dos documentos nas planilhas, foram apresentados em tabelas, explicitadas e apresentadas nas seções a seguir. Para esses seis primeiros artigos, também são apresentadas no **Anexo B** as palavras-chave mais freqüentes, para que pudéssemos ter uma base de comparação desses termos e com os SNs escolhidos como descritores. Essa discussão é apresentada no capítulo seguinte.

As subseções seguintes devotaram-se à exegese e à análise dos dados, e uma descrição dos resultados do processamento inicial pode ser conferida no **Anexo B** desta tese.

5.2.1 – Considerações sobre as freqüências de ocorrência dos SNs e a relevância semântica como descritores

Os argumentos apresentados na subseção anterior nos motivaram a oferecer metodologia flexível, que permita a escolha de certa quantidade de descritores estabelecida de acordo com a conveniência do usuário ou do sistema. Para tal propósito, elaboramos uma espécie de *ranking* indicativo de relevância associado, entre outros aspectos, às freqüências de ocorrência dos SNs nos textos. Para chegar a este relacionamento entre freqüência e relevância, analisamos nesta subseção os dados apresentados nas TAB. 8 e 9, explicitadas a seguir:

- A TAB. 8 apresenta, para cada um dos seis artigos do *corpus* de testes, os seguintes dados, assim enumerados:

  I. A quantidade total de SNs identificados e a quantidade de SNs únicos identificados (soma de todos os SNs excetuando as repetições), e a percentagem dos SNs únicos em relação aos totais;

  II. A quantidade de SNs identificados de acordo com as freqüências de ocorrência, para as freqüências de 1, 2, 3, 4 ou mais de 4 vezes, e os percentuais respectivos, relativos à quantidade de SNs únicos;

  III. A quantidade de SNs identificados com freqüência de 2 vezes, que não possuem estrutura sintática específica (explicitada na subseção 5.2.2), e seu percentual relativo à quantidade de SNs únicos (identificados por asterisco);

  IV. Os totais de SNs que aparecem mais de 1 vez, mais de 1 vez e que não possuem estrutura sintática específica (explicitada na subseção 5.2.2), e mais de 2 vezes, e seus percentuais relativos à quantidade de SNs únicos.

- A TAB. 9 apresenta, para cada um dos seis artigos do *corpus* de testes, os seguintes dados, assim enumerados:

  I. Repetindo as informações da TAB. 7, são apresentadas a quantidade total de SNs identificados e a quantidade de SNs únicos identificados (soma de todos os SNs excetuando as repetições), e a percentagem dos SNs únicos em relação aos totais - ;

II. A pontuação e a taxa de relevância (como definidas na seção 4.4), para SNs que ocorrem 3, 4 e mais de 4 vezes;

III. A pontuação e a taxa de relevância para SNs que ocorrem 2, 3, 4 e mais de 4 vezes, tendo sido expurgados os que não possuem uma estrutura sintática específica (explicitada na subseção 5.2.2).

		Artigo 1		Artigo 2		Artigo 3		Artigo 4		Artigo 5		Artigo 6		Médias	
I	**Qtd. de SNs identificados**	1673		842		783		801		1478		984		**1093,5**	
	Qtd. de SNs únicos identificados	1343	80,3%	711	84,4%	680	86,8%	688	85,9%	1252	84,7%	836	85,0%	**918,3**	**84,0%**
II	**Qtd. de SNs que aparecem somente 1 vez**	1251	93,1%	662	93,1%	645	94,9%	631	91,7%	1165	93,1%	780	93,3%	**855,7**	**93,18%**
	Qtd. de SNs que aparecem 2 vezes	66	4,9%	33	4,6%	23	3,4%	45	6,5%	72	5,8%	41	4,9%	**46,7**	**5,08%**
	Qtd. de SNs que aparecem 3 vezes	9	0,7%	5	0,7%	5	0,7%	5	0,7%	4	0,3%	8	1,0%	**6,0**	**0,65%**
	Qtd. de SNs que aparecem 4 vezes	5	0,4%	2	0,3%	2	0,3%	3	0,4%	7	0,6%	9	1,1%	**4,7**	**0,51%**
	Qtd. de SNs que aparecem mais de 4 vezes	17	1,3%	11	1,5%	7	1,0%	7	1,0%	11	0,9%	7	0,8%	**10,0**	**1,09%**
III	**Qtd. de SNs que aparecem 2 vezes, excetuando os de estrutura (D + N)**	25	1,9%	10	1,4%	7	1,0%	12	1,7%	22	1,8%	11	1,3%	**14,5**	**1,58%**
IV	**Total de SNs freqüentes (>1)**	97	7,2%	51	7,2%	37	5,4%	60	8,7%	94	7,5%	65	7,8%	**67,3**	**7,33%**
	Total de SNs freqüentes (>1) excetuando os de estrutura (D + N)	56	4,2%	28	3,9%	21	3,1%	27	3,9%	44	3,5%	35	4,2%	**35,2**	**3,83%**
	Total de SNs freqüentes (>2)	31	2,3%	18	2,5%	14	2,1%	15	2,2%	22	1,8%	24	2,9%	**20,7**	**2,25%**

Tabela 8 – Freqüências de ocorrência dos SNs nos 6 primeiros artigos do *corpus*

		Artigo 1		Artigo 2		Artigo 3		Artigo 4		Artigo 5		Artigo 6		Médias	
I	**Qtd. de SNs identificados**	1673		842		783		801		1478		984		**1093,5**	
	Qtd. de SNs únicos identificados	1343	80,3%	711	84,4%	680	86,8%	688	85,9%	1252	84,7%	836	85,0%	**918,3**	**84,0%**
II	**SNs relevantes como descritores que aparecem 3 vezes**	Pont.	TxRelev	Pont.	TxRelev	Pont.	TxRelev	Pont.	TxRelev	Pont.	TxRelev	Pont.	TxRelev	**Pont.**	**TxRelev**
		0,25	0,03	0,25	0,05	0,50	0,10	2,25	0,45	0,25	0,06	1,25	0,16	**0,79**	**0,13**
	SNs relevantes como descritores que aparecem 4 vezes	0,25	0,05	0,25	0,13	1,00	0,50	1,00	0,33	2,00	0,29	2,75	0,31	**1,21**	**0,26**
	SNs relevantes como descritores que aparecem mais de 4 vezes	6,50	0,38	5,00	0,45	3,00	0,43	4,00	0,57	2,75	0,25	3,25	0,46	**4,08**	**0,41**
III	**SNs relevantes como descritores que aparecem 2* vezes**	5,50	0,21	2,75	0,28	2,50	0,36	4,25	0,35	5,50	0,25	3,75	0,34	**4,04**	**0,28**
	SNs relevantes como descritores que aparecem 3* vezes	0,25	0,13	0,25	0,25	-	0,00	2,25	0,75	0,25	0,06	1,00	0,50	**0,80**	**0,40**
	SNs relevantes como descritores que aparecem 4* vezes	0,25	0,25	-	0,00	1,00	1,00	0,75	0,38	1,50	0,50	1,25	0,63	**0,95**	**0,63**
	SNs relevantes como descritores que aparecem mais de 4* vezes	3,50	0,58	2,25	0,56	1,00	1,00	1,00	1,00	1,25	0,63	1,00	1,00	**1,67**	**0,67**

* Excetuando aqueles que apresentam a estrutura sintática (Determinante + Nome)

Tabela 9 – Análises de correlação entre as freqüências de ocorrência e a relevância dos SNs

Dos dados da TAB. **8**, podemos perceber que, para o conjunto reduzido de seis artigos analisados, a média de ocorrência de SNs totais é aproximadamente 1093, e a média de ocorrência de SNs únicos é de aproximadamente 918, ou seja, 84% da média de SNs totais identificados. Isso significa que cerca de 16% dos SNs totais, em média, se repetem ao menos uma vez. Dentre os SNs que se repetem, a grande maioria - 10,2% dos SNs totais - se repete apenas duas vezes e cerca de 4% se repetem três ou quatro vezes. Menos de 2% se repetem mais de quatro vezes.

É importante relembrar que no cômputo das freqüências dos SNs, foram agrupados aqueles que diferiam apenas pelo determinante inicial, sendo este usualmente artigo, pronome demonstrativo, pronome possessivo, numeral ou número; ou mesmo artigo seguido por pronomes ou numerais (como apresentados na TAB. **5** do capítulo anterior); e suas freqüências foram calculadas conjuntamente. Esse tipo de processamento manual, porém, para a eficácia da metodologia, deveria ser implementado como processo totalmente automatizado.

Na TAB. **8** podemos observar as percentagens de ocorrência percebidas para as várias freqüências, em relação ao número total de SNs únicos, e a FIG. 15 ilustra essas freqüências:

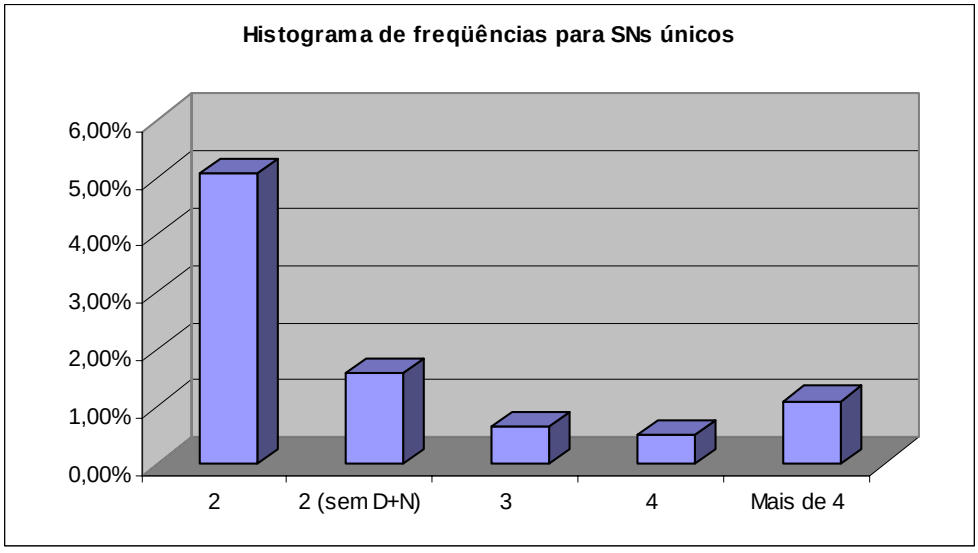


Figura 15 – Histograma de freqüência para SNs únicos

Verificou-se que as maiores relevâncias semânticas podem ser associadas às maiores freqüências de ocorrência, de modo análogo às palavras-chave, havendo

leve indicação de saturação quando a freqüência aumenta em demasia. Isso pode ser verificado no gráfico da FIG. 16, explicitada adiante.

Os resultados apresentados na TAB. **9** demonstram claramente que a taxa de relevância cresce com a freqüência, sendo de 0,13 para os SNs que ocorrem três vezes; 0,26 para os que ocorrem quatro vezes e 0,41 para os que ocorrem mais de quatro vezes. Ainda são apresentadas as taxas de relevância para freqüências (indicadas por asterisco) de SNs dos quais foram expurgados aqueles compostos por certa estrutura sintática (D + N) – objeto de discussão na próxima subseção. A FIG. 16 ilustra estes dados:

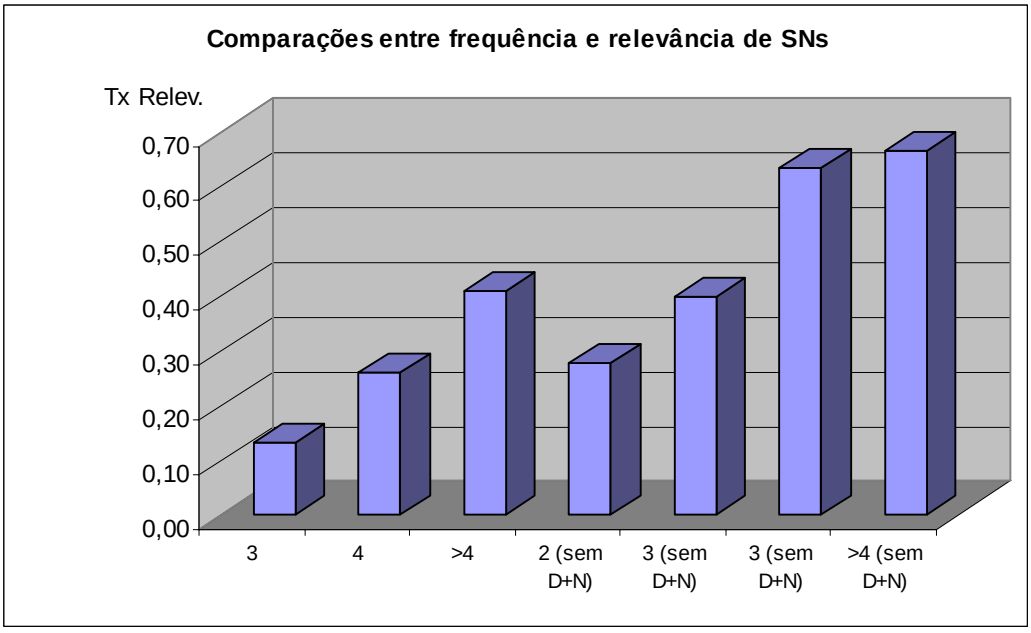


Figura 16 – Comparações entre freqüências e relevância de SNs

Pode-se perceber no histograma da FIG. 16 um comportamento notadamente assintótico, que indica que a taxa de relevância tende a aumentar cada vez menos, à medida que se aumentam as freqüências. Este fenômeno também pode ser interpretado como saturação da densidade semântica para freqüências demasiado altas, e mesmo talvez uma inflexão na curva, o que indicaria que os termos com freqüência demasiado alta podem ser descritores insignificantes; expressões por demais comuns, com pouco poder discriminatório.

Quando tratamos de palavras-chave, a lei de Zipf preconiza freqüência superior de corte, por considerar que as palavras que possuem freqüência demasiado alta

não possuem poder discriminatório e densidade informacional. Verifica-se, a partir das análises no *corpus* exemplificadas pela amostra apresentada no **Anexo B**, que os SNs que apresentam freqüências demasiadamente altas perdem em relevância, mas o descarte indiscriminado desses SNs extremamente freqüentes poderia vir a eliminar bons descritores. Quando analisamos os SNs de acordo com as freqüências totais no *corpus* reduzido de 6 documentos, verificamos que as maiores taxas de freqüência estão associadas aos SNs mais "genéricos", como por exemplo "conhecimento" e "informação". Esses SNs, se considerados isoladamente e fora do contexto de seus artigos originais, realmente revelam menor relevância. Uma possível solução para esse impasse seria estabelecer valor máximo de relevância a ser considerado a partir da análise da freqüência, para os propósitos de avaliação dos descritores. Essa sugestão será incorporada na metodologia final.

Também se pode perceber a partir da análise da aplicação da metodologia no *corpus* reduzido que o poder discriminatório dos termos diminui, se considerarmos as ocorrências dos SNs na totalidade dos textos do *corpus*, o que nos impele a considerar ponderação de valores relacionados ao **inverso da freqüência** de ocorrência no *corpus*, de modo a valorizar os SNs que ocorrem freqüentemente em poucos documentos, e penalizar os SNs extremamente freqüentes em todo o conjunto de documentos. De acordo com a "trivialidade" de alguns desses SNs, podem-se mesmo incluí-los em uma *stoplist*, caso assim seja conveniente. Ao fazê-lo, estaremos aumentando o valor discriminatório de cada termo escolhido.

Nos dados apresentados no **Anexo B**, pode-se perceber qualitativamente que a relevância cresce de forma geral com a freqüência para esse conjunto reduzido de artigos, e seria arbitrário definir *a priori* uma freqüência de corte inferior para a metodologia modificada. Um ponto a ser considerado na metodologia consolidada é a possibilidade de parametrização dos valores para as freqüências de corte, ou mesmo deixar que a quantidade desejada de descritores estabeleça esse corte.

A freqüência de corte escolhida para a aplicação prospectiva foi de 2 ocorrências, sendo que para os que ocorrem somente duas vezes, foram eliminados aqueles que possuem estrutura sintática mais simples, compostos por um determinante seguido de um nome. Na próxima subseção, vamos apresentar

considerações relativas aos níveis e às estruturas sintáticas dos SNs, e suas relevâncias relativas como descritores.

5.2.2 – Considerações sobre as estruturas sintáticas dos SNs e a relevância como descritores

Para a análise das estruturas sintáticas e a relevância dos SNs como descritores, tomaremos em conjunto os dados das TAB. 8 e 9 apresentadas anteriormente e a TAB. **10** (apresentada a seguir). A TAB. **10** apresenta, para cada um dos seis artigos do *corpus* de testes, as quantidades relativas de SNs de acordo com suas estruturas sintáticas (como exemplificadas), apresentando a taxa de relevância (como definido na seção 4.4) para cada um dos níveis de SNs;

Estrutura dos SNs freqüentes (>1)*	Artigo 1		Artigo 2		Artigo 3		Artigo 4		Artigo 5		Artigo 6		Médias	
	Qt	TxRelev	Qt	TxRelev	Qt	TxRelev	Qt	TxRelev	Qt	TxRelev	Qt	TxRelev	Qt	TxRelev
SN de Nível 1a	21	0,14	13	0,23	12	0,21	9	0,36	13	0,17	19	0,21	**14,50**	1,24
SN de Nível 1b	13	0,21	8	0,31	4	0,38	10	0,53	14	0,27	4	0,31	**8,83**	1,92
SN de Nível 2	20	0,31	6	0,38	5	0,60	5	0,40	16	0,28	11	0,50	**10,50**	2,24
SN de Nível 3 ou maior	1	0,25	1	0,50	0	-	1	1,00	1	0,25	1	0,25	**0,83**	2,70

Tabela 10 – Análises de correlação entre estrutura sintática e relevância dos SNs

Exemplos da classificação adotada para os SNs:
Nível 1a: “Os negócios”;
Nível 1b: “Os negócios internacionais”;
Nível 2: “O mundo dos negócios”;
Nível 3 : “O ambiente do mundo dos negócios”;
Nível 4: “As características do ambiente do mundo dos negócios”;
Nível 5: “As análises das características do ambiente do mundo dos negócios”.

Tabela 11 – Exemplos da classificação adotada para os SNs segundo suas estruturas sintáticas

Apesar de haverem sido tecidas algumas considerações ligeiras sobre as estruturas sintáticas dos SNs nas subseções 2.1.6 e 2.1.7 deste trabalho, tomar-se-ão emprestados alguns resultados de KURAMOTO (1999) sobre a freqüência de ocorrência dos SNs de acordo com sua estrutura, em seu *corpus* de análise de 15 textos. KURAMOTO verifica que cerca de 50% dos SNs únicos verificados são de nível 1a, ou seja, possuem a estrutura simples (D + N), sendo N uma estrutura sintática genericamente considerada como um substantivo ou nome próprio e D um determinante (artigo, pronome ou numeral), composto usualmente pelas estruturas apresentadas na TAB. **5**, na seção 4.2 deste trabalho. Esses dados corroboram o que foi verificado no *corpus* de seis textos desta tese.

Os dados apresentados nas TAB. **9** e **10** indicam que a estrutura sintática dos SNs está relacionada à sua relevância como descritores. Podemos notar que essas estruturas (D + N) sempre constituem SNs de nível 1a, como exemplificado na TAB. **11**, e não diferem muito em termos de densidade informacional das palavras-chave, que se diferenciam desses SNs apenas pela ausência dos determinantes. Quando analisamos os dados da TAB. **9,** que relaciona as freqüências e a relevância dos SNs, tendo sidos expurgados aqueles de estrutura simples (o segmento III da TAB. 9), verificamos que a taxa de relevância cresce bastante (0,13→ 0,40 para a freqüência de 3 ocorrências; 0,26→ 0,63 para a freqüência de 4 ocorrências, e 0,41→ 0,67 para freqüências maiores que 4 ocorrências).

Mesmo os SNs que apresentam freqüência de apenas 2 ocorrências conseguem a taxa de relevância de 0,28; quando são expurgados os de estrutura simples (D + N) – taxa esta maior que a relevância dos que ocorrem três vezes sem que haja esse expurgo. Entretanto, o número total de SNs selecionados para freqüências maiores que 2 decresce muito com o expurgo, de modo que a análise da estrutura sintática do SN, ao menos para as altas freqüências, deve ser critério seletivo, mas não eliminatório, a ser considerado no desenho da metodologia corrigida.

Os resultados apresentados na TAB. **10** podem ser sumarizados na FIG. 17, que considera SNs de nível “1a” aqueles que possuem a estrutura (D + N) e “1b” aqueles de nível 1 no qual foram excluídos os de estrutura (D + N):

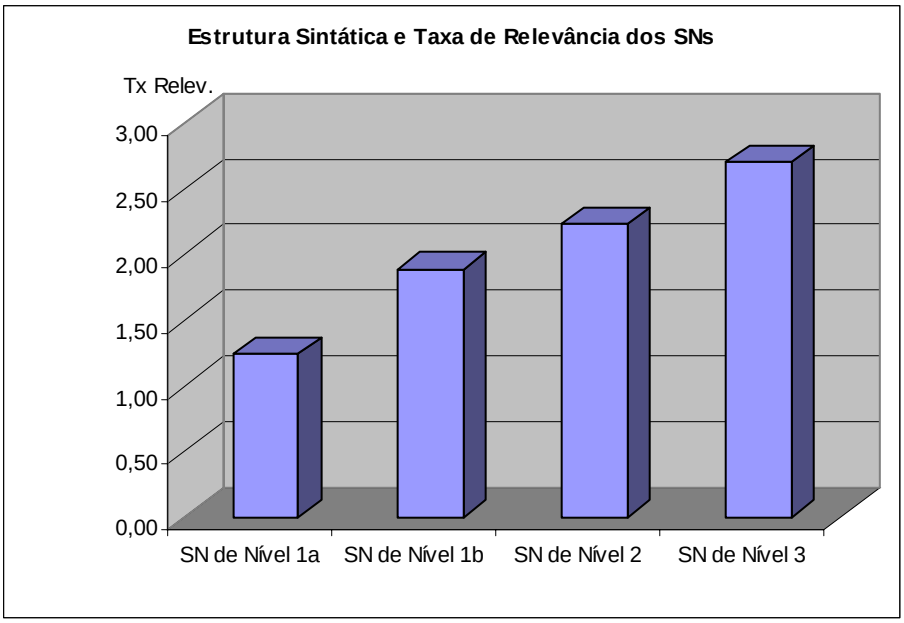


Figura 17 – Correlação entre Estrutura e Relevância dos SNs

Sumarizando os dados das TAB. **9** e **10**, e a informação apresentada na FIG. 17, podemos afirmar que:

- A densidade informacional do SN cresce com seu nível (ao menos até os de terceiro nível, que ocorrem neste *corpus* reduzido);
- A menor densidade informacional ocorre entre os SNs de estrutura (D + N).

Esses fatores, aliados às análises de relevância e freqüência, devem ser considerados no desenho da metodologia consolidada.

Na próxima subseção, vamos apresentar considerações relativas à ocorrência dos SNs freqüentes no tesauro da CI (1989).

5.2.3 – Análise integrada de freqüência, relevância semântica e ocorrência no tesauro de CI

Para a análise das estruturas sintáticas e a relevância dos SNs como descritores, tomaremos em conjunto os dados das tabelas anteriormente apresentadas e a TAB. **12** (apresentada a seguir).

- A **TAB. 12** apresenta, para cada um dos seis artigos do *corpus* de testes, os seguintes dados:

I. A quantidades de SNs que ocorrem mais de uma vez e que não possuem estrutura sintática específica (explicitada na seção 5.2.2), que são extremamente relevantes, razoavelmente relevantes, moderadamente relevantes e que não são relevantes, como descritores, para os respectivos documentos, além de seus percentuais relativos ao total de SNs que ocorrem mais de uma vez;

II. A quantidades de SNs que ocorrem mais de uma vez e que não possuem estrutura sintática específica (explicitada na seção 5.2.2), e que constam parcial ou exatamente no tesauro de CI, além de seus percentuais relativos ao total de SNs que ocorrem mais de uma vez;

III. A quantidades de SNs que ocorrem mais de uma vez e que não possuem estrutura sintática específica (explicitada na seção 5.2.2) e são ao mesmo tempo relevantes (extremamente, razoavelmente ou moderadamente), que constam parcial ou totalmente no tesauro de CI, além de seus percentuais relativos ao total de SNs que ocorrem mais de uma vez.

		Artigo 1		Artigo 2		Artigo 3		Artigo 4		Artigo 5		Artigo 6		Médias	
I	**Qtd. de SNs freqüentes (>1)* que são extremamente relevantes como descritores**	2	3,5%	3	10,3%	4	19,0%	6	23,1%	5	11,4%	6	17,1%	4,3	12,3%
	Qtd. de SNs freqüentes (>1)* que são razoavelmente relevantes como descritores	12	21,1%	5	17,2%	2	9,5%	4	15,4%	4	9,1%	5	14,3%	5,3	15,2%
	Qtd. de SNs freqüentes (>1)* que são moderadamente relevantes como descritores	17	30,4%	11	37,9%	8	38,1%	10	38,5%	14	31,8%	10	28,6%	11,7	33,2%
	Qtd. de SNs freqüentes (>1)* que não são relevantes como descritores	25	43,9%	09	32,1%	7	33,3%	6	23,1%	21	47,7%	14	40,0%	13,7	39,0%
II	**SNs freqüentes (>1)* que constam no Tesauro CI**	4	7,0%	2	7,1%	3	14,3%	3	11,1%	4	9,1%	8	22,9%	4,0	11,4%
	SNs freqüentes (>1)* que constam parcialmente no Tesauro CI	14	25%	5	17,9%	1	4,8%	11	40,7%	11	25,0%	10	28,6%	8,7	24,6%
III	**SNs freqüentes (>1)* e relevantes como descritores que constam exatamente no Tesauro CI**	1	7,1%	0	0,0%	2	33,3%	2	20,0%	0	0,0%	3	27,3%	1,3	13,8%
	SNs freqüentes (>1)* e relevantes como descritores que constam parcialmente no Tesauro CI	4	28,6%	4	50,0%	0	0,0%	6	60,0%	2	22,2%	4	36,4%	3,3	34,5%

* Excetuando aqueles que apresentam a estrutura sintática (Determinante + Nome)

Tabela 12 – Análises de correlação entre a relevância dos SNs e a ocorrência no tesauro da CI

O cômputo geral da relevância semântica dos SNs freqüentes como descritores, como apresentado na TAB. **12**, pode ser sumarizado na FIG. 18, que representa o esquema de cores do **Anexo B**. A figura mostra os percentuais de ocorrência dos SNs extremamente relevantes (SNs***), razoavelmente relevantes (SNs**) e moderadamente relevantes como descritores (SNs*), além dos SNs sem relevância como descritores (SNs-):

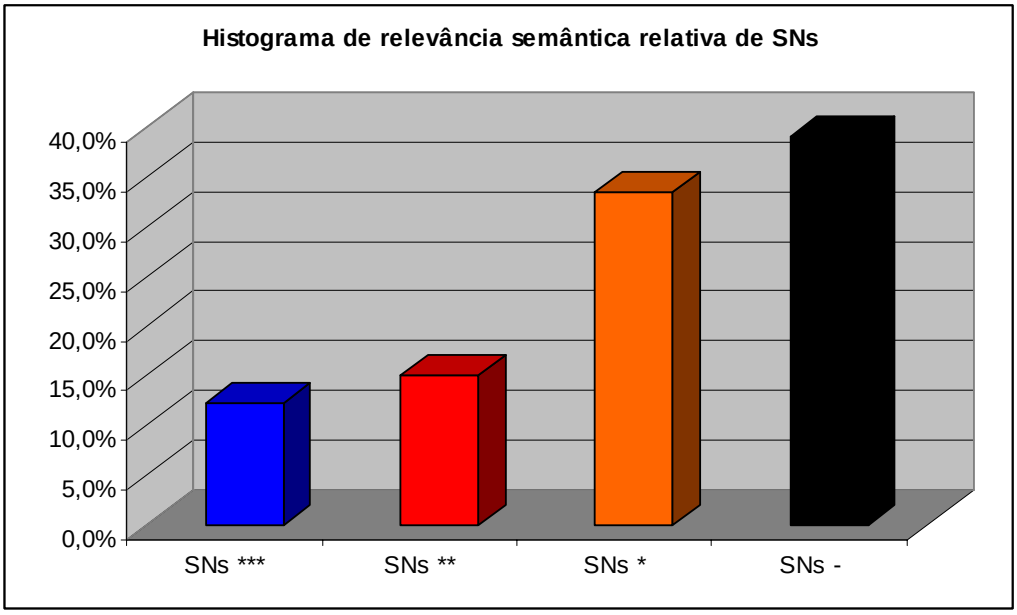


Figura 18 – Freqüências de SNs relativas à relevância semântica

Nessa aplicação da metodologia prospectiva, pôde-se perceber que, dentre os sintagmas nominais freqüentes, e excluindo os que ocorrem apenas duas vezes e possuem estrutura (D + N), 12,4% são extremamente relevantes como descritores; 15,2% são razoavelmente relevantes como descritores e 33,3% são moderadamente relevantes como descritores; o que perfaz aproximadamente 60% de SNs com algum poder de caracterização do assunto e cerca de 27,6% podem ser considerados bons descritores. Temos ainda cerca de 39% de SNs que não possuem poder de caracterização. Esses sintagmas foram escolhidos a partir apenas das análises de freqüência (aqueles que ocorriam mais de uma vez no texto) e o descarte dos que ocorrem somente duas vezes e possuem estrutura (D + N).

A metodologia prospectiva considerava que o uso do tesauro pudesse aumentar a relevância dos SNs escolhidos, por meio do descarte de parte dos SNs não relevantes – aqueles que não constassem no tesauro nem mesmo parcialmente. No entanto, contrariando esse pressuposto, o uso do tesauro específico da área de assunto dos textos escolhidos para os *corpora* – a ciência da informação – foi de pouca valia na seleção dos descritores. Como podemos perceber nos dados da TAB. **12,** apresentada anteriormente, apenas 11,4% dos sintagmas freqüentes constam de forma similar no tesauro de CI, e 24,6% constam de forma parcial, ou seja, constam apenas alguns dos morfemas dentre os lexemas componentes do SN.

Mas se analisarmos dentre aqueles que são ao mesmo tempo freqüentes e relevantes como descritores, o resultado pouco se altera, já que 13,8% constam de forma similar no tesauro da CI e 34,5% constam de forma parcial. Disso conclui-se que o fato dos SNs ocorrerem exatamente no tesauro da CI quase nada lhes confere, em termos descritores, e pouco acrescenta o fato de ocorrerem parcialmente.

Dentre os motivos para esse resultado aparentemente negativo, podemos enumerar:

- A antiguidade e falta de atualização do tesauro utilizado;
- A dinamicidade do campo da ciência da informação;
- As características interdisciplinares das temáticas da área refletidas nos artigos dos *corpora*, confrontadas com o foco do tesauro nas temáticas mais nucleares da ciência da informação;
- A dificuldade de comparar os conceitos relacionados, através de palavras-chave ou mesmo de SNs;
- A característica geral dos tesauros de focarem conceitos amplos, e genéricos – mesmo que de área específica – em oposição à necessidade de contextualização *ad hoc* dos descritores no escopo do texto, para o aumento de seu poder discriminatório e de caracterização do assunto dentre as publicações de uma área;

- E, por fim, podemos apontar o fato de que o tesauro, com seu conjunto de conceitos representados por palavras, difere qualitativamente de SNs, que, por possuírem semântica intrínseca, prescindem do contexto atribuído. No caso do tesauro, o contexto de cada termo é atribuído por notas explicativas, relacionamentos ou pelo próprio fato de fazerem parte do tesauro, mas se forem considerados isoladamente, os termos apresentam significância inferior.

Como exemplo da desatualização do tesauro, pinçamos conceitos como “gestão do conhecimento”, “sociedade da informação”, “publicações eletrônicas” e “exclusão digital”; que não ocorrem no tesauro, e são bastante freqüentes em muitos dos artigos da área. Mais uma vez, reforça-se a característica monotemática do tesauro em oposição à miríade de caminhos interdisciplinares da ciência da informação. Podem-se esperar resultados diferentes para outras áreas do conhecimento, que fossem característica interdisciplinar menos marcante, ou constituem tesauros mais atualizados; e todas as considerações demandam que sejam devotados mais estudos para o uso desse recurso em metodologias semelhantes.

Diante dessas constatações e ressalvas, decidiu-se por abandonar o uso do tesauro como fator primordial na seleção de descritores, mas talvez utilizar como um recurso acessório para melhoria da qualidade de descritores selecionados.

O próximo capítulo apresenta os conhecimentos adquiridos na aplicação da metodologia prospectiva, que serão subsídios para o desenho da metodologia consolidada. Esta será então aplicada à totalidade do *corpus*, e seus resultados avaliados.

6 A METODOLOGIA CONSOLIDADA

Neste capítulo, buscamos consolidar o aprendizado decorrente da aplicação da metodologia prospectiva ao *corpus* reduzido. As conclusões advindas desta aplicação preliminar foram enumeradas e redundaram em algumas alterações no processo de seleção de descritores, incorporados na metodologia consolidada.

Este capítulo está dividido da seguinte maneira:

- A seção 6.1 inclui os resultados apresentados na seção 5.2 para propor alterações à metodologia prospectiva, e a metodologia consolidada é apresentada;
- Na seção 6.2 são apresentados os dados provenientes da aplicação da metodologia consolidada ao *corpus* total de 60 documentos;
- Na seção 6.3 são discutidos os resultados de maneira global.

6.1 – Considerações para a alteração da metodologia

Nesta subseção, procura-se consolidar os dados apresentados anteriormente, de forma a gerar subsídios para as decisões que redundaram em alterações na metodologia. Das subseções anteriores, destacam-se as seguintes conclusões:

1. **A relevância dos SNs aumenta com a freqüência de ocorrência, sendo que para freqüências demasiadamente altas, há uma tendência de saturação;**

Na seção 5.2 foram apresentados os dados das TAB. 8 e 9, nas quais se percebem que a freqüência de ocorrência de cada SN é diretamente proporcional à relevância como descritor, com a indicação de possível saturação. Os dados empíricos analisados sugerem a adoção de freqüências de corte inferiores de 2 ou 3 ocorrências, com a possibilidade de análise concomitante de outros dos parâmetros analisados. Ambas as freqüências consideradas – de corte (inferior) e de saturação (superior) – revelaram-se dependentes do tamanho dos textos dos documentos.

2. **Embora a densidade informacional dos SNs diminua para freqüências de ocorrência muito elevada, estabelecer freqüências superiores de corte a priori pode levar ao descarte de bons descritores;**

Para este *corpus* reduzido de documentos, não foi possível estabelecer quais seriam os níveis “seguros” a considerar como freqüências superiores de corte a priori, sendo que esse quesito será observado quando da aplicação da metodologia consolidada ao *corpus* completo. A maneira de dirimir as distorções que poderiam surgir foi a consideração de uma freqüência máxima para fins de pontuação dos descritores.

3. **A densidade informacional e o poder discriminatório do SN diminuem à medida que este aparece em grande número de documentos do *corpus*;**

A metodologia consolidada deve prever alguma forma de detectar aqueles SNs que possuam baixo poder discriminatório, por serem freqüentes em documentos de todo o *corpus*, para diminuir-lhes a pontuação. Pode-se mesmo considerar a construção de uma *stoplist*, de forma a penalizar ou eliminar descritores abundantemente freqüentes no conjunto de artigos do *corpus*.

4. **A complexidade da estrutura sintática e o nível do SN são diretamente proporcionais à sua relevância como descritor;**

Como foi demonstrado na subseção 5.2.2, a complexidade da estrutura do SN e o seu nível são proporcionais à sua densidade informacional, e embora seja trabalhoso implementar essas análises em metodologias totalmente automatizadas, há que se considerar esse fator para ponderar os valores de relevância dos SNs escolhidos. Há também que se considerar que os SNs “extensos”, como os de nível 4 ou superior, e os SNs com estruturas sintáticas muito complexas não são bons candidatos a descritores por lhes faltar certa concisão, desejável nos descritores.

5. **Para o caso de SNs aninhados em outros SNs de maior nível, há a repetição de informação pela dupla ocorrência. Essa repetição poderia gerar distorções nos cálculos de freqüências e redundância de informação;**

Para remediar este problema, admitiu-se que para freqüências semelhantes, à medida que se escolhe um SN de nível 2, 3 ou 4, eliminavam-se os de nível 1, 2 ou 3 que estivessem implícitos (aninhados), respectivamente, no de maior nível, para não gerar redundância de informação (Anexo B, artigo 1 – "o valor de uma unidade de conhecimento registrada" e "uma unidade de conhecimento registrada"). Quando houve discrepância entre as freqüências, não podíamos considerar os SNs como automaticamente vinculados, embora ainda assim houvesse um aumento da freqüência dos de menor nível.

6. **O uso do tesauro não se mostrou de utilidade para ajudar na seleção primordial de descritores relevantes;**

Os motivos expostos na subseção 5.2.3 encorajaram o abandono deste recurso tal como foi imaginado na metodologia prospectiva, embora sejam demandadas mais pesquisas que levem em conta as particularidades dos dados empíricos desta pesquisa. Na metodologia consolidada, o uso do tesauro será limitado às decisões onde, por uma limitação da quantidade de descritores desejados, tenhamos que escolher o descarte de alguns dentre aqueles de igual valor, segundo os outros critérios apresentados.

Sumarizando, seguem os dados anteriormente apresentados:

	Diretamente proporcional à relevância	Inversamente proporcional à relevância	Pouca relação com a relevância
Freqüência dos SNs no documento	**X** (com saturação)		
Freqüência de ocorrência dos SNs no *corpus* de documentos		**X**	
Complexidade do nível e da estrutura do SN	**X** (com saturação)		
Ocorrência no tesauro da CI			**X**

Tabela 13 – Relacionamentos pertinentes à relevância dos SNs

Diante dessas considerações, pudemos desenhar a metodologia consolidada, sujeita a ser, permanente e sucessivamente, alterada e melhorada, de acordo com as características dos vários *corpora*, a área de conhecimento, a política de

indexação e os novos *insights* que pudessem surgir quando da aplicação da versão atual. Segue ilustração dessa metodologia:

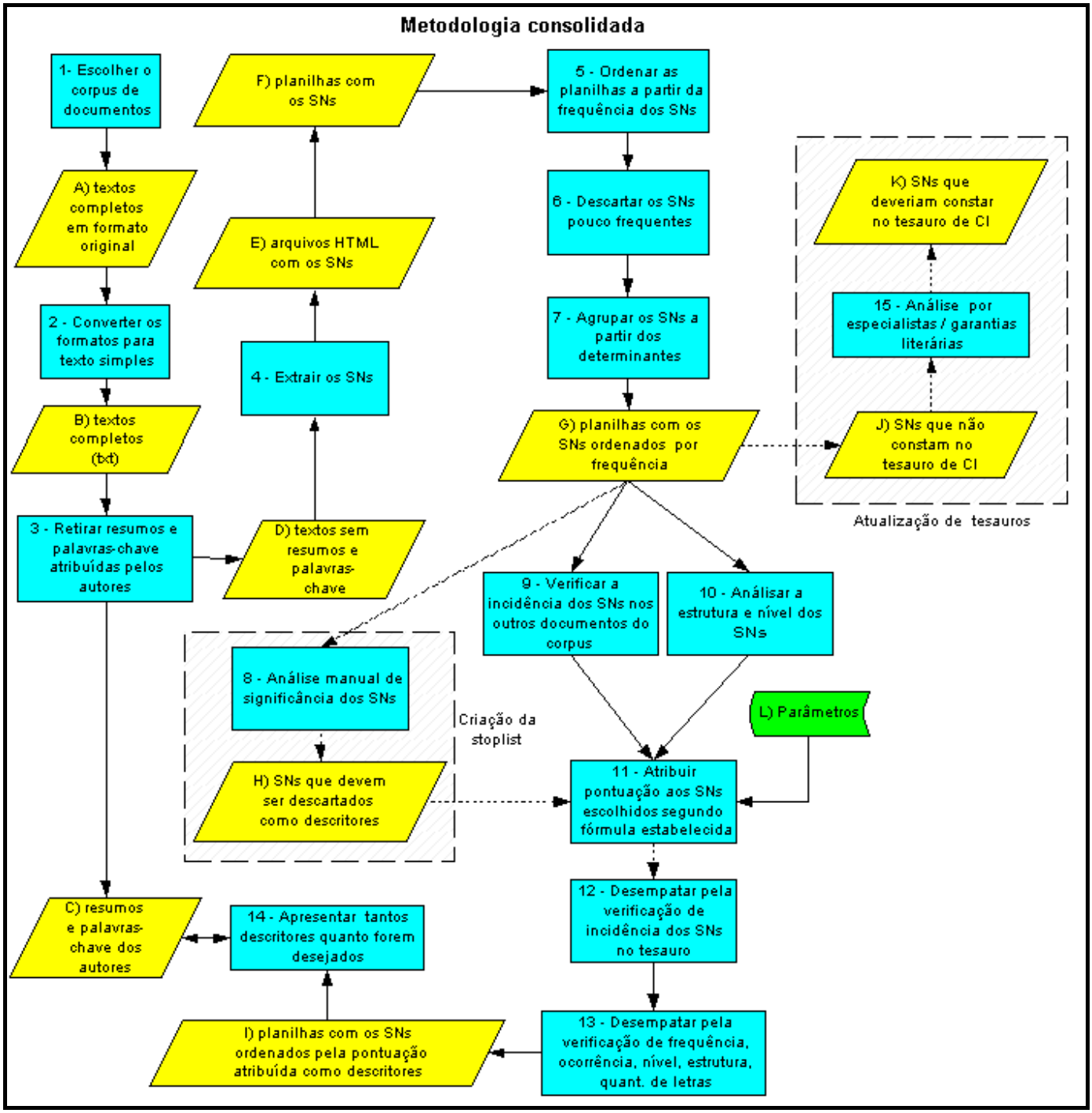


Figura 19 – Fluxograma da metodologia consolidada

E, ainda, detalhamento dessa metodologia, passo a passo, conforme executada na presente pesquisa:

1. **Escolher *corpus* significativo de documentos reconhecidamente inseridos dentro de uma área de conhecimento, como universo empírico desta pesquisa;**
2. **Converter os formatos de arquivo para texto simples;**
3. **Retirar os resumos e as palavras-chave atribuídas pelos autores;**
4. **Extrair os sintagmas nominais do corpo do texto;**
5. **Ordenar os SNs nas planilhas através da verificação da freqüência de ocorrência dos sintagmas nominais nos documentos;**
6. **Descartar os SNs que apresentam freqüências de ocorrência inferiores a um patamar preestabelecido;**
7. **Agrupar os SNs remanescentes a partir dos determinantes em suas formas "canônicas", e reordená-los;**
8. **Analisar manualmente os SNs pré-escolhidos e decidir sobre a sua relevância como descritores, para fins de construção de uma *stoplist* e verificar se algum SN escolhido consta em uma *stoplist*, dinamicamente construída, para, se for o caso, descartá-lo (em 11);**
9. **Verificar a incidência dos SNs nos outros documentos do *corpus*;**
10. **Analisar a estrutura e o nível dos SNs;**
11. **Atribuir pontuação e ranquear os SNs remanescentes de acordo com fórmula estabelecida** (explicitada a seguir)**, levando em conta a freqüência de ocorrências no texto e a freqüência de saturação definida, e a quantidade de textos do *corpus* em que ocorrem, a estrutura sintática e o nível do SN.** Esses critérios de relevância são regidos por parâmetros (representados na FIG. 19 em **L**) a serem sintonizados com a sucessiva aplicação da metodologia;
12. **Em caso de "empates" nos valores da pontuação dos SNs, considerar a ocorrência no tesauro da CI como fator de desempate;**

13. **Caso ainda ocorram “empates” nos valores da pontuação dos SNs, considerar os seguintes critérios de desempate:**
    a. **Maior valor absoluto da freqüência de ocorrência;**
    b. **Menor valor absoluto da ocorrência no *corpus*;**
    c. **Maiores nível e estrutura do SN;**
    d. **Maior quantidade de letras do SN;**
14. **Apresentar tantos descritores quanto forem desejáveis, a partir da lista ranqueada de candidatos a descritores.**

A lista ranqueada foi utilizada para a avaliação da metodologia consolidada, através da comparação com os resumos e as palavras-chave atribuídos pelos autores.

Os parâmetros customizáveis propostos, mencionados no item **11** dos passos descritos acima, possuem a característica de poderem ser alterados dinamicamente, de acordo com a *performance* dos dados de um *corpus* testado. No entanto, o dimensionamento minucioso desses parâmetros e de suas inter-relações, de modo a oferecer à metodologia performance ótima, é tarefa complexa, que demandaria muito mais tempo de análise do que a presente pesquisa se propôs a realizar. Por ora, assumiremos alguns conjuntos de valores para os quais as observações preliminares conferiram boa performance.

Para essa fase, utilizou-se uma fórmula para atribuir a pontuação, para efeitos de ranking, como apresentado a seguir:

$$Pontuação\,(SN) = \left[\left(k1 * \ freqüência\,(Xar)\right) - \left(k2 * ocorrência\,(Ytot)\right) + \left(k3 * CSN\right)\right]$$

Sendo que:

- **Pontuação(SN)**: valor atribuído ao SN de acordo com os critérios apresentados. Quanto maior for esse valor, maior a relevância esperada deste SN como descritor;
- **freqüência(Xar**) = freqüência do SN no artigo, com valor possivelmente limitado à **X** de modo a corrigir distorções;

- **ocorrência(Ytot)** = número de artigos em que o SN ocorre com freqüência maior que **Y**;
- **X, Y, k1**, **k2** e **k3** = constantes ajustadas de acordo com os testes, de modo a conseguir a performance ótima;
- **CSN** = categoria do SN, que assume um valor segundo a estrutura sintática e nível do SN, de acordo com a TAB. 14:

CSN	Estrutura e Nível do SN	Valor associado
1a	Nível 1, estrutura (D + N)	0,25
1b	Nível 1, qualquer estrutura exceto (D + N)	0,75
2	Nível 2, qualquer estrutura	1,0
3	Nível 3, qualquer estrutura	0,75
4	Nível 4, qualquer estrutura	0,5
5	Nível 5 ou superior, qualquer estrutura	0,25

Tabela 14 – Valor atribuído ao SN de acordo com sua estrutura sintática e nível

Para efeitos de otimização, testamos os resultados com alguns valores diferentes de constantes quando da apresentação dos resultados.

Espera-se que com esta metodologia alterada, possam ser obtidos resultados melhores do que os conseguidos neste teste inicial que, somando descritores excelentes, razoavelmente bons e moderadamente aceitáveis, obteve cerca de 60% de SNs relevantes semanticamente como descritores. A caracterização dos graus de relevância dos SNs como descritores foi estabelecida através da comparação com as palavras-chave e resumos atribuídos pelos autores dos documentos. Esses resultados apresentados na seção seguinte, e discutidos na seção posterior.

6.2 – A análise final dos dados

Nesta seção apresenta-se a metodologia consolidada, delineada na seção anterior, aplicada ao *corpus* completo de 60 documentos, dividido, como apresentado na seção 4.1, nos seguintes conjuntos:

- O primeiro com 30 textos, sendo que 29 provenientes da Revista *DataGramaZero*, e 1 proveniente da Revista *Ciência da Informação*, constantes no Anexo A desta tese com numeração de 1 a 30.;

- O segundo com 30 textos, todos provenientes da Revista *Ciência da Informação*, constantes no Anexo A deste documento com numeração de 31 a 60.

Os textos provenientes da revista *Ciência da Informação* apresentaram tamanho ligeiramente maior. A aplicação e análise de forma isolada da metodologia consolidada permitiram vislumbrar as diferenças decorrentes do tamanho dos documentos.

Os valores de parâmetros constantes da TAB. 15 foram escolhidos de forma arbitrária, e devem ser modificados e testados de forma exaustiva, em pesquisas posteriores, visando refinar paulatinamente a metodologia. Esses valores e parâmetros são apresentados a seguir:

- O número de descritores escolhidos para cada documento foi calculado, tendo como base 1% dos SNs únicos identificados no documento, e levando em conta os limites inferior de 8 e superior de 15 descritores por documento. Como apontado anteriormente, esse valor foi limitado apenas por uma conveniência metodológica, não havendo limitações reais para a escolha do número de descritores, excetuando o total de SNs extraídos;
- Seguindo a fórmula introduzida na seção 6.1, os valores escolhidos para as constantes **X, Y, k1**, **k2** e **k3**, nas duas aplicações da metodologia ao *corpus* final são os apresentados na TAB. 15:

Constantes	Conceituação	Conjunto de valores na primeira aplicação	Conjunto de valores na segunda aplicação
X	Valor máximo a ser considerado para a freqüência do SN no documento, para fins de pontuação.	10	7
Y	Limite inferior de freqüência do SN para o qual k2 se aplica.	3	3
k1	Ponderação da freqüência do SN no documento no cálculo da pontuação.	1	1
k2	Ponderação (negativa) da freqüência do SN no *corpus* de documentos no cálculo da pontuação.	10	15
k3	Ponderação da estrutura do SN no cálculo da pontuação.	10	15

Tabela 15 – Valores atribuídos às constantes na aplicação da metodologia

Como já se ressalvou, a manipulação intensiva das várias possibilidades, necessária para descobrir, para cada *corpus* característico, os valores ideais a serem adotados, foge ao escopo deste trabalho. Nestas duas aplicações, modularam-se os valores de forma a privilegiar a influência da freqüência (primeira aplicação) ou da estrutura do SN (segunda aplicação) no cálculo da pontuação dos SNs. Apesar de haverem sido utilizados valores para os quais foram observados resultados razoáveis, apenas foram esboçadas ínfimas parcelas da miríade de possibilidades.

As tabelas que se seguem apresentam resultados da extração de SNs do *corpus* completo e das duas aplicações da metodologia consolidada, com os valores de constantes apresentados na TAB. 15:

- A **TAB. 16** apresenta algumas informações gerais sobre o número de SNs totais, únicos e selecionados para descritores, nos 60 artigos que compuseram o *corpus*; as médias, e o percentual dos SNs únicos dentre os totais, e dos selecionados dentre os únicos, ressalvando-se o máximo de 10 descritores por documento;

- A **TAB. 17** apresenta, para os dois conjuntos de parâmetros de aplicação da metodologia, e para os dois subconjuntos de documentos do *corpus*, os seguintes dados:
    - As médias e os valores percentuais relativos de freqüência de SNs extremamente relevantes como descritores (SNs***), razoavelmente relevantes como descritores (SNs**), moderadamente relevantes como descritores (SNs*) e não relevantes como descritores (SNs-);
    - A média e o valor percentual dos "*stopwords"* (SW) em relação ao total dos SNs que foram eliminados.
    - A taxa de relevância média do conjunto, calculada através da fórmula apresentada na seção 4.4.
- A **TAB. 18** é, na verdade, um painel formado de 4 histogramas, onde são apresentados graficamente os mesmos dados da TAB. 17.

O **Anexo C** desta tese apresenta os títulos dos artigos e os descritores que foram atribuídos em cada uma das aplicações da metodologia, e o **Anexo D** apresenta a lista de SNs que foram escolhidos para compor a *stoplist* (os SW).

Artigos	Número de SNs			Artigos	Número de SNs		
DGZ	totais	únicos	selec.	**CI***	totais	únicos	selec.
1	1673	1343	13	31	1702	1528	15
2	842	711	8	32	1902	1213	12
3	783	680	8	33	1941	1290	13
4	801	688	8	34	1480	1231	12
5	1478	1252	13	35	1011	788	8
6	984	836	8	36	735	552	8
7	638	521	8	37	2054	1382	14
8	779	684	8	38	772	624	8
9	1104	932	9	39	1873	1284	13
10	1146	1035	10	40	1156	962	10
11	619	554	8	41	1008	792	8
12	791	626	8	42	1244	1002	10
13	1342	1113	11	43	1808	1325	13
14	923	747	8	44	1375	1145	11
15	1063	877	9	45	1420	1176	12
16	888	810	8	46	1829	1453	15
17	1201	1084	11	47	987	810	8
18	5686	4287	15	48	1498	1223	12
19	1094	899	9	49	884	760	8
20	1299	1039	10	50	852	677	8
21	733	616	8	51	1225	1009	10
22	1837	1368	14	52	547	483	8
23	796	699	8	53	1364	1062	11
24	2048	1434	14	54	1535	1174	12
25	1368	988	10	55	1144	840	8
26	1246	1058	11	56	1386	1119	11
27	1173	971	10	57	1702	1353	14
28	788	667	8	58	1497	1166	12
29	617	539	8	59	733	632	8
30*	633	506	8	60	1702	951	10
Médias	1212,43	985,47	9,65	Médias	1345,53	1033,53	10,69
%		81,28%	0,98%	%		76,81%	1,03%

Tabela 16 – Informações sobre os SNs dos documentos do *corpus*

		Textos de 1 a 30 do *corpus*			Textos de 31 a 60 do *corpus*		
I	**Primeira aplicação da metodologia**	SNs***	138	47,75%	SNs***	179	55,59%
		SNs**	66	22,84%	SNs**	63	19,57%
		SNs*	58	20,07%	SNs*	58	18,01%
		SNs–	27	9,34%	SNs–	22	6,83%
		SW	19	6,17%	SW	17	5,01%
		Taxa de Relevância		**0,64**	**Taxa de Relevância**		**0,70**
II	**Segunda aplicação da metodologia**	SNs***	137	47,40%	SNs***	173	52,58%
		SNs**	64	22,15%	SNs**	64	19,45%
		SNs*	56	19,38%	SNs*	64	19,45%
		SNs–	32	11,07%	SNs–	28	8,51%
		SW	5	1,70%	SW	7	2,08%
		Taxa de Relevância		**0,63**	**Taxa de Relevância**		**0,67**

Tabela 17 – Freqüências dos SNs segundo a relevância semântica

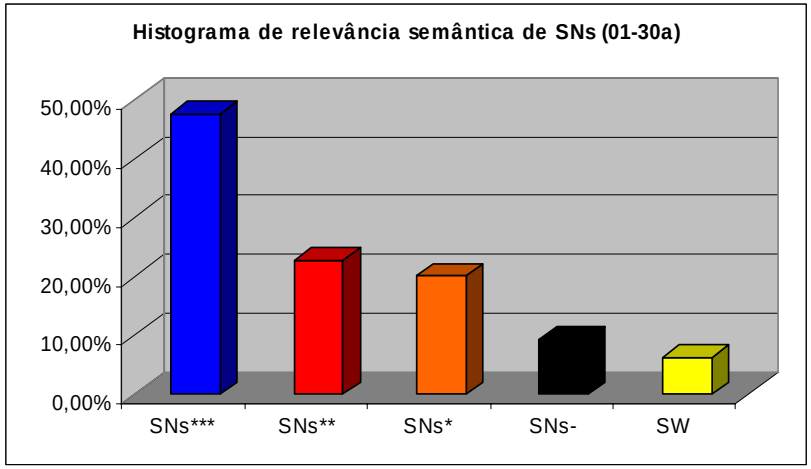
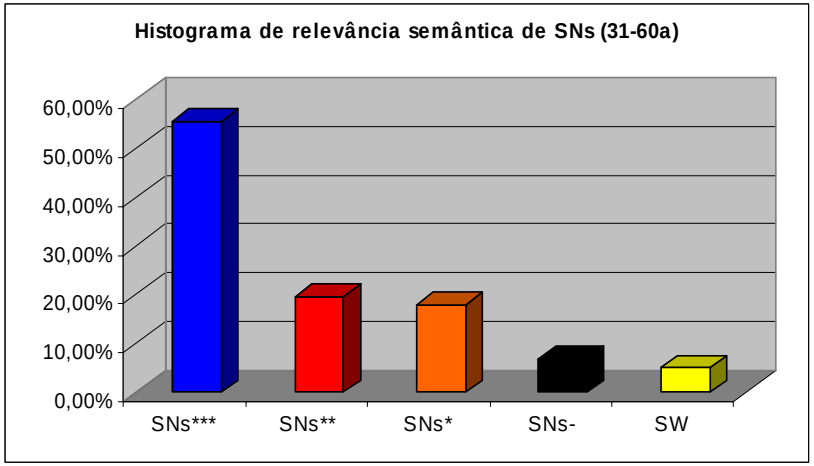
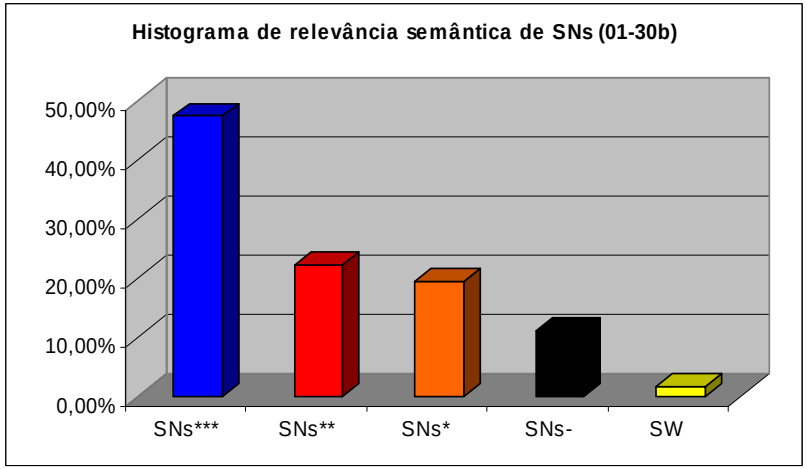
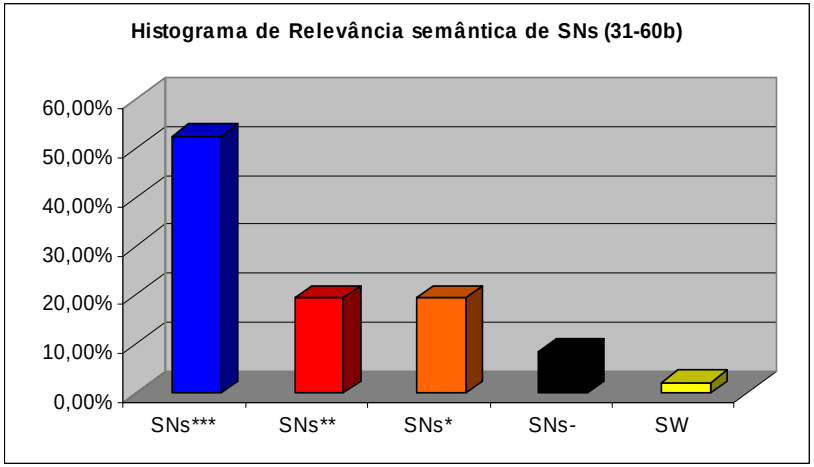
	Textos de 1 a 30 do *corpus*	Textos de 31 a 60 do *corpus*
Primeira aplicação da metodologia		
Segunda aplicação da metodologia		

Tabela 18 – Histogramas de freqüências dos SNs segundo a relevância semântica

6.3 – Discussão dos resultados

Esta seção devota-se a discutir os resultados apresentados nas tabelas anteriores e nos quadros indicativos que constam dos **Anexos B e C** deste documento. Inicialmente, são apresentadas as considerações comparativas entre o uso de SNs e palavras-chave como descritores, e posteriormente, será avaliada de maneira geral a metodologia consolidada, seus resultados e as possíveis conclusões.

6.3.1 – Comparação entre SNs e palavras-chave como descritores

Os dados apresentados no **Anexo B** desta tese permitiram realizar comparações entre as densidades informacionais e relevâncias relativas como descritores entre as palavras-chave e os SNs, mostrando de forma evidente e inequívoca que a densidade informacional dos sintagmas nominais supera em muito àquela percebida pela análise semântica das palavras-chave. Essa comparação foi possível a partir dos testes realizados com a aplicação da metodologia prospectiva ao subconjunto de seis documentos do *corpus*,

Podemos apontar as seguintes vantagens dos SNs como descritores, se comparados às palavras-chave:

1. Como característica e diferencial mais importante, verificou-se que os SNs mantêm o contexto das palavras que os compõem, permitindo a não fragmentação do discurso;

   Ex: A “quebra” de nomes próprios “Rio de Janeiro” e “São Paulo” (artigo 2), que poderia ter como conseqüência o descarte das palavras ”Rio”, “Janeiro”, “São” e “Paulo”; “o valor de uma unidade de conhecimento registrada”, em comparação com “valor”, “unidade”, “conhec*” e “registr*”, (artigo 1); “conhecimento científico”, em comparação com “conhec*” e “ciên*” (artigo 2);

2. Os SNs permitem melhor decisão sobre a relevância dos termos que, como palavras isoladas, podem ser considerados como *stopwords*;

Ex: A “quebra” de nomes próprios como “São Paulo” e “Rio de Janeiro” nas quais as partes dos nomes próprios “São” e “Rio” poderiam ser confundidas com os verbos homônimos, e descartadas (artigo 2); o caso do SN “linguagens não verbais”, em que o qualificador “não” poderia ser descartado (artigo 5);

3. Por não passarem pelo processo de *steeming*, os SNs ofereceram diferencial informacional em relação às palavras-chave, que foram armazenadas nos índices de forma indiferenciada como seus morfemas constituintes.

   Ex: Os lexemas “informação” e “informacionais”, qualitativamente bastante diferentes, seriam reduzidos ao um mesmo morfema (artigo 3); “gerenciador” e “gerenciamento” (artigo 6);

4. Para as altas freqüências, foram visíveis as diferenças qualitativas entre os SNs e as palavras-chave, mesmo com a eliminação das *stopwords*, na capacidade de descrever o tema dos documentos;

   Ex: “interface de consulta” em oposição a “interface” e “consulta” (artigo 6); “direitos autorais” em oposição a “direitos” e “autorais” (artigo 4).

Esses fatos, que corroboraram o apresentado por KURAMOTO (1996 e 1999), por si só elevariam as metodologias apresentadas a um patamar digno de consideração; mas o sucesso em descrever o assunto ou “tema” dos artigos é o maior critério de avaliação. Este assunto é discutido na subseção a seguir e nas considerações finais desta tese.

6.3.2 – Avaliação geral da metodologia consolidada

Nesta subseção são analisados os resultados da aplicação da metodologia consolidada ao *corpus* completo, segundo os dois conjuntos de parâmetros, como exposto nas seções 6.1 e 6.2. As TAB. 16, 17 e 18 sintetizam os resultados da aplicação da metodologia.

Ao analisarmos as características do *corpus*, notamos que os 30 primeiros textos apresentam média de aproximadamente 1212 SNs identificados, sendo 985 a

média dos SNs únicos – 81% do total. Os 30 textos subseqüentes apresentaram média de aproximadamente 1345 SNs, sendo 1033 a média dos SNs únicos – 76% do total. Isto indica que os textos da segunda metade do *corpus* são maiores, e que seus SNs se repetem com mais freqüência.

Os resultados, na ótica do autor, superaram em muito a expectativa inicial. As taxas de relevância dos SNs escolhidos, respectivas às duas metades do *corpus*, foram de 0,64 e 0,70 (média de 0,67) para a aplicação com o primeiro conjunto de valores para os parâmetros; e de 0,63 e 0,67 (média de 0,65) para o segundo conjunto de valores.

Ao compararmos os resultados apresentados na TAB. 18 da seção 6.2 com aqueles obtidos na aplicação da metodologia prospectiva – apresentados na FIG. 18 da subseção 5.2.3; pudemos perceber grande diferença: partindo do valor de apenas 12,4% e 15,2% para SNs extremamente relevantes e razoavelmente relevantes, respectivamente, saltamos – no pior caso de aplicação da metodologia consolidada – para os valores de 47% e 22,15%, para os SNs de mesma qualidade. Isso representou o total de quase 70% de bons descritores (extremamente relevantes + razoavelmente relevantes) e aumento de mais de 150% em comparação à aplicação da metodologia prospectiva. A TAB. 19 sintetiza esses resultados:

Relevância dos SNs	**Valor na aplicação da metodologia prospectiva**	**Piores valores na aplicação da metodologia consolidada**
SNs extremamente relevantes como descritores	12,40%	47,40%
SNs razoavelmente relevantes como descritores	15,20%	22,15%
SNs moderadamente relevantes como descritores	33,30%	19,38%
SNs não relevantes como descritores	39,00%	11,07%

Tabela 19 – Comparação dos resultados na duas aplicações da metodologia

A aplicação da metodologia prospectiva selecionou descritores com base apenas no cálculo das freqüências de ocorrência e no descarte de SNs com certa estrutura e para certas freqüências. A metodologia final adotada utilizou um algoritmo complexo e parametrizável, que levou em conta as freqüências de SNs

nos textos, no conjunto de textos, a estrutura e o nível dos SNs. Essa flexibilidade permitiu ainda que possamos melhorar os resultados a cada nova aplicação.

Dos resultados apresentados nas tabelas anteriores, pudemos destacar alguns pontos de avaliação, relativos às duas aplicações da metodologia final:

- As medidas de qualidade dos resultados não variaram em demasia com a variação dos valores dos parâmetros, sendo que as diferenças maiores dos resultados se deram em relação às aplicações nas duas metades do *corpus*. Esses dados, se analisados na perspectiva de que os artigos da *Revista Ciência da Informação,* eram sensivelmente maiores, indicando que quanto maiores os textos – e o número de ocorrências repetidas de SNs – melhores os resultados (ao menos para algumas faixas de tamanhos de documentos). Estes resultados também puderam ser interpretados à luz da variação temática de uma e de outra revista;
- A sensível piora dos resultados na segunda aplicação, quando foram escolhidos parâmetros que privilegiavam a análise estrutural em detrimento da análise de freqüência, pode indicar que os parâmetros já estavam mais bem sintonizados em relação à estrutura e à freqüência na primeira aplicação. Demandaram-se aplicações exaustivas para encontrar valores próximos ao ideal para cada tipo de *corpus*, em relação às áreas de assunto;
- Os SNs que continham palavras em inglês foram deliberadamente ignorados. Caso não o fossem, em sua grande maioria, poderiam tornar-se bons descritores, melhorando os resultados;
- A escolha de um critério que limitava a quantidade de descritores escolhidos também fez com que, por vezes, muitos bons descritores fossem eliminados;
- Pôde-se notar claramente a diminuição das *stopwords* dentre os SNs escolhidos, quando privilegiamos a estrutura em detrimento da freqüência, no cálculo da pontuação. Isso nos impeliu a privilegiar a

freqüência quando a *stoplist* estivesse disponível ou estiver sendo escolhida; ou a estrutura, quando não houve *stoplist* disponível;

- Um ponto importante a ser enfatizado é que, por vezes, a caracterização do texto através dos SNs escolhidos automaticamente é mais fidedigna, em relação ao conjunto de assuntos tratados no documento, do que a percebida através daquelas palavras-chave atribuídas pelos próprios autores, que por vezes enfatizaram ponto de vista particular e embotado;
- Num ponto certamente subjetivo, a avaliação da relevância dos SNs pelo autor desta tese foi bastante rigorosa e exigente em relação ao significado em relação ao assunto do texto. A avaliação realizada por terceiros pode apresentar resultados ainda melhores para a metodologia;
- O conjunto de SNs escolhidos para cada texto possui um grande poder de caracterização do assunto, como pode ser examinado qualitativamente no Anexo C. Deve-se considerar, a título de avaliação do sucesso da metodologia, a dificuldade de escolha de número elevado de descritores significativos no processo de indexação manual;
- O uso do tesauro, mesmo tendo sido relegado às situações de desempate, quando a pontuação dos SNs era semelhante, não se mostrou decisivo para a escolha dos melhores descritores. Os motivos podem ser aqueles apresentados na subseção 5.2.3.

Podem-se esperar resultados ainda melhores para documentos provenientes de certas áreas do conhecimento, como as ciências exatas, uma vez que a multi-temacidade é a característica marcante das ciências sociais aplicadas, nas quais se encaixa a ciência da informação.

Essas constatações apontam para a confirmação da avaliação positiva da metodologia, e apontam caminhos para sua melhoria em pesquisas futuras. Os próximos capítulos apresentam as conclusões e devotam-se à análise dos resultados à luz das teorias apresentadas e os possíveis e diversos caminhos de pesquisa que se afiguram.

7 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste capítulo, retomam-se inicialmente os principais pontos desta pesquisa, para então avaliar os resultados à luz das teorias e contextos de aplicação apresentados, de acordo com os pressupostos estabelecidos. Finalmente, são delineados alguns possíveis trabalhos futuros.

A motivação da pesquisa surgiu da constatação freqüente da impossibilidade de organização manual de grandes acervos de documentos que são continuamente produzidos, como acontece em muitos contextos digitais. Nestes contextos, observamos amiúde processos de indexação automática que buscam descrever os documentos através da análise de freqüência das palavras que neles ocorrem. O objetivo central desta investigação era propor um processo de indexação mais eficaz, que analisasse as palavras e expressões dentro de seus contextos lingüísticos.

O objetivo primordial da pesquisa era validar e apresentar metodologia de indexação automática, viabilizando o processo de atribuição de descritores a documentos digitalizados. Estes descritores foram escolhidos através da extração de SNs e da análise de fatores como a freqüência de ocorrência desses SNs nos textos dos documentos, no conjunto dos documentos; a estrutura dos SNs; o nível dos SNs e a ocorrência desses em um tesauro de um campo de conhecimento específico. A consideração desses fatores de forma conjunta permitiria a criação do ranking de candidatos a descritores, a partir dos SNs extraídos.

Para atingir esses objetivos foram analisados os *corpora*, para a) validar o processo de extração automática e b) testar e melhorar, em duas etapas, a eficácia da metodologia.

Os pressupostos de pesquisa foram positivamente confirmados, mesmo ressalvando-se as modificações introduzidas na metodologia original pelo abandono o uso do tesauro como uma das formas principais de seleção de descritores. Os resultados respectivos são comentados de forma sucinta a seguir:

1. A utilização de sintagmas nominais como descritores apresentou vantagens em relação ao uso de palavras-chave, como mostra a comparação

realizada na subseção 6.3.1. O fato de serem inerentemente mais significativos e trazerem em seu bojo o contexto semântico dos discursos faz com que sejam melhores descritores do que as palavras-chave isoladas;

2. A extração automática de sintagmas nominais com as ferramentas apresentadas na seção 4.3 se mostrou extremamente viável, para os propósitos da pesquisa – embora ainda não esteja em pé de igualdade na comparação qualitativa, com a extração manual. A melhoria das ferramentas pode fazer com que a qualidade da extração em um futuro próximo seja comparável à manual;

Além da maior velocidade, o argumento adicional favorável à extração automática advindo das teorias estudadas é o fato da extração manual incorrer em problemas típicos de falta de coerência metodológica ao longo do tempo para o mesmo indexador, fato ainda agravado se considerarmos diferentes indexadores. Esses aspectos foram apontados por O'BRIEN e CHU (1993), LANCASTER (1993, pp 61-74), PINTO MOLINA (1994), FUJITA (1999), NAVES (2001), entre outros, além de ter sido verificado para o caso específico do *corpus* de textos utilizado, por meio de trocas de mensagens e colóquios informais entre o autor desta tese e o prof. Dr. Hélio KURAMOTO.

3. E o último e principal pressuposto tergiversava sobre a possibilidade de estabelecer processo automatizado e eficaz para a escolha de descritores significativos para textos digitalizados, utilizando sintagmas nominais. Esse pressuposto central se confirmou, como se pôde verificar nos resultados apresentados ao longo do capítulo 5.

A metodologia prospectiva foi aplicada à parte do *corpus* para validação e parametrização das variáveis do algoritmo, e então a metodologia modificada e consolidada foi aplicada à totalidade do *corpus*. Nessa derradeira aplicação, dois conjuntos de valores de parâmetros foram utilizados, dentre um universo virtualmente ilimitado de possibilidades. Os testes exaustivos com outros conjuntos foram deixados como sugestões para trabalhos futuros.

Os resultados, considerados eminentemente positivos, contrariam experiências anteriores declaradamente malsucedidas, que buscavam a extração de descritores

baseando-se em estruturas sintáticas das orações [(EARL, 1970; PAICE, 1981; Fum et. al., 1982) apud LANCASTER, 1993, p. 250-251]. A bem da verdade, a inexistência, até a uma década, de ferramentas que permitissem a extração automática de SNs é um fator preponderante a ser levado em conta neste sentido.

A teoria desenvolvida por KURAMOTO (1999, 2003) e seu modelo proposto de SRI já apontava alguns caminhos possíveis, embora esses ainda estejam em estágio inicial de exploração. A pesquisa desenvolvida em sua tese de doutorado apresentou modelo de recuperação de informações baseado em sintagmas nominais, buscando a participação do usuário na definição dos contextos lingüísticos. Infelizmente, não encontramos na literatura científica nacional indício algum de continuação dessas pesquisas.

Ao que parece, a visão mais estrita de LIBERATO (1997) sobre a caracterização possível dos SNs não se confirmou como fator limitante para a avaliação do funcionamento do *parser* PALAVRAS (1996), com sua gramática de restrições, embora ainda fosse visível a diferença de performance qualitativa entre os processos automático e manual. Pode-se esperar que os *parsers* sejam continuamente melhorados e que novas pesquisas surjam.

Espera-se que a metodologia consolidada – ou qualquer metodologia que derive desta – seja utilizada em situações nas quais seja necessária a atribuição automática de descritores aos documentos, no escopo de funcionamento de SRIs. Usualmente, essa situação acontece, quando os documentos são agregados ao sistema em uma taxa que não permite a apreciação manual. Dentre os contextos de aplicabilidade, apresentados no capítulo 3, as bibliotecas digitais são grandes candidatas a terem seu acervo tratado de alguma forma automática, para que seja realizada a indexação de assuntos. Além das bibliotecas digitais, a *web*, com sua impressionante massa de documentos em várias mídias, é um dos espaços nos quais seria desejável tratamento *a posteriori* – se não for o único plausível – para fins de classificação por assunto.

Das quatro estratégias apresentadas na introdução para melhoria dos sistemas de recuperação de informações, talvez a menos explorada tenha sido a análise da semântica intrínseca aos textos dos documentos. Acreditamos que esse panorama

possa ser modificado através de outras pesquisas como a presente investigação, e que, sem ufanismo, a proposta metodológica desenvolvida nesta tese seja uma das alavancas propulsoras.

Embora se tenha constituído a partir de muitas contribuições, o presente trabalho pode ser considerado seminal, na medida em que abre caminho para aperfeiçoamento constante de metodologias de extração de descritores que levem em conta estruturas sintáticas derivadas da gramática sintagmática. Ao fazer tal afirmação, reforça-se, não estão sendo desconsideradas as diversas pesquisas anteriores e em paralelo, que procuraram acrescentar aos estudos de freqüências de palavras-chave a possibilidade de consideração de estruturas sintáticas, gramaticais, frasais e textuais, além da gama variada de novas estratégias integradas para melhoria dos processos de representação e recuperação de informações.

Cumpre ressaltar que o referencial teórico de ‘Processamento de Linguagem Natural’, utilizado para a construção desta tese advém prioritariamente da literatura da área de Ciência da informação, cujas pesquisas obtiveram maior efervescência a partir da década de 1970. Entretanto, o autor desta tese, não ignora os avanços que têm sido alcançados em áreas como a lingüística computacional aplicada, a ciência da computação e estudos interdisciplinares para a recuperação de informação, a despeito do fato dessas contribuições não haverem sido contempladas em sua totalidade no escopo desta tese. Sua consideração se constitui um imperativo para trabalhos futuros, como atualização e aproximação necessárias para a fertilização da área da Ciência da Informação.

Tendo isso posto, e a partir da teoria e dos resultados empíricos analisados anteriormente, podemos enumerar uma série de caminhos de pesquisa que poderiam redundar em melhorias metodológicas, detalhados a seguir:

1. Considerar a inclusão na metodologia de análise estrutural dos textos dos documentos, na forma que propõe KOBASHI (1994). As considerações relativas à análise da densidade informacional podem ser incorporadas à metodologia, de maneira que os *parsers* apresentem algum tipo de ponderação que leve em conta as seções mais importantes do documento;

2. Considerar os avanços que vem sendo realizados no *parser* PALAVRAS (BICK, 1996) e em outras iniciativas de estruturação de analisadores sintáticos; e, se possível, criar estrutura nacional unificada de tecnologias e ferramentas para estudos lingüísticos;
3. Considerar o desenvolvimento e a utilização de *parsers* que levam em conta a teoria advinda das gramáticas transformacionais, e incorporar outros aportes da lingüística para a recuperação de informações;
4. Considerar o poder descritivo de outras estruturas sintáticas, como os sintagmas verbais, e combinações entre as várias estruturas;
5. Considerar as construções globais *a priori* e *a posteriori* de *stoplists* de SNs freqüentes que, para uma dada área de conhecimento, apresentam reduzido valor informacional;
6. Experimentar exaustivamente a variação dos parâmetros e constantes apresentados na metodologia da presente pesquisa, até que se consigam os melhores resultados possíveis, para determinada área do conhecimento e conjunto de características dos *corpora*.

Além desses caminhos, que buscam obter maior eficácia da metodologia proposta, também podemos considerar a extrapolação do processo em uma miríade de novos caminhos, como, por exemplo:

7. Analisar as possibilidades de utilização da metodologia em outros idiomas, como o inglês e o francês, e realizar comparações;
8. Adaptar o mecanismo de indexação delineado para que se possam realizar buscas em repositórios de documentos baseadas em SNs (KURAMOTO, 1999), desta vez com a possibilidade de extração automática dos SNs;
9. Utilizar a metodologia para realizar levantamentos terminológicos em *corpora*, para diversos fins como: verificação de completude e atualização de tesauros.

E finalmente, há que se considerar as possibilidades de adaptações para usos totalmente diversos, a serem apropriados em outras áreas do conhecimento, como as exemplificadas a seguir:

10. Atividades de monitoramento ambiental de informações, como *text mining*, *clipping* de notícias, e outras;
11. A análise da qualidade literária de documentos, análise de estilos e autoria; através de estudos estatísticos de freqüências de expressões;
12. A identificação de neologismos e auxílio na tradução automática;
13. A construção e a validação de ontologias no contexto da web semântica, dentre muitas outras.

Como foi apontado na introdução desta tese, dentre os caminhos de pesquisa para melhoria de SRIs, as estratégias voltadas para a exploração da semântica intrínseca dos documentos talvez sejam as que apresentem menor volume de esforços de pesquisa. Entretanto, acreditamos que apresentem um grande campo de exploração futuro, a despeito do claudicante caminho percorrido pela pesquisa em inteligência artificial. No ato de decifrar os recônditos do discurso humano está a chave para a efetiva comunicação homem-máquina.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ABRAHÃO, P. R. Carneiro. *Modelagem e Implementação de um Léxico Semântico para o Português*. 1997. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.

2. ALVARENGA, Lídia. Representação do Conhecimento na perspectiva da Ciência da Informação em Tempo e espaço Digitais. *Encontros Bibli*, 2003 Disponível em: <http://www.encontros-bibli.ufsc.br/Edicao_15/alvarenga_representacao.pdf> . Acesso em: out. 2003.

3. ARAÚJO, Vânia M.R.H. *Sistemas de recuperação da informação*: nova abordagem teórico conceitual. 1994. Tese (Doutorado em Ciência da Informação). Universidade Federal do Rio de Janeiro, Rio de Janeiro.

4. AUSTIN, Derek. *PRECIS*: a manual of concept analysis and indexing. 1984.

5. BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval.* New York: ACM Press, 1999. 511p.

6. BARATIN, Marc e JACOB Christian (orgs.). *O Poder das Bibliotecas:* a memória dos livros no ocidente. Rio de Janeiro: Editora UFRJ, 2000. 351p**.**

7. BHATTACHARYYA, G. POPSI: its fundamentals and procedure. *Library Science with a slant to Documentation*. V.16, N.1, 1979, p.1-34.

8. BECK, U. *Risk Society:* towards a new modernity. London: Sage, 1992.

9. BERNERS-LEE, T., LASSILA, Ora. e HENDLER, James. The Semantic Web. *Scientific America*, Maio de 2001. Disponível em: <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21>. Acesso em: jun. 2003.

10. BICK, Eckhard. *Parsers and its applications*. (s/d) Disponível em: <http://www.hum.au.dk/lingvist/lineb/home_uk.htm>. Acesso em: jul. 2003.

11. _____. *Automatic parsing of Portuguese*. In: Proceedings of II Encontro para o Processamento Computacional do Português Escrito e Falado, SBIA, 1996, Curitiba. Disponível em: <http://beta.visl.sdu.dk/~eckhard/postscript/curitiba.ps>. Acesso em: jul. 2003.

12. _____. *The VISL System*: research and applicative aspects of IT-based learning. In: Proceedings of NoDaLiDa, Uppsala. 2001. Disponível em: <http://stp.ling.uu.se/nodalida01/pdf/bick.pdf>. Acesso em: jul. 2003.

13. BUCKLAND, Michel. Information as thing. *Journal of American Society of Information Science*. v.42, n.5, 1991. p. 351-360.

14. CAMPOS, Maria Luiza de Almeida. *A organização de unidades de conhecimento em hiperdocumentos*. 2001. Tese (Doutorado em Ciência da Informação) IBICT, UFRJ, Rio de Janeiro, 2001.

15. _____. *Linguagem documentária*: teorias que fundamentam sua elaboração. Niterói: EdUFF, 2001.

16. CAÑAS, A. J., LEAKE, D. B., WILSON, D. C.; Managing, Mapping, and Manipulating Conceptual Knowledge. *AAAI Workshop Technical Report WS-99-10: Exploring the Synergies of Knowledge Management & Case-Based Reasoning,* AAAI Press, Menlo Calif. Jul. 1999.

17. CASTELLS, M. *A Sociedade em Rede.* São Paulo: Paz e Terra, 1999. 617p.

18. CESARINO, Maria Augusta N., PINTO, Maria Cristina M.F. Análise de assunto. *Revista de Biblioteconomia de Brasília*, v.8, n.11, 1980, p. 33-43.

19. CHOMSKY**,** Noam. Syntactic structures. 3. ed. Paris: The Hague, 1969. 117 p

20. CHOWDHURY, G. *Introduction to modern information retrieval*. London: Library Association Publishing, 1999. 452 p.

21. CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO - CNPq / INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA - IBICT. *Tesauro de Ciência da Informação*: versão preliminar. Brasília, 1989.

22. DAHLBERG, Ingetraut. Teoria do Conceito*. Ciência da Informação*, Rio de Janeiro, v. 7, n. 2, jul./dez. 1978. p. 101-107.

23. DECKER, Stefan. et ali. *The semantic web*: the roles of xml and rdf. IEEE Expert, 15(3), October 2000.

24. ENGELBART, Douglas. *Augmenting Human Intellect*: A Conceptual Framework. 1962 Disponível em: <http://www.liquidinformation.org/engelbart/62_paper_full.pdf>. Acesso em: ago. 2003.

25. FERREIRA, Aurélio Buarque de Holanda. *Novo Aurélio Século XXI:* o dicionário da língua portuguesa*.* Rio de Janeiro: Nova Fronteira, 1999.

26. FOSKETT, A. C. *The Subject Approach to Information*. 5. ed. Londres: Library Association Publishing, 1997. 119p.

27. FUJITA, M.S.L. A leitura do indexador: estudo de observação. *Perspectivas em Ciência da Informação*, v.4, n.1, jan./jun. 1999. p. 101-116.

28. GASPERIN, Caroline Varaschin; GOULART, Rodrigo Rafael Vilarreal e VIEIRA, Renata. *Uma Ferramenta para Resolução Automática de Correferência.* In: Anais do XXIII Congresso da Sociedade Brasileira

de Computação, VI Encontro Nacional de Inteligência Artificial, Vol VII. Campinas, 2003.

29. GASPERIN, Caroline Varaschin; VIEIRA, Renata; GOULART, Rodrigo Rafael Vilarreal e QUARESMA, Paulo. *Extracting XML chunks from Portuguese corpora.* In: Proceedings of the Workshop on Traitement automatique des langues minoritaires. 2003. Batz-sur-Mer.

30. GIDDENS, A. *As Conseqüências da Modernidade.* São Paulo: Ed. Unesp, 1991.

31. GONZALEZ, M. Insaurriaga. *O Léxico Gerativo de Pustejovsky sob o enfoque da recuperação de informações*. 2000. Trabalho (Doutorado em Ciência da Computação) – Faculdade de Informática da PUC-RS – Porto Alegre.

32. _____. *Representação Semântica de sentenças em linguagem natural e sua aplicação na recuperação de informação*. 2000. Trabalho (Doutorado em Ciência da Computação) – Faculdade de Informática da PUC-RS – Porto Alegre.

33. HERMANS, B. *Inteligent Software Agents on the Internet*: an inventory of currently offered functionality in the information society & a prediction of (near) future developments. Tilburg University, Tilburg, Holanda, 1996. Disponível em: <http://www.hermans.org/agents>. Acesso em: jun. 2003.

34. HOUAISS, A. *Dicionário eletrônico Houaiss da língua portuguesa* . Rio de Janeiro: Objetiva. Versão 1.0. 1 [CD-ROM]. 2001.

35. HUTCHINS, W.J. The concept of ‘aboutness’ in subject indexing. In: JONES, Karen Spark; WILLET, Peter. *Readings In Information Retrieval.* San Francisco, Calif.: Morgan Kaufnann, 1997. p.93-97.

36. KOBASHI, Nair Yumiko. *A elaboração de informações documentárias:* em busca de uma metodologia. 1994. Tese (Doutorado em Ciência da Informação) ECA, USP, São Paulo, 1994.

37. KORFHAGE, Robert *Information Storage and retrieval*. New York: John Wiley & Sons, 1997. 349 p.

38. KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. *Ciência da Informação*, Brasília, v. 25, n. 2, 1996. Disponível em: <http://www.ibict.br/cionline/250296/25029605.pdf>. Acesso em: jul. 2003.

39. _____. *Proposition d'un Système de Recherche d'Information Assistée par Ordinateur Avec application à la langue portugaise*. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, França.

40. LAMPING, J, RAO, R. PIROLLI, P. *A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies*. 1995. Disponível em:

<http://www.acm.org/sigchi/chi95/proceedings/papers/jl_bdy.htm>. Acesso em: jul. 2001.

41. LANCASTER, F. W. *Information Retrieval Systems*. New York: John Wiley, 1968.

42. _____. *Information Retrieval Systems*: characteristics, testing and evaluation. 2nd ed. New York: John Wiley, 1979.

43. _____. *Indexação e Resumos*: teoria e prática. Brasília, Briquet de Lemos,1993.

44. LANCASTER, F. W. e WARNER, A. J. *Information Retrieval Today*. Information Resources Press, 1993.

45. LAWRENCE, Steve. Context in Web Search. *IEEE Data Engineering Bulletin*, v.23, n.3, p25-32, 2000. Disponível em: <http://citeseer.nj.nec.com/lawrence00context.html>. Acesso em: abr. 2003.

46. LÉVY**,** Pierre. *As Tecnologias da Inteligência*: o futuro do pensamento na era da informática. São Paulo: Editora 34, 1993. 203p.

47. _____. *Cibercultura*. São Paulo: Editora 34, 1999. 260p.

48. LIBERATO, Yara G. *A Estrutura do Sintagma Nominal em Português:* uma abordagem cognitiva. 1997. 203 f. Tese (Doutorado em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

49. LYONS, J. *As idéias de Chomsky.* 4ª. Edição. São Paulo: Cultrix, 1983.

50. MEADOW, Charles T. *Text information retrieval systems*. San Diego: Academic Press, 1992

51. MINISTÉRIO DA EDUCAÇÃO E CULTURA / MINISTÉRIO DA CIÊNCIA E TECNOLOGIA. *Manual de Elaboração de Tesauros Monolíngües*. Brasília: Imprensa Universitária UFSC, 1990.

52. MATTELART, Armand. *História da sociedade da informação*. São Paulo: Loyola, 2002.

53. MIORELLI, S. T. *Extração do Sintagma Nominal em sentenças em Português*. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

54. MOREIRO, José; MARZAL, Miguel Ángel; BELTRÁN, Pilar. *Desarrollo de un Método para la Creación de Mapas Conceptuales*. Anais do ENANCIB, Belo Horizonte, 2003.

55. NAVES, Madalena M. L. Análise de assunto: concepções. *Revista de Biblioteconomia de Brasília*, v.20, n.2, p. 215-226, jul./dez. 1996.

56. _____. Estudo dos fatores interferentes no processo de análise de assunto. *Perspectivas em Ciência da Informação*. Belo Horizonte, v.6, n.2, p. 189-203, jul./dez. 2001.

57. NELSON, T. H. *Literary Machines*. Sausalito, CA: Mindful Press. 1982.

58. NWANA, H.; *Software Agents*: An Overview, (1996) Knowledge Engineering Review, 11(3). p.205-244.

59. O'BRIEN, E.A. e CHU, C.M. Subject Analysis: the critical first stage in indexing. *Journal of Information Science*, v.19, 1993. p 439-454.

60. PERINI, Mário A. *A Gramática Gerativa*: introdução ao estudo da sintaxe portuguesa. 2. ed. Belo Horizonte: Vigília, 1985. 254 p.

61. _____. *Gramática descritiva do português*. 2. ed. São Paulo: Editora Ática, 1995. 380p.

62. PERINI, Mário A.; FRAIHA, Sigrid; FULGÊNCIO, Lúcia; BESSA NETO, Regina. O SN em português: a hipótese mórfica. *Revista de Estudos de Linguagem* - UFMG, Belo Horizonte, Julho / Dezembro 1996. p. 43-56.

63. PINTO MOLINA, Maria. Interdisciplinary approaches to the concept and practice of Written Documentary Content Analysis (WTDCA). *Journal of Documentation*, v.50, n.2, Jun.1994. p.111-1333.

64. PISTORI, Jeferson. *Bibliotecas digitais*. Florianópolis, UFSC, 1999. 15p.

65. POSTMAN, N. *Tecnopólio*: a rendição da cultura à tecnologia. São Paulo: Nobel, 1984.

66. RAGHAVAN, P.; BRODER, A.; HENZINGER, M. MAMBER, U.; PINKERTON, B. *Finding anything in the billion page Web*: are Algorithms the key? (Panel Abstract), WWW8, Toronto, Canada, 1999.

67. van RIJSBERGEN, C. J. *Information Retrieval*. Butterworths, 2. ed. 1979. 208 p.

68. ROBREDO, Jaime e CUNHA, Murilo. *Documentação de Hoje e de Amanhã* : uma abordagem informatizada da biblioteconomia e dos sistemas de informação. 2. ed. São Paulo : Global Ed., 1994. 400 p.

69. ROSSI, Daniela; PINHEIRO, Clarissa; FEIER, Nara e VIEIRA, Renata. *Resolução automática de Correferência em textos da língua portuguesa*. REIC Revista de Iniciação Científica da SBC, v. 1, n. 2, 2001. Disponível em: <http://www.sbc.org.br/reic/>.

70. RUWET, Nicolas *Introdução à Gramática Gerativa*. São Paulo: Perspectiva, Editora da Universidade de São Paulo, 1975. 357 p.

71. SALTON, Gerard e MCGILL, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p.

72. SANT'ANNA, V. *Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita*. 100 f. 2000. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.

73. SANTOS, Milton. *Por uma outra globalização:* do pensamento único à consciência universal. 3. ed. Rio de Janeiro: Record, 2000. 174p.

74. SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*. Belo Horizonte, v. 1, n. 1, p. 41-62, jan./jun. 1996.

75. SCHAFF, Adam. *A sociedade informática*: as conseqüências sociais da segunda revolução industrial. São Paulo: Unesp/Brasiliense, 1990.

76. SEMANTICWEB.ORG. Disponível em: <http://www.semanticweb.org/about.html>. Acesso em: jun. 2003.

77. SHANNON, C. E. *A Mathematical Theory of Communication*. The Bell System Technical Journal, Vol. 27, July, October, 1948. p. 379–423, 623–656.

78. SHERA, J. H., CLEVELAND, D. B. History and foundations of Information Science. *Annual Review of Information Science and Technology* – ARIST, v.12, p. 249-275, 1977.

79. SMEATON, A. F. Progress in the application of natural language processing to information retrieval tasks *Information Retrieval* Computer *Journa*l, v. 35, n. 3, 1994. p. 268-278.

80. SPARCK JONES, K. e WILLETT, P. (orgs.). *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997. 589p.

81. TAKAHASHI, Tadao (org). *Sociedade da informação no Brasil*: Livro Verde. Brasília: Ministério da Ciência e Tecnologia, 2000.

82. TOFFLER, A. *A Terceira Onda*. Rio de Janeiro: Record, 1980.

83. UNISIST. Princípios de indexação. *Revista da Escola de Biblioteconomia da UFMG*. Belo Horizonte, v.10, n.1, mar. 1981. p. 93-94,.

84. VIEIRA, Renata. A review of the Linguistic literature on definite descriptions. *Acta Semiotica et Lingüística*. Vol. 7, 1998. p. 219-258.

85. VIEIRA, Renata et ali. *Extração de Sintagmas Nominais para o Processamento de Co-referência*. 2000. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19-22 Novembro Atibaia SP.

86. VIEIRA, Renata e QUARESMA, Paulo. *PROJECTO DIRPI*: desenvolvimento e integração de recursos para pesquisa de informação. Cooperação Científica e Técnica Luso-Brasileira. ICCTI/GRICES-CAPES, Universidade de Évora, Universidade Nova de Lisboa, Unisinos, PUC-RS. Julho de 2001.

87. VISL. *About VISL*. Disponível em: <http://visl.hum.sdu.dk/visl/about/index.html>. Acesso em: mai.2005.

88. VYGOTSKY, L.S*. Pensamento e Linguagem*. São Paulo: Martins Fontes Ed. 1987.

89. WERSIG, Gernot. Information Science*:* the study of postmodern knowledge usage. *Information Processing & Management*, Oxford, U.K., v.29, Mar. 1993. p. 229-239.

90. WITTGENSTEIN, L. *Philosophical Investigations*. 3. ed. Oxford: Basil Blackwell, 1967.

91. WOOLDRIDGE, M. e JENNINGS, N., Intelligent Agents: theory and practice. *The Knowledge Engineering Review*, 10 (2), 1995. p.115-152.

92. _____. (ed.), *Agent Technology*: foundations, applications, and markets. Berlin, Heidelberg, New York: Springer-Verlag, 1998.

ANEXO A: O *corpus* de artigos utilizados para validação da metodologia

Neste anexo são apresentados os artigos que compõem o *corpus* total de documentos. As informações apresentadas são o título, os autores, o resumo, as palavras-chave, a edição da revista que os contêm e o endereço eletrônico aonde puderam ser encontrados na Internet, ao longo do ano de 2003.

I) Revista DataGramaZero (29 artigos)

Ia) *DataGramaZero - Revista de Ciência da Informação - v.3 n.2 abr/02*
Disponível no endereço: http://www.dgz.org.br/abr02/index.htm

Artigo 1: Transferência da Informação: análise para valoração de unidades de conhecimento
Autores: Plácida L. V. Amorim da Costa Santos e Ricardo César Gonçalves Sant'Ana
Resumo: Entender e medir o valor do conhecimento é uma das mais discutidas e menos compreendidas questões nos estudos sobre a gestão do conhecimento. Porém, se esta dificuldade fica mais explícita ao analisar o conjunto do conhecimento de uma organização em relação ao mercado, torna-se necessária a definição de parâmetros e mecanismos de avaliação de cada unidade de conhecimento disponível, principalmente aquele que, por algum processo, já foi registrado e precisa ser gerenciado, tanto em seu processo de obtenção, armazenamento, acesso e, até mesmo, descarte. Neste artigo, objetivamos descrever questões envolvidas na identificação do valor do conhecimento registrado em função de sua multidimensionalidade funcional e do processo de transmissão de informações.
Palavras chave: Transferência de informação; Gestão do conhecimento; Valor de unidades de conhecimento.

Artigo 2: Popularização do Conhecimento Científico
Autora: Suzana P. M. Mueller
Resumo: A questão da popularização da ciência é apresentada como um tema de interesse para estudos da comunicação científica pela ciência da informação. A participação da sociedade no processo de criação e divulgação da ciência por meio de pressões sociais e econômicas vêm se tornando fator importante na formulação de políticas científicas, especialmente em países com alto grau de educação. A opinião pública sobre fatos científicos, no entanto, depende em grande medida da atuação de intermediários, que traduzam a linguagem científica, especialmente a mídia imprensa e televisiva. As questões da distorção do sentido da notícia científica inerente ao processo, mas às vezes intencional, e o tratamento de notícias sobre fatos científicos que contém risco potencial à sociedade são exemplos usados para mostrar as dificuldades do processo de popularização e argumentar que o assunto, pouco estudado pela ciência da informação, é parte integrante e cada vez mais influente no processo de comunicação científica.
Palavras chave: Popularização da Ciência; Comunicação Científica.

Ib) *DataGramaZero - Revista de Ciência da Informação - v.3 n.3 jun/02*
Disponível no endereço: http://www.dgz.org.br/jun02/index.htm

Artigo 3: O Valor da Informação: um desafio permanente
Autora: Ana Lúcia Siaines de Castro
Resumo: O trabalho discute a questão da informação como uma alternativa de sobrevivência, de garantia jurídica frente a dispositivos de exceção, como ocorrido no período da ditadura militar no Brasil. A análise da informação como um valor estratégico de harmonização do indivíduo à sua capacidade de escolha, de avaliação e de reflexão. Na medida em que relatos e memórias alcançam

o espaço público, passíveis de institucionalização, a vinculação entre memória e informação revela-se confluente, interliga-se à confiabilidade dos mecanismos de produção, organização e recuperação da informação. A memória passa a representar um estoque informacional de valor social que possibilita a sociedade e os indivíduos disporem de garantias para exercitarem da melhor forma seu direito à informação.
Palavras chave: Informação; Valor Informacional; Direito à Informação; Memória Social; Estoque Informacional.

Artigo 4: Auto-arquivamento: uma opção inovadora para a produção científica
Autoras: Ligia Café e Márcia Basílio Lage
Resumo: Trata do conceito inovador de auto-arquivamento e suas implicações no sistema de publicações científicas. Esta nova filosofia procura minimizar as conseqüências provocadas pelo controle editorial, pela revisão severa entre os pares e pela reserva dos direitos autorais. A experiência da Budapest Open Access Initiative (BOAI) é relatada com o objetivo de mostrar uma ação efetiva que viabiliza o auto-arquivamento. Fundamentada no acesso livre (open access), a BOAI pretende uma reorganização dos mecanismos de produção do meio científico, baseada em conceitos mais democráticos de acesso ao conteúdo.
Palavras chave: Arquivos-abertos, Sistema de Publicação, Budapest Open Access Initiative, Acesso Livre, Auto-arquivamento.

Artigo 5: Análise Contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público
Autora: Teresinha Fróes Burnham
Resumo: Este artigo é uma reconstrução sumária e parcial da dinâmica de construção de uma metodologia de análise de processos de tradução do conhecimento científico - conhecimento privado a uma comunidade específica - já submetido a uma primeira tradução como conhecimento escolar, para acesso a um público de não-cientistas: estudantes de nível médio. Toma-se a escola como lócus de investigação, levando em conta que esta é a instituição socialmente responsabilizada pela democratização da informação científica, na perspectiva de que esta se transforme em conhecimento pessoal de indivíduos sociais, compreendidos como sujeitos do conhecimento, pela agregação de significados relevantes à formação da cidadania. O texto caracteriza-se como uma memória teórica e experiencialmente referenciada, de uma investigação realizada há mais de duas décadas, através da qual foi produzida a primeira formulação desta metodologia que, depois de várias reconstruções (que continuam a se processar contemporaneamente), vem sendo a base dos trabalhos realizados pela Rede Cooperativa de Pesquisa em (In)formação, Currículo e Trabalho - REDPECT / UFBA, dedicada a participar na construção do novo campo interdisciplinar e multirreferencial da Info-Educação.O texto limita-se a apenas um dos "componentes" do processo de tradução: a dupla dimensão de (des)construção e (re)construção de estruturas conceituais formais de um ou mais corpos teóricos da área de Biologia.
Palavras chave: Conhecimento Científico, Conhecimento Privado, Conhecimento Escolar, Democratização da Ciência, Comunicação Científica.

Ic) *DataGramaZero - Revista de Ciência da Informação - v.3 n.4 ago/02*
Disponível no endereço: http://www.dgz.org.br/ago02/index.htm

Artigo 6: O Tesauro Eletrônico do Mundo do Trabalho: produto de um esforço interdisciplinar
Autores: Marília Levacov, Nadia Vanti, Júlio César Zancan e Maria Lizete Gomes Mendes
Resumo: O presente artigo relata a implementação, de uma ferramenta para o gerenciamento do Tesauro Eletrônico do Mundo do Trabalho, criado para a Unitrabalho, uma fundação voltada a pesquisas acadêmicas sobre o trabalho, agregando 84 universidades brasileiras. A ferramenta é constituída de duas interfaces: uma para consulta e navegação e outra para gerenciamento. Esta atividade foi realizada por uma equipe interdisciplinar, do ponto de vista da Interação Humano-Computador, buscando alternativas para o diálogo entre dois universos: o dos profissionais da Ciência da Informação e o dos profissionais da Ciência da Computação.
Palavras-chave: Tesauro Eletrônico; Mundo do Trabalho; Recuperação da Informação; Interface de Consulta; Sistema de Informação; Interdisciplinaridade; Interação Humano-Computador (IHC).

Artigo 7: Inteligência Competitiva em Organizações: dado, informação e conhecimento
Autora: Marta Lígia Pomim Valentim
Resumo: O conjunto 'dados, informações e conhecimento' tem sido importante fator de competitividade em diferentes tipos de organizações. Prospectar, filtrar e transferir esse conjunto é essencial para a consolidação do processo de inteligência competitiva organizacional. Através do gerenciamento desses recursos informacionais pode-se subsidiar várias atividades para a melhoria contínua do negócio da organização. O papel do conjunto 'dados, informações e conhecimento' no processo de inteligência competitiva é fundamental para o aumento da produtividade e da qualidade da organização. Estabelecer fluxos formais e informais, bem como mapear e reconhecer os dados, informações e conhecimento estruturados, estruturáveis e não-estruturados para o negócio também são ações que contribuem para o desenvolvimento da inteligência competitiva organizacional.
Palavras chave: Inteligência Competitiva; Gestão do Conhecimento; Gestão da Informação; Fluxos Informacionais; Transferência da Informação.

Artigo 8: A conceituação de massa documental e o ciclo de interação entre tecnologia e o registro do conhecimento
Autores: Antonio Miranda e Elmira Simeão
Resumo: A polissemia do conceito de "informação" parece ser uma decorrência natural da apropriação do termo por diferentes áreas do conhecimento e está ligada ao fenômeno conhecido como "definição consuetudinária" em que diferentes especialistas se expressam conforme o estado da arte dos conhecimentos sobre determinado fenômeno. Tais definições estariam, conseqüentemente, sujeitas a reformulações e reconceitualizações pari passu com a evolução da pesquisa. A questão que se levanta constantemente é se a Ciência da Informação deveria ou não ter uma concepção única para o termo, o que parece não só impraticável, quanto inócuo.
Palavras chave: Informação; Massa Documental; Conceito de Informação; Tecnologia; Registro do Conhecimento.

Artigo 9: Informação e Universidade: os pecados informacionais e barreiras na comunicação da informação para a tomada de decisão na universidade
Autor: Claudio Starec
Resumo: O trabalho analisa o fluxo de informação nos Campi Rebouças e Nova América da Universidade Estácio de Sá. O objetivo deste trabalho é discutir as dificuldades, ruídos, os problemas e barreiras da comunicação da informação e seus efeitos no fluxo informacional numa organização voltada para o aprendizado. Duas questões levantadas pelo poeta americano T.S.Elliot retratam o viés deste trabalho:"Quanta informação perdemos devido à comunicação? e quanto conhecimento perdemos por causa da informação?" A base teórica é da Ciência da Informação, mais especificamente os conceitos de informação de Barreto, de Relevância de Saracevic, os Sistemas de Recuperação e Disseminação Seletiva de Informação de Araújo, as Barreiras de Freire. O foco da Inteligência Competitiva está em destacar a questão da informação como um dos maiores ativos de estratégias no setor e, possivelmente, aferramenta mais importante para ajudar os gestores da universidade a tomar decisões acadêmicas e administrativas à tempo e em tempo real. O modelo escolhido foi a Mandala Tibetana de Paul Carro adaptada na Mandala da Informação Universitária.
Palavras chave: Universidade; Gestão do fluxo de Informação na Universidade; Inteligência Competitiva; Barreiras na Comunicação da Informação; Pecados Informacionais.

Artigo 10: Implicações da "nova economia" para a mensuração estatística: desajustes conceituais e metodológicos
Autora: Rosa Maria Porcaro
Resumo: Este artigo discute como importantes transformações que marcam a sociedade atual se refletem na pertinência das informações estatísticas oficiais, construídas a partir de representações da realidade social. Questiona-se se tais transformações estão sendo apreendidas com o arcabouço conceitual-metodológico dos atuais levantamentos estatísticos construído e consolidado para "retratar" a sociedade capitalista industrial moderna de escopo nacional, hoje completamente modificada.
Palavras chave: Informação Estatística; Nova Economia; Mensuração Estatística; Desajuste Conceitual; Metodologia Estatística.

Disponível no endereço: http://www.dgz.org.br/out02/index.htm

Artigo 11: Por uma nova Ciência da Informação: ensino, pesquisa e formação
Autor: Luiz Carlos Brito Paternostro
Resumo: O armazenamento e a recuperação de informações incluem sua organização, classificação, proteção, difusão e transferência. O armazenamento e a recuperação dividem o mundo da experiência entre os movimentos de guardar e de tomar, inclusive sob um ponto de vista histórico e cultural. Sob a regência destes movimentos, podemos estudar qualquer coisa ligada à *informação*. Um conjunto de disciplinas interdependentes tratando especificamente do armazenamento e da recuperação de dados pode vir a compor um curso de Ciência da Informação capaz de tratar, de forma unificada, questões que variam da *propriedade intelectual* até os *fundamentos da modelagem de dados*.
Palavras-chave: Ciência da Informação, Armazenamento e recuperação, Curso em informação, Unidade e especificidade da informação.

Artigo 12: Ensino e pesquisa em ciência da informação
Autor: Eduardo Wense Dias
Resumo: Considerando-se o acesso à informação como a questão básica da ciência da informação, constata-se que é possível segmentar esse campo pelo tipo de informação a que se procura facilitar o acesso: informação publicada especializada, informação publicada não-especializada e informação não-publicada. As características peculiares desses segmentos vão deteminar a forma que os nomes dos profissionais neles atuantes podem tomar, as disciplinas importantes, a pesquisa, além de outros aspectos relacionados com a formação na área do conhecimento.
Palavras chave: Ciência da Informação, Biblioteconomia, Sistema de Informação, Arquivologia, Ensino, Pesquisa.

Artigo 13: O Profissional da Informação: O Humano Multifacetado
Autora: Kátia de Carvalho
Resumo: O profissional que na sua origem se forma no seio da biblioteca com a função de zelar pelo acervo acompanha o desenvolvimento da sociedade e se transforma em um ser humano multifacetado que além de desta primeira função citada passa a ser o responsável pela preservação da memória humana sem perder de vista o objetivo primordial que é a disseminação do conhecimento e da informação. O profissional nessa sociedade amplia as suas competências para dar conta do seu papel nos sistemas de informação. Ele, no contexto atual, deve ser um indivíduo que faz experiências e é sensível a aprendizagem sendo a sua presença insubstituível nas organizações, além de ser um mediador, entre usuário e acervos. Esse profissional representa o elemento humano nas relações com o meio em um mundo em transformação, com um modelo de economia global baseada no conhecimento.
Palavras chave: Profissional da informação, Informação organizacional, Formação e profissional da informação.

Artigo 14: Funções Sociais e Oportunidades para Profissionais da Informação
Autores: Kira Tarapanoff, Emir Suaiden, Cecília Leite Oliveira
Resumo: No contexto da sociedade em rede são discutidas funções sociais e delineados alguns perfis de atuação para profissionais da informação. Dentre as funções sociais delineadas estão as educativa e a de mediação. A educativa relaciona-se à alfabetização em informação e a segunda à animação da inteligência coletiva. Dentre os papéis profissionais emergentes são enumerados e brevemente discutidos os seguintes perfis: gestores da informação; trabalhadores do conhecimento; gestores e engenheiros do conhecimento; especialistas de informação. Conclui-se que não há um perfil único para o profissional da informação, que como um "soldado universal" atenderia a todas as demandas de informação nas organizações e na sociedade. Há papéis a serem preenchidos e demandas específicas a serem atendidas por profissionais com os mais diversos perfis, consagrados e emergentes, mas que têm como único objetivo o trabalho com a informação e o conhecimento, agregando valor à primeira e facilitando o acesso e transferindo informação e o conhecimento para todos.
Palavras chave: Profissionais da informação, Funções sociais, Perfis de profissionais da informação, Inclusão digital, Gestão da informação, Gestão do conhecimento.

Artigo 15: Relação Ensino-Pesquisa: em discussão a formação do Profissional da Informação
Autora: Mara Eliane Fonseca Rodrigues
Resumo: Tendo como referência as mudanças paradigmáticas que se avizinham para a educação, em geral, e para a universidade, em particular, discute a formação do profissional da informação no Brasil. Após, tomando por pressuposto que a formação, a prática profissional e a pesquisa, compõem a base de uma profissão e que estes três componentes devem interagir constantemente, enfoca a pesquisa como elemento capaz de permitir o repensar da formação e da prática do profissional da informação, considerando-a como um princípio também educativo.
Palavras chave: Formação profissional, Ensino e pesquisa.

Artigo 16: Educação para a Informação: desafios contemporâneos para a Ciência da Informação
Autora: Ana Maria Pereira Cardoso
Resumo: O artigo situa os desafios para a formação de profissionais de informação no contexto das mudanças no ensino superior no Brasil. Aborda a consolidação do campo da Ciência da Informação e as influências recebidas por via das literaturas americana e francesa. Discute as especificidades da Ciência da Informação em contraponto com a Biblioteconomia. Partindo destas referências apresenta o projeto de formação de analistas de informação conforme implementado na PUC Minas; destacando o perfil do profissional visado, os eixos temáticos do curso, as estratégias de ensino/aprendizagem.
Palavras chave: Ciência da Informação - Formação profissional, Educação Superior no Brasil, Sociedade da Informação - educação, Ciência da Informação e Biblioteconomia, Ciência da Informação - curso de graduação.

Ie) *DataGramaZero - Revista de Ciência da Informação - v.3 n.6 dez/02*

Disponível no endereço: http://www.dgz.org.br/dez02/index.htm

Artigo 17: Novas Tecnologias e Produção Científica: uma relação de causa e efeito ou uma relação de muitos efeitos?
Autora: Maria das Graças Targino
Resumo: Discute a relação entre novas tecnologias e o desenvolvimento da produção científica e da publicação eletrônica, enfatizando a Internet. Sem negar sua relevância como elemento interveniente da realidade contemporânea, prioriza as desvantagens trazidas pelas facilidades de produção no espaço cibernético, no caso particular da produção científica, tais como: a inconsistência, instantaneidade e efemeridade das informações; a complexidade de armazenamento; a dificuldade do controle bibliográfico; a banalização da autoria e o desrespeito à propriedade intelectual; o uso aético da informação; a invasão da privacidade x relações impessoais.
Palavras-chave: Internet e Produção Científica, Novas Tecnologias de Informação e de Comunicação, Produção Científica e Novas Tecnologias.

Artigo 18: Enfoques sobre a relação Ciência, Tecnologia e Sociedade: Neutralidade e Determinismo
Autor: Renato Dagnino
Resumo: De uma forma bastante genérica e mesmo ingênua, mas adequada à finalidade deste trabalho, é possível classificar as formas de abordar o campo dos Estudos Sociais da Ciência e Tecnologia ou, mais especificamente, a relação Ciência, Tecnologia e Sociedade, em duas grandes categorias. A primeira possui como foco privilegiado de análise, ou como elemento determinante da dinâmica da relação, o seu primeiro pólo, a C&T; enquanto que, a segunda, a Sociedade.
Palavras-chave: Estudos Sociais da Ciência, Sociologia da Ciência, Ciência e Sociedade, Tecnologia e Sociedade.

Artigo 19: Inteligência Empresarial: uma avaliação de fontes de informação sobre o ambiente organizacional externo
Autor: Ricardo Rodrigues Barbosa
Resumo: O artigo relata um estudo sobre o processo de monitoração do ambiente organizacional externo. Os 91 participantes da pesquisa registraram, dentre outros fatores, a freqüência com que utilizam diversos tipos de fontes de informação. Essas fontes foram também analisadas de acordo

com o seu grau de relevância e confiabilidade. Os resultados indicam uma elevada taxa de utilização de fontes eletrônicas de informação, porém as mesmas são vistas como pouco confiáveis e relevantes. As pessoas (colegas, subordinados e superiores hierárquicos) são vistas como as fontes mais confiáveis. As bibliotecas e centros de informação internos, embora considerados as fontes mais confiáveis, encontram-se entre as menos utilizadas e menos relevantes.
Palavras-chave: Inteligência Empresarial, Monitoração Ambiental, Fontes de Informação, Gestão do Conhecimento, Gestão da Informação

Artigo 20: Contribuição da Pós-graduação para a Ciência da Informação no Brasil: uma visão
Autores: Johanna W. Smit, Eduardo Wense Dias, Rosali Fernandez de Souza
Resumo: Síntese da avaliação continuada dos programas de pós-graduação em Ciência da Informação reconhecidos pela CAPES (PUC/CAMP, UFBA, UFMG, UFRJ/IBICT, UnB e UNESP/Marília), relativa ao ano de 2001. A partir da constituição dos corpos docente e discente, números de dissertações e teses defendidas e publicações do corpo docente, propõe-se um diagnóstico da pós-graduação na área, finalizando por uma discussão das características da pesquisa em Ciência da Informação realizada nos programas e a fragilidade da área em relação ao Sistema Nacional de Pós-Graduação. Em anexo uma tabela transcreve as áreas de concentração e linhas de pesquisa, com respectivas ementas, dos programas da área em 2001.
Palavras-chave: Ciência da Informação no Brasil, Avaliação 2001 CAPES, Pós-graduação em Ciência da Informação, Pesquisa em Ciência da Informação no Brasil.

Artigo 21: Os múltiplos aspectos e interfaces da leitura
Autora: Lígia Maria Moreira Dumont
Resumo: Este trabalho apresenta uma visão panorâmica referente às áreas do conhecimento que se entrelaçam e propiciam um melhor entendimento do ato de ler. Os estudos sobre leitura caracterizam-se pela multidisciplinaridade, portanto, estão sempre abertos à interferência de outras áreas do conhecimento, dependendo certamente de determinado recorte, dentre os múltiplos e diversos ângulos de análise possíveis na temática da leitura. Por suposto, está-se diante de um processo complexo; torna-se tarefa difícil estabelecer os limites de cada olhar, pois o ato da leitura não se efetiva em ações isoladas, lineares, mas sim em decorrência de complexa reação em cadeia de ações, sentimentos, motivações, especulações no cognóscio do leitor, suas análises e críticas. No artigo, são abordadas diversas teses e teorias sobre a temática da leitura, centradas na premissa da leitura como ação social. Primeiramente, são delineados os estudos de Mme. de Staël, Taine e Marx, pioneiros a destacarem o componente social na leitura. A seguir, são analisados os estudos sistemáticos desenvolvidos nos Estados Unidos e na França, nas décadas de 1930 e 1950, respectivamente, que se constituem nas teorias cunhadas de "sociologia da leitura": a Teoria dos fatores subjacentes de Holmes e os modelos de Carrigan e de Gray, de fundamentação organística e funcionalista. Por fim, são delineadas algumas abordagens de autores contemporâneos estrangeiros, como Escarpit, Barthes, Compagnon, Chartier, Allen e Spiro, bem como dos brasileiros Silva, Maria, Sodré e Kato. As teorias que se baseiam na área da psicolingüística e na teoria da computação (inteligência artificial) são destacadas por Kato e Spiro. As abordagens culminam com a tese de Paulo Freire, que imbrica definitivamente a vivência dos sujeitos ao aprendizado e ao desenvolvimento do ato da leitura.
Palavras-chave: Leitura-teoria, Cognóscio, Conhecimento-introjeção, Leitura e Sociedade, Informação e Sociedade.

Artigo 22: A Informação e o Paradigma Holográfico: a Utopia de Vannevar Bush
Autor: Nilton Bahlis dos Santos
Resumo: A Ciência da Informação tem dois elementos constituintes: por um lado ela nasce como acúmulo teórico e de experiências de processamento de informações, em particular da biblioteconomia e da documentação, com suas tecnologias capazes de processar volumes finitos de informação. Por outro como utopia, resultado da ampliação e alargamento do horizonte da ciência , nos esforços aliados na segunda guerra mundial e o desejo de Bush de um novo ordenamento para a Informação. O aspecto mais importante não é a "explosão informacional" como aumento quantitativo, mas a interconexão de experiências e pesquisas, que gera a necessidade de processamentos para a circulação de grandes massas de informação; utopia alimentada pela possibilidade vislumbrada de processar um volume infinito com o surgimento da tecnologia informática. Nossa reflexão é que se o primeiro aspecto está estruturado no paradigma do moderno, com sua visão determinista e racional,

resumindo-se a estudar o processo de informação em sistemas fechados, homogêneos e passíveis de serem organizados à priori, o segundo, a utopia, não consegue encontrar uma resposta no interior deste paradigma. Este segundo aspecto constituinte, isto é a busca da capacidade de processar informações em um número infinito e independente de linguagens controladas e de disciplinas, tem como marco o texto "Como nós pensamos" de Vannevar Bush. Ele aponta para a necessidade e possibilidade da Ciência da Informação enfrentar de uma maneira nova o problema da complexidade e interatividade, características cada vez mais presentes em nosso mundo, colocando em questão o próprio paradigma vigente. Esta utopia, no entanto, foi posta em segundo plano devido aos objetivos produtivistas colocados pelas opções práticas que a marcaram. Para recolocá-la na ordem do dia é necessário rever a própria definição de Ciência da Informação, seus limites como campo de conhecimento, seus métodos, suas técnicas e tecnologias. O Paradigma Holográfico apresenta determinados caminhos e opções para uma nova discussão e o hipertexto o evidencia em termos práticos.
Palavras-chave: Paradigma, Holografia, Ciência da Informação, Tecnologia da Informação, Hipertexto, Complexidade, Interatividade, Virtual, Totalidade.

If) *DataGramaZero - Revista de Ciência da Informação - v.4 n.1 fev/03*

Disponível no endereço: http://www.dgz.org.br/fev03/index.htm

Artigo 23: Informação, Memória e Espaço Prisional no Rio de Janeiro
Autora: Icléia Thiesen Magalhães Costa
Resumo: As relações entre informação, memória e espaço prisional são discutidas nessa proposta de estudo que tem por objetivo principal analisar as formações institucionais e jurídicas, direcionadas à constituição, implantação, reprodução e permanência do chamado *Panoptismo*, em especial na definição e configuração do espaço prisional, no Rio de Janeiro, no período de 1830 a 1930. A Ciência da Informação, de caráter interdisciplinar, propicia a ampliação das fronteiras da Ciência, aproximando saberes de diferentes naturezas e, por essa razão, contribuindo não apenas para a recuperação e disseminação da informação histórica contida nos escaninhos da memória, mas também para a formação de novas relações conceituais, tais como informação e história, espaço e poder, memória e documento, em suas diferentes combinatórias.
Palavras-chave: Informação, Memória Social, Espaço Prisional.

Artigo 24: O Contrato Social da Pesquisa: em busca de uma nova equação entre a autonomia epistêmica e autonomia política
Autora: Maria Nélida González de Gómez
Resumo: Consideramos próprio das modernas formações ocidentais o desenvolvimento dos conhecimentos científicos por procedimentos complementares a) de diferenciação e autonomização da atividade de pesquisa e b) de conversão da validade científica em valores econômicos ou sociais. Pergunta-se, nesse contexto, pela possibilidade de reformulação do contrato social da ciência, revisando as definições dos sujeitos e dos princípios que organizam os programas de pesquisa, em seu escopo e abrangência, tal que essa nova versão do contrato seja capaz de orientar uma ecologia política dos conhecimentos.
Palavras-chave: Contrato Social, Ciência, Pesquisa, Pesquisadores, Autonomia, Ecologia dos Conhecimentos.

Artigo 25: A Ciência da Informação no CNPq - fomento à formação de recursos humanos e à pesquisa entre 1994-2002
Autoras: Suzana Pinheiro Machado Mueller e Maria Gorette Santana
Resumo: Levantamento dos dados referentes às ações de fomento de CNPq para a área de Ciência da Informação, para o período de 1994 e 2002. Após breve introdução sobre as origens do CNPq em que é enfatizada sua vocação inicial como agência de fomento para as áreas de ciências exatas e naturais, o artigo mostra dados sobre quantidade e dispêndio do órgão com a área de Ciência da Informação. As ações do CNPq relatadas são as que se destinam à formação de recursos humanos no exterior e no país e a pesquisas no país. Os dados mostram que considerando todas as áreas financiadas pelo CNPq, uma parte muito reduzida do orçamento tem sido destinada à Ciência da Informação. Por outro lado, pode-se argumentar que dado o número de cursos pós-graduação, especialmente doutorado, existentes no período considerado e especialmente o número de pesquisa

em andamento cujos relatórios foram relatados na reunião de 2000 da ANCIB, a sociedade que congrega os pesquisadores da área, e ainda, a demanda bruta registrada no CNPq, os auxílios recebidos e vigentes, embora ainda insuficientes, parecem menos inadequados. No entanto, a estagnação no volume de bolsas concedidas entre 1994 e 2002 levanta preocupações a respeito da evolução da área.
Palavras-chave: Fomento à pesquisa - Ciência da Informação; CNPq - fomento á pesquisa em Ciência da Informação.

Ig) *DataGramaZero - Revista de Ciência da Informação - v.4 n.2 abr/03*

Disponível no endereço: http://www.dgz.org.br/abr03/index.htm

Artigo 26: Políticas de Monitoramento da Informação por Compressão Semântica dos seus Estoques
Autor: Aldo de Albuquerque Barreto
Resumo: Este artigo se orienta para o estudo da estrutura do texto escrito e sua análise morfológica com a finalidade de extrair informações para uso na gestão estratégica da informação, localizada em estoques específicos. Visa, ainda, fornecer subsídios para um processo de monitoração de conteúdos infomacionais em língua portuguesa e a realização de outros estudos de administração da informação. Procura indicar subsídios técnicos e teóricos para construção de softwares para o estudo de contextos de informação utilizando o instrumental da ciência da informação e do processamento computacional do português em linguagem natural. Ambiciona ser um instrumento estratégico para localizar e caracterizar através de palavras-chave conteúdos de famílias de textos visando a gestão e o controle de um estoque específico de informação.
Palavras-chave: Compressão Semântica, Monitoramento da Informação, Estoques de Informação, Palavras-chave.

Artigo 27: Bolsas de Pesquisador do CNPq: informações sobre política de C&T a partir da base que contém os dados cadastrais dos bolsistas
Autora: Gilda Olinto
Resumo: As bolsas de pesquisador concedidas pelo CNPq são aqui analisadas a partir das bases de dados da agência de fomento que contém o cadastro dos pesquisadores bolsistas. Destaca-se inicialmente a relevância deste objeto de estudo em função das características destas bolsas e, também, em função da riqueza de informações e possibilidades de análises que se apresentam através da transformação dessa base de dados gerada com fins administrativos para uma base com a finalidade de gerar indicadores científicos e tecnológicos. As análises aqui apresentadas focalizando apenas algumas das informações contidas nestas bases de dados – área acadêmica, estado e instituição do trabalho do bolsista – mostram que muitas informações podem ser geradas e revelar algumas características e desequilíbrios que podem ser úteis para subsidiar políticas de governo e o monitoramento da C&T no país.
Palavras-chave: Indicadores Científicos, Política Científica e Tecnológica, Gestão de Ciência e Tecnologia

Artigo 28: Arquitetura conceitual e resultados da integração de sistemas de informação e gestão da ciência e tecnologia
Autor: Roberto Pacheco e Vinícius Kern
Resumo: Iniciativas governamentais na área de gestão da informação esbarram freqüentemente na falta de integração e baixa qualidade da informação, incluindo iniciativas de governo eletrônico. Este artigo apresenta a concepção de sistemas de informação governamentais a partir da consideração dos interesses de todos os atores, configurando uma arquitetura conceitual para projetos de governo eletrônico. A Plataforma Lattes é apresentada como exemplo de implementação desta arquitetura. O papel das bibliotecas digitais de teses e dissertações é destacado, ressaltando seu papel em relação a outros provedores de informação do sistema nacional de ciência, tecnologia e inovação. A internacionalização da Plataforma Lattes é comentada à luz da oferta e demanda de informação que vem provocando.
Palavras-chave: Governo Eletrônico, Arquitetura de Sistemas de Informação, Integração de Informações, Gestão de C&T, Bibliotecas Digitais, Plataforma Lattes, Rede ScienTI.

Artigo 29: Políticas de Informação Governamental: a construção de Governo Eletrônico na Administração Federal do Brasil
Autores: Carlos Henrique Marcondes e José Maria Jardim
Resumo: Políticas de informação governamental têm sido implementadas em diversos países sob a noção de governo eletrônico, ainda pouco estruturada do ponto de vista teórico. No Brasil, a Administração Federal tem desenvolvido diversas ações desde 2000. Limitações de ordem sócio-econômica dificultam o acesso da maioria da população a sistemas de telefonia e a equipamentos de informática. Outro obstáculo ao Governo Eletrônico é a deficiência na gestão das informações governamentais. Após dois anos de implantação, o impacto do Governo Eletrônico revela-se maior na gestão interna da Administração Federal do que no atendimento ao cidadão.
Palavras-chave: Governo Eletrônico, Políticas de Informação, Informação Governamental.

II) Revista Ciência da Informação (31 artigos)

IIa) Ciência da Informação, v. 31, n. 1, jan./abr. 2002
Disponível no endereço: http://www.ibict.br/cienciadainformacao/viewissue.php?id=14

Artigo 30: Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso
Autor: Guilherme Ataíde Dias
Resumo: Este artigo apresenta uma abordagem sobre a avaliação do acesso a periódicos eletrônicos disponibilizados na World Wide Web por meio da análise do arquivo de log de acesso. O arquivo de log de acesso da revista **Informação & Sociedade: Estudos** é processado e apresentado como um exemplo de aplicação do uso de uma ferramenta automatizada de análise para arquivo de **log** de acesso. As características inerentes à análise do arquivo de **log** de acesso são apresentadas e discutidas.
Palavras-chave: Periódicos eletrônicos; Avaliação de acesso; Arquivo de log de acesso.

Artigo 31: Novos cenários políticos para a informação
Autora: Maria Nélida González de Gómez
Resumo: Poderíamos dizer que hoje, nos cenários mundiais, a economia do conhecimento é proposta, sem mais nem menos, como o novo conteúdo e referência da política da informação ou, em certa forma, da totalidade do político. Consideramos que contribui, para essa subversão de sentido, um terceiro termo, que para uns seria "infra-estrutura", e para outros, "sociedade da informação". Se o **modus operandi** dessa virada estratégica seria a transubstanciação do informacional e semiótico no econômico, através da mediação tecnológica e dos mercados, optamos por considerar as mudanças do papel do Estado – como **modus cognoscendi** dessas transformações, que afetam profundamente o que, até agora, denominara-se – em sentido restrito – "Política de informação". Nossa análise remiter-se-á à revisão do conceito "governança", adotando como apoio argumentativo o conceito de "regime de informação". A partir da consideração de alguns dos pressupostos da governança, indagaremos quais estruturas de informação poderiam sustentar os processos de formação, circulação e institucionalização do poder, em um horizonte democrático.
Palavras-chave: Política de informação; Sociedade da informação; Internet; Institucionalização da informação; Estado.

Artigo 32: Uso das linguagens controlada e natural em bases de dados: revisão da literatura
Autora: Ilza Leite Lopes
Resumo:. O trabalho tem como objetivo examinar o uso da linguagem controlada ou da linguagem natural, no planejamento da estratégia de busca em um ambiente de bases de dados em CD-ROM ou em linha. São revisados os estudos que abordam o uso das linguagens controlada e natural nas estratégias de busca, suas vantagens e desvantagens, proporcionando uma perspectiva sobre a complexidade para a busca da informação bibliográfica e referencial, incluindo a seleção de termos para as estratégias e a função do vocabulário controlado ou da linguagem natural nesse contexto.
Palavras-chave: Bases de dados; Estratégia de busca; Linguagem controlada; Linguagem natural. Recuperação da informação; Artigo de revisão.

Artigo 33: Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000)
Autora: Maria Lourdes Blatt Ohira e Noêmia Schoffen Prado
Resumo: A evolução da temática biblioteca virtual e biblioteca digital como assunto de artigos de periódicos brasileiros publicados de1995 a 2000 é o objetivo deste trabalho. Analisa 33 artigos apresentando os aspectos metodológicos adotados para o planejamento e criação de bibliotecas virtuais e digitais, o desenvolvimento de coleções diante dessa nova realidade, o impacto causado nas unidades de informação e nos profissionais da informação, as estatísticas das bibliotecas na Internet e programas institucionais, além da produção bibliográfica sobre bibliotecas virtuais e digitais. Avalia a produção no período, a produtividade dos autores e tipo de autoria, número de referências bibliográficas por artigo, tipo de documentos e o idioma dos documentos citados. Aponta, como resultados, que não há convergência sobre o conceito de biblioteca eletrônica, polimídia, digital e virtual e, para a preocupação dos autores, centrados nos aspectos metodológicos visando à implantação de bibliotecas digitais e virtuais.
Palavras-chave: Biblioteca digital; Biblioteca virtual; Produção científica: Produção bibliográfica; Periódicos.

Artigo 34: Experiência do Leaal/UFPE na produção e transferência de tecnologia
Autoras: Cecília Prysthon e Susana Schmidt
Resumo: O Laboratório de Experimentação e Análise de Alimentos (Leaal) do Departamento de Nutrição do Centro de Ciências da Saúde da Universidade Federal de Pernambuco cria/desenvolve e transfere tecnologia para o setor produtivo. Na sociedade industrial, transferência de tecnologia implica comunicação de informação tecnológica relevante para a produção de bens e serviços. Além das barreiras que interferem na comunicação final da informação, os mecanismos de visibilidade muitas vezes não são adequados ao acesso nos sistemas tradicionais e/ou automatizados de comunicação da informação tecnológica. Este trabalho trata de informação tecnológica, sua criação, desenvolvimento de bens e serviços até sua transferência e in*corpora*ção pela sociedade. Apresenta ações e mudanças necessárias ao laboratório para tornar visíveis e mais adequados os mecanismos de transferência tecnológica no âmbito da universidade e comunidade externa.
Palavras-chave: Informação tecnológica; Transferência de informação; Transferência tecnológica..

IIb) Ciência da Informação, v. 31, n. 2, maio/ago. 2002

Disponível no endereço: http://www.ibict.br/cienciadainformacao/viewissue.php?id=13

Artigo 35: Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares
Autor: Maurício Barcellos Almeida
Resumo: O HTML – Hypertext Markup Language – é uma linguagem demarcação, inicialmente concebida como uma solução para a publicação de documentos científicos em meios eletrônicos, queganhou popularidade e se tornou padrão para a Internet. Diversos tipos de aplicações, como navegadores, editores,programas de e-mail, bancos de dados etc., tornam possível atualmente o uso intensivo do HTML. Ao longo dos anos, recursos têm sido adicionados ao HTML para que ele possa atender às expectativas de usuários e sistemas computadorizados, aumentando sua complexidade. Estima-seque a versão 4.0 do HTML possua aproximadamente cem diferentes marcações fixas (conhecidas como tags), sem contar aquelas específicas para cada tipo de navegador da Internet. É comum se encontrarem páginas HTML que possuem mais marcações do que conteúdo. Uma possível solução para novas demandas nessa área é a utilização do Extended Markup Language (XML), uma linguagem de marcação que pode introduzir novas possibilidades e trazer melhor integração entre dados e usuários. Este artigo se propõe a abordar, de forma introdutória, o XML, sua utilização na Internet, alguns conceitos complementares necessários ao entendimento do assunto em apresentar vantagens no uso do XML, em relação ao HTML. Além disso, pretende apresentar o assunto como um campo fértil para discussões, proposições e estudo por profissionais da ciência da informação.
Palavras-chave: XML; HTML; Linguagens de marcação; Internet; Intranet.

Artigo 36: A Lei de Lotka na bibliometria brasileira
Autor: Rubén Urbizagástegui Alvarado

Resumo: Usando os dados reportados em artigos publicados em revistas brasileiras e trabalhos apresentados em congressos nacionais, replicaram-se as aplicações da Lei de Lotka à literatura brasileira em 10 campos diferentes. Utilizou-se o modelo do poder inverso pelos métodos do mínimo quadrado e probabilidade máxima. Das 10 literaturas nacionais analisadas, somente a literatura de medicina, siderurgia, jaca e biblioteconomia ajustaram-se ao modelo do poder inverso generalizado pelo método dos mínimos quadrados. No entanto, só duas literaturas (veterinária e cartas do Arquivo Privado de Getúlio Vargas) não se ajustaram ao modelo quando se usou o método da máxima probabilidade. Para ambas literaturas, tentaram-se modelos diferentes. A literatura de veterinária ajustou-se à distribuição binomial negativa, e as cartas do Arquivo Privado de Getúlio Vargas ajustaram-se melhor à distribuição Gauss-Poisson Inversa Generalizada.
Palavras-chave: Bibliometria; Lei de Lotka; Produtividade de autores; Brasil.

Artigo 37: Bases de dados de informação para negócios
Autora: Beatriz Valadares Cendón
Resumo: O conjunto de informações usadas por administradores para a tomada de decisão tem sido chamado de informação para negócios" e inclui informações mercadológicas, financeiras, estatísticas, jurídicas, sobre empresas e produtos e outras informações fatuais e analíticas sobre tendências nos cenários político-social, econômico e financeiro nos quais operam organizações empresariais. Este artigo categoriza e descreve algumas das principais bases de dados estrangeiras sobre informação para negócios, mostrando o universo de informações que elas disponibilizam em forma eletrônica. Para fins de discussão, as bases foram agrupadas em 10 categorias: (1) notícias em geral; (2) informações sobre empresas e setores industriais; (3) diretórios de empresas; (4) informações sobre produtos; (5) informações biográficas; (6) informações financeiras; (7) informações para investimento; (8) pesquisas de mercado; (9) informações jurídicas e (10) informações estatísticas. Agrupadas dessa forma para fins didáticos, na prática essas categorias se sobrepõem, e muitas bases de dados se enquadram em mais de uma categoria desta classificação. O artigo revê também algumas das principais empresas produtoras e distribuidoras de bases de dados sobre informação para negócios e as tendências da indústria de informação eletrônica.
Palavras-chave: *Informação para negócios; Bases de dados*

Artigo 38: Biblioteca híbrida: um novo enfoque no suporte à educação a distância
Autores: Eliane Maria Stuart Garcez e Gregório J. Varvakis Rados
Resumo: Discute-se o papel das bibliotecas híbridas no contexto atual, em face das transições pelas quais passam as bibliotecas convencionais, principalmente pelo surgimento da Internet e pela intensificação dos cursos no ensino a distância nas universidades, resultado do incremento da utilização da tecnologia da informação e comunicação no ensino. Enfoca-se a importância da flexibilização dos bens e serviços que devem ser oferecidos pelas bibliotecas híbridas para atender às necessidades de uma diversidade de tipos de usuários existentes na educação a distância.
Palavras-chave: Biblioteca híbrida; Tipos de usuários; Bens e serviços.

Artigo 39: Estratégia de busca na recuperação da informação: revisão da literatura
Autora: Ilza Leite Lopes
Resumo: Os sistemas de recuperação de informação, também denominados de bancos de dados, apresentam uma complexidade indiscutível no processo de armazenamento e busca da informação, envolvendo uma série de aspectos que são interdependentes. Dentre estes, podem ser destacados os seguintes fatores: a tecnologia eletrônica conduz os usuários ao acesso democrático à informação ampliando a busca de informação em bases de dados geograficamente distantes; o alcance da qualidade na informação recuperada requer o planejamento de estratégias de busca específicas para cada base de dados. Esse artigo revisa a literatura publicada sobre estratégias de busca abordando os tópicos relativos a seus conceitos, suas principais técnicas e etapas para sua operacionalização.
Palavras-chave: Estratégia de busca; Recuperação da informação; Técnicas de estratégia de busca; Bases de dados; Artigo de revisão.

Artigo 40: A ciência e a gestão da informação: compatibilidades no espaço profissional
Autora: Patricia Zeni Marchiori
Resumo: Apresenta o contexto contemporâneo que embasa as atividades de profissionais da informação, tendo em vista mercados de trabalhos com crescentes níveis de exigência e a necessidade de se solucionarem problemas de informação cada vez mais complexos e dinâmicos.

Define-se gestão da informação, assim como sua abrangência acadêmico operacional tendo como base os pressupostos teóricos da área de ciência da informação, em especial o núcleo de conteúdos relacionados à gestão integral dos recursos de informação de indivíduos, grupos e organizações. Ressaltase que a gestão da informação compartilha com demais profissões afins, os processos de criação, seleção e avaliação, gerenciamento, divulgação, utilização, preservação e políticas de direitos (privacidade, direitos autorais e outros) relacionados ao trinômio dado, informação e conhecimento. São descritas habilidades e conhecimentos necessários ao desempenho profissional do gestor, assim como as dificuldades inerentes à atuação no campo de atividades de informação.
Palavras-chave: Ciência da informação; Gestão da informação.

Artigo 41: Produção das literaturas "branca" e "cinzenta" pelos docentes/doutores dos programas de pós-graduação em ciência da informação no Brasil
Autoras: Dinah Aguiar Población e Daisy Pires Noronha
Resumo: Estudo cienciométrico da produção científica de docentes/doutores de programas de pós-graduação do Brasil. Objetivo: identificar o perfil dos docentes/doutores e as tendências das literaturas "branca" e "cinzenta" produzidas segundo as linhas de pesquisa dos programas. Método: dados coletados por meio de comunicação contínua e interativa com os docentes através da técnica da "Conferência de Delfos" para identificar o perfil dos docente/doutor e caracterizar a respectiva produção científica. Resultados: na análise dos 5 Programas em Ciência da Informação e na Área de Concentração do Programa de Comunicação da ECA/USP foram identificadas 22 linhas de pesquisa às quais estavam vinculados 66 docentes/doutores, sendo 54,5% titulados na área da ciência da informação. Dos 1.108 documentos produzidos no período de 1990 a 1999, 59,8% referem-se a publicações de literatura branca, na qual os artigos de periódicos detêm o maior índice, com 37,8% da produção total. Da literatura cinzenta produzida (40,2%), destacam-se as comunicações em eventos que detêm 29,8% do total da produção. Verificou-se o predomínio da autoria única (73,2%), confirmando as características dos trabalhos individuais da área de humanidades. A produção científica vinculada à linha de pesquisa influencia a formação de grupos de trabalhos e núcleos de pesquisa.
Palavras-chave: Produção científica; Literatura branca; Literatura cinzenta; Ciência da informação.

Artigo 42: Informação para negócios: os novos agentes do conhecimento e a gestão do capital intelectual
Autora: Yara Rezende
Resumo: A evolução das características e necessidades dos diferentes usuários de informação em empresas vem determinando, ao longo do tempo, não apenas a criação de diversos tipos de sistemas de informação para atendê-los, como também uma constante adaptação do perfil de formação acadêmica e de atuação dos profissionais da informação. Ao primeiro e tradicional modelo de biblioteca técnica de empresa seguiramse os centros de documentação, os centros de informação, as bibliotecas virtuais, os sistemas de inteligência competitiva e, atualmente, os programas de gestão do conhecimento. O reconhecimento da importância estratégica da administração do conhecimento e do capital intelectual das empresas configura-se como a mais recente fase de evolução na gestão da informação. Os diferentes modelos de sistemas de informação para empresas já surgidos, apesar de distintos, não se excluem e convivem, ainda que parcela significativa dos profissionais da informação não venha acompanhando e se adaptando a essa evolução e esteja perdendo espaço de atuação para profissionais de outras áreas.
Palavras-chave: Gestão do conhecimento; Capital intelectual; Informação para negócios; Sistemas de informação para negócios; Agentes do conhecimento.

Artigo 43: Análise metodológica dos estudos de necessidades de informação sobre setores industriais brasileiros: proposições
Autoras: Janete Fernandes Silva, Marta Araújo Tavares Ferreira e Mônica Erichsen Nassif Borges
Resumo: Este trabalho apresenta uma investigação sobre os diagnósticos de necessidade de informação tecnológica detectados em empresas brasileiras do setor industrial. Propõe procedimentos metodológicos que permitam orientar as futuras pesquisas sobre necessidades informacionais ditadas pelos processos de aprendizagem e da inovação tecnológica. Discutiu-se o nível de detalhamento, bem como o grau de abrangência e profundidade destes estudos. Os escolhidos para compor a pesquisa destacaram os setores com potencial de crescimento nos mercados internos e externos como possíveis participantes na geração do desenvolvimento

tecnológico e econômico. No entanto, existe uma insatisfação com estes estudos, especialmente pela sua incapacidade em reconhecer as reais necessidades de informação e tecnologia.
Palavras-chave: Necessidade de informação tecnológica; Informação tecnológica; Setor industrial; Inovação.

Artigo 44: Informação e competitividade: a contextualização da gestão do conhecimento nos processos organizacionais
Autor: Sergio Luis da Silva
Resumo: No âmbito das organizações empresariais, este artigo se propõe a discutir a visualização da gestão do conhecimento na organização em três níveis diferentes, mas fortemente inter-relacionados: o estratégico, o tático e o operacional. O primeiro nível trata da ligação entre competitividade da empresa e o trabalho com os conhecimentos para a criação de competências organizacionais. O segundo nível destaca a importância de se considerar a gestão de conhecimentos na organização como sendo parte relevante de seus processos de negócio e não somente de suas áreas departamentais. E finalmente, em um terceiro nível, está o lado operacional da gestão do conhecimento ligado à aprendizagem, aos formatos que o conhecimento assume e ao papel desempenhado pela tecnologia da informação. Este artigo, baseado principalmente em um estudo diversificado de várias referências bibliográficas, procura trazer algumas contribuições iniciais para esta discussão.
Palavras-chave: Gestão do conhecimento; Informação e competitividade; Processos organizacionais.

Artigo 45: Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento
Autora: Nadia Aurora Peres Vanti
Resumo: Este é um estudo comparativo de quatro subdisciplinas que permitem medir os fluxos da informação, a comunicação acadêmica e a difusão do conhecimento científico: a bibliometria, a cienciometria, a informetria e a webometria. Mediante a leitura de renomados autores que têm abordado estes temas, é realizada uma discussão teórico-conceitual e uma análise das semelhanças e diferenças que unem e separam os quatro métodos quantitativos no que diz respeito ao seu histórico, objeto de estudo, variáveis, técnicas, objetivos e campos de aplicação. Uma ênfase maior é dada à caracterização da webometria, por se tratar de uma área emergente dentro da ciência da informação, ainda pouco explorada no Brasil e com grandes potencialidades derivadas da expansão mundial da Internet.
Palavras-chave: Bibliometria; Cienciometria; Informetria; Webometria; Métodos quantitativos de avaliação

IIc) Ciência da Informação, v. 31, n. 3, set./dez. 2002

Disponível no endereço: http://www.ibict.br/cienciadainformacao/viewissue.php?id=12

Artigo 46: Métodos quantitativos de apoio à bibliometria: a pesquisa operacional pode ser uma alternativa?
Autor: Paulo César Rodrigues Borges
Resumo: O objetivo deste trabalho é apresentar uma forma alternativa para aplicar os métodos da Pesquisa Operacional aos fenômenos bibliométricos que surgiram no início do século XX, até hoje muito polêmicos. Dentre as várias formulações no campo da bibliometria, a chamada “lei de Bradford” foi o foco da investigação. Tentativas deste gênero podem ser uma saída para sistematizar conceitos na bibliometria, confirmando ou descartando descrições e princípios oriundos de suas formulações empíricas. Tendo por base uma linha de analogia entre fenômenos físicos da Teoria do Caos – resolvidos pela Pesquisa Operacional (PO) – e casos de oferta e procura de periódicos, é possível encontrar uma explicação para o comportamento anômalo da curva de Bradford em certas condições críticas. Para aduzir alguma evidência empírica para este ensaio, dois casos práticos na área da PO foram adaptados para a resolução de problemas bibliométricos típicos. Além disso, ao longo de todo o texto, foram assinalados alguns pontos que parecem comuns entre a bibliometria e a Teoria do Caos. Este ensaio, portanto, enseja uma nova questão: a PO poderá contribuir com a ciência da informação, suprindo-a com modelos determinísticos e bayesianos para explicar os fenômenos bibliométricos?
Palavras-chave: bibliometria; Lei de Bradford; Pesquisa operacional; Caos;
Ciência da informação; Inferência bayesiana.

Artigo 47: Periódicos eletrônicos: considerações relativas à aceitação deste recurso pelos usuários
Autor: Guilherme Ataíde Dias
Resumo: Este artigo apresenta algumas reflexões sobre a aceitação de periódicos eletrônicos disponibilizados na World Wide Web. Assuntos que freqüentemente são ignorados durante a elaboração dos mesmos são discutidos. Citam-se como exemplo alguns periódicos científicos eletrônicos brasileiros na área da ciência da informação. Analisam-se também algumas barreiras tecnológicas que impedem o uso mais amplo e irrestrito deste recurso.
Palavras-chave: Periódicos eletrônicos; Usabilidade; Novas tecnologias.

Artigo 48: Alguns aspectos do uso da informação na economia da informação
Autor: Max F. Cohen
Resumo: Se a sociedade encontra-se em uma economia da informação, como as empresas estão usando a informação para competir no mercado? Este artigo busca a estruturação do referencial teórico para a construção do modelo que permita medir o uso da informação por parte das organizações. Com base nos levantamentos realizados, entende-se que as empresas usam a informação em busca de seis estratégias genéricas: redução de custos, criação de valor, inovação, redução do risco, virtualização e diferenciação de produto. Destacam-se, na economia da informação, as firmas que conseguem criar a interação entre os atores econômicos, tirar proveito da interconectividade e sincronizar as suas operações.
Palavras-chave: Uso da informação; Economia da Informação; Modelo genérico.

Artigo 49: Ferramentas alternativas para monitoramento e mapeamento automatizado do conhecimento
Autores: Lúcia Cunha Ortiz, Wilson Aires Ortiz e Sergio Luis da Silva
Resumo: A análise da informação é uma excelente estratégia para monitoramento, pesquisa e desenvolvimento em todos os ramos
do conhecimento. O objetivo primordial deste trabalho foi consolidar um método alternativo empregando ferramentas eletrônicas na realização do monitoramento automatizado da informação e em sua análise bibliométrica. O trabalho foi desenvolvido tendo como suporte a base Web of Science, do Institute for Scientific Information (ISI), e o uso de softwar como Word, Excel, Reference Manager e Origin. A título de exemplo, aplicamos o método à área de desenvolvimento de produtos, obtendo como resultados uma lista de descritores, a relação dos periódicos mais importantes da área, os autores mais produtivos e uma indicação das parcerias mais freqüentes entre eles.
Palavras-chave: Monitoramento da informação; Biblioteconomia; Ciência da informação.

Artigo 50: A formação profissional no século XXI: desafios e dilemas
Autoras: Edna Lúcia da Silva e Miriam Vieira da Cunha
Resumo: Reflexão sobre a educação no século XXI com enfoque especial à educação dos bibliotecários. Destaca os quatros pilares básicos e essenciais, preconizados pela Unesco, a um novo conceito de educação: aprender a conhecer, aprender a viver juntos, aprender a fazer e aprender a ser. Apresenta as ponderações elaboradas por Morin , a pedido da Unesco, que poderão melhorar a educação do futuro. Com base em tais fundamentos, discute o papel e a formação do bibliotecário no século XXI. Declara que os dilemas dos educadores, nesses novos tempos, estão centrados em três questionamentos: O que ensinar? Como ensinar? Para que ensinar? Pondera que a formação do bibliotecário deverá enfatizar sua função educativa e que a base deve ser polivalente alicerçada em um conjunto de valores que possibilite alterar percepções, maneiras de pensar e instaure a cooperação e a sabedoria em detrimento do tecnicismo hoje privilegiado. Conclui que o papel mais importante do bibliotecário no século XXI parece ainda ser o de
gerenciador da informação.
Palavras-chave: Educação dos bibliotecários; Profissional da informação.

Artigo 51: A acessibilidade à informação no espaço digital
Autores: Elisabeth Fátima Torres, Alberto Angel Mazzoni e João Bosco da Mota Alves
Resumo: O trabalho aborda aspectos referentes à acessibilidade no espaço digital. Uma ênfase especial é dada às situações relacionadas à interação das pessoas portadoras de deficiência com a informação, em ambientes de bibliotecas. O texto propõe algumas adequações para a acessibilidade

ao espaço digital, conforme categorias de usuários, com o intuito de contribuir para um maior nível de acessibilidade à informação, nesse espaço.
Palavras-chave: Acessibilidade; Espaço digital; Bibliotecas; Pessoas portadoras de deficiência; Ajudas técnicas.

Artigo 52: Estudos de usuários: o padrão que une três abordagens
Autores: Isa Maria Freire, Bruno Macedo Nathanhson, Carla Tavares e Carmelita do Espírito Santo
Resumo: Trata-se de três projetos de pesquisa em andamento no Programa de Pósgraduação em Ciência da Informação – PPGCI/IBICT/UFRJ. O primeiro visa a um estudo de usuários com base em uma experiência de interatividade na rede Internet, tendo como objeto de estudo o informativo www.clippirata.com.br. O segundo aposta no papel da informação para a educação ambiental. Para tanto, objetiva demonstrar como oficinas de reciclagem artesanal de papel podem funcionar como agregados de informação para a produção do conhecimento. O último projeto tem como objetivo a construção de um instrumento digital sobre informação cultural com base na estrutura do hipertexto. A responsabilidade social da ciência da informação é a base conceitual que une as três abordagens. O fator comum aos três projetos é a participação dos usuários de informação no desenvolvimento de cada um deles, um pressuposto básico da metodologia participante adotada nas pesquisas.
Palavras-chave: Estudos de usuários; Educação ambiental; Internet; Hipertexto; Pesquisa participante.

IId) Ciência da Informação, v. 32, n. 1, jan./abr. 2003

Disponível no endereço: http://www.ibict.br/cienciadainformacao/viewissue.php?id=11

Artigo 53: Como incrementar a qualidade dos resultados das máquinas de busca: da análise de logs à interação em português
Autoras: Rachel Virgínia Xavier Aires e Sandra Maria Aluísio
Resumo: Com o intuito de avaliar a submissão de consultas em língua natural, especificamente em português, a máquinas de busca na Web, e contrastar com as consultas por palavras-chave, realizou-se um experimento com alunos, professores e funcionários de uma universidade brasileira. Particularmente, analisaram-se as consultas para verificar se os usuários expressavam bem seus objetivos em palavras-chave; como expressariam seus objetivos em língua natural, caso esta possibilidade fosse oferecida; se as consultas em língua natural forneciam informações que pudessem facilitar a recuperação de informação. O pedido de colaboração foi enviado a 440 pessoas de um instituto de computação da universidade. Foram obtidas 63 consultas, correspondentes a 42 objetivos. Observou-se que, para o item **a**, na maioria dos casos (71,43%), as consultas por meio de palavras-chave não trazem todas as informações declaradas importantes no objetivo; para o item **b** as consultas foram feitas por meio de perguntas (71,87%), afirmações (18,75%) e ordens (9,37%); e, para o item **c** todas as perguntas diretas deixavam claro o objetivo da consulta já com a primeira palavra da frase, ou com as duas ou três primeiras, com exceção das iniciadas pela palavra "qual".
Palavras-chave: Análise de **logs**; Máquinas de busca; Recuperação de informação; Comportamento de usuários; Estratégias de busca.

Artigo 54: Information literacy: princípios, filosofia e prática
Autora: Elisabeth Adriana Dudziak
Resumo: Surgida na literatura em 1974, a information literacy liga-se à necessidade de se exercer o domínio sobre o sempre crescente universo informacional. In*corpora*ndo habilidades, conhecimentos e valores relacionados à busca, acesso, avaliação, organização e difusão da informação e do conhecimento. A information literacy é a própria essência da competência em informação. O objetivo deste trabalho é definir a information literacy a partir do entendimento do conceito, objetivos e práticas relacionadas, com ênfase no papel educacional das bibliotecas e do bibliotecário. Inicialmente, apresenta-se a evolução do conceito segundo um referencial histórico. Examina-se a information literacy enquanto processo de interiorização de conhecimentos, habilidades e valores ligados à informação e ao aprendizado. Define-se a expressão, suas características e objetivos. Discutem-se diferentes concepções de information literacy, segundo três referenciais: informação, conhecimento e aprendizado. Em seguida, são elencados pontos relevantes de atuação de bibliotecas e bibliotecários na implementação de uma educação voltada para a information literacy. Explorando a information

literacy education, evidencia-se a necessidade de construção de um novo paradigma educacional ante a sociedade atual que incorpore a competência em informação.
Palavras-chave: Information literacy; Competência em informação; Alfabetização informacional; Biblioteca aprendente; Bibliotecário educador; Sociedade de aprendizagem; Habilidades informacionais.

Artigo 55: Profissional da informação: perfil de habilidades demandadas pelo mercado de trabalho
Autora: Danielle Thiago Ferreira
Resumo: Doze empresas de consultoria em recrutamento e seleção de recursos humanos foram estudadas para obter informações acerca da demanda atual do mercado de trabalho. Foram levantadas e analisadas as literaturas sobre o mercado de trabalho, as qualificações profissionais requeridas pelo mercado e as informações obtidas em depoimentos de empregadores. O estudo trouxe quatro conclusões principais: (1) os profissionais devem desenvolver continuamente suas habilidades técnicas típicas de ciência da informação, bem como suas atitudes comportamentais; (2) as potencialidades desses profissionais nem sempre são reconhecidas pelo mercado de trabalho; (3) como conseqüência, não é comum encontrar profissionais da informação ocupando posições superiores como analistas ou gerentes; (4) as causas principais das deficiências são tanto a falta de desenvolvimento dessas habilidades durante o período de formação, quanto a falta de reconhecimento do perfil dos profissionais da informação pelo mercado e da auto-imagem por eles mesmos.
Palavras-chave: Profissional da informação; Profissional da informação – habilidades; Perfil e atuação profissional; Mercado de trabalho.

Artigo 56: O olhar da consciência possível sobre o campo científico
Autora: Isa Maria Freire
Resumo: O artigo descreve o exercício de tecer, no tear da ciência da informação, uma rede para apreender e explicar um evento de comunicação da informação no campo científico. Como objeto de estudo, foi selecionado o artigo em que G. Wersig e U. Neveling propõem, em 1975, um fundamento social para a ciência da informação. A pesquisa encontrou os indícios de que os autores compartilhavam com outros cientistas uma visão socialista da ciência da informação, fundada na importância da organização da informação científica e tecnológica e de sua comunicação no campo científico. Contudo, os autores foram além da consciência real do seu grupo, ao antever a relevância da informação para todos os grupos sociais na sociedade contemporânea. Nesse contexto, a proposição de uma "responsabilidade social" é retomada como fundamento à práxis dos cientistas da informação e como 'padrão que une' ciência e ética, no campo da ciência da informação.
Palavras-chave: Teoria da ciência da informação; Sociologia da informação; História da ciência da informação; Comunicação científica; Responsabilidade social.

Artigo 57: As relações entre ciência, Estado e sociedade: um domínio de visibilidade para as questões da informação
Autora: Maria Nélida González de Gómez
Resumo: Se a origem da ciência da informação está marcada pelas alianças de pós-guerra entre ciência, Estado, sociedade, a pesquisa em questões da informação recebe hoje as demandas de articulação dos três principais eixos de integração e avaliação dos conhecimentos, no Brasil e na América latina: o eixo paradigmático, o eixo *corpora*tivo e o eixo territorial.
Palavras-chave: Recuperação da informação; Inteligência científica; Integração dos conhecimentos; Estado; Ciência; Sociedade; Informação.

Artigo 58: Interfaces entre a ciência da informação e a ciência cognitiva
Autora: Gercina Ângela Borém Lima
Resumo: Estudo panorâmico sobre aspectos da ciência da informação (CI) e da ciência cognitiva (CC), apontando recentes contribuições em quatro de suas possíveis interseções: categorização, indexação, recuperação da informação (RI) e interação homem-computador.
Palavras-chave: Ciência da informação; Ciência cognitiva; Processamento da informação; Categorização; Indexação; Recuperação da informação; Interação homem-computador.

Artigo 59: A produção científica da Anped e da Intercom no GT da Educação e Comunicação

Autores: Solange Puntel Mostafa e Luis Fernando Máximo
Resumo: Analisa as literaturas publicadas no período 1994-2001 nos grupos de trabalho da Sociedade Interdisciplinar para os Estudos da Comunicação (Intercom) e da Associação Nacional de Pesquisa em Educação (Anped) no tema da comunicação educativa, em que foram analisadas respectivamente 1.023 e 1.049 citações bibliográficas presentes nos trabalhos apresentados. O objetivo da pesquisa foi perguntar quais autores nacionais e internacionais constituem a frente de pesquisa (autores mais influentes) nas duas literaturas e, se possível, visualizar tendências epistemológicas na produção científica. Os resultados apontam o humanismo e as teorias críticas da recepção na Intercom, enquanto na Anped o pós-estruturalismo parece ser a tendência dominante.
Palavras-chave: Comunicação científica; Bibliometria; Comunicação e Educação; Estudo de citações; Cientometria.

Artigo 60: Inteligência competitiva na Internet: um processo otimizado por agentes inteligentes
Autora: Helena Pereira da Silva
Resumo: Apresenta a proposta de um processo de inteligência competitiva (IC) na Internet, utilizando agentes inteligentes na tarefa de monitoramento de fontes de informação disponíveis na rede. O processo foi aplicado como projeto-piloto no Núcleo de Estudos em Inovação, Gestão e Tecnologia de Informação (IGTI) da Universidade Federal de Santa Catarina. Em seguida, foi verificada a aplicabilidade em mais três estudos de caso. Pode-se afirmar que, pelos casos estudados, foi possível vislumbrar a possibilidade efetiva de utilização do processo proposto em diferentes tipos de organizações. Os resultados ainda confirmam, como proposta, a necessidade de formalização do uso da informação e do processo de gestão da informação nas organizações, bem como a automação do processo por meio de agentes inteligentes.
Palavras-chave: Inteligência competitiva; Internet; Monitoramento de fontes de informação; Agentes inteligentes.

ANEXO B: Resultados das análises do *corpus* inicial

Neste anexo são apresentadas tabelas com informações advindas da aplicação prospectiva da metodologia, com vistas a posterior melhoria. São apresentados, para os seis primeiros documentos do *corpus* exposto no Anexo A, as palavras-chave escolhidas pelos autores, os SNs mais freqüentes, com a freqüência em que ocorrem (agrupados dentre os vários que diferem no determinante) e a indicação de se estão presentes no tesauro de CI. Ainda são apresentadas, para fins de comparação, as palavras-chave mais freqüentes e suas respectivas quantidades. O esquema de cores adotado é o seguinte: Em azul estão grifados os SNs e as palavras-chave que foram considerados extremamente relevantes como descritores; em vermelho aqueles que foram considerados razoavelmente relevantes como descritores; em laranja os que foram considerados moderadamente relevantes como descritores e, finalmente, em preto, os que não foram considerados relevantes como descritores.

Artigo 1: **Transferência da Informação: análise para valoração de unidades de conhecimento**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Transferência de informação	o conhecimento (1a)	50	Não	conhecimento(s)	202
Gestão do conhecimento	[uma, a, as] organização(ões) (1a)	29	Sim	informação(ções)	96
Valor de unidades de conhecimento	o repositório (1a)	24	Não	valor(es)	75
	o emissor (1a)	22	Não	processo(s)	71
	o receptor (1a)	21	Não	elemento(s)	44
	[a, as] informação(ões) (1a)	13	Sim	repositório	40
	[o, os] usuários (1a)	11	Sim	conjunto	39
	[esse, o] processo (1a)	10	Não	unidade(s)	37
	[o, um] conjunto de informações (2)	10	Parcialmente	emissor	36
	[a] gestão do conhecimento (2)	08	Não	receptor	35
	[o, outros] sistema(s) (1a)	08	Parcialmente	organização(ções)	35
	a ferramenta (1a)	07	Não	registr*	27
	[o] conhecimento explícito (1b)	06	Não	contexto	25
	[o] conhecimento tácito (1b)	06	Não	transmissão	25
	o processo de transmissão (2)	05	Não	perdas	23
	o tempo (1a)	05	Não	esquema(s)	21
	o valor do conhecimento (2)	05	Não	usuário(s)	21
	as perdas (1a)	04	Não	análise(s)	20
	as pessoas (1a)	04	Não	gestão	20
	o contexto (1a)	04	Não	dado(s)	18
	o processo de interação (2)	04	Não	sistema(s)	17
	os elementos (1a)	04	Não	figura(s)	16
	[esta, a] análise (1a)	03	Parcialmente	tácito	16
	a concorrência (1a)	03	Não	tempo	16
	a produção (1a)	03	Não	forma	15
	a tona (1a)	03	Não	comunicação	14
	o contexto do receptor (2)	03	Parcialmente	explícito	14
	o saber (1a)	03	Não	função	13
	o valor (1a)	03	Não	knowledge	13
	os esquemas (1a)	03	Não	utilização	13
	os recursos tecnológicos (1b)	03	Não		
	a definição das dimensões (2)	02	Não		
	a dimensão do contexto da	02	Não		

	organização (3)				
	a mente humana (1b)	02	Não		
	a participação dos usuários (2)	02	Parcialmente		
	a transformação de dados (2)	02	Parcialmente		
	as redes informais (1b)	02	Não		
	cada unidade de conhecimento (2)	02	Parcialmente		
	conhecimento procedural (1b)	02	Não		
	o conhecimento registrado (1b)	02	Parcialmente		
	o conjunto de dados (2)	02	Parcialmente		
	o contexto da organização (2)	02	Não		
	o contexto interpretativo (1b)	02	Não		
	o processo de contextualização (2)	02	Não		
	o valor das informações (2)	02	Não		
	o valor das perdas (2)	02	Não		
	o valor de uma unidade de conhecimento registrada (2)	02	Parcialmente		
	os principais pontos de perda (2)	02	Não		
	outras unidades de conhecimento (2)	02	Parcialmente		
	repositórios de conhecimento (2)	02	Não		
	sua interação com o repositório (2)	02	Não		
	transmissão entre repositórios (1b)	02	Não		
	um agente humano (1b)	02	Parcialmente		
	um sistema informático (1b)	02	Sim*		
	uma análise mais profunda (1b)	02	Parcialmente		
	uma unidade de conhecimento registrada (1b)	02	Parcialmente		

Artigo 2: **Popularização do Conhecimento Científico**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Popularização da Ciência	**a ciência** (1a)	15	Parcialmente	científico(os,a,as)	48
Comunicação Científica	a mídia (1a)	11	Não	ciência(s)	28
	[os, alguns] cientistas (1a)	13	Parcialmente	cientistas	28
	a sociedade (1a)	08	Não	popularização	20
	o conhecimento científico (1b)	08	Parcialmente	conhecimento	18
	a informação (1a)	07	Sim	pesquisa(s)	16
	as indústrias (1a)	07	Não	processo	15
	a ciência da informação (2)	06	Sim	interesses	13
	a notícia científica (1b)	05	Não	sociedade	13
	a popularização (1a)	05	Não	comunicação	12
	a popularização da ciência (2)	05	Não	distorção	12
	a imprensa (1a)	04	Não	indústrias	12
	o processo (1a)	04	Não	mídia	11
	a dieta (1a)	03	Não	câncer	10
	o cigarro (1a)	03	Não	notícia	10
	o processo de popularização (2)	03	Não	risco	10
	o risco (1a)	03	Não	informação	09
	os adoçantes (1a)	03	Não	estimativas	08
	a camada de ozônio (2)	02	Não	fatos	08
	a comunicação científica (1b)	02	Parcialmente	Hilgartner	08
	a melhor estimativa (1b)	02	Não	dados	07
	a sociedade leiga (1b)	02	Não	intervalo	07
	estimativas aceitáveis (1b)	02	Não	notícias	07
	linguagem especializada (1b)	02	Parcialmente	resultados	07
	o meio ambiente (1b)	02	Não	vezes	07
	o processo de Lievrouw (2)	02	Não	dieta	06
	o processo de popularização do conhecimento científico (3)	02	Não	financiamento	06
	a popularização do conhecimento científico (2)	02	Não	Nelkin	06
				tabela	06
				textos	06

Artigo 3: **Valor da Informação: um desafio permanente**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Informação	**a informação** (1a)	31	Sim	Informação(ções)	67
Valor Informacional	o indivíduo (1a)	09	Não	memória(s)	34
Direito à Informação	a(s) memória(s) (1a)	06	Não	social(ais)	29
Memória Social	a sociedade (1a)	08	Não	direito	14
Estoque Informacional	**a memória coletiva** (1b)	05	Não	sociedade	14
	o passado (1a)	05	Não	Informacional(is)	14
	os relatos (1a)	05	Não	conhecimento	12
	a vida (1a)	04	Não	condição	11
	o direito à informação (2)	04	Sim	liberdade	11
	a ditadura (1a)	03	Não	processo(s)	11
	a liberdade (1a)	03	Não	depoimentos	09
	a realidade (1a)	03	Não	ditadura	09
	a sobrevivência (1a)	03	Não	indivíduo	09
	os processos (1a)	03	Não	valor	09
	a justiça militar (1b)	02	Não	acesso	08
	a liberdade de informação (2)	02	Não	militar	08
	a recuperação da informação (2)	02	Não	passado	08
	dependência de censura (2)	02	Sim	relato(s)	08
	o espaço social (1b)	02	Não	São Paulo*	08
	o valor da informação (2)	02	Não	ciência	07
	os agentes envolvidos (1b)	02	Parcialmente	comunicação	07
				função	07
				Rio de Janeiro*	07
				processos	07
				sentido	07
				tempo	07
				vida	07
				espaço	06
				forma	06
				poder	06

Artigo 4: **Auto-arquivamento: uma opção inovadora para a produção científica**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Arquivos-abertos	o autor (1a)	12	Sim	científico [os,a(s)]	35
Sistema de Publicação	os pares (1a)	09	Não	acesso	28
Budapest Open Access Initiative	**o auto-arquivamento** (1b)	08	Parcialmente	auto-arquivamento	16
Acesso Livre	**a OAI** (1a)	06	Não	autor	15
Auto-arquivamento	**a BOAI** (1a)	05	Não	*open*	15
	a Internet (1a)	05	Não	publicação	15
	os pesquisadores (1a)	05	Parcialmente	pesquisa(s)	15
	a propriedade intelectual (1b)	04	Não	arquivos	14
	os direitos autorais (1b)	04	Sim	disponível	12
	the refereed (1a)	04	Não	*access*	11
	revisão entre os pares (2)	03	Não	*eprint*	11
	o acesso livre (1b)	03	Parcialmente	*refereed*	11
	o artigo (1a)	03	Parcialmente	direitos	10
	o conteúdo (1a)	03		informação	10
	[os] arquivos abertos (1b)	03	Parcialmente	publicações	10
	a Budapest Open Access Initiative (3)	02	Não	trabalhos	10
	a coleta automática de dados (2)	02	Parcialmente	abertos	09
	a iniciativa dos arquivos abertos (2)	02	Parcialmente	artigos	09
	a legal matter (1b)	02	Não	BOAI*	09
	a literatura científica (1b)	02	Parcialmente	*initiative*	09
	as barreiras impostas (1b)	02	Não	Internet	09
	as novas tecnologias de informação e comunicação (2)	02	Não	research	09
	as publicações eletrônicas (1b)	02	Parcialmente	sistema	09
	o mercado editorial (1b)	02	Parcialmente	autorais	08
	os resultados de pesquisas (2)	02	Parcialmente	divulgação	08
	problemas relacionados (1b)	02	Não	OAI*	08
	publicação científica (1b)	02	Sim	pares	08
				pesquisadores	08
				artigo	07
				revisão	07

Artigo 5: **Análise Contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Conhecimento Científico	o processo de tradução (2)	12	Parcialmente	conhecimento	49
Conhecimento Privado	[a, essa] análise (1a)	12	Parcialmente	análise	45
Conhecimento Escolar	**o conhecimento** (1a)	10	Não	tradução	35
Democratização da Ciência	o SM (1a)	10	Não	processo(s)	47
Comunicação Científica	a(s) informação(ões) (1a)	10	Sim	Informação(ões)	31
	[o, um] conceito (1a)	09	Parcialmente	estrutura	17
	a pesquisa (1a)	08	Não	construção	16
	o conhecimento científico (1b)	07	Sim	conceito	15
	a literatura (1a)	06	Sim	pesquisa	15
	a tradução (1a)	06	Sim	termos	14
	os TE (1a)	06	Não	verbais	14
	a compreensão (1a)	04	Não	partir	13
	a escola (1a)	04	Parcialmente	campo	12
	conhecimento escolar (1b)	04	Não	forma	12
	o quadro de giz (2)	04	Não	*New York**	12
	o trabalho de campo (2)	04	Não	SM	12
	os registros (1a)	04	Parcialmente	TAs	12
	SA (1a)	04	Não	categorias	11
	a área de biologia (2)	03	Não	conceitos	11
	a sociedade da informação (2)	03	Não	diferentes	11
	cada evento registrado (1b)	03	Não	elementos	11
	dispersão de sementes ou animais (2)	03	Não	social	11
	exploração inicial (1b)	02	Não	científico	10
	a comunidade de biólogos (2)	02	Não	conceitual	10
	a construção do significado do conceito (3)	02	Parcialmente	documentos	10
	a democratização do conhecimento científico (2)	02	Não	registros	10
	a prática de ensino (2)	02	Não	tópico	10
	a situação estudada (1b)	02	Não	literatura	09
	a sociology of language (2)	02	Não	transcrito	09
	analysis of concept	02	Não	unidades	09

	learning (2)				
	as categorias analíticas (1b)	02	Parcialmente		
	concept learning (1b)	02	Não		
	o conhecimento comum (1b)	02	Não		
	o conhecimento público (1b)	02	Não		
	dispersão de sementes (2)	02	Não		
	esquemas associativos (1b)	02	Não		
	linguagens verbais e não verbais (1b)	02	Não		
	o significado de o conceito (2)	02	Parcialmente		
	símbolos não-verbais (1b)	02	Não		
	the logic of teaching (2)	02	Não		
	tradução de o conhecimento científico (2)	02	Parcialmente		
	um sistema de análise (2)	02	Parcialmente		
	unidades conceituais (1b)	02	Parcialmente		
	unpublished phd (1b)	02	Não		

Artigo 6: **O Tesauro Eletrônico do Mundo do Trabalho: produto de um esforço interdisciplinar**					
Palavras-chave atribuídas pelo(s) autor(es)	**SNs mais freqüentes**	**Qtd.**	**SNs presentes no Tesauro da CI?**	**30 Palavras-chave mais freqüentes**	**Qtd.**
Tesauro Eletrônico	**[o, um] tesauro** (1a)	17	Sim	termos	48
Mundo do Trabalho	**a interface de consulta** (2)	07	Parcialmente	tesauro	38
Recuperação da Informação	o banco (1a)	07	Não	Informação(ões)	36
Interface de Consulta	o usuário (1a)	07	Sim	Interface(s)	34
Sistema de Informação	o(s) termo(s) (1a)	07	Sim	área(s)	29
Interdisciplinaridade	o trabalho (1a)	06	Não	dados	27
Interação Humano-Computador (IHC)	o gerenciador (1a)	06	Não	ciência(s)	20
	a área (1a)	04	Não	banco	17
	a equipe (1a)	04	Não	gerenciador*	15
	a informação (1a)	04	Sim	usuário(s)	15
	a tela (1a)	04	Não	consulta	14
	o gerenciador de banco de dados (3)	04	Parcialmente	diferentes	13
	o Gerenciador do Tesauro (2)	04	Parcialmente	recuperação	13
	os dados (1a)	04	Parcialmente	sistema	12
	os documentos (1a)	04	Sim	trabalho	11
	os procedimentos (1a)	04	Não	universidade	11
	a aplicação (1a)	03	Não	nível	10
	a lista (1a)	03	Não	procedimentos	10
	a navegação (1a)	03	Não	estudos	09
	o conhecimento (1a)	03	Não	gerenciamento*	09
	o mundo do trabalho (2)	03	Não	tela	09
	o Rio Grande do Sul (1a)	03	Não	base	08
	o sistema (1a)	03	Parcialmente	federal	08
	um primeiro momento (1b)	03	Não	ferramenta	08
	a área de IHC (2)	02	Não	tecnologia	08
	a área do trabalho (2)	02	Não	Unitrabalho	08
	a Ciência da Computação (2)	02	Parcialmente	visualização	08
	a Ciência da Informação (2)	02	Sim	documentos	07
	a sua descendência de termos específicos (2)	02	Parcialmente	específicos	07
	as páginas HTML (1b)	02	Não		
	as palavras-chave (1b)	02	Sim		
	o Centro de Informação (2)	02	Sim		
	o gerenciamento do tesauro (2)	02	Parcialmente		
	o tesauro eletrônico (1b)	02	Parcialmente		
	alteração do termo (2)	02	Parcialmente		

ANEXO C: Resultados das análises do *corpus* total

Neste anexo são apresentadas tabelas com informações advindas das duas aplicações finais da metodologia. São apresentados, para os 60 documentos do *corpus* exposto no Anexo A.

Artigo 1: Transferência da Informação: análise para valoração de unidades de conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
conjunto de informações	C	conjunto de informações	C
gestão do conhecimento	A	gestão do conhecimento	A
processo de transmissão	C	processo de transmissão	C
valor do conhecimento	A	valor do conhecimento	A
processo de interação	C	processo de interação	C
unidade de conhecimento	B	unidade de conhecimento	B
conhecimento explícito	A	contexto do receptor	C
conhecimento tácito	A	conhecimento explícito	A
contexto do receptor	C	conhecimento tácito	A
interação com o repositório	B	contexto da organização	B
participação dos usuários	C	interação com o repositório	C
repositórios de conhecimento	B	participação dos usuários	C
valor de uma unidade de conhecimento registrada	A	repositórios de conhecimento	A
Taxa de Relevância	**0,60**	**Taxa de Relevância**	**0,58**

Artigo 2: Popularização do Conhecimento Científico			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
popularização da ciência	A	popularização da ciência	A
processo de popularização	A	processo de popularização	A
notícia científica	C	camada de ozônio	D
mídia	B	popularização do conhecimento científico	A
ciência da informação	C	processo de Lievrouw	B
camada de ozônio	D	notícia científica	C
popularização do conhecimento científico	A	ciência da informação	C
processo de Lievrouw	B	processo de popularização do conhecimento científico	A
Taxa de Relevância	**0,56**	**Taxa de Relevância**	**0,63**

Artigo 3: O Valor da Informação: um desafio permanente			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
direito à informação	A	direito à informação	A
memória coletiva	A	dependência de censura	B
dependência de censura	B	liberdade de informação	A
liberdade de informação	A	valor da informação	A
recuperação da informação	C	memória coletiva	A
valor da informação	A	recuperação da informação	C
agentes envolvidos	C	agentes envolvidos	D
espaço social	B	espaço social	C
Taxa de Relevância	**0,69**	**Taxa de Relevância**	**0,63**

Artigo 4: Auto-arquivamento: uma opção inovadora para a produção científica			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
auto-arquivamento	A	auto-arquivamento	A
revisão entre os pares	B	revisão entre os pares	B
Budapest Open Access Initiative	A	Budapest Open Access Initiative	A
coleta automática de dados	D	coleta automática de dados	D
iniciativa dos arquivos abertos	A	iniciativa dos arquivos abertos	A
novas tecnologias de informação e comunicação	C	novas tecnologias de informação e comunicação	C
resultados de pesquisas	D	resultados de pesquisas	D
direitos autorais	B	direitos autorais	B
Taxa de Relevância	**0,53**	**Taxa de Relevância**	**0,53**

Artigo 5: Análise Contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
processo de tradução	C	processo de tradução	C
quadro de giz	D	quadro de giz	D
trabalho de campo	D	trabalho de campo	D
conhecimento científico	A	área de biologia	C
área de biologia	C	dispersão de sementes ou animais	D
dispersão de sementes ou animais	D	comunidade de biólogos	C
SM	D	democratização do conhecimento científico	A
comunidade de biólogos	C	dispersão de sementes	D
democratização do conhecimento científico	A	prática de ensino	C
prática de ensino	C	significado de o conceito	D
significado do conceito	D	sistema de análise	D
sistema de análise	C	tradução de o conhecimento científico	A
tradução de o conhecimento científico	A	conhecimento científico	A
Taxa de Relevância	**0,33**	**Taxa de Relevância**	**0,31**

Artigo 6: O Tesauro Eletrônico do Mundo do Trabalho: produto de um esforço interdisciplinar			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
interface de consulta	A	interface de consulta	A
Gerenciador do Tesauro	A	Gerenciador do Tesauro	A
mundo do trabalho	A	mundo do trabalho	A
tesauro	A	alteração do termo	C
alteração do termo	C	área de IHC	B
área de IHC	B	área do trabalho	A
área do trabalho	A	Centro de Informação	C
gerenciamento do tesauro	A	gerenciamento do tesauro	A
Taxa de Relevância	**0,84**	**Taxa de Relevância**	**0,75**

Artigo 7: Inteligência Competitiva em Organizações: dado, informação e conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
inteligência competitiva	A	gestão da informação	A
gestão do conhecimento	A	gestão do conhecimento	A
gestão da informação	A	tomada de decisão	A
tomada de decisão	A	conhecimentos produzidos	D
tecnologias da informação	C	proximidade do seu significado	D
conhecimentos produzidos	D	inteligência competitiva	A
proximidade do seu significado	D	tecnologias da informação	C
fluxos informais de informação	A	fluxos informais de informação	A
Taxa de Relevância	**0,66**	**Taxa de Relevância**	**0,66**

Artigo 8: A conceituação de massa documental e o ciclo de interação entre tecnologia e o registro do conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
massa documental	A	massa documental	A
Ciência da Informação	A	indústria da informação	A
indústria da informação	A	necessidades de comunicação	C
necessidades de comunicação	C	tipo de documento	C
tipo de documento	C	Ciência da Informação	A
documento	C	artigo científico	C
artigo científico	C	conteúdos específicos	C
conteúdos específicos	C	novo suporte	D
Taxa de Relevância	**0,53**	**Taxa de Relevância**	**0,50**

Artigo 9: Informação e Universidade: os pecados informacionais e barreiras na comunicação da informação para a tomada de decisão na universidade			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
tomada de decisão	B	tomada de decisão	B
fluxo de informação	A	fluxo de informação	A
mercado de trabalho	B	mercado de trabalho	B
comunicação de a informação	A	comunicação de a informação	A
cultura de a organização	B	cultura de a organização	B
Mandala da Informação Universitária	A	Mandala da Informação Universitária	A
tecnologia de a informação	C	tecnologia de a informação	C
universidade	C	informação relevante	D
informação relevante	B	inteligência competitiva	A
Taxa de Relevância	**0,61**	**Taxa de Relevância**	**0,64**

Artigo 10: Implicações da "nova economia" para a mensuração estatística: desajustes conceituais e metodológicos			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
estatísticas oficiais	A	estatísticas oficiais	A
sociedade da informação	B	sociedade da informação	B
cadeia de valor	C	cadeia de valor	C
interpretação da sociedade	A	interpretação da sociedade	A
representação das atividades econômicas	A	representação das atividades econômicas	A
tecnologias de informação e comunicação	B	tecnologias de informação e comunicação	B
processo de produção	B	processo de produção	B
valor agregado	C	valor agregado	C
âmbito da representação das atividades econômicas	A	âmbito da representação das atividades econômicas	A
informação estatística	C	informação estatística	C
Taxa de Relevância	**0,65**	**Taxa de Relevância**	**0,65**

Artigo 11: Por uma nova Ciência da Informação: ensino, pesquisa e formação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
depósito	C	conceito de depósito	B
conceito de depósito	B	ocultamento de informação	A
ocultamento de informação	A	protocolos de comunicação	A
protocolos de comunicação	A	informação pervasiva	A
informação pervasiva	A	caixas-pretas	C
caixas-pretas	C	cérebro humano	D
cérebro humano	D	curso de Ciência da Informação	A
curso de Ciência da Informação	A	Ciência da Informação	A
Taxa de Relevância	**0,63**	**Taxa de Relevância**	**0,72**

Artigo 12: Ensino e pesquisa em ciência da informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
ciência da informação	A	recuperação da informação	A
recuperação da informação	A	acesso a informação	A
acesso a informação	A	campo de conhecimento	C
campo de conhecimento	C	desenvolvimento de coleções	B
desenvolvimento de coleções	B	especialidade da biblioteconomia	A
especialidade da biblioteconomia	A	necessidades de informação	A
necessidades de informação	A	ciência da informação	A
serviços de informação	A	serviços de informação	A
Taxa de Relevância	**0,84**	**Taxa de Relevância**	**0,84**

Artigo 13: O Profissional da Informação: O Humano Multifacetado			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
tecnologias da informação	C	tecnologias da informação	C
profissional da informação	A	profissional da informação	A
campo da informação	B	campo da informação	B
década de 70	D	década de 70	D
disseminação da informação	A	disseminação da informação	A
talento de seus profissionais	C	talento de seus profissionais	C
produção do conhecimento	B	produção do conhecimento	B
biblioteca	C	sociedade da informação	B
sociedade da informação	A	capital intelectual	A
capital intelectual	A	dos mais antigos sistemas de informação	C
dos mais antigos sistemas de informação	C	relações interpessoais	C
Taxa de Relevância	**0,55**	**Taxa de Relevância**	**0,50**

Artigo 14: Funções Sociais e Oportunidades para Profissionais da Informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
gestão do conhecimento	A	alfabetização em informação	A
sociedade da informação	A	gestão do conhecimento	A
profissional da informação	A	uso de ferramentas inteligentes	B
alfabetização em informação	A	profissional da informação	A
uso de ferramentas inteligentes	B	sociedade da informação	A
maio de 2000	D	maio de 2000	D
utilização da informação bibliográfica	C	utilização da informação bibliográfica	C
inteligência coletiva	A	tecnologias da informação	B
Taxa de Relevância	**0,72**	**Taxa de Relevância**	**0,66**

Artigo 15: Relação Ensino-Pesquisa: em discussão a formação do Profissional da Informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
profissional da informação	A	profissional da informação	A
educação superior	B	cursos de graduação	B
cursos de graduação	B	cursos de pós-graduação	B
cursos de pós-graduação	B	ensino de graduação	B
ensino de graduação	B	informação no Brasil	B
informação no Brasil	B	educação superior	B
cursos de graduação	B	cursos de graduação	B
produção do conhecimento	B	produção do conhecimento	B
alunos	C	articulação entre ensino e pesquisa	A
Taxa de Relevância	**0,53**	**Taxa de Relevância**	**0,61**

Artigo 16: Educação para a Informação: desafios contemporâneos para a Ciência da Informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
ciência da informação	A	sociedade da informação	A
sociedade da informação	A	formação das novas gerações	A
formação das novas gerações	A	ciência da informação	A
PUC Minas	A	PUC Minas	A
de bibliotecários	B	de bibliotecários	B
demandas locais	D	demandas locais	D
número de egressos do ensino médio	B	número de egressos do ensino médio	B
terceiro grau	B	terceiro grau	B
Taxa de Relevância	**0,69**	**Taxa de Relevância**	**0,69**

Artigo 17: Novas Tecnologias e Produção Científica: uma relação de causa e efeito ou uma relação de muitos efeitos?			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
complexidade de armazenamento	B	complexidade de armazenamento	B
dificuldade do controle bibliográfico	A	dificuldade do controle bibliográfico	A
efemeridade das informações	C	efemeridade das informações	C
evolução de Internet	C	evolução de Internet	C
novas tecnologias versus produção científica	A	novas tecnologias versus produção científica	A
produção científica	A	processo de comunicação	A
processo de comunicação	A	produção científica	A
Rede	D	controle bibliográfico	C
Internet	C	novas tecnologias	C
controle bibliográfico	C	propriedade intelectual	A
publicações eletrônicas	A	publicações eletrônicas	A
Taxa de Relevância	**0,59**	**Taxa de Relevância**	**0,68**

Artigo 18: Enfoques sobre a relação Ciência, Tecnologia e Sociedade: Neutralidade e Determinismo			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
desenvolvimento da C&T	A	desenvolvimento da C&T	A
meios de produção	A	divisão do trabalho	A
relações de produção	A	meios de produção	A
relações sociais de produção	A	política da C&T	A
divisão do trabalho	A	processo de trabalho	A
processo de trabalho	A	relações de produção	A
política da C&T	A	relações sociais de produção	A
determinismo tecnológico	A	desenvolvimento das forças produtivas	A
forças produtivas	A	idéia da neutralidade	C
não-neutralidade	A	luta de classes	B
produção capitalista	A	tese forte da não-neutralidade	B
tese forte	C	tese fraca da não-neutralidade	B
desenvolvimento das forças produtivas	A	divisão capitalista do trabalho	A
tese forte da não-neutralidade	B	modo de produção capitalista	A
tese fraca da não-neutralidade	B	relações de produção capitalistas	A
Taxa de Relevância	**0,88**	**Taxa de Relevância**	**0,85**

Artigo 19: Inteligência Empresarial: uma avaliação de fontes de informação sobre o ambiente organizacional externo			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
ambiente organizacional externo	A	fontes de informação	A
ambiente externo	A	ambiente externo	A
ambiente organizacional	A	ambiente organizacional	A
fontes de informação	A	ambiente organizacional externo	A
inteligência empresarial	A	inteligência empresarial	A
monitoração ambiental	A	sistemas de informação	C
sistemas de informação	C	monitoração ambiental	A
meio eletrônico	B	ambiente externo das organizações	A
presente estudo	D	informações sobre o ambiente organizacional externo	A
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,92**

Artigo 20: Contribuição da Pós-graduação para a Ciência da Informação no Brasil: uma visão			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
áreas de concentração	C	áreas de concentração	C
programas de pós-graduação	A	programas de pós-graduação	A
Ciência da Informação	A	área em 2001	A
área em 2001	A	construção do conhecimento	C
construção do conhecimento	C	linhas de pesquisa	B
linhas de pesquisa	B	programas de pós-graduação	A
programas de pós-graduação	A	serviços de informação	B
serviços de informação	B	circulação dos mesmos	D
corpo discente	C	docente do NRD6	A
Sistema Nacional de Pós-Graduação	A	Sistema Nacional de Pós-Graduação	A
Taxa de Relevância	**0,68**	**Taxa de Relevância**	**0,65**

Artigo 21: Os múltiplos aspectos e interfaces da leitura			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
ato de ler	A	ato de ler	A
ato da leitura	A	ato da leitura	A
temática da leitura	A	temática da leitura	A
leitor	B	leitura da palavra	A
leitura	A	leitura do mundo	A
leitura da palavra	A	sociologia da leitura	A
sociologia da leitura	A	várias áreas do conhecimento	C
várias áreas do conhecimento	C	determinado conceito	D
Taxa de Relevância	**0,84**	**Taxa de Relevância**	**0,78**

Artigo 22: A Informação e o Paradigma Holográfico: a Utopia de Vannevar Bush			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
nível do atual	D	nível do atual	D
Ciência da Informação	A	condições de tempo real	D
paradigma determinista	A	explosão de informações	B
paradigma moderno	A	grandes volumes de informações	B
hipertexto	A	informação no terreno virtual	B
condições de tempo real	D	mecanismos de busca	B
explosão de informações	B	método de análise	D
grandes volumes de informações	B	pontes entre as várias disciplinas especializadas	C
informação no terreno virtual	B	possibilidade de relações	D
mecanismos de busca	B	tecnologias de inteligência	B
pontes entre as várias disciplinas especializadas	C	trajetória do sistema	D
possibilidade de relações	D	transição de paradigmas	B
tecnologias de inteligência	B	Ciência da Informação	A
transição de paradigmas	B	processo de comunicação	C
Taxa de Relevância	**0,52**	**Taxa de Relevância**	**0,32**

Artigo 23: Informação, Memória e Espaço Prisional no Rio de Janeiro			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Rio de Janeiro	C	espaço prisional	A
espaço prisional	A	cidade do Rio de Janeiro	B
cidade do Rio de Janeiro	B	final do século	B
final do século	B	Rio de Janeiro	C
cidade	D	imagens da clausura	A
prisão	A	período de 1890 a 1930	C
imagens da clausura	A	suporte de informação	D
suporte de informação	C	tal modelo	D
Taxa de Relevância	**0,56**	**Taxa de Relevância**	**0,44**

Artigo 24: O Contrato Social da Pesquisa: em busca de uma nova equação entre a autonomia epistêmica e autonomia política			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
área do conhecimento	C	área do conhecimento	C
comunidades de pesquisa	A	comunidades de pesquisa	A
grupos de pesquisa	A	grupos de pesquisa	A
sistema de inovação	B	sistema de inovação	B
forma de vida	D	forma de vida	D
produção dos conhecimentos	A	produção dos conhecimentos	A
princípio paradigmático	D	construção de indicadores	B
pesquisadores	B	contrato social da pesquisa	A
contrato social da pesquisa	A	desenvolvimento da atividade científica	A
desenvolvimento da atividade científica	A	movimentos dos conhecimentos	B
movimentos dos conhecimentos	B	organização do conhecimento	A
organização do conhecimento	A	produção de conhecimentos científicos	A
produção de conhecimentos científicos	A	programas de pesquisa	B
sistema de ciência e tecnologia	A	sistema de ciência e tecnologia	A
Taxa de Relevância	**0,70**	**Taxa de Relevância**	**0,73**

Artigo 25: A Ciência da Informação no CNPq - fomento à formação de recursos humanos e à pesquisa entre 1994-2002			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
todas as áreas	D	cursos de pós-graduação	B
Ciência da Informação	A	quantidade de bolsas	C
área da Ciência da Informação	A	todas as áreas	D
cursos de pós-graduação	B	ano de 2002	C
quantidade de bolsas	C	atuação do CNPq	A
iniciação científica	B	formação de recursos humanos	B
ano de 2002	C	início de período	D
atuação do CNPq	A	área da Ciência da Informação	A
formação de recursos humanos	B	bolsas de produtividade	C
início de período	D	iniciação científica	B
Taxa de Relevância	**0,50**	**Taxa de Relevância**	**0,43**

Artigo 26: Políticas de Monitoramento da Informação por Compressão Semântica dos seus Estoques			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Estoques de informação	A	Estoques de informação	A
número de letras	D	número de letras	D
texto de informação	C	texto de informação	C
zonas de qualidade intensa	C	zonas de qualidade intensa	C
coeficiente de relevância	C	coeficiente de relevância	C
controle da informação	B	controle da informação	B
estoque de informação	A	estoque de informação	A
fluxos de informação	A	interesse de uma comunidade informacional	A
interesse de uma comunidade informacional	A	linguagem do pensamento	C
linguagem do pensamento	C	palavras de freqüência igual a um	C
palavras de freqüência igual a um	C	sentido de ordenação lógica	A
Taxa de Relevância	**0,52**	**Taxa de Relevância**	**0,52**

Artigo 27: Bolsas de Pesquisador do CNPq: informações sobre política de C&T a partir da base que contém os dados cadastrais dos bolsistas			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
bolsas de pesquisador	A	bolsas de pesquisador	A
área de Saúde	B	área de Saúde	B
o ano de 1998	C	o ano de 1998	C
apoio a a pesquisa	A	apoio a a pesquisa	A
área de conhecimento	A	área de conhecimento	A
mais de 100 bolsistas	C	mais de 100 bolsistas	C
site do CNPq	B	site do CNPq	B
agências de fomento	A	agências de fomento	A
bolsas de produtividade	A	bolsas de produtividade	A
alto nível	D	Resenha Estatística do CNPq	A
Taxa de Relevância	**0,65**	**Taxa de Relevância**	**0,75**

Artigo 28: Arquitetura conceitual e resultados da integração de sistemas de informação e gestão da ciência e tecnologia			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
sistema nacional de CT	A	sistema nacional de CT	A
sistemas de informação	B	sistemas de informação	B
Plataforma Lattes	A	Plataforma Lattes	A
informação em CT	A	informação em CT	A
sistemas de conhecimento	B	sistemas de conhecimento	B
unidades de informação	B	unidades de informação	B
bibliotecas digitais de teses e dissertações	A	bibliotecas digitais de teses e dissertações	A
sistemas de informação governamentais	A	sistemas de informação governamentais	A
Taxa de Relevância	**0,81**	**Taxa de Relevância**	**0,81**

Artigo 29: Políticas de Informação Governamental: a construção de Governo Eletrônico na Administração Federal do Brasil			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Governo Eletrônico	A	Governo Federal	B
Governo Federal	B	Governo Eletrônico	A
acesso à Internet	A	acesso à Internet	A
cidadão às informações	A	cidadão a as informações	A
implantação do Governo Eletrônico	A	implantação do Governo Eletrônico	A
instrumento de governança e governabilidade	A	instrumento de governança e governabilidade	A
outubro de 2000	D	outubro de 2000	D
prestação de serviços	B	prestação de serviços	B
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,75**

Artigo 30: Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
log de acesso	**A**	acesso a periódicos	**B**
acesso a periódicos	**B**	log de acesso	**A**
sessão de usuário	**B**	sessão de usuário	**B**
arquivos de log de acesso	**A**	arquivos de log de acesso	**A**
servidor web	**B**	servidor web	**B**
artigo de periódico	**A**	artigo de periódico	**A**
cache local do próprio browser	**A**	cache local do próprio browser	**A**
número de hits	**B**	número de hits	**B**
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,75**

Artigo 31: Novos cenários políticos para a informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
política de informação	A	política de informação	A
regime de informação	A	regime de informação	A
relação entre política e informação	A	relação entre política e informação	A
governança informacional	A	infra-estrutura de informação	A
infra-estrutura de informação	A	uso da Internet	C
uso da Internet	C	governança informacional	A
políticas públicas	A	atos de governo	B
década de 90	C	década de 90	C
Estado	A	intervenção do Estado	B
atos de governo	B	Política e Informação	A
intervenção do Estado	B	rede de redes	C
Política e Informação	A	serviços de Internet	C
rede de redes	C	políticas públicas	A
serviços de Internet	C	Novos cenários políticos para a informação	A
Novos cenários políticos para a informação	A	constituição comunicacional	C
Taxa de Relevância	**0,73**	**Taxa de Relevância**	**0,68**

Artigo 32: Uso das linguagens controlada e natural em bases de dados: revisão da literatura			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
estratégia de busca	A	estratégia de busca	A
bases de dados	C	termos da LN	B
recuperação da informação	A	bases de dados	C
linguagem controlada	A	recuperação da informação	A
vocabulário controlado	A	termos da LC	A
linguagem natural	A	uso da LN	A
termos da LN	B	processo de indexação	A
termos da LC	A	linguagem controlada	A
uso da LN	A	vocabulário controlado	A
processo de indexação	A	o controle do vocabulário	A
o controle do vocabulário	A	Uso das linguagens controlada	A
Uso das linguagens controlada	A	linguagem natural	A
Taxa de Relevância	**0,90**	**Taxa de Relevância**	**0,90**

Artigo 33: Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000)			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
biblioteca do futuro	A	biblioteca do futuro	A
biblioteca digital	A	desenvolvimento de coleções	B
biblioteca eletrônica	A	Grupo de Trabalho	D
biblioteca virtual	A	profissionais da informação	B
Bibliotecas virtuais e digitais	A	tipos de documentos	D
periódicos brasileiros	B	biblioteca digital	A
realidade virtual	A	biblioteca eletrônica	A
profissionais da informação	B	biblioteca virtual	A
desenvolvimento de coleções	B	Bibliotecas virtuais e digitais	A
Grupo de Trabalho	D	periódicos brasileiros	B
tipos de documentos	D	realidade virtual	A
análise de artigos de periódicos brasileiros	B	artigos de periódicos	C
artigos de periódicos	C	conceito de biblioteca virtual	A
Taxa de Relevância	**0,63**	**Taxa de Relevância**	**0,67**

Artigo 34: Experiência do Leaal/UFPE na produção e transferência de tecnologia			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
transferência de tecnologia	A	Departamento de Nutrição	B
Departamento de Nutrição	B	transferência de tecnologia	A
informação tecnológica	A	produção de tecnologia	A
setor produtivo	C	informação tecnológica	A
produção de tecnologia	A	benefício da sociedade	C
domínio tecnológico	C	desenvolvimento de alimentos funcionais	C
Leaal	A	desenvolvimento de tecnologia	C
benefício da sociedade	C	fortalecimento das capacidades	D
desenvolvimento de alimentos funcionais	C	institutos de pesquisa	A
institutos de pesquisa	A	Laboratório de Experimentação	D
referências teóricas sobre o assunto	D	referências teóricas sobre o assunto	D
transferência da informação gerada	B	transferência da informação gerada	B
Taxa de Relevância	**0,58**	**Taxa de Relevância**	**0,48**

Artigo 35: Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
linguagens de marcação	A	linguagens de marcação	A
utilização na Internet	B	utilização na Internet	B
W3 Consortium	C	W3 Consortium	C
dados semiestruturados	C	características do XML	A
SGML	B	páginas da Internet	A
XML	A	representação de dados	C
características do XML	A	um modelo de dados	D
páginas da Internet	A	uso na Internet	B
Taxa de Relevância	**0,69**	**Taxa de Relevância**	**0,56**

Artigo 36: A Lei de Lotka na bibliometria brasileira			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Lei de Lotka	A	Lei de Lotka	A
método dos mínimos quadrados	A	método dos mínimos quadrados	A
arquivo privado de Getúlio Vargas	C	arquivo privado de Getúlio Vargas	C
modelo de Lotka	A	modelo de Lotka	A
produtividade dos autores	C	produtividade dos autores	C
Lei de Lotka na bibliometria brasileira	A	Lei de Lotka na bibliometria brasileira	A
modelo do poder inverso generalizado	A	modelo do poder inverso generalizado	A
valor de n	C	valor de n	C
Taxa de Relevância	**0,72**	**Taxa de Relevância**	**0,72**

Artigo 37: Bases de dados de informação para negócios			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
bases de dados	A	dados de informação	C
setores industriais	C	informação para negócios	A
texto completo	D	acesso as bases	A
dados de informação	C	bases de dados	A
informação para negócios	A	número de empregados	D
acesso as bases	A	setores industriais	C
número de empregados	D	texto completo	D
Bases de dados de informação	A	bolsas de valores	C
empresas públicas	C	informações sobre produtos	B
texto completo	D	nomes de executivos	C
novos produtos	D	novos produtos	D
bolsas de valores	C	Bases de dados de informação	A
informações sobre produtos	B	empresas públicas	C
nomes de executivos	C	texto completo	D
Taxa de Relevância	**0,41**	**Taxa de Relevância**	**0,41**

Artigo 38: Biblioteca híbrida: um novo enfoque no suporte à educação a distância			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Biblioteca híbrida	A	acesso a informação	A
acesso a informação	A	novo enfoque no suporte	C
novo enfoque no suporte	C	Biblioteca híbrida	A
bibliotecas acadêmicas	A	processo de acesso a a informação	A
processo de acesso à informação	A	processo de atendimento	B
processo de atendimento	B	bibliotecas acadêmicas	A
expectativas de seus usuários	B	expectativas de seus usuários	B
home site das bibliotecas acadêmicas	A	home site das bibliotecas acadêmicas	A
Taxa de Relevância	**0,78**	**Taxa de Relevância**	**0,78**

Artigo 39: Estratégia de busca na recuperação da informação: revisão da literatura			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
estratégias de busca	A	estratégias de busca	A
processo de busca	A	processo de busca	A
bases de dados	B	banco de dados	B
recuperação da informação	A	bases de dados	B
planejamento da estratégia de busca	A	recuperação da informação	A
usuário final	C	programas de treinamento	C
banco de dados	B	planejamento da estratégia de busca	A
programas de treinamento	C	informação do usuário	C
informação do usuário	C	sistema de recuperação	A
sistema de recuperação	A	usuário final	C
busca na recuperação da informação	A	bases de dados textuais	A
linguagens controladas	A	resultados da busca	A
intermediários	C	termos de busca	A
Taxa de Relevância	**0,69**	**Taxa de Relevância**	**0,71**

Artigo 40: A ciência e a gestão da informação: compatibilidades no espaço profissional			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
gestão da informação	A	gestão da informação	A
gestor da informação	A	gestor da informação	A
ciência da informação	A	atividades de informação	B
atividades de informação	B	agregação de valor	D
serviços de informação	A	compatibilidades no espaço profissional	A
agregação de valor	D	serviços de informação	A
compatibilidades no espaço profissional	A	fontes de informação	B
fontes de informação	B	gerenciamento da informação	A
profissionais da informação	A	necessidades de informação e de níveis de agregação de valor	C
necessidades de informação e de níveis de agregação de valor	B	profissionais da informação	A
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,73**

Artigo 41: Produção das literaturas "branca" e "cinzenta" pelos docentes/doutores dos programas de pós-graduação em ciência da informação no Brasil			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
linhas de pesquisa	B	linhas de pesquisa	B
programas de pós-graduação	A	programas de pós-graduação	A
literatura cinzenta	A	década de 90	C
ciência da informação	A	literatura cinzenta	A
década de 90	C	crescimento da ciência	B
crescimento da ciência	B	dezembro de 1999	D
dezembro de 1999	D	literaturas branca e cinzenta pelos docentes	A
literaturas branca e cinzenta pelos docentes	A	produção dos docentes	A
Taxa de Relevância	**0,66**	**Taxa de Relevância**	**0,66**

Artigo 42: Informação para negócios: os novos agentes do conhecimento e a gestão do capital intelectual			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
capital intelectual	A	novos agentes do conhecimento	A
novos agentes do conhecimento	A	Informação para negócios	A
sistemas de informação	B	capital intelectual	A
Informação para negócios	A	gestão do capital intelectual	A
gestão do capital intelectual	A	sistemas de informação	B
gestão da informação	A	agentes criativos da empresa	A
gestão do conhecimento	A	capital intelectual da empresa	A
agentes criativos da empresa	A	capital intelectual de uma empresa	A
capital intelectual da empresa	A	gestão da informação	A
capital intelectual de uma empresa	A	gestão do conhecimento	A
Taxa de Relevância	**0,95**	**Taxa de Relevância**	**0,95**

Artigo 43: Análise metodológica dos estudos de necessidades de informação sobre setores industriais brasileiros: proposições			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
necessidade de informação	A	processo de aprendizagem	A
processo de aprendizagem	A	necessidade de informação	A
inovação tecnológica	B	informação sobre setores industriais brasileiros	A
Núcleo Especializado	C	processo de inovação	A
informação sobre setores industriais brasileiros	A	uso da informação	A
informação tecnológica	A	capacitação de recursos humanos	B
processo de inovação	A	ambiente empresarial	A
uso da informação	A	inovação tecnológica	B
ambiente empresarial	A	Núcleo Especializado	C
capacitação de recursos humanos	B	processo de inovação tecnológica	A
sistemas de informação	C	sistemas de informação	C
necessidades de informação sobre setores industriais brasileiros	A	informação tecnológica	A
setores industriais brasileiros	A	criação do conhecimento	A
Taxa de Relevância	**0,81**	**Taxa de Relevância**	**0,75**

Artigo 44: Informação e competitividade: a contextualização da gestão do conhecimento nos processos organizacionais			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
gestão do conhecimento	A	tecnologia da informação	B
tecnologia da informação	B	gestão do conhecimento	A
criação do conhecimento	A	criação do conhecimento	A
conhecimento tácito	A	administração do conhecimento	A
administração do conhecimento	A	construção do conhecimento organizacional	A
construção do conhecimento organizacional	A	inovação de produtos	B
inovação de produtos	B	melhor administração do conhecimento	A
melhor administração do conhecimento	A	conhecimento tácito	A
contextualização da gestão do conhecimento	A	contextualização da gestão do conhecimento	A
processos organizacionais	B	processos organizacionais	B
conhecimentos tecnológicos	C	conhecimentos tecnológicos	C
Taxa de Relevância	**0,80**	**Taxa de Relevância**	**0,80**

Artigo 45: Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
motores de busca	B	motores de busca	B
número de links	B	número de links	B
registro da informação	B	registro da informação	B
difusão do conhecimento	A	difusão do conhecimento	A
recuperação de informação	B	desenvolvimento de políticas científicas	C
desenvolvimento de políticas científicas	C	impacto da Web	C
impacto da Web	C	recuperação de informação	B
mecanismos utilizados para medir o registro da informação	A	campo da webometria	A
políticas científicas	C	fator de impacto	A
cienciometria	A	fluxos da informação	B
informetria	A	quantidade de resultados	D
webometria	A	resultados de uma busca	C
Taxa de Relevância	**0,65**	**Taxa de Relevância**	**0,52**

Artigo 46: Métodos quantitativos de apoio à bibliometria: a pesquisa operacional pode ser uma alternativa?			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
lei de Bradford	A	lei de Bradford	A
Teoria do Caos	A	Teoria do Caos	A
formulação de Bradford	A	formulação de Bradford	A
dispersão de artigos	A	dispersão de artigos	A
ciência da informação	A	unidade de informação	B
unidade de informação	B	campo da bibliometria	A
caos	D	Ciência do Caos	C
Po	D	decisão do problema	D
campo da bibliometria	A	dependência sensível das condições iniciais	C
Ciência do Caos	C	efeito do agrupamento	D
dependência sensível das condições iniciais	C	estabelecimento da função-objetivo	C
efeito do agrupamento	D	estudantes do 1º	D
estabelecimento da função-objetivo	C	estudantes do 2º grau	D
estudantes do 1º	D	início do século	D
estudantes do 2º grau	D	ciência da informação	A
Taxa de Relevância	**0,48**	**Taxa de Relevância**	**0,48**

Artigo 47: Periódicos eletrônicos: considerações relativas à aceitação deste recurso pelos usuários			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
meio eletrônico	B	ferramentas de busca	A
ferramentas de busca	A	meio eletrônico	B
fontes de informações secundárias	B	fontes de informações secundárias	B
tela do computador	B	tela do computador	B
uso do hipertexto	B	uso do hipertexto	B
Periódicos eletrônicos	A	endereço do periódico eletrônico	A
hipertexto	A	ferramentas de indexação e busca	A
ferramentas de indexação e busca	A	utilização do hipertexto	A
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,75**

Artigo 48: Alguns aspectos do uso da informação na economia da informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
economia da informação	A	economia da informação	A
uso da informação	A	uso da informação	A
redução de custo	A	redução de custo	A
cadeia de valor	A	cadeia de valor	A
clientes	D	cadeia de suprimentos	A
cadeia de suprimentos	A	cadeia de valor virtual	A
cadeia de valor virtual	A	comportamento das pessoas	C
comportamento das pessoas	C	Diferenciação de produto	A
Diferenciação de produto	A	fluxo de informação	B
fluxo de informação	B	gestão da informação	A
gestão da informação	A	diversas maneiras	D
diversas maneiras	D	valor virtual	C
Taxa de Relevância	**0,73**	**Taxa de Relevância**	**0,75**

Artigo 49: Ferramentas alternativas para monitoramento e mapeamento automatizado do conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
desenvolvimento de produto	C	desenvolvimento de produto	C
freqüência das respostas	C	freqüência das respostas	C
alternativas para monitoramento	A	alternativas para monitoramento	A
formação de clusters	B	formação de clusters	B
Mapa de Conhecimentos	B	Mapa de Conhecimentos	B
descritores	C	mapeamento automatizado do conhecimento	A
número de descritores	C	resultados de buscas	B
tratamento automatizado da informação	A	tratamento automatizado da informação	A
Taxa de Relevância	**0,50**	**Taxa de Relevância**	**0,63**

Artigo 50: A formação profissional no século XXI: desafios e dilemas Artigo 49: Ferramentas alternativas para monitoramento e mapeamento automatizado do conhecimento			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
sociedade do conhecimento	A	educação do futuro	A
educação do futuro	A	sociedade do conhecimento	A
condição humana	C	educação dos bibliotecários	A
educação dos bibliotecários	A	mundo do trabalho	B
mundo do trabalho	B	condição humana	C
bibliotecários	A	educação no século	B
futuro	C	ética do gênero humano	C
formação profissional no século	A	formação profissional no século	A
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,69**

Artigo 51: A acessibilidade à informação no espaço digital			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
espaço digital	A	limitações oriundas de deficiência	A
equivalentes textuais	D	leitura de tela	B
limitações oriundas de deficiência	A	equivalentes textuais	D
leitura de tela	B	espaço digital	A
acesso a informação	A	acessibilidade no espaço digital	A
acessibilidade no espaço digital	A	acesso a informação	A
pessoas portadoras de deficiência	A	pessoas portadoras de deficiência	A
usuários com limitações	A	usuários com limitações	A
ajudas técnicas	C	estrutura dos documentos	C
espaço tridimensional	B	serviços de biblioteca	B
Taxa de Relevância	**0,73**	**Taxa de Relevância**	**0,73**

Artigo 52: Estudos de usuários: o padrão que une três abordagens			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
Agregados De Informação	A	Agregados De Informação	A
Carmelita do Espírito Santo	C	Carmelita do Espírito Santo	C
produção do conhecimento	B	padrão que une três abordagens	A
padrão que une três abordagens	A	produção do conhecimento	B
transferência da informação	A	transferência da informação	A
ciência da informação	B	ciência da informação	B
hipertexto	A	hipertexto	A
oficinas	D	oficinas	D
Taxa de Relevância	**0,66**	**Taxa de Relevância**	**0,66**

Artigo 53: Como incrementar a qualidade dos resultados das máquinas de busca: da análise de logs à interação em português			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
máquinas de busca	A	máquinas de busca	A
sistemas de busca	A	sistemas de busca	A
língua natural	B	consultas em língua natural	A
consultas em língua natural	A	representação em língua natural	B
representação em língua natural	B	língua natural	B
tipo de conexão	D	tipo de conexão	D
consultas	B	análise de logs	A
Português	D	comportamento do usuário	C
recuperação de informação	A	objetivos por meio de palavras-chave	A
comportamento do usuário	A	reconhecimento sintático de padrão	B
objetivos por meio de palavras-chave	A	recuperação de informação	A
Taxa de Relevância	**0,68**	**Taxa de Relevância**	**0,70**

Artigo 54: Information literacy: princípios, filosofia e prática			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
tecnologia da informação	C	resolução de problemas	D
resolução de problemas	D	tecnologia da informação	C
programas educacionais	C	uso da informação	B
uso da informação	B	programas educacionais	C
acesso a informação	A	acesso a informação	A
aprendizado ao longo da vida	B	aprendizado ao longo da vida	B
implementação de programas educacionais	B	implementação de programas educacionais	B
profissional da informação	B	profissional da informação	B
aprendizado	D	âmbito da biblioteca	B
bibliotecário	B	bibliotecário como agente educacional	A
bibliotecário como agente educacional	A	conjunto integrado de habilidades	B
série de habilidades e conhecimentos	B	série de habilidades e conhecimentos	C
Taxa de Relevância	**0,46**	**Taxa de Relevância**	**0,48**

Artigo 55: Profissional da informação: perfil de habilidades demandadas pelo mercado de trabalho			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
profissional da informação	A	profissional da informação	A
gestão do conhecimento	A	gestão do conhecimento	A
ciência da informação	A	perfil de habilidades	A
área da ciência da informação	A	seleção de recursos humanos	A
perfil de habilidades	A	mercado de trabalho	A
recursos humanos	B	área da ciência da informação	A
seleção de recursos humanos	A	gestão da informação e do conhecimento	A
mercado de trabalho	A	organizações do conhecimento	A
Taxa de Relevância	**0,94**	**Taxa de Relevância**	**1,00**

Artigo 56: O olhar da consciência possível sobre o campo científico			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
campo da ciência da informação	A	cientistas da informação	A
ciência da informação	A	campo científico	B
campo científico	B	campo da ciência da informação	A
cientistas da informação	A	artigo de Wersig e Neveling	B
informação científica	B	evento de comunicação	C
artigo de Wersig e Neveling	B	grupo de cientistas	C
evento de comunicação	C	início dos anos 70	C
grupo de cientistas	C	problemas da informação	C
início dos anos 70	C	processo de comunicação	C
problemas da informação	C	visões do mundo	C
processo de comunicação	C	informação científica	B
Taxa de Relevância	**0,52**	**Taxa de Relevância**	**0,45**

Artigo 57: As relações entre ciência, Estado e sociedade: um domínio de visibilidade para as questões da informação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
ciência da informação	A	ações de informação	B
ações de informação	B	produção dos conhecimentos	A
produção dos conhecimentos	A	transferência de informação	A
transferência de informação	A	década de 60	C
década de 60	C	programa de pesquisa	C
conhecimentos científicos	B	regimes de informação	B
programa de pesquisa	C	áreas do conhecimento	C
regimes de informação	B	bases de dados referenciais	D
comunicação científica	A	controle de qualidade	D
inteligência científica	A	economia de Mercado	B
década de 90	C	prestação de contas	D
Estado	C	produção de conhecimentos científicos	A
bases de dados referenciais	C	questões da informação	B
produção de conhecimentos científicos	A	Um regime de informação	A
Taxa de Relevância	**0,63**	**Taxa de Relevância**	**0,48**

Artigo 58: Interfaces entre a ciência da informação e a ciência cognitiva			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
processamento da informação	A	processamento da informação	A
processo cognitivo	A	processo cognitivo	A
ciência cognitiva	A	ciência da computação	A
recuperação da informação	A	recuperação da informação	A
ciência da computação	A	ciência cognitiva	A
inteligência artificial	B	ponto de vista cognitivo	A
ponto de vista cognitivo	A	processo de indexação	A
sistemas de informação	C	inteligência artificial	B
processo de indexação	A	linguagem de indexação	A
CC	B	organização da informação	A
Ci	B	sistemas de informação	C
computação	C	tecnologias da informação	C
Taxa de Relevância	**0,75**	**Taxa de Relevância**	**0,83**

Artigo 59: A produção científica da Anped e da Intercom no GT da Educação e Comunicação			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
produção do conhecimento	A	produção do conhecimento	A
Anped	A	área de comunicação e educação	A
Intercom	A	arqueologia do saber	A
área de comunicação e educação	A	autores da Intercom	A
arqueologia do saber	A	GT da Educação	B
autores da Intercom	A	inter-relação entre comunicação e educação	A
inter-relação entre comunicação e educação	A	literatura de congressos nacionais	A
literatura de congressos nacionais	A	unidade de análise	C
Taxa de Relevância	**1,00**	**Taxa de Relevância**	**0,84**

Artigo 60: Inteligência competitiva na Internet: um processo otimizado por agentes inteligentes			
Descritores selecionados na primeira aplicação da metodologia	**Valor de Relevância atribuído**	**Descritores selecionados na segunda aplicação da metodologia**	**Valor de Relevância atribuído**
estratégia de atuação	A	estratégia de atuação	A
processo de inteligência competitiva	A	Inteligência competitiva na Internet	A
fontes de informação	A	processo de inteligência competitiva	A
agentes inteligentes	A	fontes de informação	A
Inteligência competitiva na Internet	A	informação na internet	A
inteligência competitiva	A	gestão do conhecimento	A
informação na internet	A	necessidades de informação	A
gestão da informação	A	Programa de Pós-Graduação	C
gestão do conhecimento	A	Situação a partir da abordagem	D
informações externas	A	Universidade Federal de Santa Catarina	C
Taxa de Relevância	**1,00**	**Taxa de Relevância**	**0,75**

ANEXO D: Lista de sintagmas nominais descartados

Neste anexo são apresentados os SNs que foram preteridos em todas as relações de descritores, por serem demasiado comuns e possuírem pouco poder discriminatório. Só foram eliminados quando ocorriam em estruturas simples (D + N), para qualquer determinante. Esta *stoplist* deve ser considerada apenas no contexto deste trabalho, e relativa ao *corpus* analisado. Foram descartados os seguintes SNs (apresentados sem os determinante):

Análise(s)	Expressão(ões)	Professor(a,es,as)
Autor(a,es,as)	Ferramenta(s)	Profissional(ais)
Ciência	Figura(s)	Receptor(a,es,as)
Cientista(s)	Indivíduo(s)	Rede(s)
Coleta de dados	Indústria(s)	Ser humano
Comunicação(ões)	Mesmo(s)	Si mesmo
Conceito(s)	Organização(ões)	Sistema(s)
Conhecimento(s)	País(es)	Sociedade
Curso(s)	Palavra(s)	Tabela(s)
Documento(s)	Par(es)	Trabalho(s)
Emissor(a,es,as)	Processo(s)	Usuário(a,os,as)
Empresa(s)	Produção(ões)	Valor agregado
Espaço(s)	Produto(s)	Vida(s)

A quantidade é bastante pequena porque reflete o *corpus* de apenas 60 documentos. Na medida em que mais e mais documentos forem analisados, a tendência é que sejam criadas – para cada área do conhecimento – listas específicas e extensas, que possibilitem que as representações dos documentos, através dos descritores selecionados automaticamente, sejam cada vez mais significativas.

ANEXO E: Indicações do *corpus* utilizado na comparação da extração automática e manual

Neste anexo são apresentadas as indicações dos textos utilizados pelo professor Dr. Hélio KURAMOTO em sua tese de doutorado (1999). Estes textos foram utilizados no escopo deste trabalho para realizar uma comparação entre a extração automática e a extração manual dos SNs.

Levando-se em conta que o anexo original compreende um grande número de páginas, considerou-se desnecessário reproduzi-lo na íntegra. Em vez disto, são apresentados os títulos dos artigos, um excerto do primeiro artigo na forma com que é apresentado, e um excerto do conjunto global de sintagmas nominais extraídos. Seguem os títulos dos artigos:

1. Conhecimento como recurso estratégico empresarial
2. Inteligência competititva e decisão empresarial
3. Economia da Informação
4. Informação como Insumo Estratégico
5. Informação Técnico-econômica: mais importante do que nunca
6. Perspectivas do Agente da Informação no Contexto Brasileiro
7. Sistemas de Informação: a evolução dos enfoques
8. Consultoria Informatológica em revisão: uma alternativa para serviços de informação personalizados
9. Informação para a Indústria
10. Interação entre empresas com necessidades de informação (=conhecimento) e a estrutura nacional de centros com provisão de conhecimento acumulado: referência especial à estrutura nacional de serviços de informação, documentação e de biblioteca
11. Uso da Informação na Indústria como Paradigma para o Desenvolvimento Econômico
12. A Informação Eficaz na Empresa
13. Gerência da Informação: mudanças nos perfis profissionais
14. Informação: instrumento de dominação e de submissão
15. Informação: a chave para a qualidade total

E a seguir, são apresentados pequenos excertos do primeiro artigo, na forma de tópicos, e um excerto do conjunto global de sintagmas nominais extraídos da totalidade do conjunto.

Annexe A
Le Corpus d’articles

Article nº. 1

1. Conhecimento como recurso estratégico empresarial

2. ANTECEDENTES

3. As organizações brasileiras defrontam-se hoje
com rapidez e profundas transformações (políticas, econômicas, sociais, tecnológicas) dos ambientes nacional e internacional, associadas a uma crescente competição no mundo dos negócios e ao surgimento de uma categoria de clientes conscientizados de seus direitos a produtos e serviços de alta qualidade.

4. No Primeiro Mundo, frente a idênticos desafios, a resposta das organizações-líderes tem sido um movimento de mudanças em direção à melhor sintonia com o mercado e à busca de excelencia, o que se levou à valorização da informação e da tecnologia da informação como parte de um elenco de recursos estratégicos capazes de lhes propiciar vantagem competitiva diante da concorrência.

5. Tal movimento fez com que informação, conhecimento e inteligência se incluíssem atualmente entre os termos mais freqüentes da literatura sobre gestão empresarial e que delas se tenham ocupado autores como Porter, Drucker, Toffler, Ohmae e Cronin, considerando tais elementos como recursos estratégicos e insumos para a gestão das organizações em ambiente competitivo.

6. Falando especificamente de organizações industriais, cujas atividades-fim demandam constantes insumos de informação científica e tecnológica (ICT), diferentes pesquisas desenvolvidas e divulgadas por autores diversos (Orpen, Goldhar, Koenig, Ginman), a partir de Allen, estabelecem uma relação direta entre produtividade, inovação e um livre e vigoroso fluxo de informações intra e interorganizacionais (in, out, up, down and accross the organization).

7. CONCEITOS

8. Assim, entende-se hoje como um dos mais nobres papéis do administrador aquele relacionado à preservação e ao desenvolvimento do saber específico de sua organização em todos os seus setores e níveis hierárquicos, seja esse saber codificado sob forma, de dados, documentos, informações e sistemas, ou personalizado sob forma de knowhow do especialista dotado de conhecimento teórico e experiência prática. Esse saber é hoje reconhecido como um valioso ativo empresarial que se busca maximizar (mediante a educação formal, treinamento e comunicação), registrar
(sob forma de sistemas arquivos/bibliotecas/centros-de-informação, e via tecnologia - DBMS/Data based Management System, MIS/Management Information System, EIS/Executive Information System, KBS/Knowledge-based System, KBDSS/Knowledge-based Decision Supporting System) e integrar sob uma GRI/gerência de recursos informacionais (IRM/Information Resources Management, função para cujo desempenho em ambiente tecnológico a IBM ganhou a figura do Chief Information Officer. (CIO).

9. Vistos isoladamente cada um desses recursos, dados são considerados fragmentos da realidade que, codificados/moldados para a comunicação e o uso de cliente(s) específico(s), convertem-se
em informação. Prosseguindo nessa hierarquia qualitativa, conhecimento é informação com valor agregado, produzida com pretensão de validade universal, assimilada pelo indivíduo ou pela organização e integrada a seu saber anterior. Por fim, inteligência é o conjunto de estratégias utilizadas (pelo indivíduo, pela empresa ou pelo país) para captar, avaliar, combinar e utilizar eficazmente informações em decisões e ações necessárias para sua adaptação às mudanças ambientais, tendo em vista o alcance de objetivos preestabelecidos; quando se trata de um pais, denomina-se "inteligência social", enquanto a expressão "inteligência **(continua...)**

Annexe B
Le corpus de Syntagmes Nominaux

! "1% da producao cientifica mundial"
_
! "100 mil titulos"
_
! "100 mil titulos de publicacoes tecnicocientificas"
! "15000 empresas de manufatura"
! "1914"
! "1947"
! "1950"
! "1951"
! "1960"
! "1975"
! "1987"
! "1988"
! "1990"
! "1991"
! "20 anos"
! "95% da literatura tecnico-cientifica mundial"
! "a embalagem da informacao"
! "a abordagem da qualidade total"
! "a abordagem de custos e eficacia de serviços de informacao"
! "a abordagem de economia de rede"
! "a abordagem de logistica economica"
! "a abordagem do objeto informacao"
! "a abordagem estrategica"
! "a acao"
! "a acao do sistema"
! "a acao em desenvolvimento"
! "a acao empresarial"
! "a acao governamental"
! "a acao neguentropica do conjunto instrumentos/piloto"
! "a aceitacao da informacao estrategica na definicao do futuro da empresa"
! "a aceitacao generalizada da pratica de consultoria informatologica"
! "a aceitacao generalizada do conceito de consultoria informatologica"
! "a aceitacao generalizada do conceito e da pratica de consultoria informatologica"
! "a acepcao mais ampla do conceito de ari"
! "a acumulacao de capital"
! "a acumulacao de informacao na area de automacao"
! "a acumulacao de riquezas"
! "a adaptacao a nova realidade economica"
! "a adaptacao da empresa as mudancas ambientais"
! "a adaptacao do individuo as mudancas ambientais"
! "a adequacao de novas informacoes"
! "a adequacao dos produtos as necessidades dos clientes"
! "a adequacao dos produtos ou servicos as necessidades dos clientes"
! "a adequacao dos servicos as necessidades dos clientes"
! "a adequacao entre direito do cidadao a privacidade e a necessidade de tornar eficientes os sistemas de informacao das organizacoes e do estado"
! "a administracao"
! "a administracao da empresa"
! "a administracao da informacao como recurso"
! "a administracao das empresas"
! "a administracao dos fatores de producao classicos"
! "a administracao dos recursos de informacao"
! "a administracao estrategica"
! "a administracao superior"
! "a adocao de estrategias"
! "a adocao de inovacoes tecnologicas"
! "a adocao do activity based cost"
! "a adocao do kanban"
! "a adocao generalizada do conceito de consultoria informatologica"
! "a agregacao de valor"
! "a agregacao de valor na relacao da organizacao com a sociedade"
! "a agregacao de valor nas atividades economicas"
! "a agregacao e realizacao de valor"
! "a alta administracao"
! "a alta direcao"
! "a alta gerencia"
! "a america latina"
! "a analise"
! "a analise custo-beneficio"
! "a analise da eficacia"
! "a analise da expansao industrial brasileira"
! "a analise da informacao"
! "a analise da informacao de dados prospectivos
baseados no conteudo e nos dados estatisticos agregados ou indicadores"
! "a analise da informacao pelos meios convenientes"
! "a analise das mudancas na sociedade"
! "a analise das mudancas no mundo"

(continua...)