

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL  
INSTITUTO DE INFORMÁTICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

TECNOLOGIAS DE DESCOBERTA  
DE CONHECIMENTO EM TEXTOS  
APLICADAS À INTELIGÊNCIA COMPETITIVA

por

LEANDRO KRUG WIVES

EXAME DE QUALIFICAÇÃO  
EQ-069 PPGC-UFRGS

JOSÉ PALAZZO MOREIRA DE OLIVEIRA  
Orientador – PPGC

LILIA VARGAS  
Co-orientadora - PPGA

PORTO ALEGRE, JANEIRO DE 2002.



UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitora: Profa. Wrana Panizzi

Pró-Reitor de Pós-Graduação: Franz Rainer Semmelmann

Diretor do Instituto de Informática: Prof. Philippe Olivier Alexandre Navaux

Coordenador do PPGC: Prof. Carlos Alberto Heuser

Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro



## SUMÁRIO

<b>LISTA DE ABREVIATURAS .....</b>	<b>7</b>
<b>LISTA DE FIGURAS .....</b>	<b>9</b>
<b>LISTA DE TABELAS.....</b>	<b>11</b>
<b>INTRODUÇÃO .....</b>	<b>17</b>
<b>PARTE I – ÁREA DE ABRANGÊNCIA: BUSCA E RECUPERAÇÃO DE INFORMAÇÕES.....</b>	<b>19</b>
1 BUSCA E RECUPERAÇÃO DE INFORMAÇÕES .....	21
2 RECUPERAÇÃO DE INFORMAÇÕES: HISTÓRICO E CONCEITOS BÁSICOS.....	23
2.1 <i>Documento</i> .....	24
2.2 <i>Relevância e informação</i> .....	25
3 PARADIGMA DA ÁREA DE RECUPERAÇÃO DE INFORMAÇÕES .....	27
3.1 <i>Abstração de informações</i> .....	28
3.2 <i>Descrição da necessidade do usuário</i> .....	29
3.3 <i>O processo de “casamento” ou “matching”</i> .....	31
4 SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES .....	33
4.1 <i>Tipos de sistemas de recuperação de informações</i> .....	34
4.1.1 <i>Sistemas de recuperação de informação bibliográfica</i> .....	34
4.1.2 <i>Sistemas de recuperação de informação textual</i> .....	34
4.1.3 <i>Sistemas de recuperação de informação visual</i> .....	35
4.1.4 <i>Bibliotecas Digitais</i> .....	35
5 SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES TEXTUAIS.....	37
5.1 <i>Modelos de Recuperação de Informações</i> .....	37
5.1.1 <i>Modelo booleano</i> .....	37
5.1.2 <i>Modelo espaço-vetorial</i> .....	39
5.1.3 <i>Modelo probabilístico</i> .....	40
5.1.4 <i>Modelo difuso</i> .....	41
5.1.5 <i>Modelo da busca direta</i> .....	42
5.1.6 <i>Modelo de aglomerados (clustering model)</i> .....	43
5.1.7 <i>Modelo lógico</i> .....	44
5.1.8 <i>Modelo contextual</i> .....	45
6 INDEXAÇÃO E NORMALIZAÇÃO .....	49
7 INDEXAÇÃO AUTOMÁTICA .....	51
7.1 <i>Identificação de termos</i> .....	51
7.2 <i>Identificação de termos compostos</i> .....	52
7.3 <i>Remoção de “stopwords”</i> .....	52
7.4 <i>Normalização morfológica</i> .....	53
7.5 <i>Cálculo de relevância</i> .....	54
7.6 <i>Seleção de termos</i> .....	56
7.6.1 <i>Filtragem baseada no “peso” do termo</i> .....	56
7.6.2 <i>Seleção baseada no “peso” do termo</i> .....	57
7.6.3 <i>Seleção por “Latent Semantic Indexing”</i> .....	57
7.6.4 <i>Seleção por análise de linguagem natural</i> .....	57
8 ESTRUTURAS DE ARMAZENAMENTO.....	59
8.1 <i>Arquivos invertidos</i> .....	59
8.2 <i>Árvores TRIE</i> .....	59
8.3 <i>Método da assinatura</i> .....	60
8.4 <i>Árvores PAT</i> .....	61
9 BUSCA E RECUPERAÇÃO.....	65
9.1 <i>Formulação de consulta</i> .....	66
9.2 <i>Identificação de itens relevantes (técnicas de “casamento”)</i> .....	68
9.3 <i>Visualização e análise dos resultados</i> .....	68
10 BIBLIOMETRIA .....	71
10.1 <i>Recall</i> .....	72

10.2	<i>Precision</i> .....	72
10.3	<i>Fallout</i> .....	72
10.4	<i>Effort</i> .....	73
11	EXEMPLOS DE SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES.....	75
11.1	<i>Sistemas clássicos</i> .....	75
11.2	<i>Motores de busca na WEB</i> .....	76
11.3	<i>Sistemas de meta-busca</i> .....	77
<b>PARTE II – PROFUNDIDADE: TECNOLOGIAS DE DESCOBERTA DE CONHECIMENTO APLICADAS À INTELIGÊNCIA COMPETITIVA .....</b>		<b>79</b>
12	INTELIGÊNCIA COMPETITIVA.....	81
12.1	<i>Definição da inteligência competitiva</i> .....	82
12.2	<i>Objetivos</i> .....	83
12.3	<i>Etapas do processo de inteligência</i> .....	83
13	DESCOBERTA DE CONHECIMENTO .....	85
13.1	<i>Etapas do processo de descoberta de conhecimento</i> .....	86
13.2	<i>Mineração de dados</i> .....	87
14	DESCOBERTA DE CONHECIMENTO EM TEXTOS .....	89
14.1	<i>Etapas do processo de descoberta de conhecimento em textos</i> .....	89
14.2	<i>Tipos de descoberta de conhecimento em textos</i> .....	91
14.2.1	<i>Extração de informações</i> .....	91
14.2.2	<i>Sumarização</i> .....	92
14.2.3	<i>Clustering (agrupamento)</i> .....	93
14.2.4	<i>Classificação e categorização</i> .....	94
14.2.5	<i>Filtragem de informação</i> .....	95
15	SISTEMAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOS .....	97
15.1	<i>Phoaks</i> .....	97
15.2	<i>Referral Web</i> .....	97
15.3	<i>Fab</i> .....	98
15.4	<i>Siteseer</i> .....	98
15.5	<i>GroupLens</i> .....	98
15.6	<i>Umap</i> .....	99
15.7	<i>Sphinxs</i> .....	99
15.8	<i>Leximine (Sampler)</i> .....	100
15.9	<i>GrapeVine</i> .....	100
16	MÉTODOS DE MINERAÇÃO DE TEXTOS APLICADOS À INTELIGÊNCIA COMPETITIVA.....	103
16.1	<i>Análise lexicométrica</i> .....	103
16.2	<i>Extração de informações</i> .....	104
16.3	<i>Identificação de clusters (clustering)</i> .....	104
16.4	<i>Classificação</i> .....	105
16.5	<i>Análise de ferramentas versus métodos e etapas de inteligência competitiva</i> .....	105
17	CONCLUSÕES .....	107
BIBLIOGRAFIA.....		109

## **LISTA DE ABREVIATURAS**

- BD – Banco de Dados
- DBMS – Database Management System
- DM – DataMining (mineração de dados)
- IE – Information Extraction
- EI – Extração de Informações
- IC – Inteligência Competitiva
- KDD – Knowledge Discovery from Data (descoberta de conhecimento em dados)
- NI – Necessidade de Informação
- KDT – Knowledge Discovery from Texts (descoberta de conhecimento em textos)
- PLN – Processamento de Linguagem Natural
- RI – Recuperação de Informações
- SEI – Sistema de Extração de Informações
- SGBD – Sistema de Gerenciamento de Bancos de Dados
- SRI – Sistema de Recuperação de Informações



## LISTA DE FIGURAS

Figura 3-1 – Componentes básicos do modelo.....	28
Figura 3-2 – O processo de abstração .....	29
Figura 3-3 – Problema do processo de descrição de uma consulta .....	30
Figura 3-4 – Diferença entre domínios de vocabulário .....	31
Figura 3-5 – O processo de "casamento" de características .....	32
Figura 5-1 – O modelo espaço-vetorial .....	39
Figura 5-2 – Um exemplo de recuperação utilizando lógica.....	44
Figura 5-3 – Recuperação contextual .....	46
Figura 7-1 – Etapas do processo de indexação automática .....	51
Figura 7-2 – Identificação de termos válidos .....	51
Figura 7-3 – Identificação de <i>Stop-Words</i> .....	53
Figura 8-1 – Estrutura de uma Lista Invertida.....	59
Figura 8-2 – Nodo de uma árvore trie .....	60
Figura 8-3 – Estrutura exemplo de uma árvore trie.....	60
Figura 8-4 – Exemplo de assinatura .....	61
Figura 8-5 – Exemplo de árvore PAT .....	62
Figura 10-1 – Efeitos de uma busca no banco de dados textual.....	71
Figura 13-1 – O processo de KDD (simplificado) .....	86
Figura 13-2 – O processo de KDD .....	87
Figura 14-1 – Tipos de agrupamento .....	93



## **LISTA DE TABELAS**

Tabela 1 – Técnicas de KDT que podem ser utilizadas em cada etapa de IC .....	105
Tabela 2 – Comparação entre as ferramentas de text-mining estudadas .....	106



## **RESUMO**

Com a globalização as empresas necessitam cada vez mais obter vantagem competitiva sobre seus concorrentes a fim de manter ou conquistar novos mercados. A vantagem competitiva é obtida pelo oferecimento de um produto ou serviço diferenciado que os concorrentes não possuam. Para que esse produto ou serviço seja criado ou identificado, torna-se necessária a coleta e análise de informações relevantes sobre o ambiente interno e externo da empresa. Como a quantidade de informações disponível é muito grande, coletar as mais relevantes e analisa-las a fim de tomar uma decisão, em prol da obtenção de uma vantagem competitiva, torna-se uma tarefa complexa. Essa monografia apresenta um conjunto de técnicas, métodos e ferramentas computacionais, provenientes da área de recuperação de informações, descoberta de conhecimento em textos e inteligência competitiva empresarial, capaz de facilitar essa tarefa.



## **ABSTRACT**

Globalization is a factor that increases enterprises' needs for competitive advantage over its concurrency. The competitive advantage helps an enterprise to preserve and to attain new markets. The advantage is obtained by the offering of a product or service that the concurrency does not have. To achieve this differential it is necessary to collect and to analyze all relevant information about the internal and external enterprise's environment. The problem is that the amount of information available is very large and to analyze it in order to obtain competitive advantage is a complex task that can be aided by the use of computational tools. This work presents a set of techniques, methods and computational tools, within the fields of information retrieval, knowledge discovery from texts and competitive intelligence, capable of aiding the overall process of competitive intelligence.



## INTRODUÇÃO

A cada dia o mundo dos negócios possui mais e mais dados e informações que podem ser analisados, resumidos e transformados em ação rapidamente. Em parte, essa quantidade de dados e informações é sustentada pelo fenômeno da globalização. A globalização e os novos avanços tecnológicos, principalmente na área da comunicação, proporcionam que uma empresa atue em mercados posicionados fisicamente em locais distantes do mundo. Isso significa que a localização física de uma empresa já não importa mais, pois ela pode atuar em qualquer lugar do planeta como se estivesse em todos eles ao mesmo tempo.

O que acontece, na verdade, é uma “virtualização” da empresa, que passa a existir em um site Internet que pode ser acessado por todos. Com isso, uma empresa comum, que tem seu mercado fixo e ativo, pode, de um momento para outro, perder seu mercado para um concorrente cuja localização está em um ponto completamente oposto do mundo, mas que oferece melhor preço e melhores produtos. Dentro deste contexto, somente as empresas que possuem alguma vantagem competitiva poderão se manter ativas.

Devido a isso, as empresas devem estar constantemente inovando, oferecendo novos produtos e utilizando novas tecnologias. São esses fatores que oferecem vantagem competitiva. A vantagem competitiva é portanto adquirida através do conhecimento de seu ambiente interno e, principalmente, externo, ou seja: seus clientes (o mercado), fornecedores e concorrentes.

Como os concorrentes também podem coletar as mesmas informações, já que muitas fontes são públicas, eles e outros possíveis concorrentes (empresas distantes mas do mesmo ramo ou de ramos similares que podem mudar de ramo em busca de um novo nicho de mercado) devem ser constantemente monitorados para que possíveis ataques (invasões de mercado) sejam prevenidos ou, similarmente ao que eles fariam, para que novos nichos de mercado possam ser identificados.

Todos esses fatores fazem com que a análise dos dados e informações tenha de ser cada vez mais rápida para que decisões e ações sejam tomadas antes que a concorrência o faça. Isso significa, também, que já não basta simplesmente coletar informações e armazená-las em grandes bases de dados. A *concorrência* faz com que as informações possuam uma *vida-útil* cada vez menor (em termos de vantagem competitiva). Se as ações que elas instigam não são colocadas em prática antes dos concorrentes, a vantagem competitiva não ocorre. Nesse momento, a informação e o conhecimento adquirido com ela viram *commodities*, e, neste momento, a sua aplicação torna-se necessária (e básica) para que a empresa mantenha-se no nível das concorrentes e não perca seu próprio mercado.

Devido a grande quantidade de informações existente sobre os elementos do ambiente externo da empresa, e da necessidade de análise e ação rápida (antes dos concorrentes), torna-se necessária a utilização de metodologias, ferramentas e softwares de apoio que auxiliem esse processo.

Pelo fato desse processo lidar com informações (em sua maioria textuais) em todo o seu percurso, torna-se necessário conhecer as tecnologias de indexação, manipulação e análise desse tipo de informação. Portanto, na primeira parte desse exame de qualificação serão abordadas justamente as técnicas de manipulação de informações (textuais) utilizadas nos sistemas de recuperação de informações e Internet.

Iniciar-se-á com um breve histórico do surgimento dessas técnicas, a definição dos conceitos necessários a sua compreensão e a apresentação do paradigma utilizado por elas. A seguir, serão abordados os sistemas de recuperação de informação que buscam implementar o paradigma na prática. Será dada maior ênfase ao tipo textual, que até o momento foi o mais abordado na área da computação. Nesse momento, todos os detalhes técnicos relacionados ao desenvolvimento e utilização desse tipo de sistema serão apresentados. Para finalizar, serão apresentados alguns sistemas de recuperação de informação textual existentes no mercado.

Após, na segunda parte, correspondente ao exame de profundidade, serão abordados os métodos e ferramentas capazes de analisar as informações do ambiente empresarial (disponíveis no formato textual). Nesse momento, o conhecimento abordado na parte anterior torna-se extremamente importante, já que estas ferramentas e métodos de análise de informações textuais são baseadas em muitos conceitos, modelos e técnicas oriundas de lá. Pode-se, inclusive, dizer que elas são a evolução natural dos sistemas de recuperação de informações, oferecendo técnicas e métodos complementares e mais refinados de recuperação e análise de busca.

Essas técnicas, métodos e ferramentas são relativas a uma nova área denominada *descoberta de conhecimento em textos (knowledge discovery from texts)* ou, simplesmente, *mineração de textos (text-mining)*. Essa área possui uma metodologia de aplicação, ainda não bem formalizada, que também é apresentada nessa parte desse exame.

Como este exame foca-se na análise de informações com o objetivo de aumentar a competitividade em empresas, a área de inteligência competitiva empresarial é rapidamente abordada.

Finalmente, são apresentadas algumas direções de como as técnicas de descoberta de conhecimento em textos podem ser agregadas ao processo de inteligência competitiva empresarial a fim de que o processo de transformação de dados e informações em inteligência ou ação seja efetivado de forma mais produtiva.

PARTE I – ÁREA DE ABRANGÊNCIA:  
BUSCA E RECUPERAÇÃO DE INFORMAÇÕES



# 1 Busca e recuperação de informações

O *Ser Humano* necessita de conhecimento para viver. A fim de adquirir conhecimento o *homem* necessita obter informações, e o faz interagindo com outros homens e objetos. Esse é um fato implantado em nossa cultura. A prova disso é que historicamente os homens vêm acumulando grandes quantidades de objetos capazes de transmitir informação. Grandes bibliotecas foram criadas e até hoje continuam sendo grandes centros de informação populares.

Técnicas de suporte para que os ambientes propostos pelos grandes centros de informação se tornassem viáveis foram elaboradas. Isso deu origem à grande maioria dos métodos de armazenamento, localização e manipulação de informações que existem atualmente.

Com o surgimento da informática, os computadores passaram a ser utilizados no auxílio à busca e manipulação de informações, dando origem aos primeiros Sistemas de Recuperação de Informações (SRI). Com os demais adventos da computação, o homem foi capaz de criar a Internet, uma rede eletrônica de computadores desenvolvida com a finalidade de facilitar o intercâmbio de informações. Depois de alguns anos de implantação a Internet mostrou-se capaz de oferecer mais opções do que as inicialmente previstas, além de atingir proporções mundiais, interligando pessoas de diversas culturas.

Dentro desses ambientes de manipulação e troca de informações automatizados, novas técnicas de suporte se tornaram necessárias.



## 2 Recuperação de informações: histórico e conceitos básicos

Durante algum tempo a atividade de produção e gerenciamento de literatura em geral foi chamada de *bibliografia* (*bibliography*) [BUC 97]. Bibliografia é um termo de origem grega que denomina a atividade de escrever livros. Por isso, até a metade do século XVIII, uma pessoa que escrevia ou copiava livros era conhecido por *bibliógrafo* (*bibliographer*) [KOC 74].

Porém, já por volta do século XVII, o número de publicações técnicas e científicas começou a crescer rapidamente. Esse crescimento deveu-se à *revolução científica*<sup>1</sup>, iniciada entre os séculos XVI e XVIII, e, principalmente, ao surgimento da imprensa. Isso fez com que novas técnicas de gerenciamento de literatura fossem necessárias.

Neste momento, gerenciamento literário passou a envolver técnicas eficientes de coleta, preservação, organização, representação (descrição), seleção (recuperação), reprodução (cópia) e disseminação dos documentos. Com isso, o termo *bibliografia* tornou-se impróprio, pois muitas destas técnicas de gerenciamento já não faziam parte do tradicional.

Durante esse período, *bibliografia* passou a denominar a atividade de escrever *sobre* livros, e *documentação* (*documentation*)<sup>2</sup> tornou-se o termo empregado para denotar o conjunto de técnicas necessárias ao gerenciamento de documentos [BUC 97] (principalmente na Europa).

Rapidamente *documentação* tornou-se um termo mais genérico que passou a envolver a bibliografia e serviços de informação eruditos, gerenciamento de registros e trabalho de arquivamento [BUC 97]. Com o tempo, após 1950, termos mais elaborados passaram a substituí-lo: “Ciência da Informação” (*Information Science*), “Armazenamento e Recuperação de Informações” (*Information Storage and Retrieval*) e “Gerenciamento de Informações” (*Information Management*).

Particularmente, para a área da computação, o termo “Armazenamento e Recuperação de Informações” ou simplesmente “Recuperação de Informações” (RI), é o mais utilizado. Academicamente, esse termo foi cunhado por Calvin Moores, um empresário que trabalhava nessa área, por volta de 1950. Moores foi um dos primeiros a utilizar o termo “recuperação de informações” em um artigo científico.

Moores define “recuperação de informações” como sendo uma atividade que envolve os aspectos de descrição de informação (indexação, padronização) e sua especificação para busca, além de qualquer técnica, sistema ou máquina empregada para realizar ou auxiliar essas tarefas<sup>3</sup>.

Na época Moores trabalhava somente com informações textuais. Durante algum tempo essa foi a abrangência da área: um pequeno mercado onde a maior parte das aplicações eram os bancos de dados bibliográficos. No entanto, com o advento das tecnologias de obtenção (digitalização) e armazenamento de dados, novos tipos de dados surgiram, fazendo com que a noção inicial de Moores fosse estendida e passasse a envolver outros *tipos de*

<sup>1</sup> Maiores informações sobre a Revolução Científica podem ser obtidas em HENRY, John. **A Revolução Científica e as Origens da Ciência Moderna**. Rio de Janeiro: Jorge Zahar, 1998.

<sup>2</sup> Atualmente, costuma-se denominar a tarefa de gerenciamento de coleções de *biblioteconomia* (*librarianship*), enquanto que a *bibliografia* fica encarregada do processo de descrição de documentos.

<sup>3</sup> “Information Retrieval embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, or machines that are employed to carry out the operation” [MOO 51].

*informação*<sup>4</sup> (imagens, por exemplo). Obviamente, recuperar informações já não significa mais buscar textos, mas sim, também, sons, imagens, vídeos e outros tipos de dados que surgem a cada momento.

Atualmente, *recuperar informações*, no contexto de SRI, significa recuperar *documentos* que (supostamente) contenham *informações relevantes* para o usuário. Essa definição, aparentemente simples, envolve todos os aspectos discutidos anteriormente. Essa afirmação é mais bem esclarecida através da leitura das subseções seguintes, que contextualizam e definem os termos *documento*, *informação* e *relevância*.

## 2.1 Documento

Inicialmente *documento* foi um termo utilizado para denotar um registro textual (um texto). Isso porque os primeiros documentos eram na verdade informações elaboradas (descritas) na forma de um texto escrito em linguagem natural. Porém, existem outros objetos que podem conter e transmitir informações.

Em um estudo realizado por Michael Buckland, da universidade da Califórnia, foram coletadas as seguintes definições de documento:

- a) “Qualquer expressão do pensamento humano” [BUC 97];
- b) “Qualquer base material capaz de estender nosso conhecimento, que seja disponível para estudo ou comparação, pode ser um documento” [SHU 35];
- c) “Qualquer fonte de informação, em formato material, capaz de ser utilizada para referência ou estudo ou como uma autoridade”, conforme o *International Institute for Intellectual Cooperation* (uma agência da Liga das Nações) [BUC 97];
- d) “Um documento é uma evidência que suporta um fato. [...] qualquer signo físico ou simbólico, preservado ou registrado, com a intuição de representar, reconstruir ou demonstrar um fenômeno físico ou conceitual é um documento” [BRI 51].

Deste modo, um documento não necessita estar limitado a textos impressos. O documento pode ser um uma pintura, uma figura, um gráfico, uma escultura, um filme ou outro objeto qualquer, desde que ele transmita informação. Até mesmo um animal pode ser considerado um documento. Um animal capturado e colocado em um zoológico e transformado em objeto de estudo passa a ser um documento [BRI 51]. O que dá o *status* de documento a um *objeto* ou *ser* é a qualidade de ser relacionado com outras evidências [BUC 97].

As técnicas apresentadas neste exame de qualificação foram elaboradas para a manipulação de documentos textuais. Isso porque durante muito tempo essa foi a abrangência da área de recuperação de informações (e mesmo as bibliotecas atuais contêm basicamente informações textuais). Cabe salientar que essas técnicas apresentadas podem ser utilizadas para outros tipos de documentos com pequenas adaptações, já que a metodologia por trás delas não condiciona o objeto que está sendo manipulado a um determinado formato ou tipo.

---

<sup>4</sup> Os termos *informação*, *dado*, *documento* e até mesmo *conhecimento* são muitas vezes utilizados de forma errônea (aplicados como sinônimos), até mesmo nos documentos da área.

## 2.2 Relevância e informação

Relevância é um conceito fundamental para as áreas relacionadas com informação. Um sistema de recuperação de informações só pode retornar informações relevantes para o usuário se ele for capaz de compreender e identificar o que *ser relevante* significa para o usuário.

Pode-se dizer que informação relevante é aquela informação capaz de satisfazer determinada necessidade de informação do usuário. Porém, antes de entrar nos detalhes técnicos dessa definição, torna-se necessário contextualizar o significado do termo *informação* para a área de recuperação de informações.

Dentro da área de RI, a informação possui um significado similar ao dado pela teoria da comunicação de Shannon [RIJ 79]. Dentro dessa teoria, informação é aquilo que um agente receptor (uma pessoa) recebe de um agente emissor em um processo de comunicação (em uma mensagem). Porém, a mensagem necessita ser primeiramente compreendida e, em seguida, identificada como contendo alguma coisa nova (para o receptor). Caso o receptor receba uma mensagem contendo algo que já conhece ou que não compreende, essa não transmite informação.

Um fator extremamente importante é que a informação faz com que o agente modifique seu estado de conhecimento<sup>5</sup> atual.

No escopo de RI, as informações estão contidas nos documentos (que pode ser considerado um conjunto de dados). Quando um determinado dado (que é ou representa uma entidade física do mundo) consegue, de alguma forma, fazer com que uma pessoa modifique seu estado de conhecimento atual é dito que esse dado contém (ou carrega) informação.

O fato de o dado carregar ou conter informação é um pouco subjetivo. Determinado dado pode transmitir informações para uma pessoa e não transmitir para outra. O dado pode ainda carregar diferentes informações em diferentes quantidades e em diferentes tempos. Há também a possibilidade de dados diferentes transmitirem uma mesma informação. Tudo isso vai depender da pessoa que recebe e interpreta os dados, e, para tal, de seu estado de conhecimento atual<sup>6</sup>.

O conceito de informação relevante vai além, pois a mensagem precisa ser primeiramente considerada informação para, depois, poder ser considerada relevante. Caso a mensagem enquadre-se nos princípios e qualificações de informação ela *pode* ser (não quer dizer que seja) considerada relevante. Isso porque informação relevante é aquela informação que o usuário necessita em determinado momento para a realização de alguma coisa, ou seja, ela deve estar no contexto que o usuário quer e no momento certo.

Por exemplo, caso uma pessoa interessada em obter informações sobre o ex-prefeito de São Paulo (por exemplo, Celso Pita) receba *informações* sobre shoppings virtuais em São Paulo (supondo que, neste caso, a pessoa já saiba sobre os shoppings virtuais de São Paulo), essas chamadas *informações* não podem ser assim consideradas, já que o indivíduo já as possui. Por outro lado, caso o usuário receba informações sobre *Paulo Maluf* (considerando

---

<sup>5</sup> A definição de conhecimento envolve uma série de questões filosóficas e fisiológicas que fogem do escopo desse texto. Dentro do contexto desse documento, conhecimento é a forma com que uma pessoa percebe o mundo [MIZ 96], isto é, a pessoa processa dados e informações que recebe e transforma-os em conhecimento. Esse conhecimento passa então a ser um potencial para certos tipos de ações [KOC 74]. O conhecimento de uma pessoa pode mudar, variar com o tempo. O conhecimento de uma pessoa em um determinado momento é denominado *estado de conhecimento* [MIZ 96].

<sup>6</sup> Decorrente disso, algumas pessoas definem informação como “a diferença entre dois estados de conhecimento” [MIZ 96, p241], e, neste caso, pode até ser medida.

que o usuário não tenha algum conhecimento sobre esse e sua relação com o estado de São Paulo), não há como negar o fato de que ele recebeu informação. Essa, porém, não é relevante pois não faz parte do contexto desejado (que seria o de política, mas durante o mandato de Celso Pita). A relevância está no fato da informação pertencer ao contexto do que usuário deseja naquele momento.

Em um levantamento recente [MIZ 97] foi identificado que existe uma extensa literatura sobre este tema em diversas áreas (filosofia, psicologia, lingüística, entre outras), e, mesmo assim, relevância ainda não tem um conceito bem compreendido. Mesmo limitando o escopo de sua definição às áreas de documentação, ciência da informação e recuperação de informações, uma quantidade grande de documentos relevantes à relevância é encontrada (no levantamento citado, 160 artigos foram analisados).

Stefano Mizzaro afirma que a informação relevante está diretamente relacionada com o usuário, com a sua necessidade de informação (o contexto que é expresso na sua consulta) e com o momento que isso ocorre. Segundo ele, a relevância pode ser vista como o relacionamento entre duas entidades, uma de cada um dos seguintes grupos [MIZ 97]:

- a) Grupo constituído das entidades *documento* (o objeto que o usuário vai obter depois de sua busca), sua *representação* ou *surrogate* (as palavras-chave, por exemplo) e *informação* (o que o usuário recebe quando lê um documento);
- b) Grupo contendo as entidades *problema* (que necessita de informação para ser resolvido), *necessidade de informação* (o que o usuário entende ou percebe do problema – representação do problema na mente do usuário), *solicitação* ou *request* (representação da necessidade de informação do usuário em uma *linguagem humana* – geralmente *linguagem natural*) e *consulta* ou *query* (a representação da necessidade de informação do usuário na *linguagem do sistema*).

Este relacionamento pode ser visto de diferentes aspectos [MIZ 97]. Há o aspecto *tópico*, que se refere ao assunto que o usuário deseja; *tarefa*, que se refere ao que o usuário vai fazer com os documentos retornados; e *contexto*, que inclui tudo aquilo que não pertence aos tópicos anteriores, mas, de alguma forma, afeta a busca e a avaliação dos resultados. Assim, um *surrogate* (ou um documento ou uma informação) é relevante a uma *consulta* (ou uma *solicitação*, necessidade de informação ou problema) com respeito a um ou mais aspectos.

Além disso, ainda há o fator tempo [MIZ 97], pois determinado documento pode não ser relevante a uma consulta em determinado momento e tornar a ser em outro.

Disso tudo, conclui-se que a relevância é um conceito estritamente relacionado com o usuário.

### 3 Paradigma da área de recuperação de informações

Nesta seção explicar-se-á de forma abstrata o processo de recuperação de informações e os agentes envolvidos nele. Há todo um contexto, um paradigma por assim dizer, onde as informações *são recuperadas*.

Nesse paradigma há um modelo genérico que demonstra o que cada elemento do paradigma faz, além de demonstrar a interação entre estes elementos. Os componentes básicos do modelo são: *usuário, sistema e documento*.

No contexto de recuperação de informação o usuário é qualquer pessoa que possua uma *necessidade de informação* e que se disponha a buscar informações a fim de satisfazer essa necessidade. Cabe salientar que o usuário pode não ter certeza sobre qual é a sua real necessidade de informação, mas nem por isso ele deixa de ser um usuário. Por outro lado, podem existir pessoas que não sabem que têm necessidade de alguma informação ou que realmente não tenham necessidade de informação. Em qualquer um destes casos a pessoa não irá buscar informações e, portanto, não pode ser considerada como sendo um usuário. A necessidade de informação pode ser definida como a falta de conhecimento que o usuário tem para realizar determinada tarefa [MIZ 96]. Para solucionar determinado problema, o usuário necessita de determinado conhecimento. Se ele não possui esse conhecimento ele passa a ter, então, uma necessidade de informação. Essa necessidade de informação deve ser satisfeita de alguma forma. Geralmente isso é obtido através de um *Sistema de Recuperação de Informações*, o próximo elemento do modelo.

O Sistema de Recuperação de Informações (SRI) é o “miolo” do modelo. Ele é a “interface” entre o usuário e os documentos de uma coleção. Em um sistema automatizado, o SRI é encarregado de receber a consulta do usuário e compara-la com os documentos ou descrições de documentos presentes em seu banco de dados e retornar uma lista de documentos relevantes. Em muitos casos o sistema também fica encarregado de analisar os documentos que chegam (são adicionados) ao sistema, padronizando-os e indexando-os (ou seja, ele é encarregado de criar uma descrição para os documentos e coloca-la em seu banco de dados). Dependendo do SRI, o documento (em si) também pode estar armazenado no sistema. Existem diversos tipos de SRI, cada um com características específicas e capazes de manipular um tipo de documento específico. Os diferentes tipos de SRI existentes, assim como uma definição mais formal, são apresentados na seção 4. É importante salientar que o paradigma aqui em questão independe do tipo de SRI. Porém, cabe avisar que o tipo de SRI abordado nesse trabalho é o textual automatizado (ou seja, com o auxílio de um computador). O objetivo maior de um SRI é fazer com que o usuário encontre a informação de que necessita rapidamente, de modo que esse usuário não precise analisar ele próprio as informações existentes na base de informações.

Um documento, pela definição proposta anteriormente (seção 2.1), é um objeto do mundo real ou uma abstração de um objeto do mundo real que contém um grande potencial para transmitir informação para alguém. Por esse motivo, costuma-se utilizar os termos informação e documento como sinônimos. Mesmo na área de RI isso acontece. O próprio nome *Sistema de Recuperação de informações* pode não ser condizente com a realidade, já que o quê um sistema recupera é na verdade um documento. Esse documento pode ou não conter informações para o usuário. Isso é um pouco amenizado pelo fato do SRI ser construído com a intenção de recuperar somente os documentos contendo informações relevantes para o usuário. Porém, nem sempre isso acontece e mesmo assim o termo é utilizado na área.



FIGURA 3-1 – COMPONENTES BÁSICOS DO MODELO

Os elementos da Figura 3-1 representam os objetos do paradigma e sua interação. Nesta figura existem três pontos-chave que devem ser trabalhados com atenção.

O primeiro é o processo de abstração de informações, determinado pela modelagem do SRI. O segundo é decorrente da abstração que o usuário faz ao descrever sua necessidade de informação através de algum formalismo (linguagem de consulta do SRI). O último é o processo de *casamento* (*matching*) que o sistema faz entre a consulta do usuário e as informações do sistema, a fim de determinar quais informações são relevantes.

É nestes pontos que os problemas ocorrem e, conseqüentemente, onde a recuperação pode falhar. Os estudos na área de Recuperação de Informações (RI) buscam resolver estas falhas, desenvolvendo técnicas específicas para cada um destes pontos-chave.

### 3.1 Abstração de informações

É através das características de um documento (de um objeto) que o SRI é capaz de localiza-lo e identifica-lo como relevante para o usuário. Portanto, uma das primeiras interações se dá entre o documento e o SRI. O SRI de alguma forma deve poder identificar as características de um objeto e descreve-lo (criar uma representação dele) através delas.

Essa descrição do objeto nada mais é do que uma modelagem (uma abstração) do documento através de algum formalismo. Porém, como em todo processo de modelagem, deve-se agir com extrema cautela, pois se o formalismo adotado não representar corretamente o documento ou se uma de suas características não for considerada, o usuário pode não conseguir localizar e recuperar esse documento.

Cada objeto (tipo de documento) possui atributos (características) que são mais apropriados para a sua descrição e caracterização. A escolha dos atributos mais relevantes para o tipo de documento que se está trabalhando está diretamente relacionada com a capacidade delas poderem caracterizá-lo, ou seja, distingui-lo dos outros objetos do mesmo tipo.

Depois de determinar os atributos mais relevantes para a descrição dos documentos<sup>7</sup>, torna-se necessário analisar cada um dos documentos, selecionar essas características e armazená-las<sup>8</sup>.

<sup>7</sup> No caso de imagens, por exemplo, as características ou atributos relevantes poderiam ser as cores, formas e texturas.

<sup>8</sup> O formalismo escolhido para armazenar as características também é extremamente dependente do tipo de documento em questão. Uma descrição textual de uma figura, por exemplo, pode não ser capaz de representá-la corretamente.

Logo, escolher um formalismo que consiga armazenar o conteúdo da informação como um todo é tarefa complexa e difícil, mas de extrema importância. Vários problemas podem surgir em decorrência de uma modelagem incorreta da informação. Porém, depois de definidas as características da informação, o processo de modelagem pode ser realizado manualmente ou automaticamente.

A Figura 3-2 demonstra o processo de abstração, onde as informações são analisadas manualmente ou automaticamente. Após a análise as características são armazenadas, conforme o modelo, em uma representação interna.

A indexação é a técnica de abstração utilizada na área de RI. Ela é responsável por analisar os documentos (objetos) do mundo real e representa-los dentro do sistema (ver seção 6).

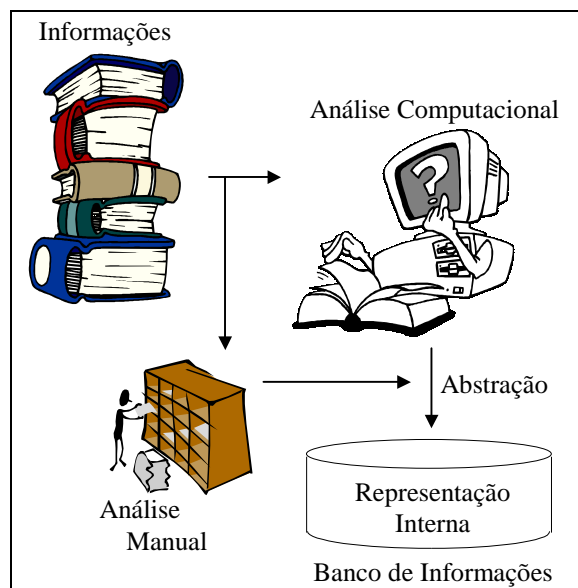


FIGURA 3-2 – O PROCESSO DE ABSTRAÇÃO

### 3.2 Descrição da necessidade do usuário

As interfaces homem-máquina existentes ainda não permitem com que o SRI (computacional) obtenha informações diretamente da mente do usuário. Logo, o usuário precisa descrever sua necessidade de informação utilizando uma linguagem formal de consulta específica do SRI.

É através das características que o usuário fornece na consulta que o SRI vai ser capaz de determinar quais informações são mais relevantes para este usuário. Porém, quando o usuário descreve sua necessidade de informação através de uma expressão na linguagem do sistema, uma série de problemas pode surgir. Por isso, o usuário deve expressar corretamente sua necessidade, detalhando-a e fornecendo a maior quantidade possível de características. Do contrário o resultado não será satisfatório.

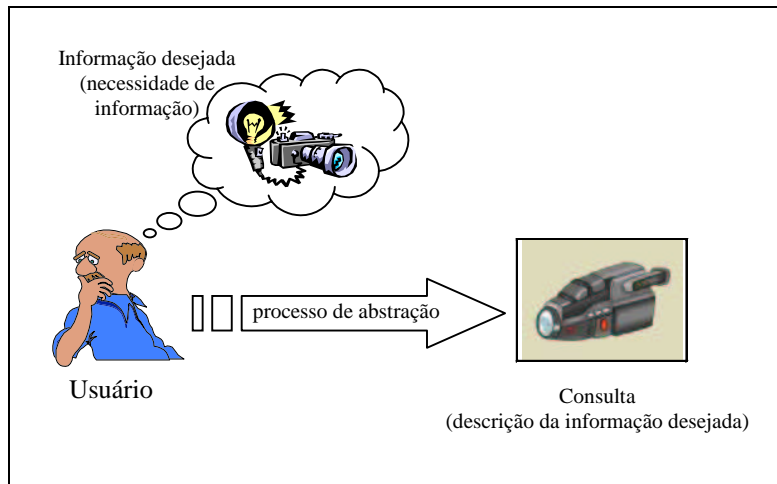


FIGURA 3-3 – PROBLEMA DO PROCESSO DE DESCRIÇÃO DE UMA CONSULTA

O primeiro dos problemas está diretamente relacionado com o conhecimento que o usuário possui no momento. O usuário pode não saber como descrever sua necessidade de informação, pois não sabe realmente qual ela é. Ele também pode não ser capaz de descrever sua necessidade de informação por não entender direito como funciona o formalismo de consulta oferecido pelo SRI ou por não saber como descrever a informação de que necessita. Nestes casos, a descrição pode não corresponder exatamente com o que o usuário quer ou pensa querer (ver Figura 3-3).

Pode ainda ocorrer o fato do formalismo do SRI não permitir com que o usuário descreva ou expresse corretamente sua necessidade. Isso pode ocorrer, por exemplo, em sistemas cujo documento seja do tipo imagem e a forma de consulta não permita com que o usuário desenhe imagens, mas sim, descreva-as através de uma forma textual (imagens são difíceis de serem descritas textualmente).

Neste caso, o problema provavelmente está relacionado com o modelo utilizado pelo sistema para modelar informações. Isso significa que o modelo de abstração utilizado não é bom para o tipo de informação com que o sistema trabalha. Quando isso ocorre, mesmo os usuários mais experientes costumam ter problemas [GUP 97a].

Porém, mesmo que o usuário consiga descrever corretamente sua necessidade de informação pode não ser recuperado exatamente o que ele espera. Isso porque cada pessoa descreve um mesmo objeto de diversas formas. Este problema, onde vários usuários descrevem o mesmo objeto de formas diferentes é conhecido por *Problema do Vocabulário (Vocabulary Problem)* [CHE 94; CHE 96].

Isso dificulta muito a localização de informações, porque o vocabulário (a forma de descrição) utilizado por uma pessoa que utiliza um SRI para consultar informações pode ser diferente da forma utilizada pelas pessoas que criaram os documentos nele contidos. A Figura 3-4 (adaptada de [KOW 97, p8]) ilustra essa diferença entre os *domínios* de vocabulário que dificulta a localização de informações.

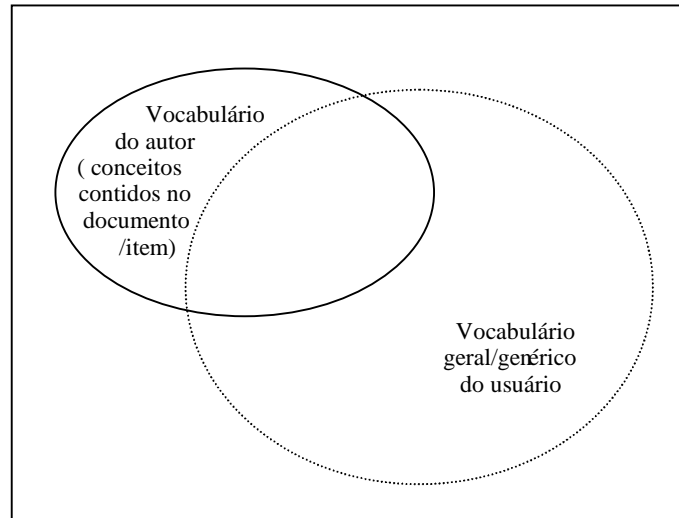


FIGURA 3-4 – DIFERENÇA ENTRE DOMÍNIOS DE VOCABULÁRIO

Por todos esses motivos os sistemas devem oferecer recursos capazes de auxiliar o usuário no processo de descrição da sua necessidade de informação. Eles devem informar ao usuário como funciona seu formalismo de consulta e qual é o modelo utilizado pelo sistema para descrever os atributos de um documento (nem todos os usuários são obrigados a conhecer o sistema). Devem, ainda, permitir com que o usuário familiarize-se com a necessidade de informação de que necessita (caso não a compreenda bem) ou que ele possa muda-la caso identifique que ela não está correta (muito ou pouco abrangente, ou fora de contexto). Esse problema será novamente abordado na seção 9, onde podem ser encontradas algumas soluções para ele.

### 3.3 O processo de “casamento” ou “matching”

A única forma que o sistema tem de saber se recuperou informação relevante para a necessidade do usuário é comparando-a com a expressão de consulta. Essa comparação pode ser problemática, como já comentado anteriormente, já que os documentos podem ser relevantes à consulta do usuário mas não serem relevantes para o usuário (pois ele pode ter expresso incorretamente sua necessidade de informação).

Esse processo de identificação de informações relevantes é denominado *processo de casamento (matching)*. Ele é o último mas não menos importante ponto-chave da recuperação de informações. Este processo é responsável pelo mecanismo que faz a identificação de quais informações são relevantes para a consulta do usuário, identificando a similaridade entre as informações armazenadas no sistema e a necessidade de informação descrita pelo usuário na expressão de consulta.

Por mais bem realizado que seja esse processo sempre podem haver distorções na identificação de relevância de um documento. Isso porque tanto o usuário quanto o sistema efetuam alguma forma de abstração, o que pode acarretar em falhas de descrição. O usuário o faz quando descreve a necessidade de informação e o sistema quando modela o documento em sua representação interna.

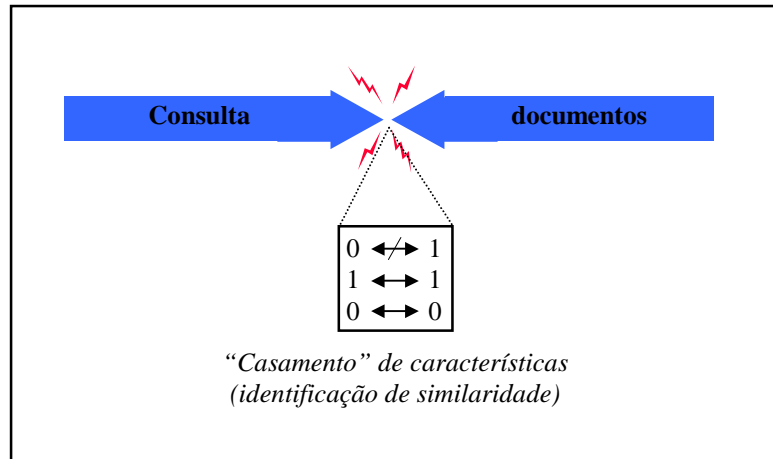


FIGURA 3-5 – O PROCESSO DE "CASAMENTO" DE CARACTERÍSTICAS

Muitos métodos de casamento utilizam um processo de comparação direta entre as características da consulta e as características das informações na base de dados. Aqui novamente o problema do vocabulário incide, já que as pessoas que descrevem a informação podem não utilizar as mesmas características que o usuário. Existem várias técnicas que buscam minimizar esse problema. Essas técnicas variam de acordo com o modelo de recuperação utilizado pelo sistema (os diferentes modelos são discutidos na seção 5.1).

## 4 Sistemas de recuperação de informações

Conforme Kowalski, “Um sistema de recuperação de informações – SRI (Information Retrieval System) é capaz de armazenar, recuperar e gerenciar informação. Informação, neste contexto, pode ser composta de textos (incluindo o formato numérico e datas), imagens, áudio, vídeo e outros objetos multimídia. Apesar da forma de um objeto em um SRI ser diversa, o aspecto textual tem sido o único tipo de dados [até o momento] que permite um processamento funcional completo.” [KOW 97].

Essa definição pode ser refinada com os elementos já definidos nas seções anteriores, onde um SRI poderia ser definido como um sistema capaz de catalogar e recuperar documentos (os diferentes *tipos de informação* existentes) relevantes à consulta do usuário<sup>9</sup>.

Os SRI são classificados de acordo com o tipo de documento que eles manipulam. Para cada tipo de informação (documento) há um tipo de SRI correspondente. Há informações textuais, visuais e multimídia. Há ainda sistemas que recuperam dados [LAN 68], porém, esses não fazem mais parte do contexto de RI, pois se enquadram na área de Banco de Dados (BD).

A diferença entre a área de RI e a área de BD está no paradigma adotado pelos sistemas construídos nessas áreas.

Em um *Sistema de Gerenciamento de Bancos de Dados – SDBD (Database Management System – DBMS)* cada consulta possui somente uma resposta correta, ou seja, ou existe um conjunto de elementos que corresponde exatamente ao que o usuário está requisitando, ou não.

Já em um SRI, devido à incerteza associada ao tipo de documento (diferenças de vocabulário), pode não haver uma resposta exata à consulta do usuário. Esses sistemas, segundo Fraques, são difusos e probabilísticos [FRA 92a], não trabalhando com uma teoria de conjuntos exata como acontece em um SDBD. Um SRI consegue, na maioria dos casos, recuperar somente uma aproximação, encontrando diversas respostas possíveis e elaborando um *ranking* onde os documentos são listados de acordo com sua estimativa de relevância.

Isso ocorre porque não há uma classificação binária onde se poderia dizer que determinado documento é relevante ou não à consulta do usuário (a função de relevância é contínua). Esse comportamento, considerado *difuso (fuzzy)* por muitos, já que não há um padrão ou controle de linguagem, é uma consequência da automação do processo de entendimento da *linguagem natural* [HEA 99].

Portanto, um SRI manipula informações (genéricas), enquanto que um SDBD é otimizado para manipular dados estruturados.

O objetivo de um SRI é de minimizar o custo (overhead) de um usuário localizar uma informação necessária (pois o usuário poderia localizar uma informação analisando todos os documentos de uma coleção). O custo pode ser entendido como o tempo gasto pelo usuário para completar a tarefa de recuperação de uma informação relevante [KOW 97].

O sucesso disso é bem subjetivo e dependente do usuário, e há uma grande possibilidade do usuário não encontrar documentos que satisfaçam totalmente sua necessidade de informação. O SRI pode não encontrar ou não conter todos os itens necessários para isso. O usuário deve interagir com o sistema, refinando sua consulta através da adição ou eliminação de características a fim de localizar novos documentos relevantes.

---

<sup>9</sup> Conforme já discutido anteriormente, um SRI não modifica o conhecimento do usuário, mas sim, informa sobre a existência ou não de documentos relacionados (relevantes) a sua consulta [LAN 68].

## 4.1 Tipos de sistemas de recuperação de informações

Há diversas classificações de SRI que acabam gerando taxonomias diferentes de sistemas [CHA 97; FRA 92a]. Algumas são bem simples, fazendo somente distinção entre sistemas manuais e automáticos, enquanto que outras levam em conta o tipo de aplicação dada ao sistema, classificando-os entre *sistemas de recuperação de documentos*, *sistemas de recuperação de referências*, *sistemas de recuperação de dados* e *sistemas de recuperação de fatos* (ou “*query answering systems*”) [LAN 68].

A classificação adotada neste documento foi elaborada buscando envolver as diferentes classificações encontradas na literatura, porém, levando em conta o contexto paradigmático definido anteriormente. Isso significa que os sistemas de recuperação de dados e de fatos, por exemplo, não devem enquadrar-se na classificação adotada.

Nessa classificação também não se torna relevante o fato de um sistema poder ser realizado manualmente ou automaticamente. Porém, pelo fato do autor estar interessado na área computacional, além de seguir uma tendência tecnológica, tende-se a abordar somente os sistemas dentro do contexto automatizado (o que não significa que a mesma classificação não possa ser utilizada dentro de um contexto manual).

Essa classificação foi elaborada visando basicamente os tipos de informações existentes. Desse modo, há um tipo de SRI para cada tipo de informação. A classificação elaborada envolve os tipos de documentos para os quais já existem sistemas. Tipos de documentos não manipulados atualmente na área de RI (como esculturas e outros objetos) não entram nessa classificação.

### 4.1.1 Sistemas de recuperação de informação bibliográfica

Os sistemas de recuperação de informação bibliográfica foram os primeiros a surgir. Atualmente eles também são conhecidos por *catálogos*. Neste caso, faz-se uma descrição do objeto a ser *catalogado* no sistema. Para tanto, escolhe-se os atributos mais descritivos do mesmo (por exemplo: título, autor, data, resumo-descrição, palavras-chave) e estes são adicionados ao sistema. Quando o usuário consulta o sistema, apenas a referência bibliográfica (os atributos) ao(s) objeto(s) relevante(s) é retornada. Em alguns casos o sistema pode indicar onde o documento pode ser encontrado. O sistema ALEPH (<http://143.54.1.5:4505/ALEPH>), que controla a biblioteca da Universidade Federal do Rio Grande do Sul, e o DBLP (<http://www.acm.org/sigmod/dblp/db/>) podem ser considerados sistemas desse tipo. Esses sistemas podem trabalhar com qualquer tipo de documento, já que não necessariamente os armazenam, mas sim, mantêm sua referência (em um índice).

### 4.1.2 Sistemas de recuperação de informação textual

Os sistemas de recuperação textual são sistemas que manipulam basicamente informações do tipo texto (ASCII). Apesar disso, com a utilização de *filtros*, outros formatos que contenham textos, figuras, tabelas e imagens, mas que possuam um aspecto de documento textual (tais como o PDF, o PS e o DOC), também podem ser manipulados.

O sistema de busca Altavista™, por exemplo, pode ser considerado um sistema de recuperação de informações pois utiliza tecnologias da área. Costuma-se, porém, denominar os sistemas similares a ele de *motores de busca* (*search engines*).

A diferença entre os SRI textuais e os do tipo anterior (bibliográfico) é a de que os sistemas de recuperação textual costumam trabalhar com o texto em si e todo o seu conteúdo, enquanto que os outros trabalham apenas com a sua descrição. Nesse caso (dos textuais), não

há campos ou estruturas que possam ser consultados (tais como autor e título) e a busca deve ser realizada no documento inteiro.

### 4.1.3 Sistemas de recuperação de informação visual

Esses sistemas são desenvolvidos para trabalhar com um tipo especial de documento – as imagens. Portanto, o termo visual não diz respeito à forma como o sistema é construído (sua interface), mas sim quanto às informações que ele manipula. É claro que, por trabalhar com informações visuais, esses sistemas geralmente possuem interfaces visualmente mais elaboradas.

Os primeiros sistemas de recuperação de informações visuais utilizavam um modelo textual para descrever essas informações. Porém, pelo fato de diferentes pessoas compreenderem uma figura de maneiras diferentes (dependendo do contexto) [CHA 97], a descrição de uma imagem pode variar de uma pessoa para outra. Além, disso, todos os problemas inerentes à diferença de vocabulário já comentados surgem. Portanto, deve-se utilizar um modelo visual de representação e descrição da informação.

Em um sistema que manipule imagens o ideal é que o usuário possa descrever sua consulta utilizando imagens. Assim, o modelo perderia menos em termos de abstração e seria capaz de recuperar informações muito mais relevantes (isso exige um sistema que utilize técnicas específicas, cuja maioria pode ainda não existir ou não ser funcional).

Exemplos desse tipo de sistemas são o *Visual SEEK*, o *VideoQ*, *Virage* e *QBIC* [CHA 97]. Mais informações sobre esse tipo de sistema podem ser obtidas nos artigos de Gupta [GUP 97a] e de Yeo [YEO 97].

### 4.1.4 Bibliotecas Digitais

*Bibliotecas digitais* nada mais são do que a *virtualização* das bibliotecas tradicionais [FOX 95]. Para a computação, uma biblioteca digital, tecnicamente falando, nada mais é do que um sistema de informação distribuído cujas informações estão interconectadas [FOX 95].

A biblioteca digital não possui uma dimensão física. Ela pode utilizar toda a infraestrutura de comunicação existente (Internet, por exemplo) para que funcione.

Decorrente disso, uma biblioteca digital não necessita necessariamente conter o conteúdo das informações, mas sim, prover acesso até elas [KOW 97].

Uma biblioteca digital real teria como benefícios o acesso facilitado à informação por qualquer pessoa em qualquer parte do mundo, a eliminação do papel e um custo baixo de manutenção e armazenamento. Porém, ainda existem muitos problemas a serem resolvidos: nem todas as pessoas possuem acesso a informação digital, nem todas as informações estão e podem estar em formato digital, as diferenças entre línguas e culturas devem ser eliminadas, enfim, há uma série de desafios que exigem o aperfeiçoamento e o desenvolvimento de novas tecnologias, tais como recuperação inteligente, protocolos eficientes de comunicação cliente-servidor, tecnologias de aquisição de dados e novos paradigmas de interação.

Bibliotecas digitais prontas ainda não existem, mas existem vários projetos em andamento: Stanford (<http://www.diglib.stanford.edu/diglib>), Berkeley (<http://http.cs.berkeley.edu/~wilensky>), Illinois (<http://www.grainger.uiuc.edu/dli>), Michigan (<http://www.sils.umich.edu/umdl/homepage.html>), a biblioteca digital da *Association for Computing Machinery* (<http://www.acm.org/dl>) e o projeto Alexandria (<http://Alexandria.sdc.ucsb.Edu>).



## 5 Sistemas de recuperação de informações textuais

Um sistema de recuperação de informações textuais é um sistema desenvolvido para indexar e recuperar documentos do tipo textual, ou seja, documentos cujas informações estão descritas através da linguagem natural. Os dados não estão dispostos de forma tabular como ocorre em um SGBD.

Um documento textual é composto de palavras, logo, a palavra é a menor unidade de análise e de acesso nesse tipo documento. As palavras (ou termos) são os atributos ou características de um texto. São elas que conseguem distinguir um documento de outro.

Decorrente disso, em um SRI textual as consultas do usuário são descritas através de palavras. O usuário deve escolher os termos mais adequados para caracterizar sua necessidade de informação. Os documentos relevantes a essa consulta são então selecionados de acordo com a quantidade de palavras semelhantes que eles possuem com a consulta [SAL 83].

Quem faz essa análise de relevância é uma função denominada função de *similaridade*. Essa função busca identificar uma relação entre os termos da consulta e os termos dos documentos. Teoricamente pode ser feita uma comparação direta entre esses termos, mas, devido a problemas de sinonímia<sup>10</sup>, polissemia<sup>11</sup> e outros relacionados ao vocabulário [CHE 94], essa simples comparação nem sempre oferece resultados satisfatórios e os documentos recuperados acabam sendo de assuntos variados.

Buscando solucionar isso, vários métodos de cálculo de similaridade foram criados. Além disso, diversos modelos conceituais de recuperação foram elaborados. Esses modelos são apresentados a seguir.

### 5.1 Modelos de Recuperação de Informações

Os modelos de recuperação são modelos conceituais ou abordagens genéricas para a recuperação de informações [BAE 92b].

Apesar de terem sido desenvolvidos dentro do escopo de documentos textuais, os modelos conceituais de recuperação de informações podem ser utilizados em qualquer tipo de documento. Para tanto, basta modificar o tipo de atributo, que em documentos textuais são as palavras, pelo tipo de atributo adequado ao tipo de documento em questão.

Na literatura da área é possível definir uma taxonomia de modelos que inclui o *booleano*, o *espaço-vetorial*, o *probabilístico*, o *difuso (fuzzy)*, o da *busca direta*, o de *aglomerados (clusters)*, o *lógico* e, mais atualmente, o *contextual* ou *conceitual*.

#### 5.1.1 Modelo booleano

O modelo conceitual booleano [RIJ 79, cap5; SAL 83] considera os documentos como sendo conjuntos de palavras e possui esse nome justamente por manipular e descrever esses conjuntos através de conectivos de boole (*and*, *or* e *not*). As expressões booleanas são capazes de unir conjuntos, descrever intersecções e retirar partes de um conjunto.

<sup>10</sup> O problema de sinonímia (*synonymy*) ocorre porque o significado de uma palavra pode existir em uma variedade de formas [DOY 75, p294].

<sup>11</sup> Polissemia ou homografia (*homography*) é o nome dado a característica de linguagem que um termo (a mesma forma ortográfica) possui de poder conter vários significados [DOY 75, p294].

Em uma busca, por exemplo, o usuário indica quais são as palavras (elementos) que o documento (conjunto) resultante deve ter para que seja retornado. Assim, os documentos que possuem interseção com a consulta (mesmas palavras) são retornados. Os documentos podem ser ordenados pelo grau de interseção, onde o mais alto é aquele que contém todas as palavras especificadas na consulta do usuário e o mais baixo o que contém somente uma.

Há uma série de operadores que o usuário pode utilizar para especificar sua consulta. Os operadores mais comuns são o “*and*” (união), “*or*” (interseção) e o “*not*” (exclusão ou negação). Neste caso, o modelo pode ser compreendido de outra forma, onde o conjunto de documentos relevantes à consulta é o conjunto de documentos que satisfaz as restrições especificadas na consulta.

Como exemplo, caso o usuário necessite de informações sobre onde realizar compras na Internet, ele poderia utilizar uma consulta do tipo “virtual *and* store”. A fim de diminuir a abrangência para livros, ele poderia refinar a consulta para “virtual *and* book *and* store”. É possível aumentar a abrangência do resultado adicionando os tipos de itens desejados: “virtual *and* (book *or* cd) *and* store”. Poderia-se ainda excluir alguns locais não desejados: “virtual *and* (book *or* cd) *and* store *and not* (Amazon *or* Submarino)”. Nesses casos os parênteses são utilizados para estabelecer a precedência de operadores e, em alguns casos, o conjunto final pode variar de acordo com a ordem utilizada (como em matemática).

Em teoria o modelo booleano é um dos que apresenta melhores resultados, pois permite que o usuário especifique consultas complexas, detalhadas e bem definidas. Por outro lado, todo esse *poder* pode acabar atrapalhando o usuário, pois ele deve conhecer bem o que necessita para poder especificar a consulta. Além disso, existem estudos que mostram que muitos usuários sentem dificuldades em expressar suas necessidades de informação através de operadores booleanos.

Um outro problema do modelo booleano é o dele não ser capaz de identificar a importância de um termo em um documento, ou seja, se um termo está presente em um documento ele é considerado muito importante (não há valores intermediários). Porém, o simples fato de uma palavra aparecer em um documento não significa que ela seja significativamente importante para ele. Existem palavras que são utilizadas para o encadeamento de idéias e frases (como as conjunções) que não são relevantes para a busca de informações. Há outras palavras, ao contrário, que facilitam a compreensão das idéias básicas do texto e que são extremamente importantes (palavras em títulos e resumos, por exemplo). O modelo booleano não leva em conta nenhum destes casos, considerando relevantes os documentos que possuem as mesmas palavras (mesmo que estas tenham importâncias diferentes em cada documento).

A fim de minimizar esse problema foi desenvolvido o modelo *booleano estendido* [FOE 92], que leva em conta a importância das palavras nos documentos (a identificação de importância é feita com a utilização de uma das técnicas apresentadas na seção 7.5). Nesse modelo há ainda a possibilidade do usuário especificar o quanto cada termo da consulta é importante para ele, recuperando assim documentos mais relacionados com os termos considerados mais relevantes por ele.

Essa solução, porém, oferece mais um problema, já que os usuários, além de tudo, têm que indicar o grau de importância de cada termo de consulta, o que pode não ser uma tarefa fácil para aqueles que não conhecem bem sua necessidade de informação.

### 5.1.2 Modelo espaço-vetorial

O modelo espaço-vetorial (*vector-space model*) foi desenvolvido por Gerard Salton, que durante anos contribuiu com estudos relevantes para a área. Salton desenvolveu esse modelo para poder ser utilizado em um SRI chamado SMART que ele criou enquanto trabalhava na universidade de Cornell.

Nesse modelo cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica o grau de importância (denominado *peso*) desse no documento. Portanto, cada documento possui um vetor associado que é constituído por pares de elementos na forma  $\{(palavra\_1, peso\_1), (palavra\_2, peso\_2) \dots (palavra\_n, peso\_n)\}$ .

Nesse vetor são representadas todas as palavras da coleção e não somente aquelas presentes no documento. Os termos que o documento não contém recebem grau de importância zero e os outros são calculados através de uma fórmula de identificação de importância. Isso faz com que os pesos próximos de um (1) indiquem termos extremamente importantes e pesos próximos de zero (0) caracterizem termos completamente irrelevantes (em alguns casos a faixa pode variar entre -1 e 1).

O peso de um termo em um documento pode ser calculado de diversas formas [SAL 83] (a seção 9.2 apresenta algumas destas formas ou métodos). Esses métodos de cálculo de peso geralmente se baseiam no número de ocorrências do termo no documento (frequência).

Cada elemento do vetor é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados em um espaço euclidiano de  $n$  dimensões (onde  $n$  é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso.

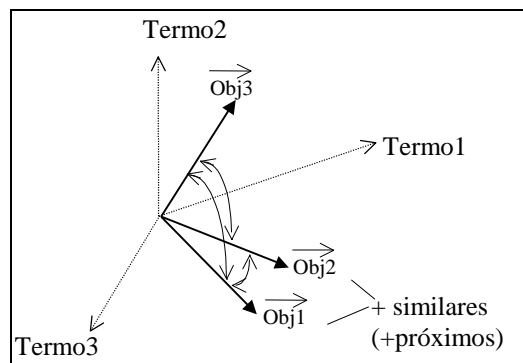


FIGURA 5-1 – O MODELO ESPAÇO-VETORIAL

As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos que possuem os mesmos termos acabam sendo colocados em uma mesma região do espaço e, em teoria, tratam de um assunto similar (essa característica é que dá o nome de espaço-vetorial ao modelo).

A consulta do usuário também é representada por um vetor. Dessa forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado. Os documentos mais similares (mais próximos no espaço) à consulta são considerados relevantes para o usuário e retornados como resposta para ela.

Uma das formas de se calcular a proximidade entre os vetores é testar o ângulo entre estes vetores. É exatamente isso o que Salton faz no seu modelo original; ele utiliza uma

função que ele batizou de *cosine vector similarity* [SAL 87b] que calcula o produto dos vetores de documentos através da seguinte fórmula:

$$\text{similaridade (Q,D)} = \frac{\sum_{k=1}^n w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^n (w_{qk})^2 \cdot \sum_{k=1}^n (w_{dk})^2}}$$

Nessa fórmula, Q representa o vetor de termos da consulta; D é o vetor de termos do documento;  $w_{qk}$  são os pesos dos termos da consulta e  $w_{dk}$  são os pesos dos termos do documento.

Depois dos graus de similaridade terem sido calculados, é possível montar uma lista (um ranking) de todos os documentos e seus respectivos graus de relevância à consulta. Essa lista é ordenada por ordem de similaridade mais alta à mais baixa para que o usuário identifique os documentos relevantes.

### 5.1.3 Modelo probabilístico

O modelo probabilístico possui essa denominação justamente por trabalhar com conceitos provenientes da área de probabilidade e estatística. Nesse modelo, busca-se saber a probabilidade de um documento  $x$  ser relevante a uma consulta  $y$  caso os termos especificados por ela apareçam nele.

Tal informação pode ser obtida assumindo-se que a distribuição de termos na coleção seja capaz de informar a relevância provável para um documento qualquer da coleção [RIJ 79].

Existem diversas formas de se obter estatisticamente essa informação. Porém, a base matemática comumente adotada para esse modelo é o método bayesiano (teorema de Bayes). Devido a isso, muitas vezes, esse modelo é chamado de modelo bayesiano [KEI 97].

No modelo probabilístico a função de similaridade pode aproveitar-se das informações estatísticas de distribuição dos termos contidos no índice. Com isso, determinados parâmetros podem ser ajustados de acordo com a coleção em questão, obtendo assim resultados mais relevantes.

Cada documento é modelado novamente como um vetor de características na forma  $x = (x_1, x_2, \dots, x_n)$ , onde cada  $x$  representa um termo e contém informação sobre sua ausência ou presença no documento (forma binária).

A identificação dos documentos relevantes a uma consulta é determinada pelo cálculo de probabilidade de cada um dos documentos da coleção ser relevante à consulta dada. Os documentos são então listados de acordo com o seu grau provável de relevância (na forma de um ranking).

A probabilidade de relevância de um documento é calculada através da identificação de sua relevância ou não à determinada consulta. Para cada termo da consulta seu grau de relevância é identificado no documento. A informação de relevância de um termo é calculada estatisticamente com bases na frequência desse termo nos documentos da coleção.

$$P(w_i/x) = \frac{P(x/w_i)P(w_i)}{P(x)} \quad i = 1,2$$

Nessa fórmula,  $P(w_i)$  é a probabilidade *a priori* de relevância (quando  $i=1$ ) ou de irrelevância (quando  $i=2$ );  $P(x/w_i)$  é a aparência de relevância ou irrelevância, dado um termo  $x$ . Já o  $P(x)$  é calculado pela fórmula seguinte.

$$P(x) = \sum_{i=1}^2 P(x/w_i)P(w_i)$$

Essa fórmula calcula a probabilidade de observação aleatória de  $x$  que pode ser tanto relevante quanto irrelevante. Verifica-se que essas fórmulas levam em conta ambos os fatos: ser relevante e não ser relevante. A teoria de Bayes auxilia a identificar para cada termo da consulta o grau de relevância e de irrelevância do documento, selecionando o mais adequado (o que produz menor erro) para o somatório final, já que o grau final de probabilidade de relevância é dado pelo somatório dos graus de relevância de cada termo.

#### 5.1.4 Modelo difuso

No modelo difuso (fuzzy) [CRO 94] os documentos também são representados por vetores de palavras com seus respectivos graus de relevância. A diferença está no conceito relacionado à relevância.

Na teoria de conjuntos difusos (fuzzy), introduzida por Laft Zadeh [ZAD 65], todas as características de determinado universo estão presentes em todos os conjuntos. A diferença é que a *presença* pode ser medida e pode não ser exata, ou seja, pode haver incerteza. Logo, não há conjunto vazio, mas sim um conjunto cujos elementos possuem uma relevância (importância) muito baixa (próxima de zero).

Tradicionalmente um objeto pertence ou não a um conjunto, isto é, não há possibilidade do objeto pertencer parcialmente ao conjunto. Esta afirmação é muito rígida, e, na prática, as pessoas utilizam raciocínios onde o objeto pode pertencer parcialmente ao conjunto. É o que ocorre quando são utilizadas as expressões do tipo “mais ou menos”, “muito”, “pouco” e “talvez”.

A teoria difusa permite trabalhar com esses valores intermediários que indicam o quanto determinado objeto pertence e o quanto não pertence ao conjunto, pois ela foi construída com a finalidade de tratar incertezas e imprecisões [ZAD 65; ZAD 73].

É como que se a teoria fuzzy trabalhasse com valores entre zero (0) e um (1). Quanto maior o grau de pertinência do objeto ao conjunto, mais próximo de 1 é o valor correspondente, e quanto menor o grau de pertinência, mais próximo de 0 ele é. Decorrente disso é possível medir a distância entre o grau de pertinência e qualquer um dos extremos (0 ou 1), e dizer *quanto* um objeto pertence a um conjunto. Por exemplo, se  $x$  pertence 0.4 graus ao conjunto  $C$ , este mesmo  $x$  está 0.6 graus de distância do grau máximo 1 e, portanto, *não pertence* 0.6 graus ao conjunto  $C$ .

Já que a teoria difusa permite manipular com facilidade incertezas e imprecisões [CRO 94] ela pode ser utilizada para modelar documentos textuais cuja linguagem e o processo de recuperação são imprecisos.

Por ter como base uma teoria que manipula conjuntos, o modelo difuso é considerado por alguns uma extensão do modelo booleano [CRO 94; FOE 92]. Portanto, os documentos são representados por conjuntos (vetores) e operadores lógicos difusos podem ser utilizados para especificar as consultas.

Formalmente, um documento é descrito nesse modelo pela seguinte notação [CRO 94]:

$$F_d = \{\mu_{FI}(d, w)/(d, w) \mid d \in D \text{ e } w \in W\}$$

Nesse caso,  $F_d$  é o conjunto fuzzy de termos ( $w$ ) que descrevem o documento  $d$ . Cada par  $(d, w)$  possui um grau de relação (entre zero e um) que é a relação entre o documento  $d$  e o termo  $w$ , e é expresso pela função  $\mu_{FI}(d, w)$ .

Os pesos (graus de importância) de um termo podem ser determinados por qualquer um dos métodos existentes para isso (ver seção 7.5), já que não há uma função *fuzzy* específica para isto.

O usuário pode especificar o grau de importância de cada um dos termos descritos em sua consulta. Os operadores lógicos tradicionais (and, or e not) podem ser utilizados. Para cada um desses operadores há uma série de funções difusas que podem ser utilizadas. As funções correspondentes mais comuns são as seguintes [BAR 82]:

- a) **Função máximo** – retorna o maior valor entre dois valores. Substitui o operador de intercessão *and*;
- b) **Função mínimo** – retorna o menor valor. Substitui o operador *or*;
- c) **Função complemento de um** – retorna o complemento de um para o termo seguinte ao operador *not*.

A consulta do usuário é analisada e as funções correspondentes são aplicadas, gerando um único conjunto difuso resultante. Esse conjunto representa a necessidade de informação do usuário.

O cálculo de relevância de um documento é baseado no grau de pertinência do documento ao conjunto especificado pela consulta. Quanto maior o grau de pertinência maior é o grau de relevância.

Dependendo da função adotada para cada operador, alguns termos da consulta podem acabar sendo prejudicados pelos outros (a função mínimo, por exemplo, desconsidera o termo mais importante). A adoção de um operador deve ser realizada com cautela. Cada uma obtém um resultado diferente e, portanto, análises devem ser feitas a fim de identificar a que proporciona melhores resultados (ou menos perdas).

Existem alguns operadores difusos capazes de compensar os efeitos de perda quando os conjuntos são agregados (em uma operação *and*, por exemplo). Henry Oliveira [OLI 96] apresenta um estudo que demonstra as diferenças que ocorrem a partir da utilização de uma ou outra função para determinado operador. Lá podem ser obtidas técnicas que auxiliam a escolha da função que mais se adapta a determinado contexto.

### 5.1.5 Modelo da busca direta

O modelo de busca direta é denominado modelo de busca de padrões (*pattern search*) e utiliza métodos de busca de *strings* para localizar documentos relevantes. Na verdade, quando esse modelo é utilizado, não se tem a idéia de localização de documentos relevantes, mas sim, de localização da string, em si, no documento.

Nesse modelo, geralmente, os documentos não são representados ou modelados internamente pelo sistema e, portanto, não há índices. Assim, as buscas são realizadas diretamente nos dos textos originais, em tempo de execução. O resultado da busca é a

localização de todas as ocorrências do padrão de consulta em um documento ou conjunto de documentos. Esse padrão é geralmente uma palavra ou expressão regular descrevendo os caracteres que devem ser encontrados.

Essa técnica é denominada de *string search* e sua utilização é aconselhada em casos onde a quantidade de documentos é pequena (geralmente coleções pessoais de documentos) [FRA 92a]. Ela é muito utilizada em softwares de edição de documentos para que o usuário possa localizar de palavras ou expressões no texto que está editando (o comando *grep* do *Unix* também é um exemplo de *sistema* que utiliza esse modelo).

Há diversos algoritmos de busca que variam entre a força bruta, analisando um documento do início ao fim; e outros que utilizam técnicas heurísticas [BAE 92b].

No caso da utilização de índices há uma série de etapas de normalização e padronização que podem ser utilizadas. O tempo de processamento necessário para a criação do índice é compensado nas buscas que se tornam mais rápidas. Além do que, por algum motivo, se o termo não estiver presente no índice, a palavra não será encontrada e o documento não será retornado, mesmo que o documento original a contenha [BAE 92b].

Em pequenas quantidades de textos que mudam constantemente ou que são adicionados/excluídos (como em um sistema de arquivos do tipo DOS/UNIX) o melhor é utilizar o método de *string search* que realiza uma busca seqüencial ou linear do texto.

### 5.1.6 Modelo de aglomerados (clustering model)

O modelo de *aglomerados (clustering model)* [FRA 92a] utiliza técnicas de *agrupamento* (ou *clustering*) de documentos.

A idéia básica envolvida nesse modelo consiste em identificar documentos de conteúdo similar (que tratem de assuntos similares) e armazená-los ou indexá-los em um mesmo grupo ou *aglomerado (cluster)*. A identificação de documentos similares em conteúdo dá-se pela quantidade de palavras similares e freqüentes que eles contêm.

Quando o usuário especifica sua consulta e essa é remetida ao sistema, que, por sua vez, identifica um documento relevante (e isso pode ser feito através de técnicas de “casamento” tradicionais ou através de técnicas específicas para grupos) e retorna para o usuário todos os documentos pertencentes ao mesmo grupo.

Teoricamente os documentos pertencentes a um mesmo grupo também são relevantes à consulta. A base disso é a *hipótese de agrupamento (cluster hypothesis)*, que diz que objetos semelhantes e relevantes a um mesmo assunto tendem a permanecer em um mesmo grupo e possuem atributos em comum [RIJ 79].

Pelo fato das buscas tradicionais ignorarem o co-relacionamento entre documentos, o modelo de aglomerados tende a aumentar a qualidade dos resultados (pois retorna todo o grupo coeso e relevante à consulta) e o tempo de processamento (já que os grupos de documentos tentem a ser armazenados em um mesmo bloco do dispositivo de armazenamento).

Se o agrupamento for realizado de forma hierárquica é possível oferecer um sistema de busca por navegação, aonde o usuário vai selecionando os ramos que considera mais adequados até encontrar o grupo de documentos mais relevantes.

Um dos maiores problemas desse modelo é justamente identificar os grupos de documentos mais coesos e mantê-los assim durante a utilização do sistema. Todo documento inserido ou modificado deve ser re-analisado a fim de ser colocado no grupo correto.

Não há muitas informações sobre *clustering* como um modelo, mas sim como técnica de descoberta de conhecimento (ver seção 14.2.3).

### 5.1.7 Modelo lógico

O modelo lógico baseia-se em métodos e teorias provenientes da lógica matemática para modelar o processo de recuperação de documentos. Não se tem conhecimento de SRI comerciais e práticos que utilizem esse modelo. As aplicações existentes são aparentemente de âmbito acadêmico e teórico.

Para que o modelo lógico funcione torna-se necessário modelar os documentos através de lógica predicativa, o que exige um enorme trabalho de modelagem, incorporando semântica ao processo de recuperação. Com isso o sistema passa a “ter uma noção” do conteúdo dos documentos, podendo julgar melhor a relevância desses para o usuário [CRE 95].

Nesse contexto, para que o SRI possa *inferir* a consulta a partir dos documentos, esses são modelados ou expressos por assertivas (preposições) lógicas, no formato  $d \leftarrow \{ termo_1, termo_2, \dots, termo_n \}$ . O documento ( $d$ ) é considerado um modelo ou *framework* onde a consulta (expressa por  $q$ ) pode ser interpretada.

Deste modo, para que o processo de recuperação ocorra, a assertiva  $d \rightarrow q$  deve ser válida. Essa assertiva deve ser compreendida como:  $q$  é válido na situação  $d$  (*holds in situation*), ou seja,  $d$  implica em  $q$ . Isso significa que o processo busca determinar se a informação contida na situação  $d$  (o documento) trata da informação especificada na consulta ( $q$ ). Se essa assertiva for verdadeira, pode-se dizer que  $d$  trata de  $q$ , ou seja,  $d$  é relevante a  $q$  e pode ser retornado ao usuário.

Para que esta assertiva seja verdadeira a informação contida em  $d$  deve ser suficiente para que o sistema infira a informação representada por  $q$ . Cabe salientar que  $d$  não é na verdade o documento em si, mas sim uma representação sua (abstração). Essa representação deve conter o máximo de informações possíveis sobre o objeto.

Consulta:	Documentos:	Base de Conhecimento:	
$q = \{\text{recuperação, informação}\}$	$d_1 = \{\text{recuperação, informação}\}$ $d_2 = \{\text{busca, informação}\}$ $d_3 = \{\text{busca, dados}\}$ $d_4 = \{\text{organização, memória}\}$	$K = \{\text{sinônimo}\{ \{\text{busca, recuperação}\}, \{\text{dado, informação}\}, \dots \}$	
$d_1 = \{\text{recuperação, informação}\}$ $d_1 = q$ $d_1$ é 100% relevante	$d_2 = \{\text{busca, informação}\}$ $d_2 \neq q$ busca $\rightarrow$ recuperação (em K) $d_2 = \{\text{recuperação, informação}\}$ $d_2 = q$ (com 1ª transformação) $d_2$ é 85% relevante	$d_3 = \{\text{busca, dados}\}$ $d_3 \neq q$ busca $\rightarrow$ recuperação (em K) $d_3 = \{\text{recuperação, dados}\}$ $d_3 \neq q$ dados $\rightarrow$ informação (em K) $d_3 = \{\text{recuperação, informação}\}$ $d_3 = q$ (com 2ªs transformações) $d_3$ é 70% relevante	$d_4 = \{\text{organização, memória}\}$ $d_4 \neq q$ (não há transformação possível em K) $d_4$ não é relevante

FIGURA 5-2 – UM EXEMPLO DE RECUPERAÇÃO UTILIZANDO LÓGICA

Informações sobre a estrutura do objeto e palavras com seu respectivo grau de importância (peso) são informações básicas. Como o vocabulário contido nos documentos pode variar, deve existir ainda uma base de conhecimento de apoio que contenha informações semânticas e relacionamentos (sinônimos, termos mais genéricos e termos mais específicos) para que o sistema possa realizar transformações e inferir (alcançar) o objetivo ( $q$ ). Esse conhecimento adicional, expresso por predicados, é denominado conhecimento de domínio

(*background knowledge*) [CRE 95] e é geralmente armazenado em um *thesaurus* (uma espécie de dicionário de variações morfológicas – ver seção 9.1).

Já que o vocabulário utilizado em documentos pode ser ambíguo e conter incertezas, essas incertezas devem ser consideradas durante o processo de inferência. Essa incerteza pode ser medida [CRE 95], e quanto mais transformações (passos de inferência) o sistema tiver que realizar para chegar em  $q$ , maior é o grau de incerteza<sup>12</sup>. Isso porque cada troca de palavras (substituição por um sinônimo ou palavra mais ou menos abrangente) possui uma incerteza associada que acaba afetando a idéia geral de determinada informação.

Logo, quanto maior o grau de incerteza menor o grau de relevância. Aqueles documentos que necessitarem de menos transformações no processo de inferência da consulta são considerados os mais relevantes para a necessidade de informação do usuário.

### 5.1.8 Modelo contextual

A grande maioria dos modelos anteriores leva em conta a presença dos termos nos documentos e realizam o “casamento” entre um documento e a consulta somente se as palavras contidas no documento forem exatamente iguais (casando os padrões similares) às palavras especificadas na consulta. Logo, os documentos que possuem as palavras identificadas na consulta são considerados relevantes e os que não as possuem (mesmo que os termos tenham o mesmo sentido) são considerados irrelevantes por possuírem uma *morfologia* diferente.

Esse tipo de “casamento” é muito restritivo, pois, como já salientado, a linguagem natural possui ambigüidade e incerteza inerentes, causando problemas de sinonímia (onde vários termos podem denotar um mesmo objeto) e polissemia (onde um termo possui vários significados). Esses são os *problemas do vocabulário*<sup>13</sup> (ou problemas da diferença de vocabulário) já discutidos anteriormente. Com isso, se um documento trata do assunto especificado pela necessidade do usuário, mas seu autor não utiliza os mesmos termos que o usuário, esse documento não é considerado relevante.

Existe uma série de ferramentas e técnicas desenvolvidas com o intuito de minimizar esses problemas, porém, nem todos os SRI as fornecem e nem todos os usuários as utilizam. Algumas dessas técnicas sugerem que a compreensão do conteúdo dos documentos e da consulta oferece melhorias significativas ao processo de recuperação, já que o significado dos termos da consulta pode ser identificado.

O modelo contextual (ou conceitual) [LIN 93; LOH 97] é desenvolvido a partir do princípio de que todo documento possui um contexto, pois a pessoa que escreve um texto o faz desenvolvendo um assunto específico e utiliza frases interconectadas ou encadeadas que fazem sentido dentro assunto (o contexto).

<sup>12</sup> Afirmação proveniente do *princípio da menor transformação*, onde “dada uma representação de documento ( $d$ ), uma representação de consulta ( $q$ ) e um conjunto de conhecimento  $K$ ; a medida de relevância, denotada por  $r(d \rightarrow q)$  e relativa à  $K$ , é determinada pela mínima transformação aplicada em  $d$  para algum  $d'$  que contenha  $q$ ” [LAL 96]. Nesse princípio, o símbolo  $d$  representa a informação explícita do documento e contém tanto sua informação estrutural quanto sua informação de significância. O conjunto de conhecimento  $K$  contém informações sobre todos os relacionamentos os quais a transformação baseia-se. Esses relacionamentos contêm informações semânticas e pragmáticas sobre o domínio e podem conter incertezas. O documento transformado,  $d'$ , representa a informação explícita e implícita (inferida pelas transformações). Esse último pode conter incertezas (pois algumas transformações são incertas) e pode ser parcial. Essa incerteza é utilizada para calcular a medida de relevância  $r(d \rightarrow q)$ .

<sup>13</sup> Apesar de serem consideradas problemas, pois dificultam a localização de informações, essas características da linguagem é que a tornam expressiva e elástica [DOY 75, p294].

A consulta do usuário também possui um contexto que é definido por sua necessidade de informação. Uma vez identificado o contexto dessa necessidade de informação e os contextos dos documentos de uma coleção (base de documentos), o processo de recuperação e de identificação de informações relevantes pode ser feito ao nível de contextos e não mais ao nível de palavras isoladas. Espera-se com isso que os resultados em relação à relevância dos documentos retornados sejam melhores.

A princípio essa idéia parece prática e fácil, porém, identificar e modelar os contextos dos documentos não é uma tarefa trivial. Os processos de cognição humana ainda não são completamente compreendidos e, portanto, não é possível saber que elementos são necessários para modelar um contexto. Atualmente isso é feito selecionando algumas palavras que em conjunto (e estando correlacionadas) podem definir (dar uma idéia de) esse contexto.

Como cada palavra pode estar presente em mais de um contexto, deve haver um grau (peso) indicando quanto uma palavra é relevante (importante) em cada contexto. Esse conjunto de palavras é então utilizado para representar o contexto [LOH 97].

Os documentos são então indexados de acordo com os contextos existentes e definidos. Isso é feito através de uma espécie de classificação onde as características (palavras ou termos) que descrevem determinado contexto são localizadas nos documentos [RIL 94]. Cada característica encontrada ativa seu contexto correspondente (palavras com graus elevados de relevância podem definir um contexto pelo simples fato de aparecerem em um documento). O valor de relevância (importância) dessa palavra é adicionado ao grau de pertinência do documento aos contextos que ela representa. Logo, quanto mais palavras localizadas (ativas) um contexto possuir, maior é o grau de relação do documento com esse contexto. Nota-se com isso que um documento pode pertencer a mais de um contexto com graus diferentes de relevância (relação ou pertinência).

A identificação do contexto da consulta do usuário também é feita da mesma forma: as palavras existentes ativam o contexto mais relevante. Dependendo do sistema o usuário pode “navegar” pelos contextos existentes e selecionar um para a busca, encontrando assim todos os documentos que estejam relacionados com o contexto escolhido [LOH 97].

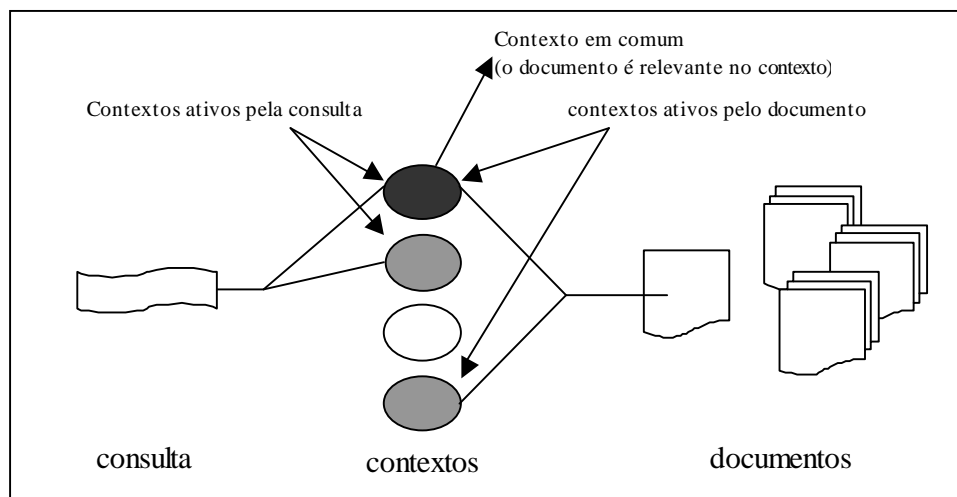


FIGURA 5-3 – RECUPERAÇÃO CONTEXTUAL

Esse modelo não elimina o problema do vocabulário, mas pode minimizá-lo se o conjunto de palavras utilizado na descrição dos contextos for bem escolhido. Várias palavras podem ser utilizadas nessa descrição. Porém, muitas delas certamente são encontradas em vários contextos. Devem ser escolhidas aquelas que caracterizam bem cada contexto sem que

indiquem ou apareçam em outros (ou muitos outros) contextos. Ou seja, as palavras devem ter um alto grau de discriminação.

Isso induz que os contextos sejam descritos manualmente por uma pessoa que conheça muito bem o assunto (contexto) que está descrevendo. O ideal é que essa pessoa seja um especialista na área (contexto) em questão. Essa pessoa deve selecionar os termos mais relevantes do contexto, adicionando também sinônimos e palavras específicas. Pelo fato de muitas dessas palavras poderem aparecer em mais de uma descrição de contexto, além do fato de umas poderem ser mais relevantes (descritivas ou discriminantes) do que outras, um grau (peso) de relevância (ou importância) deve ser atribuído a cada palavra.

Essa descrição manual de contextos pode ser auxiliada (ou substituída, dependendo do caso) por um processo automatizado de identificação de palavras relevantes. Existe uma série de técnicas desenvolvidas com o propósito de identificar palavras importantes e correlações entre palavras [CHE 96; KOW 97; SAL 83], que podem ser utilizadas. Dicionários de sinônimos, thesaurus e redes semânticas também podem ser utilizadas nesse processo.

Um dos maiores problemas desse modelo está no fato dos descritores poderem ser elaborados incorretamente, o que ocasionaria uma busca incorreta onde os documentos retornados provavelmente seriam irrelevantes para a necessidade do usuário. Ou seja, a descrição dos contextos deve ser elaborada cuidadosamente para que a recuperação contextual funcione de forma a oferecer resultados relevantes e coerentes.



## 6 Indexação e normalização

A primeira tarefa que um SRI deve realizar para que o usuário possa disparar buscas é a *catalogação* dos documentos. Todo documento adicionado ao sistema deve ser analisado ou descrito para que possa ser recuperado. Nessa fase as características dos documentos são identificadas e adicionadas ao sistema.

Para que o sistema possa encontrar rapidamente um documento a partir de um conjunto de características descritas em uma consulta, deve existir um índice. Esse índice é construído através de um processo de indexação. Indexar significa, justamente, identificar as características de um documento e colocá-las em uma estrutura denominada *índice*.

O índice pode ser compreendido como uma espécie de filtro que é capaz de selecionar os documentos relevantes e manter de fora os documentos irrelevantes [LAN 68]. Quando a indexação é realizada manualmente, a pessoa encarregada de fazê-la deve analisar o conteúdo de cada documento e identificar palavras-chave que o caracterizem. Essas palavras, quando adicionadas ao índice, passam a ser chamadas de *termos* de índice<sup>14</sup>.

Os termos de índice podem variar, ou seja, dependendo da área das pessoas que irão utilizar o SRI, um documento pode ser indexado por termos diferentes que são correspondentes ao vocabulário utilizado na área. Nesse caso, geralmente, há um conjunto de termos predefinidos e específicos para cada assunto da área em questão. A pessoa encarregada de indexar os documentos deve identificar a que assunto cada um deles pertence e utilizar então os termos adequados. Essa técnica, denominada *vocabulário controlado* [LAN 68], facilita muito a localização de informações pois os usuários estão acostumados a utilizar os termos comumente utilizados na sua área de interesse [IIV 95]. Por outro lado, se o SRI for utilizado em uma área diferente da área para a qual foi indexado ele não será tão eficiente porque os problemas relacionados à diferença de vocabulário serão mais frequentes.

Pelo fato dos SRI atuais possuírem um enfoque muito mais abrangente, tanto em termos de conteúdo (documentos variados) quanto em termos de audiência (usuários de diversas áreas e localizados em diversas partes do mundo), não é recomendado utilizar essas técnicas de controle de vocabulário. Nesse caso, recomenda-se que o SRI utilize todas as palavras possíveis de um documento como termos de índice e ofereça ao usuário ferramentas de apoio à elaboração de consultas capazes de auxiliá-lo na escolha dos termos mais adequados.

Esse processo de descrição onde as palavras dos documentos são colocadas em um índice é denominado *indexação*. O objetivo da indexação é identificar e construir pontos de acesso para um documento. Os SRI podem permitir a coordenação (combinação ou relacionamento) de termos durante o processo de indexação ou depois dele, durante a consulta.

No primeiro caso, onde os termos são combinados e correlacionados no momento da indexação, diz-se que o índice utiliza uma linguagem pré-coordenada. Nesse momento, as variações morfológicas (número, grau) são eliminadas e as palavras são agrupadas em classes que acabam sendo mapeadas para um único termo (geralmente pertencente ao vocabulário controlado).

---

<sup>14</sup> Palavra e termo diferem no sentido de a primeira ser menos formal do que a segunda. As palavras são encontradas nos documentos. Quando uma palavra é identificada como importante, ela é *mapeada* para um termo que vai então caracterizar o documento. Nesse processo, diversas palavras (sinônimos) podem ser mapeadas para um único termo.

No segundo caso, quando não é feita uma análise à priori dos termos e seus relacionamentos, costuma-se dizer que o índice possui uma linguagem pós-coordenada. Esse nome deve-se ao fato de a pessoa que busca as informações ter que selecionar e relacionar os termos durante o processo de busca.

Ambos os casos possuem suas vantagens e desvantagens. Quando o índice é pré-coordenado os documentos são organizados de maneira a facilitar sua localização em topologias (classes ou assuntos) conhecidas pelas pessoas de determinada área. Quando uma dessas pessoas deseja algum documento ela sabe que topologias retornam determinadas informações. Por outro lado, pessoas que não dominam a área ficam sem saber que topologias (ou termos correspondentes a elas) utilizar e podem pensar que não existem documentos relevantes à sua necessidade de informação. Já no caso pós-coordenado, os documentos não são organizados em classes de assuntos similares. Assim, para que um usuário localize um documento de determinado assunto ele deve utilizar todos os termos e suas variações morfológicas e ortográficas, pois o vocabulário não é controlado. Apesar desse fato, a princípio, parecer uma desvantagem, o usuário tem maior liberdade de escolha, podendo aumentar ou diminuir a abrangência de sua consulta. Para tanto, porém, ele deve conhecer o domínio da informação de que necessita e os termos empregados para descreve-la. Atualmente existem muitas ferramentas de auxílio à especificação de consultas que podem facilitar esse processo (ver seção 9).

Os índices possuem também o fator exaustividade [LAN 68], que mede a quantidade de assuntos distintos que um índice é capaz de reconhecer. Quanto maior a exaustividade, maior a abrangência e menor a precisão, já que mais palavras podem levar ao mesmo item (e uma consulta acaba retornando muitos documentos). Muito relacionado com a exaustividade está o fator especificidade [LAN 68], que é a capacidade dos termos de índice descreverem corretamente os tópicos de um documento. Quanto mais específico for um índice, maior a precisão e menor a abrangência. Esses dois fatores podem ser manipulados por uma indexação pré-coordenada, e é possível encontrar um nível de equilíbrio para os dois em uma população fechada de usuários [LAN 68].

O processo de indexação pode ser realizado manualmente ou automaticamente. O processo manual de elaboração de índices é muito abordado pela área da biblioteconomia. Já para a área da computação os índices criados pelo processo automatizado são os mais relevantes. Esse processo automatizado é o assunto abordado na próxima seção.

## 7 Indexação automática

O processo de indexação automática busca identificar palavras relevantes (descritores) nos documentos de uma coleção de documentos e armazená-las em uma estrutura de índice. As fases normalmente encontradas nesse processo são a *identificação de termos* (simples ou compostos), a *remoção de stopwords* (*palavras irrelevantes*), a *normalização morfológica* (*stemming*) e a *seleção de termos*. Para cada uma dessas etapas existem diversas técnicas. Dependendo da situação a ordem de aplicação dessas etapas pode variar ou alguma delas pode não ser utilizada [RIL 95].

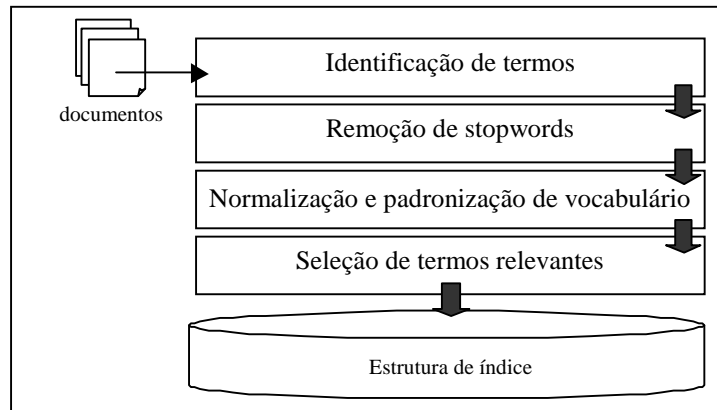


FIGURA 7-1 – ETAPAS DO PROCESSO DE INDEXAÇÃO AUTOMÁTICA

### 7.1 Identificação de termos

Essa fase nada mais é do que a aplicação de um *parser* (analisador léxico<sup>15</sup>) que identifique as palavras presentes nos documentos, ignorando os símbolos e caracteres de controle de arquivo ou de formatação.

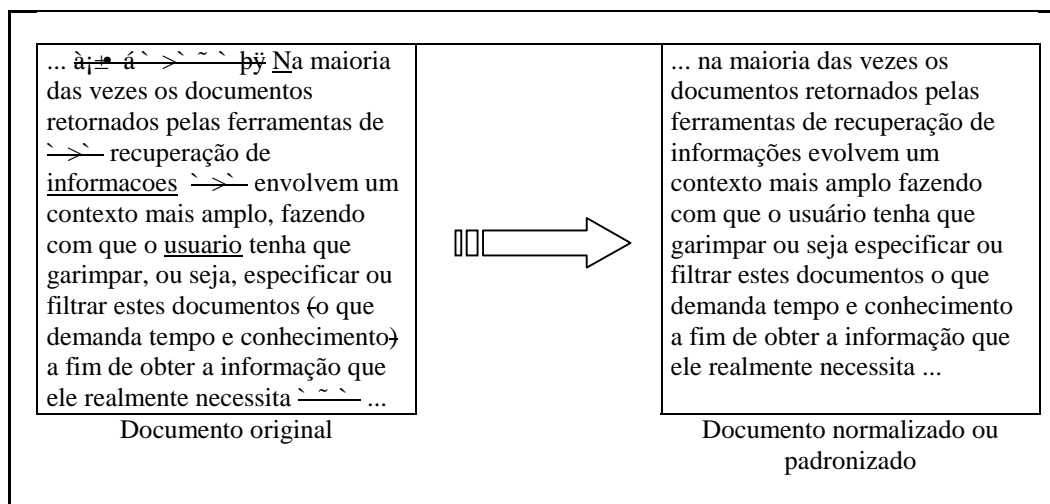


FIGURA 7-2 – IDENTIFICAÇÃO DE TERMOS VÁLIDOS

Pode-se utilizar um dicionário a fim de fazer a validação das seqüências de caracteres identificadas a fim validar sua existência e corrigir possíveis erros ortográficos (*dictionary*

<sup>15</sup> A análise léxica consiste na conversão de uma cadeia de caracteres de entrada em uma cadeia de palavras ou *tokens* [FOX 92].

*lookup*) [SAL 83]. Um thesaurus ou um dicionário de sinônimos pode auxiliar na normalização do vocabulário, caso deseje-se trabalhar com um vocabulário controlado.

Diversas técnicas adicionais de padronização podem ser aplicadas [BAE 92b; FOX 1992]: a passagem de todos os caracteres para a forma maiúscula (ou minúscula); a substituição de múltiplos espaços e tabulações por um único espaço; a padronização de datas e números; a eliminação de hífen. Se uma técnica for adotada, ela também deve ser utilizada em cima da consulta do usuário.

A utilização de uma técnica de padronização não oferece somente vantagens. Se a transformação de caracteres maiúsculos para minúsculos for adotada, por exemplo, não é possível diferenciar substantivos próprios de comuns nas buscas.

## 7.2 Identificação de termos compostos

Muitas palavras têm um significado diferente quando utilizadas em conjunto. Isso costuma acontecer porque existem conceitos que só podem ser descritos pela utilização de duas ou mais palavras adjacentes. Algumas vezes uma palavra é combinada com outra a fim de modificar ou refinar seu significado (exemplo: processo *judicial*, processo *computacional*). Quando isso ocorre, essas duas ou mais palavras não podem ser separadas quando indexadas. Caso sejam separadas, o conceito ou idéia perde-se.

A fase de identificação de termos compostos (denominada *Word-phrase formation*) busca identificar essas expressões compostas de dois ou mais termos [CRO 82].

Existem basicamente duas formas de identificação de expressões. A primeira é feita com base na identificação de termos que co-ocorrem com frequência em uma coleção de documentos. Nesse caso torna-se interessante que o sistema apresente ao usuário as expressões identificadas e solicite a ele que decida quais são as corretas. A segunda consiste na utilização de um dicionário de expressões que indique então quais palavras devem ser combinadas.

Esse tipo de técnica torna a busca mais precisa, já que os termos compostos costumam aparecer em um número menor de documentos, tornando a consulta menos abrangente. Porém, esses termos são geralmente armazenados no índice de forma composta e, nesse caso, o usuário não pode localizá-los de forma separada. Uma solução para esse problema consiste em armazenar ambas as formas: combinada e separada.

Caso a técnica de identificação de termos compostos não seja aplicada, o usuário ainda pode especificar em sua consulta a informação que eles representam. Isso pode ser feito indicando que dois ou mais termos devem aparecer no mesmo documento. Em alguns sistemas é possível especificar a distância máxima que esses termos devem ser encontrados (uma ou duas palavras de distância, por exemplo) [SAL 83]. Essa especificação deve ser cuidadosa, pois se não especificada corretamente aumenta a abrangência de uma consulta ao invés de torna-la mais precisa (podem ser retornados documentos que contenham uma das palavras, e não todas).

## 7.3 Remoção de “stopwords”

Existem algumas palavras presentes em um documento textual que são utilizadas com o intuito de conectar as frases. Essas e outras palavras, pertencentes a classes de palavras cuja finalidade é auxiliar a estruturação da linguagem (tais como conjunções e preposições), não necessitam ser incluídas na estrutura de índice.

Além dessas, existem também palavras cuja frequência na coleção de documentos é muito alta. Palavras que aparecem em praticamente todos os documentos de uma coleção não são capazes de discriminar documentos e também não devem constar na estrutura de índice [FOX 92].

Todas essas palavras consideradas sem valor para a busca devido a sua natureza freqüente ou semântica são denominadas palavras negativas (ou *stopwords*) [KOR 97; KOW 97; SAL 83]. Essas palavras dificilmente são utilizadas em uma consulta, pois sua indexação somente tornaria o índice maior do que o necessário.

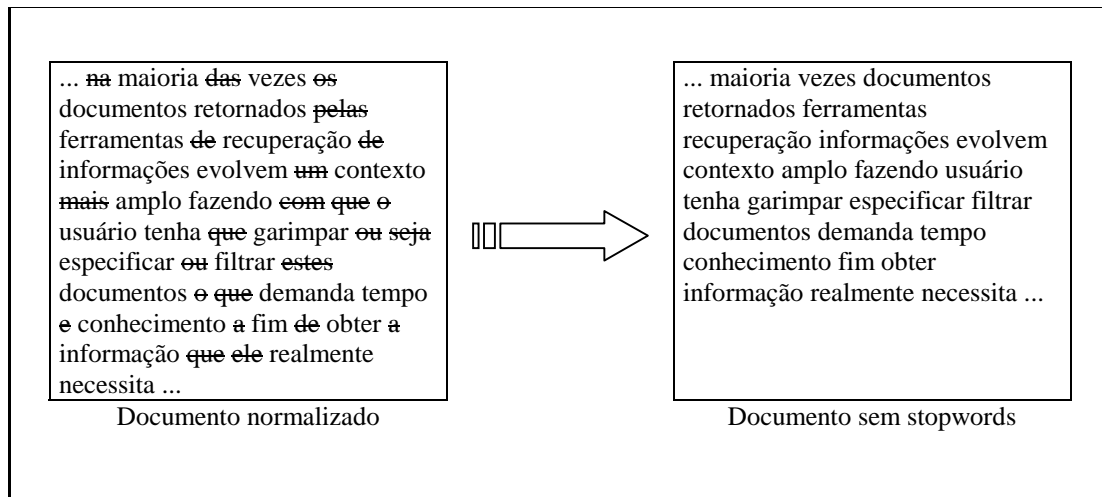


FIGURA 7-3 – IDENTIFICAÇÃO DE *STOP-WORDS*

Existem estudos que oferecem listas de stopwords (denominadas *stoplists* ou *dicionários negativos*) que podem ser livremente utilizadas na elaboração de ferramentas que realizem o processo de remoção de stopwords.

## 7.4 Normalização morfológica

Durante o processo de indexação, dependendo do caso, torna-se interessante eliminar as variações morfológicas de uma palavra. As variações morfológicas são eliminadas através da identificação do radical de uma palavra. Os prefixos e os sufixos são retirados e os radicais resultantes são adicionados à estrutura de índice. Essa técnica de identificação de radicais é denominada *lematização* ou *stemming* [FRA 92b; KRA 96], que em inglês significa reduzir uma palavra ao seu radical (ou raiz).

As características de gênero, número e grau das palavras são eliminadas. Isso significa que várias palavras acabam sendo mapeadas para um único termo, o que aumenta a abrangência das consultas. Com essa técnica o usuário não necessita preocupar-se com a forma ortográfica com a qual uma palavra foi escrita no texto. Assim, uma idéia, independente de ter tendo sido escrita através de seu substantivo, adjetivo ou verbo, é identificada por um mesmo (e único) radical. Essa aparente vantagem ocasiona uma diminuição na precisão, já que o usuário não consegue mais procurar por uma palavra específica.

Existem várias formas de identificação do radical de palavras [FRA 92b].

Uma delas consiste na definição de uma lista de prefixos e ou sufixos comumente encontrados no vocabulário de uma língua. Após, toda vez que um desses prefixos ou sufixos é encontrado, ele é retirado da palavra (o padrão mais comprido encontrado é que deve ser retirado). Um problema dessa técnica é que, dependendo da língua ou do contexto, o padrão

encontrado nem sempre corresponde a um prefixo ou sufixo, pois ele pode fazer parte do radical da palavra (o sufixo *ual* deve ser retirado de *fatual*, mas não de *igual*).

Outra solução consiste na utilização de um dicionário morfológico onde o radical de cada palavra poderia ser identificado corretamente [KOW 97, p71]. Porém, esses dicionários não costumam ser completos e são específicos da língua para a qual foram construídos.

É possível também identificar os padrões (seqüências de caracteres) que co-ocorrem com freqüência nas palavras. Para tanto se deve considerar o texto como uma seqüência de caracteres sem sentido semântico (sem significado) e segmentar essa seqüência em *strings* de tamanho predefinido. Essas strings de tamanho fixo são denominadas *anagramas* ou *n-agramas* (*n-grams*) [KOW 97, p79], onde *n* indica o tamanho dos strings. Costuma-se trabalhar com *bigramas*, *trigramas* e *pentagramas*<sup>16</sup>. Apesar de ser considerada uma forma de stemming, a técnica de anagramas não costuma ser utilizada para fins de recuperação, já que os termos tornam-se incompreensíveis. Suas aplicações incluem a criptografia e a detecção de erros ortográficos [KOW 97, p79].

Além de eliminar as variações morfológicas das palavras e aumentar a precisão das consultas, o método de stemming também é capaz de reduzir o tamanho de um índice em até 50% [FRA 92b].

Essas vantagens, dependendo da aplicação, podem acabar se transformando em problemas. Na classificação de documentos, por exemplo, a variação morfológica é extremamente importante, pois aumenta a discriminação entre documentos [RIL 95].

Devido a isso, sugere-se que as palavras sejam indexadas utilizando a forma ortográfica encontrada nos documentos e que o usuário encarregue-se de especificar que variações morfológicas ele deseja durante o processo de consulta. Em alguns sistemas é possível que o usuário especifique uma máscara. O usuário pode se aproveitar dessa característica, utilizando o radical da palavra seguido do símbolo adotado pelo sistema para representar a de máscara (geralmente o asterisco).

Nesse caso, o sistema considera que todas as palavras que iniciam com o radical especificado são o mesmo termo, e todos os documentos em que elas aparecem acabam sendo retornados. O sistema pode ainda identificar as variações morfológicas durante a consulta e mostrá-las para que o usuário decida quais são as de seu interesse, eliminando assim variações incorretas. Essa técnica é denominada *conflação* (*conflation*).

A normalização do vocabulário também compreende técnicas de tratamento de anáforas, ou seja, a localização de pronomes e a identificação dos substantivos a quem eles referem-se. Essas técnicas, porém, exigem processamento de linguagem natural e nem sempre são aplicadas.

## 7.5 Cálculo de relevância

Nem todas as palavras presentes em um documento possuem a mesma importância. As palavras utilizadas mais freqüentemente (com exceção das *stopwords*) costumam ter um significado mais importante. Palavras constantes em títulos ou em outras estruturas também possuem uma importância maior, já que o autor do documento deve tê-las colocado lá por considera-las como sendo muito relevantes e descritivas para a sua idéia. Os substantivos e complementos também podem ser considerados mais relevantes que os demais termos de uma oração.

<sup>16</sup> Bigrams: re cu pe ra çã od ei nf or ma çõ es. Trigrams: rec upe raç ãod ein for maç ões

Logo, o cálculo de relevância de uma palavra pode basear-se na frequência das palavras, na análise estrutural do documento ou na posição sintática de uma palavra.

As técnicas mais comuns são baseadas na frequência das palavras na coleção de documentos, pois as outras necessitam de métodos adicionais (análise de linguagem natural, por exemplo) que exigem maior complexidade (conhecimento).

Existe uma série de fórmulas foram desenvolvidas ou aplicadas com o intuito de calcular a importância de uma palavra baseando-se em sua frequência [RIJ 79; SAL 87a]. Essa importância costuma ser chamada de *peso* e indica o grau de relação entre a palavra e os documentos em que ela aparece.

Algumas fórmulas de identificação de peso são as mais simples outras. As mais simples são baseadas em cálculos simples de frequência: *frequência absoluta*, *frequência relativa*, *frequência inversa de documentos*; outras, tais como *information gain* [MLA 2000], *signal weighting* e *noise* [SAL 83, p63], são baseadas na teoria da informação, e, nesse caso, quanto maior a probabilidade de um termo transmitir informação, maior é o seu peso. Ou probabilidade.

Há ainda algumas mais complexas, que envolvem análise de correlação entre documentos ou termos e podem envolver técnicas de agrupamento de documentos ou palavras [YAN 97]. São exemplos o *term strength* [WIL 92], *discrimination value* [SAL 83, p66] e Qui-quadrado [WIE 95; YAN 97].

As fórmulas mais simples e comuns, que servem para praticamente todo o tipo de aplicação, são apresentadas a seguir.

A frequência absoluta, também conhecida por frequência do termo ou *term frequency* (TF) nada mais é do que a medida da quantidade de vezes que um termo aparece em um documento. Essa é a medida de peso mais simples que existe, mas não é aconselhada porque não é capaz de fazer distinção entre os termos que aparecem em poucos documentos e os termos que aparecem em muitos documentos. Em alguns casos esse tipo de análise poderia ser extremamente importante, pois os termos que aparecem em muitos documentos não são capazes de discriminar um documento de outro.

Além disso, a frequência absoluta não leva em conta a quantidade de palavras existentes no documento. Com isso, uma palavra pouco frequente em um documento pequeno pode ter a mesma importância de uma palavra muito frequente de um documento grande.

A frequência relativa busca solucionar esse último problema, levando em conta o tamanho do documento (quantidade de palavras que ele possui) e normalizando os pesos de acordo com essa informação. Sem essa normalização, os documentos grandes e pequenos acabam sendo representados por valores em escalas diferentes. Com isso os documentos maiores possuem melhores chances de serem recuperados, já que receberão valores maiores no cálculo de similaridades [SAL 87b].

A frequência relativa ( $F_{rel}$ ) de uma palavra  $x$  em um documento qualquer, é calculada dividindo-se sua frequência absoluta ( $F_{abs}$ ) pelo número total de palavras no mesmo documento ( $N$ ):

$$F_{rel}x = \frac{F_{abs}x}{N}$$

Para solucionar o outro problema da frequência absoluta, onde a quantidade de documentos em que um termo aparece não é considerada, torna-se necessário obter essa

informação. A freqüência de documentos é quem indica a quantidade de documentos em que um termo aparece.

De posse da freqüência absoluta e da freqüência de documentos é possível calcular a freqüência inversa de documentos (*inverse document frequency* – IDF), capaz de aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos documentos [ROB 97], justamente pelo fato dos termos de baixa freqüência de documentos serem, em geral, mais discriminantes [SAL 83].

Existe mais de uma maneira de se identificar o peso através da freqüência inversa de documentos. Uma das mais conhecidas é obtida pela aplicação da fórmula seguinte [SAL 83, p63]:

$$Peso_{td} = \frac{Freq_{td}}{DocFreq_t}$$

Nessa fórmula,  $Peso_{td}$  é o grau de relação entre o termo  $t$  e o documento  $d$ ,  $Freq_{td}$  corresponde ao número de vezes que o termo  $t$  aparece no documento  $d$  e  $DocFreq_t$  corresponde ao número de documentos que o termo  $t$  aparece.

Uma consideração sobre os pesos identificados em uma coleção de documentos é a de que eles são válidos por determinado período de tempo [KOW 97]. Isso porque a coleção pode variar devido à adição de novos documentos ou devido a mudanças no conteúdo dos documentos (que podem ser modificados). Nesse caso, torna-se necessário recalcular periodicamente os pesos das palavras ou adotar alguma outra solução.

Até o momento não se tem um estudo que indique a superioridade de uma técnica sobre outra *de forma significativa*. Porém, algumas são mais adequadas do que outras para certas aplicações ou modelos conceituais.

## 7.6 Seleção de termos

Os arquivos de índice de um SRI geralmente consomem muito espaço, podendo chegar a 300% do espaço correspondente aos documentos originais [FAL 92, p45]. Esse tamanho pode ser diminuído excluindo-se alguns termos de menor importância (pouco discriminantes) dos documentos. Assim, há uma redução no espaço de dimensões que modelam os documentos.

As técnicas de seleção de termos relevantes podem ser baseadas no peso dos termos ou na sua posição sintática.

É importante salientar que a seleção de termos deve ser realizada com cautela. Algumas aplicações são influenciadas pelos termos de menor importância (clustering, classificação e sumarização, por exemplo) [YAN 97]. Cabe portanto ao desenvolvedor da aplicação ou ao usuário decidir se esses termos são relevantes ou não para o seu experimento. Além disso, existem técnicas de compactação que podem ser aplicadas aos índices, permitindo que os termos menos importantes também sejam utilizados sem que o tamanho do índice ocupe muito espaço de armazenamento.

### 7.6.1 Filtragem baseada no “peso” do termo

A determinação da importância de um termo geralmente é dada pelo seu *peso* (grau de correlação com o documento identificado com base na sua freqüência). A técnica mais simples de redução de dimensões é a filtragem baseada no peso de um termo [YAN 97], e

consiste em eliminar todos os termos abaixo de um *limiar* (*threshold*) estabelecido pelo usuário ou pela aplicação.

### 7.6.2 Seleção baseada no “peso” do termo

Mesmo depois de filtrados, o número de termos resultantes ainda pode ser alto. Esse número pode ser reduzido pela seleção dos  $n$  termos mais relevantes. Essa técnica de seleção é denominada *truncagem* [SCH 97], pois se estabelece um número máximo de características a serem utilizadas para caracterizar um documento e todas as outras são eliminadas.

Para tanto, é necessário que as características estejam ordenadas de acordo com seu grau de relevância (ou importância). Assim, somente as primeiras  $x$  características são utilizadas.

A truncagem pode ser aplicada em técnicas de descoberta de conhecimento a fim de aumentar a performance dos algoritmos, já que quanto maior o número de características a ser comparado, mais demorado se torna o processo.

Um dos maiores problemas dessa técnica consiste justamente em estabelecer a quantidade mínima de palavras necessária para uma boa descrição dos documentos, sem que suas características mais relevantes sejam perdidas no processo. Alguns experimentos indicam que na grande maioria dos casos um número de 50 características é suficiente [SCH 97], mas esse valor pode variar dependendo da coleção.

### 7.6.3 Seleção por “Latent Semantic Indexing”

A técnica de indexação *semântica latente* (*latent semantic indexing – LSI*) foi desenvolvida com o intuito de reduzir o número de dimensões utilizadas pelo modelo espaço-vetorial. Ela busca transformar os vetores de documentos originais para um espaço dimensional pequeno e significativo, fazendo uma análise da estrutura co-relacional de termos na coleção de documentos [WIE 95].

A redução é feita identificando-se as dimensões mais similares (próximas). Uma vez identificadas, elas são aproximadas por um processo matemático de rotação. Isso faz com que as palavras mais similares acabem em uma mesma dimensão. Em alguns casos, sinônimos e outras palavras de forte correlação acabam sendo colocados na mesma dimensão, o que minimiza um pouco os problemas relacionados à diferença de vocabulário.

A LSI pode ser aplicada de maneira global (no vocabulário de todos os documentos) ou local (em um conjunto de documentos pertencentes ao mesmo grupo, por exemplo) [SCH 97]. Geralmente a truncagem é realizada localmente [SCH 97].

Esse tipo de análise requer um pré-processamento que identifique o vocabulário presente em todos os documentos da coleção.

### 7.6.4 Seleção por análise de linguagem natural

É possível aplicar algumas das técnicas de análise de linguagem natural [SAL 83, cap7] [BAK 98] para identificar as palavras mais importantes de um documento. Essas técnicas incluem a análise sintática e a análise semântica dos documentos.

Com uma gramática bem definida para um domínio específico (um *lexicon* [GUT 96]), é possível realizar uma análise sintática em orações não muito complexas. Os objetos que compõem uma oração (uma frase) costumam ter posições sintáticas definidas. É

possível influenciar o peso dos termos encontrados nessas posições a fim de torná-los mais ou menos relevantes. Pode-se, também, simplesmente selecionar os termos mais importantes, de acordo com sua categoria sintática (sujeitos e complementos, por exemplo), e ignorar os outros. Com isso, somente os termos mais importantes são adicionados a estrutura de índice.

Porém, esse tipo de técnica exige uma base de conhecimento contendo todas as combinações sintáticas possíveis (uma gramática).

Atualmente, principalmente para a língua portuguesa, esse tipo de análise *não funciona em 100% dos casos* mas pode ser aplicado (esse documento, por exemplo, foi verificado por um software de correção gramatical que *minimizou* o número de erros contidos nele).

Por outro lado, apesar de existirem vários experimentos realizados [SAL 88], a grande maioria dos SRI não costuma utilizar esse tipo de análise. A análise lingüística é complexa em termos de implementação e não existem estudos que indiquem que sua utilização oferece resultados estatisticamente melhores do que os obtidos pelas técnicas que são baseadas em frequência [RIJ 79].

Além disso, a aplicação de uma técnica desse tipo envolve mais uma etapa de processamento, o que pode não ser prático em algumas aplicações (pelo fato de tornar o processo de indexação ainda mais demorado).

Já a análise semântica baseia-se no princípio de que as partes mais relevantes de um documento já estão de alguma forma demarcadas por estruturas de formatação específicas para isso.

A análise semântica parece estar ganhando força, principalmente agora que os documentos têm sido elaborados com marcas *HTML* e *XML*. Esse tipo de marca (*tag*) serve para estruturar o documento, facilitando a identificação das estruturas mais relevantes. Cada uma dessas marcas possui uma definição, um significado semântico. Quando a pessoa que elaborou o documento utiliza-as, ela o faz justamente para atribuir essa semântica às partes do documento que considera relevante.

Termos encontrados em títulos e subtítulos, por exemplo, geralmente identificam o assunto tratado em um documento e podem ser utilizados para indexá-lo.

Infelizmente para que essa técnica funcione os documentos necessitam conter essas marcas. Isso exige com que a pessoa que elabora o documento as coloque ou que haja uma pessoa encarregada para tal no momento em que o documento é inserido no sistema. Nesse último caso pode ocorrer da pessoa encarregada de demarcar o documento não fazer essa tarefa corretamente (demarcando posições incorretas ou não demarcando o que deveria).

Em determinados domínios as marcas são definidas por palavras-chave ou símbolos que podem ser utilizados na identificação de termos importantes. Nesse caso, as marcas não são estruturais e devem ser reconhecidas por seu padrão léxico. Em prontuários médicos, por exemplo, os nomes de substâncias e remédios sempre vêm acompanhados de um valor e uma sigla que indicam sua quantidade (por exemplo: **10 mg** de penicilina).

Para detectar esse tipo de marca é necessário estudar bem o domínio dos documentos e dispor de um especialista na área. Todas as marcas devem ser especificadas e armazenadas em uma espécie de dicionário de marcas. Podem ser definidas regras simples [SCA 97], do tipo *se palavra x então selecionar y*, capazes de tratar essas marcas rapidamente. Essas regras costumam ser específicas para o domínio e não podem ser aplicadas corretamente em outras coleções de documento. Esse tipo de regra é muito utilizado em técnicas de extração de informações (ver seção 14.2.1).

## 8 Estruturas de armazenamento

Ao longo dos anos diversas estruturas de armazenamento foram desenvolvidas. As estruturas mais comuns e eficazes para a área de recuperação de informações textuais são aquelas que utilizam técnicas lexicográficas, ou seja, são baseadas nos caracteres e em sua ordenação alfabética. A estrutura de *arquivos invertidos*, árvores *TRIE* e *PAT* pertencem a esse grupo e são descritas a seguir. O método da assinatura, que utiliza uma estrutura baseada em acesso direto (hash), também vem ganhando atenção na área e é apresentado.

### 8.1 Arquivos invertidos

A estrutura de arquivo invertido [CLA 83, p152] [HAR 92a] nada mais é do que uma lista ordenada de palavras onde cada palavra contém apontadores (*links* ou ponteiros) para os documentos onde ela aparece. Logo, quando um termo é localizado na lista, o registro correspondente contendo a lista de todos os documentos em que ele aparece é retornada.

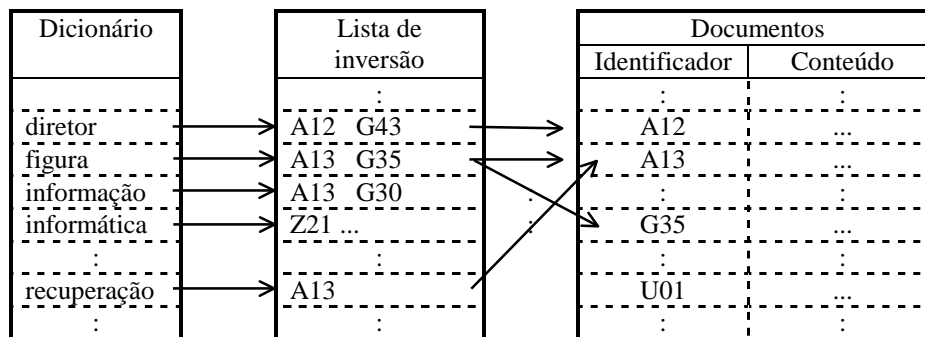


FIGURA 8-1 – ESTRUTURA DE UMA LISTA INVERTIDA

Essa estrutura é geralmente composta de três arquivos: o dicionário ou lista de palavras, a lista de inversão e os documentos. A entrada para o índice é o dicionário – uma lista que contém todas as palavras da coleção de documentos indexada. Ao ser localizada a palavra no dicionário, identifica-se sua lista invertida de documentos correspondentes. O dicionário pode ser implementado em alguma estrutura mais eficiente, tal como uma *TRIE* ou árvore-B, e pode conter qualquer tipo de informação necessária ao sistema, tal como a frequência ou relevância das palavras nos documentos.

Essa estrutura é muito eficiente em termos de acesso, porém, consome muito espaço (variando entre 10% e 100% ou mais do tamanho do documento indexado). Existem algumas variações que podem minimizar esse problema [HAR 92a].

Devido à sua rapidez de acesso e à sua facilidade de identificação de documentos relevantes a um termo, essa estrutura é uma das mais utilizadas em SRI [KOW 97, p76].

### 8.2 Árvores *TRIE*

A árvore *TRIE*<sup>17</sup> [CLA 83, p110] é uma estrutura em árvore criada especialmente para indexar palavras. Sua principal utilização se dá em arquivos cujo objetivo é armazenar palavras (dicionários, por exemplo). Nessa estrutura, cada nodo é um vetor contendo 27 componentes que correspondem às letras do alfabeto mais um componente em branco adicional.

<sup>17</sup> O nome “trie” é derivado da palavra *retrieval* [BAK 98].

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
--	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

FIGURA 8-2 – NODO DE UMA ÁRVORE TRIE

O conteúdo de cada componente de um nodo pode ser um número, indicando o número do nodo seguinte, ou um conjunto de caracteres que indica o conteúdo do nodo.

A estrutura é construída um nível por vez e cada nodo contém uma letra do termo sendo armazenado. O primeiro nodo, o nodo raiz, contém a primeira letra de todas as palavras indexadas pela estrutura. Essa técnica facilita a identificação de palavras inexistentes, já que com um único acesso é possível descobrir se determinada letra possui ou não conteúdo. Se determinada letra não possuir conteúdo é porque não existem palavras indexadas que iniciem com aquela letra.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	2	6				FUGA																					
2		ABACATE																		3							
3					4																						
4	ATÉ													ATENÇÃO													
5																											
6	BALEIA														BOLO												

FIGURA 8-3 – ESTRUTURA EXEMPLO DE UMA ÁRVORE TRIE

Não é necessário que existam nodos para todas as letras de uma palavra. Nodos são adicionados toda vez que existirem palavras com letras iguais. Os nodos param de ser adicionados no momento em que as palavras diferenciam-se. Por exemplo, na Figura 8-3 a palavra “fuga” é localizada logo no primeiro nodo, pois não existem mais palavras que comecem com “f” indexadas. As palavras que começam com a letra “b” se diferenciam na segunda letra, sendo necessário assim dois nodos para elas. A palavra “até” deveria ser armazenada no nodo de número 3, porém, por existir outra palavra cujas letras iniciais também são “ate”, outro nível de nodos foi adicionado. Nesse caso a palavra que termina no nível anterior deve ser alocada na região vazia do vetor, indicando que ela acaba com a letra que aponta para esse nodo.

As estruturas TRIE podem ser compactadas através de representações binárias que indicam a presença ou não de determinada letra no índice [BAE 92a, p21]. Essas estruturas são denominadas *c-trie* e sua implementação é mais complexa [CLA 83].

### 8.3 Método da assinatura

Os arquivos de assinatura (signature files) utilizam métodos similares aos empregados nos arquivos de acesso *hash*<sup>18</sup> (direto). O objetivo do método da assinatura é prover um teste que indique rapidamente quais são os arquivos mais (provavelmente) relevantes à consulta do usuário, o que elimina a maioria dos itens irrelevantes (pois, por ser um método inexato, alguns dos itens irrelevantes podem não ser descartados pelo teste) [FAL 92].

Os itens que passam pelo teste podem então ser passados diretamente para o usuário, ou ainda, serem avaliados por algum outro método de filtragem que identifique os itens mais relevantes [FAL 92]. Estes últimos não necessitam mais analisar toda a coleção, mas sim, somente os itens recomendados pelo método da assinatura.

Nessa estrutura, os documentos costumam ser divididos em blocos a fim de evitar que as assinaturas sejam muito densas (o que ocasionaria uma grande quantidade de colisões, ou seja, palavras com assinaturas similares). Quanto maior for a assinatura, menor é a possibilidade de colisões.

<sup>18</sup> Informações sobre funções hash podem ser obtidas no artigo de S. Wartik [WAR 92].

Dentro de um bloco, cada palavra é mapeada para um código com tamanho pré-fixado de bits – a assinatura [KOW 97, p87]. Esse código é estabelecido por uma função hash que determina quais posições desse código devem ser *setadas* para 1. Depois de determinados os códigos de todas as palavras de um bloco, eles são combinados (geralmente por uma função estilo OR) a fim de criar a assinatura do bloco.

Os blocos de cada arquivo são armazenados de modo contíguo no arquivo de assinaturas. Com isso, os documentos são armazenados de uma forma altamente comprimida, ocupando menos espaço do que uma lista invertida. O arquivo não possui uma ordenação interna, logo, novos documentos podem ser concatenados ao final da estrutura.

Computer	0001	0110	0000	0110
Science	1001	0000	1110	0000
Graduate	1000	0101	0100	0010
Students	0000	0111	1000	0100
Study	0000	0110	0110	0100
Assinatura do bloco:	1001	0111	1110	0110

FIGURA 8-4 – EXEMPLO DE ASSINATURA

No exemplo apresentado na Figura 8-4 (fonte [KOW 97]), o tamanho do bloco é de 5 palavras, o tamanho da assinatura é de 16 bits e o número máximo de dígitos “1” permitidos é 5.

As palavras de uma consulta também são mapeadas para sua assinatura correspondente. A busca é realizada através de comparação direta (*template matching*) entre os bits da assinatura da consulta e as assinaturas de documentos (ou blocos de documentos) dos bits especificados pelas palavras da consulta (através de uma leitura linear no arquivo de assinaturas).

O arquivo de assinaturas pode ainda ser compactado por meio de técnicas de compressão, obtendo reduções de armazenamento ainda maiores. No artigo de [FAL 92] podem ser encontradas diversas técnicas empregadas na construção e utilização de arquivos que utilizam esse método.

## 8.4 Árvores PAT

As estruturas *PAT* [GON 92] são estruturas de arquivo parecidas com a estrutura *TRIE* descrita anteriormente (seção 8.2). O termo *PAT* significa *PA*tricia *TR*ie, sendo que *PATRICIA* é a abreviação de “*Practical Algorithm To Retrieve Information Coded In Alphanumeric*” [KOW 97, p83], ou seja: algoritmo prático para recuperar informações codificadas em formato alfanumérico.

Nesse tipo de estrutura o documento é visto como uma cadeia de caracteres (uma grande string) onde cada uma de suas posições pode ser um ponto de entrada (um ponto de busca). Cada posição da string define uma *substring* que começa nesse ponto e vai até o final do documento, incluindo todo o texto intermediário.

As substrings são comumente denominadas de *strings semi-infinitas* (semi-infinite string) ou simplesmente *sistrings* [GON 92, p68]. O nome é uma analogia às linhas semi-infinitas, definidas na geometria como sendo linhas que possuem um ponto de origem mas seguem indefinidamente em uma direção.

Como exemplo, algumas sistrings do parágrafo anterior seriam assim definidas:

Sistring1: as substrings são comumente denominadas...

Sistring2: s substrings são comumente denominadas...

Sistring5: ubstrings são comumente denominadas...

Sistring10: ings são comumente denominadas...

Sistring20: omumente denominadas...

A árvore PAT armazena todas as sistrings possíveis de um documento [GON 92]. Porém, a estrutura PAT é baseada em árvores PATRICIA, que são árvores digitais cujos nodos são binários. Devido a isso, cada caractere de uma sistring deve ser representado por um código binário correspondente.

A fim de exemplificar<sup>19</sup>, escolhendo-se o valor 100 como representação binária da letra “h”, 110 o da letra “o”, 001 o da letra “m” e 101 o da letra “e”, a palavra “home” seria codificada pela cadeia 100110001101. Nesse caso suas primeiras oito sistrings seriam as seguintes:

Sistring 1: 100110001101 Sistring 2: 00110001101 Sistring 3: 0110001101

Sistring 4: 110001101 Sistring 5: 10001101 Sistring 6: 0001101

Sistring 7: 001101 Sistring 8: 01101

A árvore PAT correspondente é construída com base nessas sistrings identificadas. A Figura 8-5 (retirada de [KOW 97, p85]) apresenta um exemplo de árvore criada utilizando as sistrings do exemplo anterior (definidas a partir da palavra *home*).

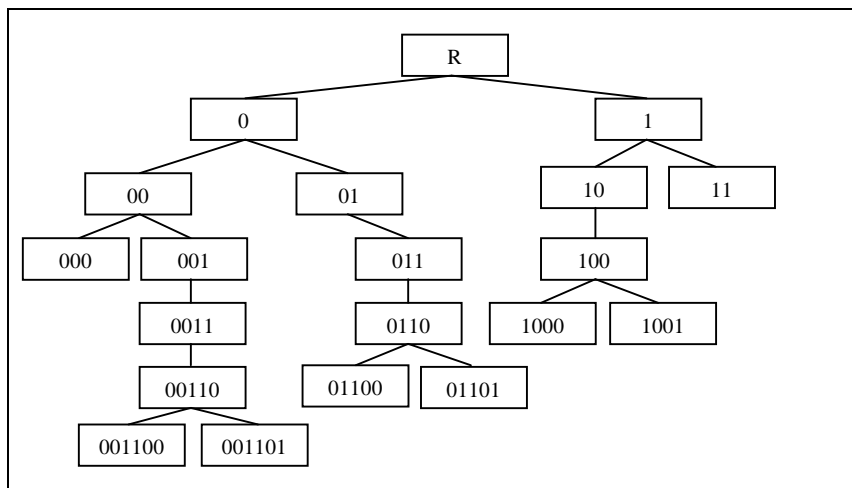


FIGURA 8-5 – EXEMPLO DE ÁRVORE PAT

Para que essa estrutura possa ser “consultada” a expressão de busca (geralmente um trecho a ser localizado) também deve ser mapeada para seu código binário correspondente. Os bits individuais do código são utilizados para definir o caminho a ser percorrido, onde “0” indica a seleção do ramo da esquerda e “1” o ramo da direita.

Esse tipo de estrutura é adequado para buscas de palavras a partir de seus prefixos, já que toda sub-árvore contém as palavras ou trechos que iniciem com o prefixo definido pelo

<sup>19</sup> Adaptado de [KOW 97].

caminho percorrido até o início da sub-árvore. Uma vez identificada a sub-árvore do prefixo desejado, todos os nodos folha dessa sub-árvore contêm palavras que iniciem com ele. Essa característica permite a busca de qualquer palavra ou frase, mesmo que incompleta, ou seja, os strings “recup” e “uperação” podem obter os mesmos resultados que a string “recuperação”.



## 9 Busca e recuperação

Quando o usuário passa a interagir com um SRI diz-se que ele está buscando informações (*information seeking*). Nesta fase o usuário tem que mapear sua necessidade de informação para uma linguagem abstrata, a linguagem utilizada pelo SRI, a fim de descrevê-la. Essa descrição é a única forma de especificação que o SRI tem da necessidade de informação do usuário. É através dela que o SRI vai poder identificar os itens de que o usuário necessita e analisar se esses itens são relevantes para ele.

Durante esse processo de especificação diversos problemas podem ocorrer<sup>20</sup>. Esses problemas variam desde a inabilidade do usuário para com o sistema (por não conhecê-lo), passando por problemas de especificação ou abstração (pois é difícil descrever uma necessidade através de um formalismo) e finalizando com o fato do usuário não ter certeza sobre sua real necessidade de informação (se o usuário necessita informação sobre algum assunto é porque ele não sabe muito sobre ele).

Se esses problemas não forem tratados ou minimizados, o SRI poderá interpretar incorretamente a necessidade de informação do usuário e retornará documentos irrelevantes para ele.

Para que esses problemas sejam minimizados, o processo de busca<sup>21</sup> deve ser realizado de forma interativa e iterativa. Durante as primeiras interações o usuário geralmente especifica incorretamente sua necessidade de informação e dificilmente recupera documentos relevantes. Porém, aos poucos, o usuário vai compreendendo o sistema, o formalismo, a interface e o modelo adotado por ele, e começa a especificar melhor sua consulta.

Na fase inicial costuma-se dizer que o usuário está coletando meta-informações, ou seja, compreendendo melhor o domínio de sua necessidade de informação e o vocabulário empregado nele. O usuário pode analisar a influência que cada termo da consulta tem nos resultados e pode modificá-la adicionando termos mais adequados caso ela não descreva o que ele exatamente necessita. Tudo isso a fim de identificar melhor quais termos são realmente importantes em uma coleção de documentos, descartando-se os termos menos cruciais [SAL 87b].

Após sucessivas iterações o usuário aprende a utilizar o sistema (caso já não o saiba) e consegue identificar melhor quais são as informações realmente necessárias para ele. Nesse momento o usuário pode descobrir que sua necessidade de informação inicial estava errada<sup>22</sup> ou que necessita de informações adicionais.

Finalmente o usuário, após receber a informação relevante, parte para a resolução de seu problema. Nesse ponto, o usuário já deve ter adquirido o conhecimento que pode ser utilizado para resolver seu problema.

Todo esse processo pode ser dividido em três etapas que são a formulação de consultas, a identificação de itens relevantes e a visualização e análise dos resultados. Para

---

<sup>20</sup> Uma série de obstáculos, alguns naturais, limita a habilidade do usuário em descrever sua necessidade de informação: a ambigüidade inerente à linguagem natural (homógrafos, acrônimos); a diferença de vocabulário entre o usuário e a pessoa que elaborou um documento; a diferença de vocabulário entre documentos; a inexperiência do usuário em determinada área, resultando em um desconhecimento do vocabulário específico da área; a dificuldade que muitas pessoas têm em abstrair ou especificar os conceitos que necessitam; a dificuldade de fazer isso usando a linguagem do sistema, geralmente limitada.

<sup>21</sup> Busca é o nome dado ao processo de identificação de correlações entre a consulta e o conjunto de itens presentes na base de documentos.

<sup>22</sup> Para compreender o teorema de Pitágoras, por exemplo, é necessário que o usuário tenha noções de geometria euclidiana [MIZ 96].

cada uma dessas etapas há uma série de técnicas e ferramentas de auxílio que podem ser utilizadas. Essas técnicas e ferramentas são descritas nas subseções seguintes.

## 9.1 Formulação de consulta

A consulta é o formalismo com o qual o usuário comunica-se com o sistema. É nela que o usuário especifica sua necessidade de informação, definindo a que assuntos os documentos devem pertencer quando retornados. Devido a isso, ela deve ser especificada corretamente para que os documentos relevantes sejam retornados.

Por ser um processo delicado, o processo de especificação ou formulação de consultas pode ser auxiliado por uma série de ferramentas. Essas ferramentas são denominadas *ferramentas de auxílio à elaboração de consultas*, e guiam o usuário na seleção do vocabulário mais adequado para a sua especificação.

As ferramentas de auxílio à elaboração de consulta mais comuns são o *thesaurus*, a *retro-alimentação por relevância (relevance feedback)* e a *expansão semântica*. Essas ferramentas são abordadas superficialmente a seguir. Maiores detalhes podem ser obtidos em [WIV 97].

- a) Thesaurus: o thesaurus é uma ferramenta similar a um dicionário, só que, ao invés de informar o significado das palavras, ela informa o relacionamento entre elas [KOC 74, p119]. Nessa ferramenta os termos estão estruturados em grupos ou classes de termos correlacionados [SAL 83]. Essas classes são dispostas de forma hierárquica de acordo com a relação entre elas. Essas relações podem ser do tipo *palavra mais abrangente versus lista de palavras mais específicas* (descrevendo sub-assuntos ou sub-ramos); *palavra-padrão e lista de palavras de significado comum* (para descrever sinônimos ou para especificar como padronizar um vocabulário, indicando que qualquer palavra da lista deve ser substituída pela palavra padrão); ou algum outro tipo de relação necessária à aplicação para qual o thesaurus for construído. Devido a essa característica de informar diferentes palavras de um mesmo assunto o thesaurus se mostra bastante útil para pessoas que não sabem que palavras utilizar em uma consulta inicial. O usuário pode navegar pelo thesaurus e descobrir como o vocabulário do domínio está organizado, compreendendo melhor esse domínio e selecionando termos mais específicos ou mais abrangentes dependendo do que ele quer retornar. O thesaurus também pode ser utilizado para mostrar o relacionamento entre diferentes vocabulários [CHE 96] (como os termos utilizados em uma área e os seus correspondentes em outra) ou para realizar a normalização de vocabulário. Os thesaurus podem ser construídos manualmente ou automaticamente, a partir de um corpus (conjunto de documentos) específico de alguma área [SAL 83].
- b) Retro-alimentação por relevância (relevance feedback): A retro-alimentação por relevância [ALL 95; MOR 82; SAL 87a] é um processo automático de refinamento de consultas que utiliza informações fornecidas pelo usuário para torná-las mais precisas, retornando somente os documentos relevantes e descartando os irrelevantes [BUC 95]. Essa técnica só pode ser aplicada após o usuário ter realizado sua primeira consulta. Para tanto, o SRI pede para que o usuário selecione dentre os documentos retornados os documentos que ele considera mais relevantes. Após, o SRI altera a consulta do usuário, modificando os termos de acordo com o conteúdo dos documentos selecionados. As consultas subsequentes passam a retornar cada vez mais documentos relevantes ao usuário, pois ele vai *contextualizando* melhor o assunto que deseja, utilizando novas

palavras e retirando as palavras que *desvirtuam* sua consulta recuperando documentos fora de seu interesse. Assim, são produzidas novas consultas teoricamente mais precisas e mais úteis (mais documentos relevantes são retornados). O refinamento de consultas pode ser feito de duas formas: na primeira, denominada retro-alimentação positiva, o sistema adiciona à consulta termos que apareçam nos documentos selecionados (ou aumenta a importância desses termos e diminui a dos que não aparecem nos documentos selecionados). Na segunda, denominada retro-alimentação negativa, os termos que não aparecem nos documentos são excluídos da consulta ou seu valor de importância é diminuído. O refinamento mais comum utiliza a retro-alimentação positiva, pois, segundo Kowalski, ela consegue levar o usuário para o conjunto de documentos mais similares aos itens já recuperados e, portanto, em direção aos itens relevantes. Já a negativa, consegue afastar-se dos itens mais irrelevantes, mas nada garante que esse movimento será em direção aos itens relevantes [KOW 97]. Aconselha-se que o sistema solicite para usuário conferir os refinamentos feitos na consulta, antes que ela seja realmente efetivada. Assim, ele modificar os pesos sugeridos de acordo com sua intuição.

- c) Expansão semântica: A expansão semântica é uma técnica utilizada para refinar a consulta do usuário, adicionando a ela termos correlacionados aos termos especificados pelo usuário [CHA 95]. Expandir semanticamente uma palavra nada mais é do que encontrar outras palavras relacionadas a ela e utilizá-las no processo de recuperação. Dependendo do caso, a expansão semântica modifica os termos da consulta tornando-a mais eficiente. Para que essa técnica seja aplicada é necessário que o SRI possua alguma espécie de dicionário contendo termos e relacionamentos possíveis para eles. Esse dicionário pode ser um thesaurus, uma rede semântica ou um *hiperdicionário* [WIV 98b]. Em SRI que utilizem o modelo contextual, essa técnica é utilizada automaticamente, já que os termos de entrada são comparados com os contextos existentes e os contextos (termos) mais significativos são selecionados para a recuperação dos documentos [LOH 99]. Um dos maiores problemas da técnica está em identificar as palavras relacionadas mais adequadas, caso exista mais de um relacionamento possível para uma palavra ou conjunto de palavras.

Essas ferramentas aliadas a estratégias de consulta [LAN 68, p198] permitem que o usuário elabore sua consulta de diversas formas a fim de obter os melhores resultados. O objetivo de uma estratégia é variar a variabilidade e especificidade de sua busca [LAN 68, p198].

As estratégias mais comuns consistem na utilização de termos compostos, sinônimos, termos abrangentes ou termos específicos e sua combinação através de expressões booleanas. O usuário pode, por exemplo, utilizar um conceito mais genérico (quando não tem uma definição exata do que necessita) que resulte em uma consulta mais abrangente e, após identificar o contexto onde se encontram as informações de que necessita, utilizar um termo específico. Ele pode ainda partir para uma consulta específica, correndo o risco de perder alguma informação.

Em alguns sistemas há a possibilidade do usuário utilizar máscaras<sup>23</sup> (\*, ?) que especificam padrões de variação morfológica de uma palavra, aumentando a abrangência da

---

<sup>23</sup> Esse processo de identificação de palavras a partir de uma expressão com máscaras é denominado *conflação* (*conflation*).

consulta. Em outros SRI é possível utilizar a linguagem natural que permite que o usuário descreva o que quer de forma mais natural e simples [KOW 97].

## 9.2 Identificação de itens relevantes (técnicas de “casamento”)

O SRI faz a identificação dos documentos relevantes a uma consulta comparando as características da consulta com as características dos documentos presentes na base de documentos.

Essa análise de similaridade de características geralmente depende do modelo conceitual adotado pelo SRI e é feita por uma classe de funções conhecidas por funções de similaridade<sup>24</sup> [KOW 97, p152]. Essas fórmulas não servem somente para a identificação de documentos relevantes. Elas também podem ser utilizadas para identificar a similaridade entre documentos, o que é extremamente útil para a área de descoberta de conhecimento em textos.

As técnicas mais comuns são baseadas no modelo de vetores. As fórmulas mais conhecidas são a *cosine* (ver seção 5.1.2), o coeficiente *dice* e o coeficiente *jaccard* [RAS 92]. Uma particularidade destas fórmulas é que o grau de similaridade sempre é normalizado (valores entre 0 e 1) e esse valor fica próximo de um (1) quando os itens são muito similares ou próximo de zero (0) quando eles são muito diferentes.

## 9.3 Visualização e análise dos resultados

Os algoritmos de busca não conseguem obter informações relevantes com 100% de abrangência e precisão [KOW 97]. Devido a isso, técnicas de visualização e refinamento são de grande valia para um SRI, reduzindo o *overhead* que o usuário tem em encontrar e compreender a informação de que necessita.

Existem diversas formas do SRI apresentar o resultado para o usuário e existem diversas informações que podem ser mostradas para o usuário enquanto ele interage com o SRI. Elas têm a habilidade de melhorar a habilidade do usuário em minimizar recursos para localizar a informação necessária [KOW 97, cap8].

Uma dessas formas é a navegação (*browsing*), que consiste em mostrar para o usuário uma lista de documentos com seus respectivos sumários. Essa lista pode ser *navegada* e os documentos relevantes selecionados e expandidos (*zooning*<sup>25</sup>) de modo que o usuário possa ver seu conteúdo. Os documentos podem possuir elos (links) entre eles, indicando sua similaridade. Os documentos também podem ser organizados em grupos ou hierarquias de grupos de documentos similares (*clustering*), onde o usuário pode *navegar* pelos grupos e *expandir* os grupos mais relevantes.

A lista de documentos deve estar organizada de alguma forma. Geralmente essa lista é ordenada em uma espécie de ranking onde os documentos mais relevantes são mostrados primeiro [HAR 92b]. O SRI pode permitir que o usuário ordene os documentos da lista de acordo com algum critério (título, autor, data, conteúdo ou resumo) a fim de facilitar a localização e análise dos documentos mais relevantes.

Além de retornar os documentos relevantes torna-se necessário que o SRI informe para o usuário o porquê de eles terem sido recuperados. Uma forma de fazer isso é selecionar

<sup>24</sup> As funções de similaridade são desenvolvidas em uma sub-área da RI que estuda técnicas de “casamento” (*matching*) de características.

<sup>25</sup> No *zooning* o item é inicialmente mostrado com o mínimo de informações possíveis para não sobrecarregar o usuário (só o título, por exemplo). Uma vez determinada a relevância de um item o usuário pode expandi-lo e receber informações cada vez mais detalhadas, conforme o nível de *zoom*.

os trechos do documento que contenham as palavras da consulta e mostrá-los para o usuário. Essa técnica é conhecida por seleção (*highlight*) [KOW 97]. Os trechos podem conter elos que permitam que o usuário navegue entre eles e podem ser identificados por cores que indiquem sua importância no documento.

É possível também se aproveitar das novas capacidades gráficas da WEB e dos sistemas operacionais e apresentar os resultados em forma de gráficos que indiquem a distribuição de frequência das palavras na coleção de documentos. Essa técnica é extremamente útil na identificação do vocabulário utilizado no corpus e, por consequência, do assunto neles tratado.

Todas essas técnicas indicam uma tendência de que os SRI serão todos visuais no futuro [JAI 97] (tanto em relação a sua interface quanto em relação às informações). As informações visuais são mais complexas mas conseguem resumir uma grande quantidade de dados em um pequeno espaço (gráficos tridimensionais, por exemplo), sem contar com o fato de que o ser humano consegue captá-las mais facilmente do que as dispostas no formato alfanumérico [JAI 97].



## 10 Bibliometria

O resultado de uma busca realizada em um SRI pode ser avaliado através de métricas provenientes de uma área chamada *bibliometria*<sup>26</sup>, que é uma sub-área da biblioteconomia [ZAN 98] encarregada de estudar e aplicar métodos matemáticos e estatísticos em documentos e outras formas de comunicação.

Essas métricas, em teoria, poderiam ser utilizadas nos SRI para que o usuário ficasse sabendo se sua consulta funcionou como deveria. Nesse caso, as métricas poderiam informar para o usuário quantos e quais documentos lhe são relevantes, além de quanto cada um deles é relevante. Porém, para que essas métricas funcionem corretamente, é necessário que a coleção de documentos manipulada pelo sistema seja muito bem conhecida, ou seja, para cada documento deveria-se saber, a priori, para quais consultas (ou assuntos) eles são relevantes.

Tendo-se uma coleção de documentos conhecida, pode-se adotar a seguinte estratégia: quando o usuário de um SRI dispara uma busca o conjunto de documentos é dividido em quatro segmentos lógicos: o primeiro contendo os documentos relevantes à consulta (que são recuperados), o segundo contendo os documentos relevantes que não foram recuperados, o terceiro contendo os documentos irrelevantes e recuperados e o último contendo os documentos irrelevantes que não foram recuperados.

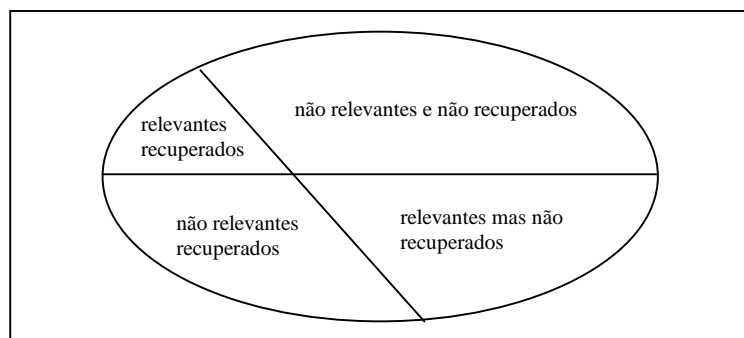


FIGURA 10-1 – EFEITOS DE UMA BUSCA NO BANCO DE DADOS TEXTUAL<sup>27</sup>

A eficiência e a eficácia de um SRI é então avaliada de acordo com a sua capacidade em recuperar o máximo possível de itens relevantes ao mesmo tempo em que filtra o maior número de itens irrelevantes. É em cima dessa estratégia que as métricas são desenvolvidas e aplicadas.

Dentro da computação as métricas mais importantes para a avaliação do resultado e do desempenho de um SRI são: *recall*, *precision*, *fall-out* e *effort* [KOR 97; KOW 97; SAL 83]. Dessas, *recall* e *precision* são as mais utilizadas.

Porém, pelo fato destas métricas necessitarem de informações relativas ao número de documentos relevantes à uma consulta e os SRI não terem como fornecê-las (um SRI não tem como saber que existem itens relevantes que não foram recuperados, até mesmo porque, se soubesse, teria-os retornado), essas métricas não são aplicadas na prática (em SRI funcionais).

Por outro lado, para efeito de comparação de sistemas (principalmente acadêmicos), existem coleções públicas de documentos preparadas especialmente para o processo de avaliação de sistemas e técnicas de recuperação. As coleções do Text REtrieval Conference (TREC<sup>28</sup>), uma conferencia internacional anual sobre a análise de desempenho de SRI, são

<sup>26</sup> A. Pritchard, 1969.

<sup>27</sup> Fonte: [KOW 97].

<sup>28</sup> <http://trec.nist.gov/>

exemplos de coleções que foram desenvolvidas com esse fim específico. Elas oferecem uma série de consultas pré-definidas e conjuntos de documentos que são relevantes a cada uma delas. Assim, o resultado de uma busca realizada em um sistema pode ser comparada com o conjunto de documentos que ela deveria retornar. Com isso os sistemas podem ser avaliados e comparados.

Em situações reais, o SRI pode pedir para que o usuário solicite quais documentos recuperados são relevantes, e fazer uma estimativa dos documentos relevantes não recuperados identificando em sua base de documentos aqueles que são parecidos com eles.

A seguir são apresentadas as medidas mais comuns para a área de recuperação de informações.

## 10.1 Recall

O *recall* (abrangência ou revocação) mede a habilidade do sistema em recuperar os documentos mais relevantes para o usuário [LAN 68]. Ele mede a quantidade de itens relevantes, dentre os existentes na base de dados, que foram recuperados. Porém, o sistema ou usuário que está avaliando o resultado deve saber, a priori, quantos documentos relevantes a sua consulta existem na base de dados, e, geralmente, essa informação não é conhecida e só pode ser estimada estatisticamente.

$$\text{recall} = \frac{n_{\text{recuperados\_relevantes}}}{n_{\text{possíveis\_relevantes}}}$$

Para que sistemas diferentes possam ser comparados deve-se adotar uma coleção específica para testes e comparações, onde as classes de documentos são conhecidas *a priori*. Deste modo, identificam-se quantos documentos o sistema conseguiu recuperar para cada classe e quantos ele deveria ter recuperado. Assim a métrica pode ser aplicada.

## 10.2 Precision

A *precision* (precisão) mede a habilidade do sistema manter os documentos irrelevantes fora do resultado de uma consulta [LAN 68]. O usuário indica quantos documentos recuperados são relevantes para ele e o SRI pode calcular, então, a quantidade de documentos recuperados que são relevantes.

$$\text{precision} = \frac{n_{\text{recuperados\_relevantes}}}{n_{\text{total\_recuperados}}}$$

A precisão é capaz de indicar o overhead (esforço) que o usuário teria para analisar uma determinada busca. Isso significa que, se 60% dos itens retornados fossem relevantes, o usuário teria desperdiçado 40% de seu esforço analisando itens irrelevantes. Logo, quanto maior a precisão, menor o esforço do usuário em analisar itens.

Portanto, a precisão pode ser utilizada nas interações do usuário com o sistema para indicar o quanto ainda o usuário necessita iterar para conseguir filtrar os itens irrelevantes e retornar itens mais relevantes [DOY 75].

## 10.3 Fallout

A quantidade de informações que um sistema possui pode influenciar diretamente nas métricas de abrangência e precisão. Basta imaginar uma base de documentos que possua 10 documentos relevantes para uma consulta  $q$ , e, que além desses, sejam recuperados mais 5

documentos irrelevantes. Essa consulta consegue uma medida de 100% de recall e 60% de precision. Supondo-se que a base de dados seja modificada e que 100 novos documentos tenham sido adicionados, é aceitável que a quantidade de documentos irrelevantes retornados aumente (é uma situação possível, apesar de não ser obrigatória). Nesse caso, o grau de precision diminuiria, apesar do sistema ainda estar recuperando os únicos documentos relevantes existentes na base de dados.

Devido à isso, verifica-se que essas duas métricas, de recall e precision, não são suficientes para avaliar a qualidade de um sistema em recuperar documentos relevantes.

A métrica *Fallout* foi desenvolvida levando em conta o fato da quantidade de documentos irrelevantes poder ser modificada pelo crescimento ou decrescimento da base de dados [DOY 75]. É exatamente isso que ela mede: a quantidade de documentos irrelevantes, permitindo que se identifique se a quantidade de documentos relevantes permanece a mesma quando o número de documentos varia. Ela é calculada dividindo-se o número de documentos irrelevantes recuperados pelo número possível de documentos irrelevantes presentes em toda a coleção [CLV 97].

## 10.4 Effort

O *effort* (esforço) mede o esforço gasto pelo usuário durante o processo de busca [LAN 68]. Ele envolve a preparação da consulta, o processo de análise de resultados e a reformulação da consulta (ou seja, toda a interação do usuário com o sistema).

A taxa de precisão é uma medida do esforço que o usuário deve realizar para obter determinada precisão. Consegue-se saber se o usuário está caminhando na direção certa se a cada iteração o grau de precisão aumenta. Esse grau de imprecisão (o complemento da precisão) pode ser considerado como sendo o esforço que ele ainda tem que realizar para localizar todos os itens relevantes.

É interessante salientar que o esforço necessário deve levar em conta os documentos relevantes existentes na base de dados e que ainda não foram recuperados (ou seja, não basta analisar somente o complemento da precisão). Como já comentado, o grau de precisão também está relacionado com o tamanho da base de dados e não indica exatamente quantos documentos ainda devem ser recuperados.



# 11 Exemplos de sistemas de recuperação de informações

## 11.1 Sistemas clássicos

As principais teorias e modelos da área de sistemas de recuperação de informação surgiram entre os anos 50 e 60. Durante esse tempo diversos sistemas foram implementados. Esses sistemas foram a base de muitos dos SRI atuais.

Um desses sistemas é o SMART [BUC 96]. Ele foi projetado por Gerard Salton em 1960 para validar seu modelo espaço-vetorial. Esse sistema realiza a indexação de termos identificando as palavras presentes nos documentos e descartando as stopwords. O índice pode conter informações sobre a frequência e o tipo (*ctype*) do documento [BUC 96]. Esse tipo pode ser uma palavra-chave, parte de um título, nome de autor, número, data ou nome próprio.

O SMART faz ainda a normalização de termos (por lematização e lista de sinônimos), identifica termos compostos, processa consultas em linguagem natural e faz análise de conteúdo, armazenando os documentos similares em um mesmo grupo (*cluster*) [LAN 68].

Esse sistema é acadêmico e flexível, criado para ser genérico, sem ter uma aplicação ou contexto específico. Sua versão atual funciona no UNIX e seu código fonte é aberto, podendo ser modificado e adaptado para outras plataformas e aplicações específicas. Atualmente ele é mantido no departamento de ciência da computação da universidade de Cornell em *Ithaca* – Estados Unidos, sob a responsabilidade de Chris Buckley.

Outro sistema que se destaca é o INQUERY [BRO 9?], desenvolvido na Universidade de Massachusetts (<http://ciir.cs.umass.edu/>). Atualmente ele é propriedade da DataWare Company (<http://www.dataware.com>). Ele indexa termos simples, compostos e números, além de remover stopwords, aplicar lematização e reconhecer palavras dependentes de domínio, tais como nomes próprios, datas e locais.

A recuperação pode ser feita através inferência probabilística (modelo bayesiano) ou booleana. Ele consegue fazer expansão semântica de consultas, utilizando um thesaurus que pode ser construído manualmente ou automaticamente (por análise de co-ocorrência de palavras no corpus). É possível utilizar a técnica de retro-alimentação, que adiciona termos ou modifica os valores dos termos de uma consulta de acordo com o *feedback* do usuário.

Há também o DIALOG (<http://www.dialog.com>), que é um sistema de recuperação de referências bibliográficas onde os documentos são catalogados manualmente e seus atributos (nome, autor, data...) são colocados em campos específicos (um desses campos pode conter até mesmo o documento ou seu resumo).

No DIALOG as buscas são construídas através da especificação dos termos desejados e dos campos que devem ser vasculhados. É possível utilizar operadores booleanos, o que dá maior poder de expressividade ao usuário.

O DIALOG foi desenvolvido por uma empresa chamada Lockheed, que já não existe. Até há pouco tempo ele era comercializado pela empresa Knight-Ridder Information (<http://www.onlineinc.com/>) que cataloga informações bibliográficas de diversas áreas do conhecimento e vende acesso a seus bancos de dados. Atualmente seus direitos pertencem a empresa Thomson Corporation.

Além desses sistemas, uma série de empresas desenvolve softwares de catálogo e busca de informações que podem ser instalados em qualquer computador pessoal.

O software DtSearch (<http://www.dtsearch.com>) é um desses sistemas. Ele é capaz de catalogar e recuperar todos os documentos armazenados em um computador pessoal, e pode ser executado em qualquer da linha IBM-PC que tenha o sistema operacional Windows 95 ou posterior instalado.

O DTSearch manipula diversos tipos de arquivos (ASCII/ANSI, HTML, PDF), provenientes de diversos editores (WordStar, WordPerfect, MSWord), bancos de dados (MSAccess) e planilhas (MSExcell), além de trabalhar com arquivos compactados (do tipo ZIP).

Esse software permite ainda a realização de buscas em linguagem natural, booleanas ou *fuzzy*. Além de conseguir identificar palavras com fonemas (através de stemming) ou conceitos similares (definidos em um thesaurus). Ele também é capaz de diferenciar palavras maiúsculas de minúsculas (se assim o usuário desejar), reconhecer palavras acentuadas, limitar a distância entre os termos e permite a utilização de *máscaras*.

## 11.2 Motores de busca na WEB

Apesar de existirem a algum tempo, foi somente com o desenvolvimento da WEB que os SRI tornaram-se populares e de acesso ao grande público. Atualmente a WEB oferece uma série de motores de busca (search engines) que podem ser utilizados para a localização de documentos na Internet (no caso, páginas WEB).

O Altavista™ (<http://www.Altavista.com>) é um dos motores de busca mais conhecidos e foi um dos primeiros a conquistar espaço e fama na WEB<sup>29</sup>. O objetivo do Altavista™ é indexar a WEB por inteiro. Atualmente ele possui cerca de 250 milhões de páginas indexadas [ALT 2000] e é capaz de indexar mais de 225 tipos de arquivos, incluindo bancos de dados, planilhas, documentos WEB e documentos textuais diversos.

A forma de indexação adotada por esse sistema é simples. Ele seleciona todas as palavras de uma página, inclusive as palavras geralmente consideradas stopwords.

Ele permite busca em linguagem natural, porém, aparentemente não realiza nenhuma análise sintática ou léxica. Ele consegue selecionar os substantivos e palavras mais relevantes de uma consulta em LN através de informações estatísticas, ou seja, ele desconsidera palavras de alta frequência. Logo, preposições, conjunções e advérbios são retirados da consulta e a busca é realizada com as palavras restantes.

É possível buscar trechos de documentos (frases ou termos compostos) ou palavras soltas. Pode-se utilizar operadores booleanos, máscaras ou limitar a busca a uma categoria específica de documento (imagens, vídeos, grupos de notícias ou WEB) ou a uma localização específica (site). O Altavista™ identifica também *variantes ortográficas* (*spelling variations*) e sinônimos, além de conseguir reconhecer erros de digitação (*typos*) e suportar múltiplas línguas.

O resultado da consulta é ordenado, onde as páginas que contém o maior número de palavras dentro das especificações da consulta são mostradas primeiro. Ele informa a quantidade de documentos que satisfaz o critério de busca, porém, essa informação não corresponde a um valor real, mas sim, a uma estimativa estatística (por razões de performance).

---

<sup>29</sup> Na verdade, um dos primeiros a serem implementados foi o WAIS (<http://www.w3.org/Gateways/WAISGate.html>), porém, ele é utilizado para indexar arquivos locais ou de *Intranets*, não indexando arquivos remotos.

A busca nem sempre é feita em todo o índice. Logo, dependendo do horário em que é realizada uma consulta, o resultado pode variar.

O Altavista™ possui agentes *spiders* que estão sempre vasculhando os *links* das páginas já indexadas em busca de novas páginas. Há ainda agentes que ficam monitorando as páginas já indexadas em busca de modificações.

Há uma versão comercial do Altavista™ [ALT 2000], que pode ser utilizada em corporações: o Altavista™ search engine 3.0. Ele só funciona em plataformas robustas (inclusive arquiteturas 64bit), cujo sistema operacional seja o windows NT, Linux ou Unix.

Um dos concorrentes do Altavista™ é o TodoBr. O TodoBr<sup>30</sup> é um motor de busca desenvolvido no Brasil cujo objetivo é o de indexar toda a WEB brasileira. Ele não oferece todos os recursos que o Altavista™ oferece mas permite a localização de páginas por palavras simples, compostas ou frases.

Além dos sistemas que utilizam motores de busca similares ao Altavista™, que indexam automaticamente todo o conteúdo dos documentos que encontram na WEB, há os sistemas que utilizam o modelo de agrupamento ou classificação de documentos, onde as páginas são geralmente indexadas manualmente em categorias ou hierarquias de assuntos predefinidos.

O Yahoo!<sup>31</sup> é um desses sistemas. Nele os usuários podem navegar pelas categorias existentes iniciando por temas ou assuntos mais abrangentes que vão se tornando mais específicos conforme a profundidade. Ao encontrar uma categoria de interesse (uma especificidade de interesse), o usuário pode então analisar as páginas nela catalogadas. Se o usuário desejar, ele também pode realizar uma busca tradicional por palavras-chave, só que essa busca é feita em cima das categorias, títulos e resumos de páginas lá catalogados.

A indexação em classes facilita a busca, já que o conjunto total de documentos diminui com a profundidade (ou especificidade) da categoria e as categorias são geralmente bem coesas e definidas. Porém, devido ao processo manual de categorização, esses sistemas contêm uma quantidade muito menor de páginas indexadas.

### 11.3 Sistemas de meta-busca

Os sistemas de *meta-busca* (*meta-search/meta-tools*) são sistemas que utilizam outras ferramentas ou sistemas de busca para realizar a localização de informações e então combinam os resultados para apresentar ao usuário uma resposta mais refinada.

A vantagem desses sistemas sobre os outros é que eles não necessitam indexar as informações nem armazená-las. Eles preocupam-se simplesmente repassar a consulta do usuário para outros sistemas (mapeando-as para o formato de cada sistema) e coletar rapidamente os resultados da busca que eles realizam. Após coletar os resultados, esses devem ser filtrados (pois cada sistema possui uma forma de apresentar os resultados) e completados (alguns oferecem informações complementares sobre os mesmos resultados).

Por utilizarem-se de vários sistemas para realizar a busca, as ferramentas de meta-busca acabam tendo uma cobertura bem maior da WEB (nem todos os sistemas indexam as mesmas páginas). Além disso, eles são capazes de avaliar melhor os resultados, considerando mais relevantes àqueles que são retornados por mais de um sistema.

---

<sup>30</sup> <http://www.todobr.com.br>

<sup>31</sup> <http://www.yahoo.com>

Na WEB podem ser encontradas várias ferramentas de meta-busca: MetaSEEK ([www.ctr.columbia.edu/metaseek](http://www.ctr.columbia.edu/metaseek)), Miner (<http://miner.bol.com.br/index.html>), MetaCrawler (<http://www.metacrawler.com/index.html>), WEBCrawler (<http://www.webcrawler.com>), entre outras.

Algumas, como o Copernic, podem ser instaladas na máquina do usuário, oferecendo uma série de vantagens. O Copernic (<http://www.copernic.com>) é um software que pode ser instalado em um computador pessoal da linha IBM-PC com o Windows 95 ou posterior instalado. O software submete a busca para vários motores de busca e armazena os resultados para análise e visualização posterior. Ele pode realizar atualizações periódicas, mantendo os resultados atualizados constantemente.

PARTE II – PROFUNDIDADE:  
TECNOLOGIAS DE DESCOBERTA DE CONHECIMENTO  
APLICADAS À INTELIGÊNCIA COMPETITIVA



## 12 Inteligência competitiva

O fenômeno da globalização, aliado a ambientes de troca de informações globais como a Internet, minimiza as fronteiras entre países permitindo que qualquer empresa atue em qualquer mercado, independente de sua localização física [ZAN 98]. Isso porque é possível obter informações sobre clientes, legislação e possíveis concorrentes de mercados que a empresa ainda não atua (o que eles produzem e como). Essas informações podem indicar se uma possível invasão de mercado pode ser feita e se ela ofereceria alguma vantagem competitiva para a empresa.

Além disso, em teoria, qualquer pessoa pode implementar uma empresa virtual na Internet, sendo assim capaz de concorrer com grandes empresas localizadas em um local fisicamente oposto do mundo.

Nesse cenário as empresas estão expostas em inúmeras situações e devem ser capazes de conquistar novos mercados ou até mesmo manter os atuais. Para tanto, os empresários devem estar sempre bem informados a fim de minimizar riscos, antecipar crises e tornar seus produtos mais competitivos. As empresas devem ser capazes de obter rapidamente informações (antes dos seus concorrentes) sobre todos os elementos internos e externos à empresa. Esses elementos, que incluem produtos, clientes, fornecedores, tecnologias, concorrentes e mercados, devem ser constantemente monitorados.

As informações internas são aquelas informações produzidas pela própria empresa [NAS 97] e que podem ser obtidas nos seus Sistemas de Informação – SI (cadastro de clientes, controle de estoque, relatórios de vendas...). Esses sistemas fornecem as informações básicas para o funcionamento da empresa.

Porém, em um estudo realizado no M.I.T, foi identificado que a grande maioria dos SI não está preparada para fornecer as informações necessárias ao processo de tomada de decisão [ROC 79]. Em muitos casos, os SI geram uma enorme quantidade de relatórios que são dificilmente digeridos ou contêm informações irrelevantes.

Com isso, torna-se necessário reformular os sistemas atuais, identificando as necessidades de informação (NI) que cada funcionário da empresa possui e fazendo com que esses sistemas supram essas necessidades. Uma vez identificadas as NI, torna-se necessário identificar as fontes de informação apropriadas (subsistemas internos ou bancos de dados externos) e integrá-las (muito possivelmente em um *Data-Warehouse*<sup>32</sup>).

Uma vez que essas fontes estejam integradas, podem ser aplicadas ferramentas de descoberta de conhecimento, capazes de analisar os dados e informações explícitas com o intuito de identificar co-relacionamentos entre eles, ou ainda, identificar informações implícitas (não passíveis de serem informadas diretamente pelos meios tradicionais da empresa).

Porém, grande parte das fontes de informação necessária à tomada de decisão (entre 80 e 90% [CLE 98]) não consta nos bancos de dados da empresa, mas sim, em fontes externas à empresa. Nesse caso, torna-se necessário utilizar técnicas provenientes da área de

---

<sup>32</sup> Data warehouses são coleções muito grande de dados, provenientes de diversas fontes e de diversas formas, cujo processamento não automatizado é completamente inviável. Eles são boas fontes de conhecimento e informação. *Data-warehouses* também podem ser considerados bibliotecas digitais, já que ambos são repositórios de informação e o seu objetivo primário é satisfazer as necessidades de informação do usuário. A construção de *data-warehouses* vem da necessidade das organizações controlarem a proliferação de informação digital conhecida e recuperável. Geralmente são focados em dados estruturados (próprios para descoberta de conhecimento) [KOW 97].

inteligência competitiva, que é a área da administração que estuda e provem métodos de coleta e análise de informações externas e as transforma em informações úteis para a tomada de decisão.

## 12.1 Definição da inteligência competitiva

A *Inteligência Competitiva* (*competitive intelligence / veille stratégique*) ou *inteligência dos negócios* (*business intelligence / intelligence économique*) é o nome dado ao conjunto de ações coordenadas de busca, tratamento, distribuição e proteção de informações utilizado para fins de vigilância e proteção do patrimônio, e que envolve diversos fatores externos à empresa (tecnológicos, comerciais, financeiros, econômicos, jurídicos, sócio-culturais e políticos) [CLE 98].

Os processos de inteligência foram muito utilizados durante a segunda guerra mundial e pelos departamentos de defesa americanos durante a guerra fria. Com o tempo, os profissionais que trabalhavam exclusivamente na área militar passaram a ser contratados pelas grandes empresas anglo-saxônicas, britânicas e americanas para atuar, principalmente, nas áreas de marketing, pesquisa e desenvolvimento.

Por isso, essa atividade consiste quase que exclusivamente na aplicação de métodos de vigília ao ambiente externo de uma empresa, buscando monitorar concorrentes, tecnologias e produtos.

Como a maioria das informações necessárias ao processo de inteligência está em fontes públicas, esse processo baseia-se na exploração de fontes abertas e na utilização de informações provenientes de produtores específicos de informação para negócios (tais como o SEBRAE, a RNIT/PADCT, o CNI/DAMPI, e os *trade-points*) [NAS 97]. Portanto, é possível obter muitas informações sem o risco de ser acusado de espionagem industrial.

A decisão de quem executa a atividade de inteligência varia de empresa para empresa. Em alguns casos pode haver um departamento específico para a coleta e análise de informações. Outra opção é utilizar um profissional da empresa que já trabalhe com a manipulação de informações, que possua um certo conhecimento de informática e que possa tomar decisões ou influenciar no processo de tomada de decisões (geralmente alguém da área do marketing ou da alta gerência). É possível, também, solicitar o serviço à empresas especializadas nesse tipo de processo [ZAN 98].

O mais importante nisso tudo é a empresa dispor de dados e fatos oportunos e confiáveis sobre o ambiente externo para que a administração possa prever as intenções dos concorrentes, antecipar medidas governamentais, desenvolver estratégias e tomar decisões. Todos esses dados contêm informações de grande valor que podem ser utilizadas para melhorar as decisões e otimizar o sucesso [GOE 99]. Essas informações podem se tornar conhecimento empresarial e patrimônio da empresa<sup>33</sup>.

---

<sup>33</sup> O mundo dos negócios nos últimos cinquenta anos tem sofrido uma transição do domínio do capital para o do conhecimento [INZ 2000]. O conhecimento (a respeito de produtos, mercados, tecnologias e organização) é um ativo que pode ser obtido em bases de conhecimento, arquivos e nas pessoas da empresa. Esse conhecimento, conforme Macintosh, citado em [INZ 2000], tornou-se um fator crucial para a promoção do crescimento econômico e para o êxito dos negócios. Logo, as empresas devem *saber que sabem* (ter meta-conhecimento) e ser capazes de maximizar o uso desse conhecimento.

## 12.2 Objetivos

Os objetivos da inteligência competitiva são, portanto, promover o saber-fazer tecnológico e científico de uma empresa, país ou região; detectar riscos e oportunidades no mercado exterior e interior; monitorar as ações dos concorrentes (modos de pensar, técnicas, cultura, intenções e capacidades) e definir estratégias e ações que devem ser tomadas a fim manter sua estabilidade [CLE 98].

As técnicas de inteligência competitiva devem ser capazes de sanar as necessidades básicas de informação das pessoas que tomam a decisão na empresa. Essas necessidades geralmente são expressas pelas seguintes questões [ZAN 98]:

- a) “Que áreas específicas de pesquisa e desenvolvimento precisam ser aprimoradas?”
- b) “Quais as tendências do mercado?”
- c) “Em que setores os nossos concorrentes estão preparando novos produtos para serem lançados no mercado?”
- d) “Quando o farão?”
- e) “Que setor eles abandonarão nos próximos anos?”
- f) “Temos segurança, considerando nossa experiência, para entrar em um novo mercado?”
- g) “Quais os “tijolos” tecnológicos essenciais para determinado setor?”

Essas questões auxiliam na identificação da direção em que o mercado se direciona, evitando que a empresa invista tempo e dinheiro no desenvolvimento de produtos que tenham pouca aceitação no mercado ou que tenham pouca vida útil.

## 12.3 Etapas do processo de inteligência

As etapas do processo de inteligência podem variar de empresa para empresa. Tudo depende de sua situação atual. Caso a empresa já conheça suas necessidades de informação ela pode partir diretamente para a coleta e monitoração da informação. Caso contrário, torna-se necessário identificar essas necessidades e, muito provavelmente, remodelar os SI da empresa.

Clerk comenta que o processo de inteligência inicia pela etapa de *coleta* de dados, passando pela *exploração*, *distribuição* e finalizando com a aplicação de mecanismos de *segurança* das informações e conhecimentos descobertos [CLE 98]. Já Zanasi aborda a questão da identificação das necessidades de informação, identificando as etapas de *compreensão do problema*, *definição de fontes*, *identificação de dados relevantes (pesquisa estratégica)*, *análise* e *interpretação* [ZAN 98].

Analisando-se as diferentes etapas existentes, propõe-se a utilização da seguinte metodologia, que é composta das seguintes etapas: Identificação da necessidade de informação, identificação de fontes, coleta, filtragem, distribuição, exploração e segurança.

- a) Identificação da necessidade de informação – nessa etapa deve-se identificar quais são as necessidades de informação de cada pessoa da empresa (principalmente dos tomadores de decisão), quais dessas necessidades a própria empresa pode *sanar* e quais necessitam de dados externos;

- b) Identificação e análise de fontes de informação – uma vez identificadas as necessidades de informação, torna-se necessário identificar onde essas informações podem ser recuperadas (as fontes). Essas fontes podem ser internas ou externas. No caso de fontes externas, deve-se descobrir o formato, o tempo de acesso e o custo das informações, assim como deve ser identificado como elas podem ser agregadas às informações já existentes na empresa;
- c) Coleta – é a busca, em si, da informação ou dos dados nas fontes identificadas;
- d) Filtragem – Devido à grande quantidade de dados e informações que podem ser coletadas, muitas podem não ser específicas às necessidades identificadas inicialmente. As informações irrelevantes devem ser descartadas e as relevantes selecionadas;
- e) Distribuição – os dados ou informações selecionadas devem ser encaminhadas às pessoas que as necessitam (que expressaram sua necessidade).
- f) Exploração – corresponde à transformação dos dados em informação e conhecimento. Para tanto podem e devem ser utilizadas ferramentas computacionais e métodos estatísticos de análise.
- g) Segurança – Depois de adquiridos os conhecimentos e informações, esses devem ser, obviamente, postos em prática (utilizados na tomada de decisão) e armazenados em algum local seguro para que não caiam nas mãos dos concorrentes.

Nessa metodologia sugerida, espera-se que a pessoa que recebeu a informação ou os dados saiba tratá-los, ou seja, é ela quem deve realizar a etapa de *exploração*. Eventualmente, essa etapa pode vir antes da distribuição, onde os dados seriam então analisados por um especialista ou departamento de inteligência e seus resultados é que seriam repassados para o tomador de decisão.

Uma consideração importante é a de que as fontes (os ambientes) devem ser constantemente monitoradas, mesmo que as necessidades de informações sejam sanadas. Isso porque toda vantagem competitiva é momentânea: uma mesma tecnologia está igualmente acessível a todos, não sendo possível sustentar vantagem competitiva por muito tempo [WEB 98]. Novos processos e produtos surgem a todo o momento e as necessidades de informação podem mudar. Logo, o processo de inteligência deve ser executado constantemente.

Nos capítulos seguintes serão abordados os métodos e ferramentas computacionais que podem ser utilizados no processo de análise (exploração) de dados e informações para a inteligência. Esses métodos são estudados, pesquisados e elaborados por uma área da computação que é denominada *Descoberta de Conhecimento*.

## 13 Descoberta de conhecimento

A tomada de decisão e o planejamento estratégico de negócio necessitam de muita informação [FEL 98]. Essa informação, muitas vezes, não está presente de forma clara no SI de uma empresa, mas sim, de forma implícita. Esse tipo de informação implícita ou até mesmo escondida não é obtida pelos métodos de recuperação tradicionais oferecidos pelos SI. Para que esse tipo de informação seja encontrado torna-se necessário aplicar algum método ou ferramenta de *Descoberta de Conhecimento (Knowledge Discovery)*.

Os métodos de descoberta de conhecimento surgiram dentro da área de Inteligência Artificial (IA). Porém, esses métodos necessitam de uma grande quantidade de dados para que sua utilização compense. Logo, os cientistas da área foram buscar em outras áreas da computação aplicações que dispusessem de dados em grande quantidade.

Já que os bancos de dados possuem uma grande quantidade de dados os métodos de descoberta de conhecimento passaram a ser aplicados em cima deles. Com isso, pouco-a-pouco esses métodos foram sendo adotados pelo pessoal da área de Sistemas de Informação, que descobriu que suas coleções de dados poderiam ser fontes valiosas de conhecimento.

Assim surgiu a área de *Descoberta de Conhecimento a partir de Dados (Knowledge Discovery from Data – KDD)*, que busca descobrir co-relacionamentos e dados implícitos nos registros de um banco de dados, estudando e desenvolvendo um processo de extração de conhecimento novo, útil e interessante, e apresenta-lo de alguma forma acessível para o usuário [FEL 97a].

Esse tipo de análise resolve em parte o problema da sobrecarga de informações, já que oferece meios automatizados para analisar e processar a quantidade superabundante de informações.

A descoberta de conhecimento foi primeiramente utilizada em dados estruturados. Esse tipo de informação é muito importante para que os empresários consigam identificar novos dados e conhecimentos que estejam, de alguma forma, implícitos ou escondidos nos seus SI e que não possam ser recuperados pelos meios tradicionais de recuperação oferecidos por eles.

Apesar de serem importantes para os empresários, as informações internas e estruturadas não são as únicas necessárias para eles. Para o processo de IC grande parte das informações é obtida em fontes externas à empresa. Essas fontes oferecem, na maioria dos casos, informações dispostas num formato sem estrutura ou semi-estruturado (informações textuais).

Esse tipo de informação textual não é tratado pelas ferramentas tradicionais de descoberta de conhecimento, pois possuem características que tornam sua análise complexa [ZAN 98]. Para que as etapas do processo de IC sejam aplicadas corretamente são necessárias técnicas e ferramentas computacionais desenvolvidas especificamente para tratar esse tipo de informação [GOE 99]. Essas técnicas e ferramentas são encontradas dentro da área de recuperação de informações e da área de descoberta de conhecimento em textos (*Knowledge Discovery from Text – KDT*).

Pelo fato da grande maioria das informações necessárias ao processo de IC ser de natureza textual, aliado ao fato de já existirem no Instituto de Informática da UFRGS muitos estudos sobre o processo de KDD [FEL 97b; NOG 00; PRA 97; PRA 98] (entre outros), esse exame de profundidade buscará abordar somente os métodos e ferramentas de KDT.

Porém, devido a esse processo estar muito relacionado com o processo de KDD (o KDT utiliza-se de muitas técnicas e metodologias de KDD), além do fato de muitas empresas possuírem informações estruturadas e necessitarem também das ferramentas de KDD para parte do processo de IC, iniciar-se-á apresentando as etapas básicas do processo de KDD, para após aprofundar-se na metodologia de KDT.

### 13.1 Etapas do processo de descoberta de conhecimento

O KDD é composto de uma série de etapas ou passos que buscam transformar os dados de baixo nível em conhecimento de alto nível [GOE 99].

O processo de KDD não é linear e é orientado à aplicação. Esse processo é iterativo e interativo [FAY 96], e composto de uma série de atividades ou etapas que requerem a intervenção do usuário.

Basicamente, as etapas (ou o ciclo de vida) do processo de KDD são o *pré-processamento*, a *mineração de dados (data-mining)* e o *pós-processamento* (que se constitui na seleção e ordenação das descobertas, elaboração de mapeamentos de representação de conhecimento e na geração de relatórios).

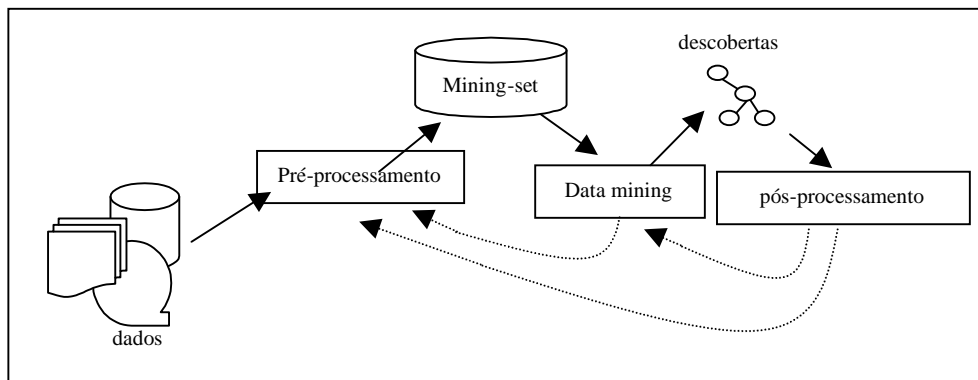


FIGURA 13-1 – O PROCESSO DE KDD (SIMPLIFICADO)<sup>34</sup>

Uma observação interessante é a de que o termo “mineração de dados<sup>35</sup> é comumente utilizado como sinônimo para todo o processo de descoberta de conhecimento. Porém, na verdade, ele é somente o núcleo do processo, correspondendo de 15 à 25% do processo total de descoberta [GOE 99], e é encarregado de realizar a extração de padrões e modelos nos dados observados. Muito provavelmente por isso, mineração de dados tornou-se o termo adotado comercialmente para o processo, enquanto que os acadêmicos preocupam-se mais em especificá-lo corretamente por KDD.

O pré-processamento inclui tudo o que é feito antes da mineração e inclui processos de análise, integração, transformações e limpeza dos dados existentes. Já o pós-processamento compreende a aplicação de filtros de estruturação e ordenação para que o conhecimento possa ser apresentado ao usuário de alguma forma mais simples e compreensível.

Existem muitos trabalhos que detalham cada uma destas etapas. O trabalho de Hércules Prado [PRA 97], por exemplo, contém uma descrição minuciosa de cada uma destas etapas, incluindo objetivos, problemas e vantagens do processo de KDD em geral.

<sup>34</sup> Fonte: [FEL 98, p936]

<sup>35</sup> O termo tem origem na área da Estatística, onde, por volta de 1960, os estatísticos davam o nome de *pescaria* ou *peneiramento* de dados ao processo de exploração descontrolada de dados – o que já não acontece atualmente, pois os processos estão bem mais definidos [PRA 97].

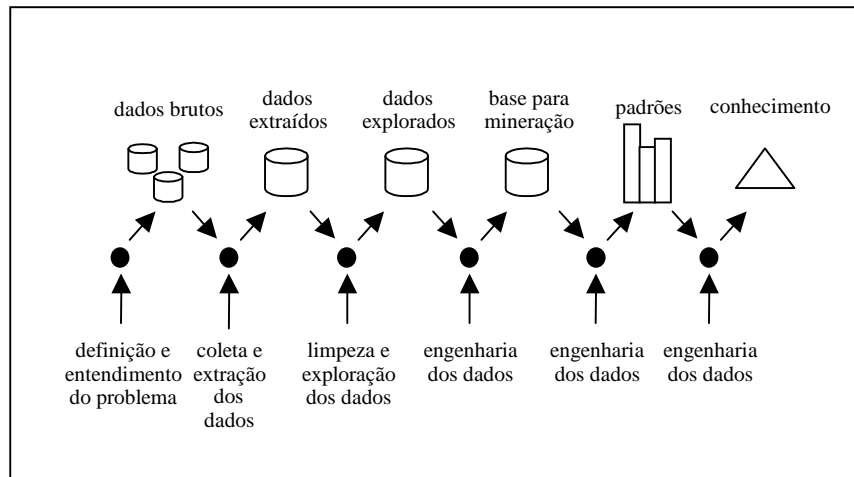


FIGURA 13-2 – O PROCESSO DE KDD<sup>36</sup>

Estas etapas podem ser refeitas ou retornadas, pois as descobertas realizadas (ou não) até o momento podem levar a novas hipóteses que podem modificar todo o processo. Existem muitas ferramentas que trabalham em cada uma dessas etapas, geralmente de forma separada. Isso requer muito trabalho do usuário a cada iteração, já que a necessidade por novas informações pode aparecer ao final de cada processo ou interação.

## 13.2 Mineração de dados

A mineração de dados (*Data-Mining* – DM) é a etapa mais importante do processo de descoberta de conhecimento. O objetivo dessa etapa constitui em descobrir pequenas informações úteis – os chamados *nuggets*. Um *nugget* é um pedaço de conhecimento que pode ser utilizado pela empresa no processo decisório ou estratégico [FEL 98].

Para que a DM funcione, é extremamente importante que a empresa ou pessoa encarregada de aplicar esse processo saiba exatamente o que deseja descobrir ou o que pode ser descoberto com os dados que possui. Os métodos de DM costumam apresentar uma série de nuggets, e é muito importante que a empresa saiba identificar quais deles podem ser realmente relevantes para a empresa.

Existem diferentes métodos que podem ser utilizados na mineração. São eles: o *processamento de dados*, a *predição*, a *regressão*, a *classificação*, o *agrupamento* ou *clustering*, a *análise de associações* e a *visualização* [GOE 99].

- a) Processamento de dados – são aplicados com o objetivo de selecionar, filtrar, agregar, exemplificar, limpar e transformar dados;
- b) Predição – buscam prever o valor de um atributo específico;
- c) Regressão – analisam a dependência de valores de alguns atributos em relação a outros no mesmo item, gerando um modelo capaz de prever os valores de novos registros;
- d) Classificação – determinam a que classe determinado dado pertence;
- e) Clustering (agrupamento) – particionam ou dividem um conjunto de itens em grupos de itens com características similares;
- f) Análise de associações – identificam relacionamentos entre atributos e itens (a fim de detectar se a presença de um padrão implica na presença de outro);

<sup>36</sup> Fonte: [PRA 97, p12]

- g) Visualização – tornam o conhecimento entendível (gráficos e animações 2D e 3D);
- h) Análise de dados exploratória (EDA) – que corresponde à exploração interativa de um conjunto de dados sem modelos ou hipóteses a priori, a fim de identificar padrões.

Esses métodos podem utilizar as mais diferentes abordagens (implementações). As abordagens mais comuns utilizam técnicas *estatísticas, raciocínio baseado em casos, redes neurais, árvores de decisão, indução de regras, redes de bayes, algoritmos genéticos, conjuntos difusos (fuzzy) ou conjuntos aproximados (rough sets)* [GOE 99].

Pode ser necessário aplicar os mesmos métodos repetidas vezes e com diferentes parâmetros até que obtenham a melhor performance ou os melhores resultados (resultados esperados). Esses métodos podem inclusive ser aplicados em conjunto e em diversas ordens, dependendo do que o usuário espera descobrir.

Cada um desses métodos possui um objetivo diferente, que pode ser:

- a) Descobrir dependência entre os dados, identificando os atributos que influenciam uns aos outros;
- b) Descrever conceitos, a fim de identificar os atributos comuns nos membros de uma classe, facilitando a construção de regras;
- c) Detectar desvios, para identificar os elementos ou objetos que se encontram fora das regras já definidas ou identificadas;
- d) Identificar clusters (aglomerados), para saber quais são os itens com características similares (ou com o mesmo perfil);
- e) Descobrir de fórmulas, (também) com o objetivo de descrever conceitos, porém identificando alguma função ou modelo matemático que descreva o domínio em questão;

Uma colocação importante é a de que o padrão encontrado em uma base de dados pode ser válido somente para ela, não sendo possível transpor seu comportamento ou conhecimento para outro domínio. É necessário identificar e estabelecer métricas a fim de saber o quanto do “modelo” extraído é acurado e pode ser transposto para outros domínios ou dados [PRA 97].

## 14 Descoberta de conhecimento em textos

A *Descoberta de Conhecimento a partir de Textos (Knowledge Discovery from Text – KDT)*, também conhecida por *Mineração de Textos (Text Mining)*, é relativamente nova como área, porém muitas de suas técnicas e métodos já existem há algum tempo.

Pode-se dizer que KDT é uma evolução natural da recuperação de informações, já que os SRI passaram a adotar algumas técnicas de análise de informações e de aprendizado de máquina, muitas das quais provenientes da área de descoberta de conhecimento em bases de dados. Assim, ao invés do usuário ter que analisar quais dos documentos retornados são realmente relevantes, o próprio sistema faria essa análise e retornaria as informações de forma condensada e resumida.

Descoberta de conhecimento em textos pode ser entendida como a aplicação de técnicas de KDD sobre dados extraídos de textos [FED 97]. Entretanto, cabe salientar que KDT não inclui somente a aplicação das técnicas tradicionais de KDD, mas também qualquer técnica nova ou antiga que possa ser aplicada no sentido de encontrar conhecimento em qualquer tipo de texto. Com isso, muitos métodos foram adaptados ou criados para suportar esse tipo de informação semi-estruturada ou sem estrutura, que é o texto.

A KDT vem solucionar grande parte dos problemas relacionados à busca, recuperação e análise de informações. A sobrecarga de informação, um dos maiores problemas enfrentados pelos usuários da Internet, é um desses problemas. Sistemas de informação que ofereçam características de KD podem beneficiar os empresários e os países, auxiliando-os a coletar e analisar os dados necessários à tomada de decisão e permitindo com que se posicionem melhor no mercado.

### 14.1 Etapas do processo de descoberta de conhecimento em textos

Por ser relativamente nova como área, ainda não existem livros que tratem especificamente da descoberta de conhecimento em textos, o processo de KDT e suas etapas. No entanto, muitos métodos e ferramentas que fazem alguma espécie de KDT, e que podem ser enquadrados na área, realizam alguns procedimentos que são similares. Fazendo-se uma análise técnica nos artigos e revistas correlacionados com o KDT consegue-se identificar que há uma metodologia para o processo de KDT. Essa metodologia é análoga à metodologia de KDD e muito similar à metodologia de IC.

As etapas que compõem essa metodologia são descritas a seguir. Apesar de ter sido identificada como metodologia, nem todos os modelos de descoberta de conhecimento a seguem a risca. Em alguns casos, determinadas etapas simplesmente não são realizadas. Em outros, etapas adicionais são utilizadas. Metodologia, portanto, provavelmente não seja o nome mais adequado. Uma denominação mais adequada seria guia, já que indica que caminho deve ser seguido para a elaboração de um método ou software de KDT.

Basicamente, as etapas mais importantes do processo de descoberta são as seguintes:

- a) Definição de objetivos - compreensão do domínio, identificação de o quê deve ou o quê pode ser descoberto;
- b) Seleção de um subconjunto de dados - nem todas as informações disponíveis devem ser utilizadas. A utilização de muitas informações pode influenciar negativamente no resultado de uma descoberta, além de torna-lo mais demorado;
- c) Pré-processamento ou limpeza dos dados - com o intuito de remover ruídos e preparar o dados. Em textos, principalmente, vários métodos podem e devem ser

aplicados. Esses métodos vão desde a limpeza de caracteres indesejados, passando pela correção ortográfica e morfológica e indo até a análise semântica e normalização de vocabulário. Esse é um ramo muito amplo e aberto para pesquisas e softwares;

- d) Redução ou projeção dos dados (escolha de características relevantes para a análise) – Nem todas as características (palavras) são importantes. Dependendo do objetivo da análise, partes de um documento ou conjunto de documentos podem ser mais importantes do que outras. Somente estas devem ser analisadas para que o resultado não seja inútil e para que o processamento seja mais eficiente (rápido);
- e) Escolha da técnica, método ou tarefa de mineração – Existem vários métodos de descoberta (ver seção seguinte), cada um deles capaz de descobrir algo diferente. Dependendo do que se deseja descobrir, alguns métodos não são necessários;
- f) Mineração – A aplicação do(s) método(s) escolhido(s);
- g) Interpretação dos resultados (podendo se retornar aos passos anteriores);
- h) Consolidação do conhecimento descoberto (documentação ou incorporação dos dados em um sistema) e aplicação prática do mesmo (por exemplo, na definição ou redefinição da estratégia de uma empresa).

Assim como no KDD, o processo de descoberta de conhecimento em textos é interativo e iterativo, correspondendo à aplicação repetida de métodos de mineração e interpretação dos resultados pelo usuário.

Essas etapas não necessitam necessariamente estarem relacionadas com a manipulação de textos. É possível, em qualquer etapa (mas geralmente nas iniciais), extrair características (ou conhecimentos) dos textos, colocá-las em *templates* e aplicar algum método tradicional de KDD.

Assim, até mesmo diferentes tipos de dados e técnicas poderiam ser combinados em um processo (mais genérico) de descoberta de conhecimento.

Seguindo esse princípio, alguns pesquisadores sugerem fazer o KDD nos seguintes moldes [DIX 97]:

- a) Recuperação de Informações – localização e recuperação dos textos relevantes para o que o usuário necessita descobrir;
- b) Extração de Informação – identificação de itens (características, palavras) relevantes nos documentos (geralmente o usuário deve, de alguma forma, estabelecer quais são estes itens e como eles podem ser identificados). Esses itens devem ser *extraídos* e convertidos em dados (tabelas ou *templates*) que possam ser utilizados pelos métodos KDD tradicional.
- c) Mineração – aplicação de uma técnica ou método de mineração que identifique padrões e relacionamento entre os dados;
- d) Interpretação: A interpretação e a aplicação dos *nuggets* identificados.

## 14.2 Tipos de descoberta de conhecimento em textos

A seguir, são discutidas várias abordagens ou técnicas para descoberta de conhecimento em textos, identificadas na literatura. As abordagens de descoberta discutidas permitem extrair conhecimento tanto na forma de informações (por mecanismos de dedução) quanto na forma de regras (por indução).

Algumas utilizam o aprendizado supervisionado e outras o não supervisionado [PRA 97]. Diz-se que o aprendizado é supervisionado quando o processo de aprendizado baseia-se em exemplos. Os exemplos e os resultados esperados são apresentados ao algoritmo de aprendizado. Esse tenta adaptar-se a fim de produzir o resultado esperado para cada exemplo. O processo é repetido até que o algoritmo apresente os resultados esperados com o mínimo de erro possível. O aprendizado é chamado de não supervisionado quando se possuem os dados, mas não se possuem modelos ou exemplos que possam ser ensinados ao algoritmo. Nesse caso, o algoritmo fica encarregado de identificar alguma espécie de relacionamento entre os dados. Os relacionamentos identificados são apresentados a um especialista que deve então validar a relação e encontrar algum significado para ela.

O aprendizado supervisionado costuma apresentar resultados melhores e mais refinados do que o não supervisionado. Por outro lado, ele só funciona se existirem exemplos e resultados esperados conhecidos.

Os métodos/tipos mais comuns de descoberta de conhecimento em texto, encontrados na literatura, são: extração de informações, sumarização, clustering e classificação ou categorização. Porém, qualquer um dos métodos de descoberta tradicional pode ser aplicado nos textos, principalmente se for utilizado o método de extração de informações, que identifica informações relevantes nos documentos e coloca-as em um formato estruturado.

### 14.2.1 Extração de informações

As técnicas de *Extração de Informações (EI)* não possuem uma classificação muito bem definida. Elas podem ser enquadradas na área de recuperação, pois são compreendidas algumas vezes como técnicas especiais de indexação ou por extraírem de um texto ou conjunto de textos somente as informações mais relevantes para o usuário. Por outro lado, se não fossem extraídas, talvez essas informações não fossem facilmente identificadas pelo usuário (poderiam estar implícitas ou passar despercebidas). Vistas dessa forma, elas são enquadradas na área de descoberta de conhecimento.

A EI surgiu, na verdade, dentro da área de *Processamento de Linguagem Natural (PLN)* [SCA 97]. Ela possui muitos componentes de um sistema de PLN. Muitos desses componentes também são utilizados por SRI para indexar documentos.

Por esse motivo o processo de extração torna-se muito parecido com o processo de indexação de informações, mas há algumas diferenças. A indexação busca identificar palavras capazes de caracterizar o documento e coloca-las em um índice. Já a extração, que também identifica palavras importantes, difere-se da indexação pelo fato de focar conceitos específicos e conter um processo de transformação que modifica a informação extraída em um formato compatível com o de um banco de dados alvo [KOW 97].

Colocado assim dessa forma, pode-se afirmar que o objetivo de um processo de extração é o de transformar dados semi-estruturados ou desestruturados (os textos) em dados estruturados (geralmente registros que possuem *slots* predefinidos que devem ser preenchidos) que serão armazenados em um banco de dados [KOW 97]. Mas na verdade a EI

possui um objetivo mais amplo, que é o de extrair tipos específicos de informações a partir dos textos [SCA 97].

Uma vez estruturadas essas informações podem ser utilizadas para outros fins (tornando-se simples dados de entrada), inclusive em processos tradicionais de descoberta de conhecimento. Por isso, a extração de informações é geralmente utilizada como um processo anterior à etapa de mineração (que é o miolo da descoberta de conhecimento), sendo considerada uma etapa de pré-processamento.

O processo de extração é um pouco mais simplificado do que o PLN, pois não utiliza todos os módulos normalmente necessários para fazer uma análise de língua natural completa [SCA 97]. Apesar disso, os *sistemas de extração de informações (SEI)* também possuem os mesmos problemas dos sistemas de PLN.

Para exemplificar, os SEI exigem muito conhecimento do domínio e acabam se tornando muito dependentes da aplicação, não sendo possível, na maioria dos casos, aplica-los em outros domínios.

Isso porque para que a extração se realize é necessário definir que informações (palavras) devem extraídas e como. Esse processo é feito através da identificação de *tags* (marcas) sintáticas ou semânticas que indicam a presença de uma informação importante e que deve ser extraída.

Na área da saúde, por exemplo, números seguidos de constantes do tipo “mg” indicam a quantidade (dosagem) de medicamento que deve ser ministrada a um paciente. Fazendo-se uma análise dos prontuários também é possível descobrir que o nome do medicamento geralmente encontra-se à direita ou nas proximidades de sua dose.

Da mesma forma, datas são identificadas pelo formato “99 / 99 / 99”, onde o padrão “99” indica um número de dois dígitos, sendo o primeiro correspondente ao dia, o segundo ao mês e o terceiro ao ano (valores entre 0 e 31 para o dia, 0 e 12 para o mês e 0 e 99 para o ano).

Isso indica a possibilidade de construção de regras de extração. Algumas destas regras são genéricas, enquanto que outras são específicas do domínio (as datas, por exemplo, podem ser representadas de forma diferente em alguns países).

Existem alguns sistemas utilizam essas regras como entrada e realizam a extração de acordo [SCA 97]. Porém as linguagens de definição de regras não costumam ser completas e não permitem que o usuário especifique regras muito complexas.

Essas linguagens e ferramentas estão em constante desenvolvimento. Espera-se que um dia elas sejam capazes de trabalhar com estruturas complexas de documentos, de forma quase que automática e com um baixo custo computacional.

### 14.2.2 Sumarização

A *sumarização* [FAY 96] é uma técnica que identifica as palavras, palavras e frases mais importantes de um documento ou conjunto de documentos e gera um resumo ou sumário. Esse sumário pode dar uma visão geral do conjunto de documentos ou pode ainda salientar as partes mais importantes e interessantes. Desta forma o usuário pode identificar rapidamente o assunto abordado por um documento ou conjunto de documentos sem ter que lê-lo(s) na íntegra.

A sumarização pode utilizar técnicas de extração de informações para gerar os sumários [KOW 97]. Neste caso, porém, os documentos devem pertencer a um domínio específico para que as regras de extração possam ser identificadas e funcionem corretamente.

Isso porque textos de domínios diferentes variam muito, possuindo estruturas e vocabulários diferentes, dificultando a identificação de regras que tenham validade em todos os domínios.

Um exemplo disso é o experimento de extração realizado por Bernard Moulin [MOU 92]. No seu caso, os textos eram específicos da área jurídica (leis e decretos) e obedeciam a certos critérios. Assim, foi possível que ele extraísse regras de legislação e relações entre elas. Ao invés de apresentá-las de forma resumida, ele as armazenava em uma base de conhecimento para consultas futuras.

Outra forma de sumarização que é muito empregada após processos de clustering (ver a seção seguinte) é a análise de centróide. O *centróide* é um termo proveniente da física que indica o centro de gravidade ou de forças de um objeto. Em uma figura geométrica o centróide corresponde ao seu ponto de equilíbrio. No caso de clustering de documentos o centróide corresponde ao conjunto de palavras estatisticamente mais importantes de um cluster (grupo). Esse centróide é geralmente utilizado para representar o grupo. Essas palavras mais importantes dariam para o usuário uma visão geral do assunto tratado no documento ou conjunto de documentos.

### 14.2.3 Clustering (agrupamento)

O *clustering* (ou *agrupamento*) é um método de descoberta de conhecimento utilizado para identificar co-relacionamentos e associações entre objetos, facilitando assim a identificação de classes [WIV 98a]. No caso de documentos, o clustering identifica os documentos de assunto similar e aloca-os em um grupo, gerando grupos de documentos similares. Esse método é extremamente útil quando não se tem uma idéia dos assuntos (das classes) tratados em cada documento e deseja-se separá-los por assunto.

Essa técnica é geralmente utilizada antes de um processo de classificação, facilitando a definição de classes, pois o especialista pode analisar os co-relacionamentos entre os elementos de uma coleção de documentos e identificar a melhor distribuição de classes para os objetos em questão. Isso significa que não há a necessidade de se ter conhecimento prévio sobre os assuntos dos documentos ou do contexto dos documentos. Os assuntos e as classes dos documentos são descobertos automaticamente pelo processo de agrupamento.

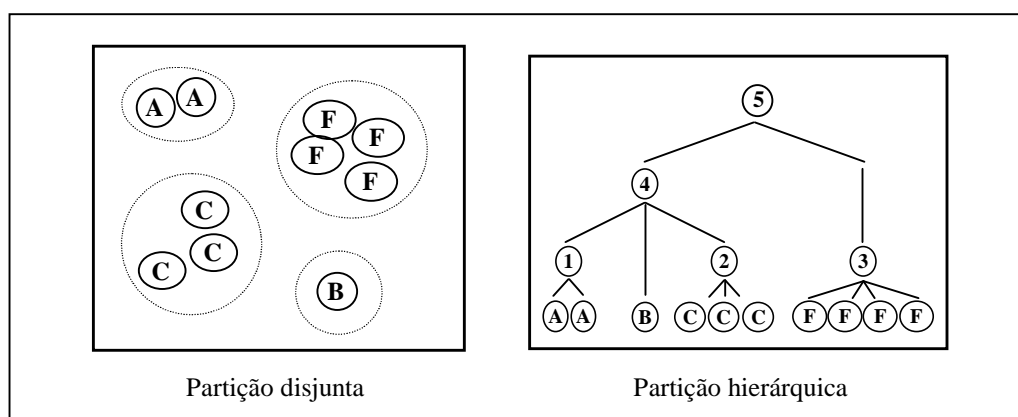


FIGURA 14-1 – TIPOS DE AGRUPAMENTO

O clustering pode gerar topologias de grupos *isolados* ou *hierárquicos* [CUT 92]. No primeiro caso, um algoritmo de *partição* é aplicado à coleção de documentos e estes são colocados em grupos distintos, geralmente não havendo espécie alguma de relacionamento entre os grupos identificados. Já no segundo, pode haver algum relacionamento ou ligação entre os grupos. Nesse caso, o processo de identificação de grupos é aplicado recursivamente

e acaba gerando uma espécie de árvore onde as folhas representam os grupos mais específicos e os nodos intermediários representam grupos mais abrangentes.

Cada uma destas topologias possui suas vantagens e desvantagens. No primeiro caso, disjuncto, não há estruturas que indiquem o co-relacionamento entre os grupos, impossibilitando o usuário de identificar os assuntos mais específicos e os mais abrangentes. Esse problema é solucionado pelo segundo caso, que oferece estruturas de navegação hierárquica entre os grupos, facilitando a localização de informações. Essa vantagem exige um tempo de processamento maior, já que o algoritmo de clustering deve passar analisar os grupos identificados várias vezes, tornando-se uma desvantagem. Outra desvantagem do método hierárquico diz respeito à manutenção dos clusters, que é mais complexa [KOW 97].

Em geral (hierárquico ou não) o clustering possui diversas aplicações. Ele pode ser utilizado para facilitar a organização e a recuperação de informações ou em outros processos de análise textual que visam a descoberta de conhecimento a partir de textos.

A recuperação de informações é facilitada porque o método desenvolvido consegue processar uma grande quantidade de documentos (de assuntos diversos) e agrupá-los em *clusters* de documentos de assuntos similares. Os grupos de documentos similares são armazenados em um mesmo local no arquivo de dados e indexados de forma que todo um cluster seja recuperado quando um dos documentos que fazem parte dele for considerado relevante a uma consulta.

Já na área de descoberta de conhecimento em textos o agrupamento é comumente utilizado no processo de descoberta de associações entre palavras, facilitando o desenvolvimento de dicionários e thesaurus. Esses dicionários podem ser utilizados em ferramentas de busca ou editoração de documentos, expandindo consultas ou padronizando o vocabulário dos documentos em edição.

Os grupos identificados também podem ser utilizados em alguns processos de identificação de características relevantes (espécie de sumarização), capazes de identificar o padrão e, em diferentes períodos de tempo, as tendências dos grupos (ou seja, as características que mudam com o decorrer do tempo).

#### 14.2.4 Classificação e categorização

A Classificação [LEW 91; YAN 99] é uma técnica empregada para identificar a que classe ou categoria determinado documento pertence, utilizando como base o seu conteúdo [NOR 96]. Para tanto, as classes devem ter sido previamente modeladas ou descritas através de suas características, atributos ou fórmula matemática [RIJ 79, cap3].

Classificação e categorização podem ser consideradas processos análogos, porém, alguns autores preferem distingui-los [LOH 2000a]. Esses autores consideram a categorização como sendo um processo que identifica as categorias (nesse caso consideradas como sendo um assunto ou tema) que um documento contém ou se enquadra. Esse processo é basicamente similar ao de classificação, mas sua aplicação é um pouco diferente, já que o primeiro identifica a classe a que o documento pertence e o segundo identifica quais são os assuntos de que o documento trata (esse último também é conhecido por *topic spotting* [WIE 95]).

Os processos de classificação/categorização também têm diversos propósitos. Eles podem ser utilizados em SRI para auxiliar o processo de indexação, identificando os tópicos nos quais os documentos devem ser alocados/indexados. Podem ser ainda utilizados por qualquer sistema ou técnica que necessite de uma pré-filtragem das informações, tais como

sistemas de extração de informações, sistemas de recomendação de informações e leitores de e-mails ou de notícias eletrônicas.

Os sistemas de classificação de objetos geralmente utilizam uma das seguintes técnicas:

- a) Regras de inferência – baseadas em um conjunto de características que devem ser encontradas no documento para que esse seja identificado como pertencendo a determinada categoria. Necessitam de muito tempo para serem elaboradas (esse processo é geralmente manual) e devem ser adaptadas caso o domínio mude. Geralmente são desenvolvidas para uma tarefa e domínio específico. O conhecimento modelado é facilmente compreendido (por estar em formato de regra) e seus resultados são, na maioria dos casos, melhores do que os apresentados pelos outros métodos (maiores informações em [APT 94; LEW 94]);
- b) Modelos conexionistas (redes neurais artificiais) – Esses sistemas induzem automaticamente um modelo matemático ou um conjunto de regras a partir de um *corpus* (conjunto de documentos) de treinamento. Podem ser colocados em prática rapidamente e são capazes de se adaptar as mudanças do ambiente de dados. Eles não necessitam de um especialista ou pessoa perdendo tempo na análise do domínio. Por outro lado, necessitam do conjunto de treinamento e seu modelo ou regras não são tão compreensíveis (maiores informações em [WIE 95]);
- c) Método de similaridade de vetores ou de centróides – nesse caso as classes são representadas por vetores (conjuntos) de palavras (denominados centróides). O documento é comparado com o vetor descritivo de cada classe. A classe que apresentar maior similaridade com o documento é tomada como classe do documento [LOH 2000a].
- d) Árvores de decisão – Uma abordagem parecida com a primeira, porém, utiliza técnicas de aprendizado de máquina para induzir as regras. Para cada classe uma árvore de decisão é criada [APT 94; LEW 94; YAN 99].
- e) Classificadores de Bayes – Parecidos com os conexionistas, porém têm como base a teoria da probabilidade. Eles conseguem informar a probabilidade de determinado documento pertencer a uma determinada classe [KOW 97].

#### 14.2.5 Filtragem de informação

A filtragem de informação é uma espécie de classificação de informações. Apesar disso, optou-se por dar-lhe uma atenção especial pelo fato de haver um grande interesse da comunidade científica da computação em desenvolver sistemas específicos de filtragem, encaminhamento e recomendação de informações. Esses estudos, apesar de terem os mesmos fundamentos da classificação, não costumam ser enquadrados por seus autores na área de classificação.

Dentro da área de filtragem podem ser identificados dois tipos de sistemas: os sistemas de recomendação e os sistemas de filtragem colaborativa. Ambos<sup>37</sup> possuem o

---

<sup>37</sup> Também é possível chamá-los de SRI, já que também retornam informações relevantes para o usuário. A diferença é que em um SRI tradicional o usuário recebe a informação no momento em que ele a solicita (ou seja, o funcionamento ocorre sob demanda e de forma específica), enquanto que nos sistemas de recomendação ou filtragem a informação é constantemente enviada para o usuário (assim que um novo item chega, o sistema o

objetivo de selecionar as informações e enviar para o usuário somente as que ele possui interesse (Kowalski denomina esse método de *disseminação seletiva de informação* [KOW 97]).

Os Sistemas de Recomendação (*Recommender/Recommendation Systems*) [RES 97] são sistemas capazes de analisar uma série de alternativas e escolher somente aquelas que são de alguma forma úteis para o usuário. Um Sistema de Recomendação poderia, por exemplo, analisar a sinapse de filmes que estão em cartaz em determinado local e indicar para o usuário o filme mais adequado para ele. Da mesma forma, os sistemas de recomendação podem analisar e recomendar restaurantes, bares, livros, revistas ou os candidatos mais adequados a um determinado cargo de uma empresa.

Os sistemas de recomendação também podem se basear na análise de outros usuários que já tenham formado alguma opinião sobre o objeto de análise. Ou seja, se um documento, um local ou um objeto é bem recomendado pelos outros usuários, é bem provável que ele também seja interessante para o usuário para o qual o sistema está realizando a análise. Da mesma forma, se um objeto, local ou documento é muito procurado, isso pode significar que esse objeto, local ou documento é muito interessante e deve ser recomendado.

Decorrente dessas duas propostas apresentadas, existem dois modelos ou tipos de sistemas de recomendação: *por conteúdo* e *colaborativo* [BAL 97]. O primeiro, denominado *por conteúdo*, busca recomendar itens similares aos itens que o usuário gostou no passado. O segundo, *colaborativo* propriamente dito, identifica usuários cujos gostos sejam similares ao do usuário atual e recomenda itens que eles tenham gostado.

Os Sistemas de Filtragem Colaborativa (*Collaborative Filtering*) são sistemas da mesma classe dos Sistemas de Recomendação, porém, mais simples no sentido em que não são eles que fazem a análise do objeto em questão. O que estes sistemas fazem é analisar as recomendações postas por outros usuários do sistema, filtrando as recomendações adequadas e encaminhando-as aos usuários interessados.

Alguns sistemas recebem recomendações de usuários que já tenham utilizado determinado serviço ou objeto e tentam agrupa-las e encaminha-las aos usuários que estejam necessitando destas recomendações. Esses são os sistemas mais simples. Há outros (ReferralWeb, GroupLens, PHOAKS) capazes de analisar mensagens de *news groups* ou arquivos de *Bookmark* (SiteSer) em busca de URLs interessantes. Há ainda os sistemas que empregam técnicas de análise de conteúdo (como é o caso do FAB). Nestes sistemas, recomendações negativas são excluídas ou filtradas (ver próxima seção).

---

analisa e o compara com uma espécie de *profile* que indica quais são seus interesses. Se ela for enquadrado em um dos tipos especificados no *profile*, ele é enviada para o usuário.

## 15 Sistemas de descoberta de conhecimento em textos

A seguir encontram-se alguns sistemas que utilizam alguma técnica de descoberta de conhecimento para o seu funcionamento interno, ou que podem ser utilizados, de alguma forma, durante o processo de KDT para a identificação de conhecimento útil.

### 15.1 Phoaks

O sistema PHOAKS (People Helping One Another Know Stuff) [TER 97] é um sistema desenvolvido com o intuito de facilitar a localização de informações de alta qualidade na WEB. Ele utiliza a abordagem de filtragem colaborativa em cima de mensagens da Usenet. O sistema vasculha as mensagens em busca de URLs. Ao ser encontrada uma URL o sistema tenta identificar se ela faz parte de uma recomendação através de uma série de testes pré-definidos (segundo os autores ela obtém sucesso em 90% dos casos). Quanto mais mensagens recomendarem o mesmo endereço (URL), mais significantes elas são para o sistema. Naturalmente as recomendações devem ser de pessoas distintas para terem mais valor. A premissa do sistema é que os *newsgroups* estão cheios de pessoas que recomendam recursos (objetos, páginas WEB, etc) para outras pessoas. Essas recomendações nada mais são do que análises sobre o referido recurso indicando sua utilidade.

### 15.2 Referral Web

Referral Web [KAU 97] é um sistema de Recomendação capaz de indicar pessoas *experts* em determinada área ou assunto, além de analisar co-relacionamentos entre pessoas. O software baseia-se na premissa de que as redes informais de troca de informação, constituídas de amigos e colegas de trabalho, são os canais mais efetivos de disseminação de informação.

Esse sistema utiliza portanto essas redes de informação como fonte para descobrir quais são as pessoas mais experientes em determinado assunto. Ele parte da premissa de que as pessoas mais experientes são aquelas mais citadas ou requisitadas na rede informal.

Qualquer rede informal *on-line* (e-mail, *newsgroups*, página WEB) poderia ser utilizada pelo Referral Web como fonte de informação. Algumas destas são mais confiáveis e específicas do que outras (a troca de e-mails constante entre duas pessoas pode indicar um forte relacionamento). Porém, em alguns casos, há o problema da invasão de privacidade. Logo, o sistema utiliza páginas WEB como fontes de informação.

O funcionamento do sistema é simples: um usuário registra-se no sistema e, a seguir, todas as páginas que citam ou mencionam esse usuário são identificadas (através de uma busca no Altavista™). Nessas páginas, todos os nomes citados em *links*, listas de autores em artigos ou citações, ou ainda, organogramas, são adicionados à rede particular do usuário. Essa rede, recém construída, é adicionada a rede global do sistema.

A rede global pode ser utilizada no processo de localização de pessoas ou documentos. Podem ser realizadas consultas do tipo “quem se relaciona com o *fulano*?” ou “com quem eu estou relacionado?”, que indicam o relacionamento entre pessoas. É possível consultar também quem são as pessoas que conhecem determinado assunto ou ainda construir consultas mais complexas, do tipo: “quais são os documentos que pertencem ao tópico *x* e que se relacionam com a pessoa *y*?”

### 15.3 Fab

O sistema FAB tenta combinar recomendação colaborativa com a baseada em conteúdo (ver seção 14.2.5). Os usuários recebem recomendações sobre itens quando o sistema encontra algum item que “case” (*match*) com o seu *profile* ou com o profile de um usuário similar. Esse sistema faz parte do projeto de biblioteca digital da universidade de Stanford. Conhecendo as relações (similaridades) entre os usuários e os seus gostos (profiles) o sistema consegue identificar clusters (grupos) de usuários. De tempos em tempos o sistema realiza buscas na WEB por assuntos de interesse de cada um dos grupos existentes na população de usuários. Depois o sistema encaminha as páginas para os usuários interessados. Depois de receber as recomendações os usuários são convidados a avalia-las. A avaliação é utilizada para adaptar (update) o profile do usuário e os profiles de coleções de usuários (usuários de um mesmo cluster). O sucesso da ferramenta depende da construção correta do profile dos usuários [BAL 97].

### 15.4 Siteeer

O Siteeer [RUC 97] é um sistema que recomenda páginas WEB baseando-se no arquivo de *bookmarks* do usuário. Ele considera o *bookmark* como sendo uma declaração implícita dos interesses do usuário, pois contém as páginas que ele considera mais interessantes. Da mesma forma, a disposição dos links no bookmark do usuário (a classificação em subfolders) reflete o relacionamento entre as páginas, e, portanto, a semântica que o usuário atribui a seus assuntos de interesse.

Baseando-se na análise do bookmark do usuário, o Siteeer consegue classificar novas páginas em uma das subcategorias (subfolders) do usuário. Com isso, o sistema pode recomendar páginas que se enquadrem em uma das categorias de interesse do usuário (expressas e definidas pelos seus subfolders).

Apesar dos links do bookmark estarem (muito bem) organizados de acordo com os interesses do usuário (pois os bookmarks não são construídos ao acaso, mas sim, intencionalmente pelo usuário), nem sempre os usuários adicionam um link interessante ao seu bookmark. Isso pode ocorrer, entre outros fatores, pelo fato do usuário crer que o link seja facilmente acessível. Esses fatores podem ocasionar uma má recomendação ou simplesmente recomendação alguma, onde um site interessante pode não ser recomendado.

O Siteeer também é capaz de identificar comunidades de usuários, comparando interesses (bookmarks) de usuários diferentes. Com isso, é possível identificar que determinados usuários podem ser utilizados como fonte de recomendação de outros. Esse fato permite com que o sistema utilize novas inclusões de páginas em bookmarks dos usuários da vizinhança virtual como fontes de recomendação.

O sistema é recomendado para usuários com um bookmark já pronto e não traz benefícios para usuários que estejam iniciando um bookmark (pois não consegue obter informações para isso). Além disso, ele só funciona a partir do momento que identifica uma *comunidade* a qual o usuário pertença.

### 15.5 GroupLens

GroupLens é um sistema que analisa informações de newsgroups e faz filtragem colaborativa. Isso é particularmente interessante em grupos de discussão (newsgroups) porque a quantidade de mensagens diária é muito grande. Um usuário experiente pode não estar

interessado em mensagens básicas. E o usuário iniciante pode não estar interessado em mensagens complexas.

O sistema foi desenvolvido de maneira tal que pudesse ser integrado com os sistemas de news existentes. Os usuários são requisitados a darem uma nota às mensagens que recebem (0 – ruim ... 5 – excelente). Com o tempo, o sistema consegue *rankear* as mensagens, identificar as mensagens mais relevantes para cada usuário ou grupo de usuários. O sistema coloca ao lado do cabeçalho da mensagem um valor correspondente ao quanto ela pode ser interessante para o usuário [KON 97].

## 15.6 Umap

UMAP (<http://www.trivium.fr>) é um software capaz de analisar um documento ou um conjunto de documentos e identificar seus tópicos (as palavras, na verdade) mais relevantes. Com isso espera-se que o usuário reconheça rapidamente e facilmente o assunto tratado neste(s) documento(s).

O UMAP constrói gráficos onde cada palavra é uma *ilha*. O *tamanho* de cada ilha é determinado pela relevância de cada palavra. Ilhas maiores representam palavras mais importantes. Geralmente o gráfico mostra arquipélagos (grupos) de ilhas, significando que as palavras do arquipélago estão correlacionadas de alguma forma. As distâncias entre as ilhas também indicam seu grau de co-relacionamento. As ilhas posicionadas mais ao centro indicam os tópicos principais do documento (ou conjunto de documentos).

O software possui diversas versões: *UMAP for WEB*, *UMAP for Outlook* e *UMAP Universal*. O primeiro trabalha com páginas recuperadas na WEB. Nesse caso o usuário especifica palavras-chave e o UMAP localiza as páginas na WEB que as possui. Ele faz isso realizando buscas em motores de busca da WEB (Altavista, por exemplo). Após coletar as páginas ele monta o gráfico de ilhas (denominado mapa), contendo todos os tópicos principais e secundários relacionados com a busca.

Da mesma forma, o UMAP for Outlook serve para gerar mapas de tópicos encontrados em mensagens eletrônicas (e-mails). O UMAP universal também gera mapas, porém de documentos localizados no computador do usuário (formato MS-Word).

Ao se selecionar uma ilha (uma palavra) o UMAP gera uma lista de todos os documentos que contém aquela palavra. O inverso também é possível: ao se selecionar um documento o software realça as ilhas (palavras) que ele contém.

O UMAP também consegue listar as palavras mais freqüentes e as menos freqüentes de um mapa. As palavras mais freqüentes são consideradas os tópicos principais e as palavras menos freqüentes os tópicos específicos (em se tratando de um único documento).

O UMAP pode ser utilizado para realizar a análise rápida de um conjunto de documentos. Assim, uma série de documentos pode ser analisada e seus tópicos (opiniões e idéias) podem ser identificados rapidamente, sem que os documentos tenham que ser lidos na íntegra (o que levaria muito tempo).

## 15.7 Sphinxs

Um dos objetivos da aplicação de técnicas de Inteligência Competitiva é o de conhecer melhor o mercado – o consumidor. Uma forma de se obter esse conhecimento é requisita-lo diretamente ao consumidor. Isso pode ser feito através da aplicação de

questionários. Nos questionários são colocadas todas aquelas questões que, a princípio, conseguem obter do consumidor todas as informações que a empresa necessita saber.

O Sphinx (<http://www.sphinxbr.com.br>) é um software desenvolvido especialmente para a elaboração, tratamento e a análise de questionários. Ele oferece uma série de testes estatísticos de validação e análise dos questionários nele construídos.

A princípio o Sphinx foi construído para trabalhar com questões fechadas e objetivas (do tipo escalar ou de múltiplas respostas). Porém, existem muitas questões importantes que são abertas, onde o usuário pode expressar sua opinião através da língua escrita (o formato textual).

Atualmente o Sphinx oferece algumas técnicas interessantes de lexicografia (análise da frequência léxica das palavras), que são específicas para a análise das questões abertas.

Basicamente o que ele faz é identificar a frequência dos diferentes lexemas (palavras) que aparecem nas respostas das questões. Como nem todas as palavras são relevantes para esse tipo de análise, os usuários podem definir listas de stopwords (palavras instrumentais) que são aquelas palavras que não devem ser levadas em conta no processo de análise. Com esse tipo de análise, torna-se possível identificar o foco (o assunto) abordado em cada resposta.

Por outro lado, a análise desse tipo de questão envolve uma série de problemas já discutidos anteriormente (variações morfológicas, sinonímia...). O Sphinx, possui algumas funcionalidades que minimizam esses problemas, mas, mesmo assim, os testes estatísticos (que são o forte e o interessante do programa) não podem ser realizados nas questões abertas.

## 15.8 Leximine (Sampler)

O Leximine (<http://www.leximine.com>) é um software sucessor do Sampler, um sistema desenvolvido para a análise de textos ou páginas da WEB. O Sampler é um software que analisa as palavras mais relevantes de um documento e identifica correlações entre elas. As palavras são então listadas na forma de um grafo onde linhas que ligam as palavras indicam quais palavras estão correlacionadas. A espessura de cada linha indica o grau de relação entre as palavras.

Atualmente o Sampler foi dividido em três softwares: Lexiguide, Lexisez e Leximine. Os três trabalham com análise lingüística que extrai os termos candidatos via um dicionário eletrônico. Esse dicionário indica quais são as expressões, palavras compostas, nomes de pessoas, produtos ou empresas. Após, uma série de algoritmos estatísticos identifica as palavras associadas e co-ocorrentes (uma técnica capaz de detectar os sinais fracos e auxiliar a construção de thesaurus). Os mapas léxicos gerados podem ser comparados e salvos para análises futuras.

## 15.9 GrapeVine

O GrapeVine (<http://www.grapevine.com>) é um software que serve para indexar, priorizar e disseminar informações de forma seletiva. As informações que passam por ele podem armazenadas para recuperação posterior. Ele trabalha de forma integrada com o sistema de e-mail e de newsgroups de uma empresa, indexando todas as idéias e opiniões (relativas à cooperação e ao trabalho em grupo) que os funcionários injetam no sistema.

Antes de entrar em funcionamento as diferentes categorias de informação que o sistema vai manipular devem ser definidas. Assim, cada usuário pode estabelecer um grau

peçoal de prioridade (grau de interesse) para cada uma das categorias de informação existentes.

Quando uma informação chega ou é colocada no sistema, ela é categorizada em uma das categorias predefinidas (isso pode ser feito manualmente ou automaticamente, através de um thesaurus predefinido de palavras-chave). A seguir, o sistema envia as mensagens para todas as pessoas que desejam receber mensagens daquela categoria. Se o grau de prioridade for mais alto do que o limiar estabelecido pelo usuário, esse usuário recebe a mensagem. Caso contrário, não a recebe. Para que o sistema funcione, deve haver na empresa ao menos um funcionário encarregado por cada categoria. Esse funcionário pode aumentar o grau de prioridade das mensagens, fazendo com que outras pessoas também recebam essa mensagem.

Deste modo, os funcionários de nível mais elevado não recebem mensagens sem importância. Somente aquelas mensagens realmente relevantes passam para a alta diretoria da empresa, sobrando mais tempo para que essas pessoas preocupem-se com outros problemas.



## 16 Métodos de mineração de textos aplicados à inteligência competitiva

As atividades relacionadas com a área de inteligência empresarial estão fortemente ligadas com a descoberta de conhecimento (para qualquer tipo de dado ou informação). Dentro da área de inteligência empresarial podem ser destacadas as áreas de inteligência competitiva (competitive intelligence) e de inteligência do negócio (business intelligence) [LOH 2000a].

A última (business intelligence) preocupa-se mais com as informações do ambiente interno da empresa. Estas informações estão geralmente dispostas na forma de dados estruturados, o que facilita a aplicação de métodos de mineração tradicionais. Isso, aliado ao fato de muitas empresas não possuírem informações textuais disponíveis no formato eletrônico (pois analisa-las automaticamente nunca havia sido pensado) ou, quando as possuem, não dispõem de técnicas capazes de analisá-las, fazendo com que a mineração de textos não seja muito difundida nas empresas.

É na área de Inteligência Competitiva que a mineração de textos tem maior aplicação ou aplicação mais imediata. Isso porque os empresários estão tendo como principal preocupação a análise da concorrência, além do que, grande parte dos dados e informações necessários para esse tipo de análise é geralmente encontrado em formato textual.

Independente da área de aplicação, os métodos e técnicas de mineração de textos utilizados são os mesmos. A seguir estes métodos e técnicas encontrados nas ferramentas de mineração de textos são listados. Cada um é acompanhado de uma breve descrição de sua utilidade ou aplicação dentro das áreas de inteligência.

### 16.1 Análise lexicométrica

A análise lexicométrica<sup>38</sup> é uma das técnicas de descoberta de conhecimento em textos mais utilizada. Dentre as existentes ela é a mais simples, e consiste na identificação da frequência de palavras (características) presentes nos documentos.

Esse tipo de análise serve para que o usuário identifique o conteúdo tratado em um documento ou conjunto de documentos. A listagem de palavras por ordem de frequência (partindo da mais frequente para a menos frequente) permite a identificação das palavras mais relevantes de um documento e, conseqüentemente, seu conteúdo ou assunto mais importante. Com essa análise também é possível (rapidamente) identificar novos produtos e concorrentes que, eventualmente, apareçam nas listagens. Do mesmo modo, são identificados os centros de interesse, tópicos mais relevantes e os objetos envolvidos (pessoas, institutos, países).

Aplicando-se essa técnica em diferentes conjuntos de documentos relativos a períodos ou épocas diferentes é possível realizar uma análise de tendências, identificando, por exemplo, que determinado concorrente está entrando em determinado mercado ou que determinado produto está passando a ser utilizado em alguma área ou ramo de atividade.

Essas análises servem para que o empresário possa modificar ou comprovar o plano de negócio de sua empresa.

Praticamente todas as ferramentas de text-mining a utilizam. Mesmo que ela não seja oferecida para o usuário final, internamente ela é utilizada como base para a aplicação de outros métodos mais complexos (como o clustering, por exemplo).

---

<sup>38</sup> Um exemplo de aplicação desta técnica com sucesso pode ser visto em [LOH 2000b].

## 16.2 Extração de informações

A técnica de extração (ver seção 14.2.1) é extremamente útil quando aliada a outras técnicas e métodos. A extração tem por objetivo identificar determinados objetos (dados ou informações) considerados relevantes por algum motivo. Geralmente os dados extraídos são colocados em alguma base de conhecimento ou banco de dados para que possam ser utilizados futuramente (esse banco de dados pode até mesmo ser utilizado em um processo de mineração tradicional).

No caso abordado por essa monografia, os dados extraídos podem ser utilizados para a geração de um resumo (sumarização), centróide (lista de palavras que indica os temas ou centros de interesse em torno de uma mesma informação) ou para a identificação de trechos relevantes (uma espécie de recuperação por passagem). Todas essas variações têm o propósito de auxiliar na identificação do conteúdo de um documento ou conjunto de documentos.

## 16.3 Identificação de clusters (clustering)

A técnica de identificação de clusters pode ser aplicada em dois níveis: ao nível de palavras e ao nível de documentos. O clustering de palavras consegue identificar relacionamentos entre palavras, já o clustering de documentos identifica relacionamento entre documentos.

O clustering pode ser utilizado por empresários para que eles possam fazer o mapeamento de elementos do ambiente externo (concorrentes, tecnologias e produtos identificados por palavras-chave específicas) e suas relações e co-relações. Assim, é possível identificar quais empresas estão fabricando determinados produtos, por exemplo.

Havendo um conjunto de documentos cujo assunto sejam as empresas concorrentes, a análise de clusters (por documentos) poderia identificar aqueles que possuem alguma relação (aliança, objetivos ou mercados comuns). Após, uma análise de centróides ou uma sumarização poderia ser utilizada a fim de identificar o porquê da relação (*quais seriam os objetivos da aliança? Quais seriam os mercados em comum?*) e isso então poderia gerar alguma ação na empresa (de *atacar* um novo mercado, por exemplo).

Seguindo o mesmo raciocínio, o clustering auxiliado pela técnica de centróides também poderia ser utilizado para identificar o posicionamento de uma empresa em relação a outras empresas. O fato de uma empresa estar posicionada em um cluster periférico (distante dos outros clusters) poderia ser interpretado como uma mudança de mercado.

A fim de facilitar essas análises e interpretações, sugere-se que os clusters e suas relações sejam visualizados em forma de grafos ou árvores (ver software Leximine e UMAP – seção 15). Nesse caso, os nodos geralmente representam os documentos ou palavras relevantes e as linhas que os conectam representam seus co-relacionamentos (clusters). Diferentes cores e espessuras podem ser utilizadas para indicar o grau e o tipo de relacionamento. Esse tipo de visualização oferece uma forma rápida de análise, dando uma visão geral dos objetos e elementos tratados nos documentos.

## 16.4 Classificação

Dentro da área de IC a classificação pode ser utilizada para realizar a filtragem das informações que chegam na empresa. Assim, todo documento ou mensagem eletrônica que chega pode ser analisado por uma ferramenta de classificação. Aqueles que não se encaixaram em alguma das categorias predefinidas podem ser descartados ou colocados em um local (de menor prioridade) para que sejam analisados futuramente ou quando houver necessidade. Os

demais documentos podem ser armazenados em um banco de documentos de uso corrente (para a aplicação das demais técnicas de mineração) ou enviados para setores ou pessoas específicas (nesse último caso o que ocorre é a *disseminação* de informações).

A disseminação é uma técnica auxiliada pela classificação capaz de enviar as informações certas para as pessoas certas. Cada departamento ou pessoa determina sua necessidade de informação que, neste caso, é a descrição de uma ou mais categorias relevantes para eles. Assim, depois de realizada a classificação ou categorização de um documento, ele é automaticamente enviado para os setores que indicaram seu interesse. Assim, somente os documentos realmente relevantes são recebidos pelas pessoas da empresa, minimizando o problema de sobrecarga de informações.

Em atividades de inteligência, determinada pessoa pode ficar encarregada de monitorar determinado concorrente ou produto. As características desse concorrente ou produto podem acabar gerando um descritor de categoria. Assim, toda vez que um documento for categorizado nessa categoria, ele é automaticamente enviado para a pessoa encarregada de analisá-lo. Eventualmente, poderiam ser criados agentes *spies* encarregados de coletar informações em fontes específicas ou até mesmo *sniffers* capazes de monitorar o tráfego de informações e capturar a informação caso ela enquadre-se na(s) categoria(s) especificada(s).

## 16.5 Análise de ferramentas versus métodos e etapas de inteligência competitiva

Para concluir, são apresentadas duas tabelas: a primeira apresenta os processos básicos do processo de inteligência e os métodos de KDT mais adequados (que poderiam ser utilizados) em cada um deles. A segunda contém uma listagem das ferramentas analisadas na seção anterior e os métodos utilizados por elas (e que foram discutidos nessa seção). Assim, o usuário pode identificar aquelas ferramentas mais adequadas para o tipo de análise que necessitar.

Etapa	Técnica						
	Normalização	Recuperação por passagem	Lexicometria	Clustering	Classificação	Análise de centróides	Sumarização
Coleta	X						
Disseminação				X	X		
Filtragem				X	X		
Análise de tendências			X	X		X	
Análise de conteúdo	X	X	X			X	X
Resumo	X		X	X	X	X	X

TABELA 1 – TÉCNICAS DE KDT QUE PODEM SER UTILIZADAS EM CADA ETAPA DE IC

Ferramentas	Técnicas											
	Recuperação			Vocabulário		Classificação		Clustering		Extração		
	Coleta	Indexação	Recuperação	Normalização	Lexicometria	Filtragem	Disseminação	Palavras	Documentos	Centróide	Sumarização	Passagem
Phoaks	X						X					
Referral Web	X		X			X						
Fab	X						X					
Siteseer						X						
GropupLens	X	X				X						
Umap	X	X			X			X	X			X
Sphinxs	X			X	X			X				
Leximine					X			X				
GrapeVine		X	X				X					

TABELA 2 – COMPARAÇÃO ENTRE AS FERRAMENTAS DE TEXT-MINING ESTUDADAS

Cabe salientar que essa tabela não cobre todo o espectro de ferramentas existentes, até mesmo porque novas ferramentas surgem a cada momento. Espera-se, porém, que, ao vê-la, o empresário ou a pessoa encarregada de realizar algum processo de inteligência saiba avaliar<sup>39</sup> que métodos são necessários para cada etapa e que ferramentas ele poderia utilizar para realiza-los.

<sup>39</sup> Algumas outras características também podem ser relevantes na escolha de uma ferramenta. Pode ser importante, em alguns casos, que a ferramenta possua a habilidade de acessar uma grande variedade de fontes e tipos de dados e informações; Da mesma forma, pode ser relevante a quantidade de atributos que ela pode manipular, o tamanho do documento que ela pode analisar, a granulosidade de sua análise (documento, parágrafos, palavras) e se ela é capaz de acessar e trabalhar com informações on-line ou off-line (em batch ou não). Demais informações, principalmente sobre comparações entre ferramentas de mineração de dados tradicionais, podem ser encontradas no estudo de Goebel [GOE 99].

## 17 Conclusões

O objetivo desse exame foi mostrar como as técnicas de descoberta de conhecimento em texto (e recuperação de informações) existentes podem ser utilizadas no auxílio ao processo de inteligência competitiva empresarial. Para atingir esse objetivo foram identificadas as técnicas e métodos de descoberta de conhecimento, assim como a metodologia e as técnicas de inteligência empresarial. Após uma análise comparativa entre as diferentes técnicas e etapas existentes em cada uma das áreas, foi possível criar um guia que indica quais técnicas e métodos de descoberta de conhecimento são mais indicadas (e em que ordem) para cada uma das etapas e objetivos do processo de inteligência.

Isso significa que a integração entre essas áreas é possível. Além de possível, essa integração é necessária, já que as empresas possuem uma grande quantidade de informação disponível para análise e essa análise torna-se inviável caso não seja realizada com o auxílio de técnicas e ferramentas computacionais.

Devido a essa facilidade, as empresas estão cada vez mais sujeitas aos riscos e as vantagens da competitividade, e não podem deixar de realizar constantemente as atividades de inteligência. Dentro deste contexto, algumas considerações devem ser levadas em conta. A consideração mais importante, tanto para a área da inteligência quanto para o processo de descoberta de conhecimento, é de que essas atividades devem ser realizadas com um objetivo para que o resultado seja coerente e útil.

Nas empresas, em especial, esse objetivo está geralmente atrelado ao plano de negócios da empresa ou a alguma necessidade de informação que a empresa possua. Logo, a identificação do problema e da necessidade de informação é uma atividade que exige cuidados especiais, pois, se estas forem mal identificadas, o processo de inteligência pode não apresentar resultados corretos.

Da mesma forma, o local onde as informações são coletadas (a fonte) também merece atenção especial. Primeiramente, ela deve ser confiável. A Internet está cheia de informações disponíveis, mas, em muitas delas, não há como garantir a veracidade e a confiabilidade da informação. Esse tipo de informação que pode não ser verdadeiro deve ser evitado. Recomenda-se que as fontes utilizadas, além de confiáveis, sejam respeitáveis. As fontes mais confiáveis são aquelas provenientes de bases de dados (on-line ou não) cujas informações são compiladas. Exemplos desse tipo de fonte são as bases de patentes, bases bibliográficas e artigos de jornais ou revistas. Por outro lado, as informações menos confiáveis podem, de alguma forma, dar alguns indícios. O analista das informações deve possuir um grande senso crítico, e conhecer muito bem o contexto das informações que está analisando para saber identificar o que é relevante e importante.

Essas fontes de informação devem ser variadas, pois uma mesma informação pode ser vista de modos diferentes em fontes diferentes. Além disso, algumas fontes podem conter informações complementares.

Devido à grande quantidade de informações disponíveis torna-se necessário aplicar algum mecanismo de filtragem. Mesmo depois de filtrada, a quantidade restante ainda pode ser muito grande. Nesses casos, técnicas de clustering capazes de separar uma coleção em sub-coleções, aliadas a técnicas de extração (sumarização e centróide) podem ser muito úteis, pois auxiliam na identificação de sub-categorias de informação mais relevantes. Se essas sub-categorias puderem ser modeladas e utilizadas em sistemas de categorização (filtragem e difusão), a quantidade de informações a ser analisada diminui muito, assim como sua qualidade e relevância podem aumentar consideravelmente.

Além da seleção (filtragem) da informação, pode ser interessante separar os dados (documentos) em diferentes épocas (temporalidades) de diferentes granularidades (ano, meses, semanas ou dias). Esse tipo de separação permite a análise de tendências e de mudanças de uma época para outra, gerando informações (conhecimento) mais específicas e interessantes.

Para concluir, de modo algum, pelo menos no momento atual de desenvolvimento tecnológico, o processo de inteligência e/ou de mineração deve ser realizado de modo totalmente automático. A perícia humana é indispensável para determinar o que deve ser vigiado e buscado, assim como para avaliar e validar as informações coletadas, interpretá-las e analisá-las [CLE 98].

## Bibliografia

- [ALL 95] ALLAN, James. Relevance feedback with too much data. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'95), 1995, Seattle, USA. Proceedings... New York: ACM Press, 1995. p.337-343.
- [ALT 2000] ALTAVISTA. Altavista Whitepaper. [s.l.:s.n.], 2000. Disponível em: <[http://doc.altavista.com/business\\_solutions/search\\_products/search\\_engine/av3white.doc](http://doc.altavista.com/business_solutions/search_products/search_engine/av3white.doc)>. Acesso em: 11 Nov. 2000.
- [APT 94] APTÉ, Chidanand et al. Automated learning of decision rules for text categorization. **ACM Transactions on Information Systems**, New York: ACM Press. v.12, n.3, p.233-251, 1994.
- [BAE 92a] BAEZA-Yates, Ricardo A. Introduction to data structures and algorithms related to information retrieval. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.13-27.
- [BAE 92b] BAEZA-Yates, Ricardo A. String Searching Algorithms. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.219-240.
- [BAK 98] BAKEL, Bas van. Modern classical document indexing: a linguistic contribution to knowledge-based IR. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'98), 1998, Melbourne, AU. **Proceedings...** New York: ACM Press, 1998. p.333-334.
- [BAL 97] BALABONIVIC, Marko; SHOHAM, Yoav. FAB: Content-Based, Collaborative Recommendation. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.66-72, 1997.
- [BAR 82] BARTSCHI, M.; FREI, H. P. Adapting a data organization to the structure of stored information. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1982, Berlin. **Proceedings...** New York: ACM Press, 1982.
- [BRI 51] BRIET, Suzanne. **Qu'est-ce que la documentation**. Paris: [s.n.], 1951.
- [BRO 9?] BROGLIO, John et al. **INQUERY System overview**. Amherst: University of Massachusetts, [199?], 20p. (Technical report for TIPSTER Project). Disponível em: <<http://ciir.cs.umass.edu/info/inqueryrep.ps>>. Acesso em: 08 Jul. 2000.
- [BUC 97] BUCKLAND, Michael K. What is a "document"? **Journal of the American Society for Information Science**, New York: John Wiley & Sons. v.48, n.9, p.804-809. 1997.
- [BUC 96] BUCKLEY, Chris. **SMART System Overview**. Ithaca, New York: Cornell University, 1996. 50p. (Technical Report). Disponível em: <<ftp://ftp.cs.cornell.edu/pub/fielding/lupOverview.ps>>. Acesso em: 08 Jul. 2000.

- [BUC 95] BUCKLEY, Chris; SALTON, Gerard. Optimization of relevance feedback weights. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'95), 1995, Seattle, USA. **Proceedings...** New York: ACM Press, 1995. p.351-357.
- [CHA 95] CHAKRAVARTHY, Anil S.; HAASE, Kenneth B. NetSerf: using semantic knowledge to find Internet information archives. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'95), 1995, Seattle, USA. **Proceedings...** New York: ACM Press, 1995. p4-11.
- [CHA 97] CHANG, Shih-Fu et al. Visual Information Retrieval from Large Distributed Online Repositories. **Communications of the ACM**, New York: ACM Press. v.40, n.12, p.63-71, 1997.
- [CHE 94] CHEN, Hsinchun. The vocabulary problem in collaboration. **IEEE Computer: Special Issue on CSCW**, Los Alamitos: IEEE Computer Society. v.27, n.5, p.2-10, 1994. Disponível em: <<http://ai.bpa.arizona.edu/papers/cscw94/cscw94.html>>. Acesso em: 22 Mai. 2000.
- [CHE 96] CHEN, Hsinchun. A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. **Journal of the American Society for Information Science**, London: ASLIB. v.47, n.8, 1996. Disponível em <<http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html>>. Acesso em 22 Mai. 2000.
- [DIX 97] DIXON, Mark. **An overview of document mining technology**. [s.l.: s.n.], 1997. Disponível em: <[http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97\\_dm.ps](http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps)>. Acesso em: 10/07/2000.
- [DOY 75] DOYLE, Lauren B. **Information Retrieval and Processing**. New York: John Wiley & Sons, 1975. 410p.
- [FEL 97a] FELDENS, Miguel Artur. Descoberta de Conhecimento em Bases de Dados e Mineração de Dados. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL, I, 1997, Pelotas, RS. **Proceedings...** Pelotas: EDUCAT, 1997. p.51-59.
- [FEL 97b] FELDENS, Miguel Artur; CASTILHO, José Mauro Volkmer de. **Engenharia da descoberta de conhecimento em bases de dados: estudo e aplicação na área da saúde**. Porto Alegre, 1997. 90p. Dissertação de mestrado – Instituto de Informática, UFRGS.
- [FEL 98] FELDENS, Miguel Artur et al. Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization. In: CONFERENCIA LATINOAMERICANA DE INFORMATICA (CLEI'98), XXIV, 1998, Quito, Ecuador. **Proceedings...** [s.l]: PUCE-XEROX, 1998. 2V. v.2, p.935-947.
- [FED 97] FELDMAN, Ronen; HIRSH, Haum. Exploiting background information in knowledge discovery from text. **Journal of Intelligent Information Systems**, Netherlands: Kluwer Academic Publishers. v.9, n.1, p.83-97. 1997.

- [FOX 92] FOX, Christopher. Lexical analysis and stoplists. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.102-130.
- [FOX 95] FOX, Edward A. et al. Digital Libraries. **Communications of the ACM**, New York: ACM Press. v.38, n.4, p.23- 28, 1995.
- [FOE 92] FOX, E. et al. Extended Boolean Models. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.393-418.
- [FRA 92a] FRAKES, William B. Introduction to information storage and retrieval systems. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.1-12.
- [FRA 92b] FRAKES, William B. Stemming Algorithms. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.131-160.
- [GOE 99] GOEBEL, Michael; GRUENWALD, Le. A survey of data mining and knowledge discovery software tools. **SIGKDD Explorations**, ACM. v.1, n.1, p.20-33. 1999. Disponível em: <<http://www.acm.org/sigkdd/explorations>>. Acesso em: 10 Jul. 2000.
- [GON 92] GONNET, G. H.; BAEZA-Yates, Ricardo A. New indices for text: PAT trees and PAT arrays. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.66-82.
- [GUP 97a] GUPTA, Amarnath et al. In Search of Information in Visual Media. **Communications of the ACM**, New York: ACM Press. v.40, n.12, p.35-42, 1997.
- [GUP 97b] GUPTA, Amarnath; JAIN, Ramesh. Visual information retrieval. **Communications of the ACM**, New York: ACM Press. v.40, n.5, 1997.
- [GUT 96] GUTHRIE, Louise et al. The role of lexicons in natural language processing. **Communications of the ACM**, v.39, n.1, p.63-72, 1996.
- [HAR 92a] HARMAN, Donna et al. Inverted Files. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.28-43.
- [HAR 92b] HARMAN, Donna. Ranking algorithms. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.363-392.
- [HEA 99] HEARST, Marti. When Information technology "goes social". **IEEE Intelligent Systems**, Los Alamitos: IEEE Computer Society. v.14, n.1, p.10-15, 1999.

- [IIV 95] IIVNEN, Mirja. Searches and Searches: Differences Between the Most and Least Consistent Searches. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'95), 1995, Seattle, USA. **Proceedings...** New York: ACM Press, 1995. p.149-157.
- [INZ 2000] INZUNZA, Víctor Saúl. Las TI's en una economía basada en el conocimiento. **TI Magazine**. 2000. Disponível em: <<http://www.timagazine.net/timagazine/1a2b3c/0499/ti.cfm>>. Acesso em: 10 Jul. 2000.
- [JAI 97] JAIN, Ramesh. Visual Information Management. **Communications of the ACM**, New York: ACM Press. v.40, n.12, p.31-32, 1997.
- [KAU 97] KAUTZ, Henry et al. Referral Web: Combining Social Networks and Collaborative Filtering. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.63-65, 1997.
- [KEI 97] KEIM, Michelle et al. **Bayesian Information Retrieval: Preliminary Evaluation**. 1997. Disponível em: <<http://www.research.att.com/~lewis/papers/keim97.os>>. Acesso em: 10 Nov. 2000.
- [KOC 74] KOCHEN, Manfred. **Principles of Information Retrieval**. New York: John Wiley & Sons, 1974. 203p.
- [KON 97] KONSTAN, Joseph a. et al. GroupLens: Applying Collaborative Filtering to Usenet News. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.77-87, 1997. Disponível em <<http://www.cs.umn.edu/Research/GroupLens>>. Acesso em: 23 Mai. 2000.
- [KOR 97] KORFHAGE, Robert R. **Information Retrieval and Storage**. New York: John Wiley & Sons, 1997. 349p.
- [KOW 97] KOWALSKI, Gerald. **Information Retrieval Systems: Theory and Implementation**. Boston: Kluwer Academic Publishers, 1997. 282p.
- [KRA 96] KRAAIJ, Wessel; POHLMANN, Renée. Viewing stemming as recall enhancement. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'96), 1996, Zurich, Switzerland. **Proceedings...** New York: ACM Press, 1996. p.40-48.
- [LAL 96] LALMAS, Moulina. Modelling information retrieval with Dempster-Shafer's Theory of Evidence: A Study. In: ECAI WORKSHOP ON "UNCERTAINTY IN INFORMATION SYSTEMS: QUESTIONS OF VIABILITY", 1996, Budapest. **Proceedings...** [s.l.:s.n.], 1996. Disponível em: <<http://www.dcs.gla.ac.uk/ir/publications/papers/Postscript/lalmas96c.ps.gz>>. Acesso em: 16 Ago. 2000.
- [LAN 68] LANCASTER, F. Wilfrid. **Information Retrieval Systems: Characteristics, Testing and Evaluation**. New York: John Wiley & Sons, 1968. 222p.

- [LEW 91] LEWIS, David D. Evaluating text categorization. In: SPEECH AND NATURAL LANGUAGE WORKSHOP, 1991, Defense Advanced Research Projects Agency. **Proceedings...** San Francisco, CA: Morgan Kaufmann, 1991. p.312-318. Disponível em <<http://www.research.att.com/~lewis>>. Acesso em: 20 Abr. 2000.
- [LEW 94] LEWIS, David D.; RINGUETTE, Marc. Comparison of two learning algorithms for text categorization. In: ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL, III, 1994, Las Vegas, NY. **Proceedings...** ISRI, University of Nevada: Las Vegas, 1994. p.81-93. Disponível em: <<http://www.research.att.com/~lewis/papers/lewis94b.ps>>. Acesso em: 12 Nov. 1999.
- [LIN 93] LIN, Chung-hsin; CHEN, Hsinchun. An automatic indexing and neural network approach to concept retrieval and classification of multilingual (chinese-english) documents. **IEEE Transactions on Systems, Man and Cybernetics**, v.26, n.1, 1993. Disponível em: <<http://ai.bpa.arizona.edu/papers/chinese93/chinese93.html>>. Acesso em: 10 Set. 1999.
- [LOH 97] LOH, Stanley et al. Uma abordagem para busca contextual de documentos na Internet. **Revista de Informática Teórica e Aplicada (RITA)**, v.4, n.2, p.79-92, 1997.
- [LOH 99] LOH, Stanley et al. Recuperação semântica de documentos textuais na Internet. In: CONFERENCIA LATINOAMERICANA DE INFORMÁTICA, XXV (CLEY'99), 1999, Assunción, Paraguay. **Proceedings...** Assunción: Universidad Autónoma de Asunción, 1999. 2V. v.2, p.827-836.
- [LOH 2000a] LOH, Stanley et al. Concept-based knowledge discovery in texts extracted from the WEB. **ACM SIGKDD EXPLORATIONS**, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining. v.2, n.1, p.29-39. 2000. Disponível em: <<http://www.acm.org/sigs/sigkdd/explorations/issue2-1/loh.pdf>>. Acesso em: 20 Jul. 2001.
- [LOH 2000b] LOH, Stanley et al. Descoberta proativa de conhecimento em textos: aplicações em inteligência competitiva. In: INTERNATIONAL SYMPOSIUM ON KNOWLEDGE MANAGEMENT/DOCUMENT MANAGEMENT, ISKM/DM, 2000, Curitiba, Brasil. **Proceedings...** Editora Universitária Champagnat: Curitiba, 2000. p.125-147.
- [MIZ 96] MIZZARO, Stefano. A Cognitive Analysis of Information Retrieval. In: INFORMATION SCIENCE: INTEGRATION IN PERSPECTIVE - CoLIS2, 1996, Copenhagen, Denmark. **Proceedings...** The Royal School of Librarianship, 1996. p.233-250. Disponível em: <<http://ten.dimi.uniud.it/~mizzaro/papers/colis.ps.gz>>. Acesso em 01 Jun. 2000.
- [MIZ 97] MIZZARO, Stefano. Relevance: The Whole History. **Journal of the American Society for Information Science**, New York: John Wiley & Sons. v.48, n.9, p.810-832. 1997.

- [MLA 2000] MLADENIC, Dunja; GROBELNIK, Marko. Feature Selection for Classification Based on Text Hierarchy. In: CONFERENCE ON AUTOMATED LEARNING AND DISCOVERY (CONALD-98), 2000, **Proceedings...** Pittsburg: Carnegie Mellon University, 2000. p.6p. Disponível em: <<http://www.cs.cmu.edu/afs/cs/user/dunja/www/pww.html>>. Acesso em: 24 Out. 2000.
- [MOO 51] MOORES, Calvin N. Datacoding applied to Mechanical Organization of Knowledge. **American Documentation**, Apud [GUP 97b]. v.2, p.20-32. 1951.
- [MOR 82] MORRISSEY, Joan. An intelligent terminal for implementing relevance feedback on large operational retrieval systems. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 1982, Berlim. **Proceedings...** New York: ACM Press, 1982.
- [MOU 92] MOULIN, Bernard; MOULIN, Daniel. Automated knowledge acquisition from regulatory texts. **IEEE Expert**, v7, n5, 1992.
- [NAS 97] NASSIF, Mônica Erichssen Borges; CAMPELLO, Bernadete Santos. A organização da informação para negócios no Brasil. **Perspectivas em Ciência da Informação**, Minas Gerais: Escola de Ciência da Informação. v.2, n.2, p.149-161. 1997.
- [NOG 2000] NOGUEZ, José Hiram S. **Técnicas de mineração de dados no processo de descoberta de conhecimento em banco de dados**. Porto Alegre, 2000. 47p. Trabalho Individual (TI-874) – Instituto de Informática, UFRGS.
- [NOR 96] NORMAN, Benjamin et al. **A Learning Subject Field Coder**. Syracuse, NY: NPAC, Syracuse University, 1996. 6p. (Project REU'96 Report). Disponível em: <[http://www.npac.syr.edu/REU/reu96/project\\_reu.html](http://www.npac.syr.edu/REU/reu96/project_reu.html)>. Acesso em: 23 Mai. 2000.
- [OLI 96] OLIVEIRA, Henry M. **Seleção de entes complexos usando lógica difusa**. Porto Alegre, 1996. Dissertação de mestrado – Instituto de Informática, PUC-RS.
- [PRA 97] PRADO, Hércules Antonio do. **Conceitos de Descoberta de Conhecimento em Bancos de Dados**. Porto Alegre, 1997. 43p. Trabalho Individual (TI-709) – CPGCC, UFRGS.
- [PRA 98] PRADO, Hércules Antônio do. **Abordagens híbridas para mineração de dados**. Porto Alegre, 1998. 87p. Exame de Qualificação – Instituto de Informática, UFRGS.
- [RAS 92] RASMUSSEN, Edie. Clustering Algorithms. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.419-442.
- [RES 97] RESNICK, Paul; VARIAN, Hal R. Recommender Systems. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.56-58, 1997.
- [RIJ 79] RIJSBERGEN, C. van. **Information Retrieval**. 2ed. London: Butterworths, 1979. 147p.

- [RIL 95] RILOFF, Ellen. Little words can make big difference for text classification. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'95), 1995, Seattle, USA. **Proceedings...** New York: ACM Press, 1995. p.130-136.
- [RIL 94] RILOFF, Ellen; LEHNERT, Wendy. Information extraction as a basis for high-precision text classification. **ACM Transactions on Information Systems**, v.12, n.3, 1994.
- [ROB 97] ROBERTSON, S. E.; WALKER, S. On relevance weights with little relevance information. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'97), 1997, Philadelphia, USA. **Proceedings...** New York: ACM Press, 1997. p.16-24.
- [ROC 79] ROCKART, John. Chief executives define their own data needs. **Harvard Business Review**, v.57, n.2, p.81-92, 1979. Disponível em: <[http://www.hbsp.harvard.edu/hbsp/prod\\_detail.asp?79209](http://www.hbsp.harvard.edu/hbsp/prod_detail.asp?79209)>. Acesso em: 11 Nov. 2000.
- [RUC 97] RUCKER, James; POLANCO, Marcos J. Site-seer: Personalized Navigation for the Web. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.73-75, 1997.
- [SAL 87a] SALTON, Gerard; BUCKLEY, Chris. **Improving Retrieval Performance by Relevance Feedback**. Ithaca, New York : Department of computer science, Cornell University, 1987. (Technical Report).
- [SAL 87b] SALTON, Gerard; BUCKLEY, Chris. **Term weighting approaches in automatic text retrieval**. Ithaca, New York : Department of computer science, Cornell University, 1987. (Technical Report).
- [SAL 83] SALTON, Gerard; MACGILL, Michael J. **Introduction to Modern Information Retrieval**. New York: McGRAW-Hill, 1983. 448p.
- [SAL 88] SALTON, Gerald; SMITH, Maria. **On the application of syntactic methodologies in automatic text analysis**. Ithaca, New York: Department of Computer Science, Cornell University, 1988. (Technical Report).
- [SCA 97] SCARINCI, Rui Gureghian. **SES: Sistema de Extração Semântica de informações**. Porto Alegre, 1997. 165p. Dissertação de mestrado – Instituto de Informática, UFRGS.
- [SCH 97] SCHÜTZE, Hinrich; SILVERSTEIN, Craig. Projections for efficient document clustering. In: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR'97), 1997, Philadelphia, USA. **Proceedings...** New York: ACM Press, 1997. p.60-66.
- [SHU 35] SHÜRMEYER, W. Aufgaben und Methoden der Dokumentation. **Zentralblatt für Bibliothekswesen**, APUD [BUC 97]. v.52, n.1, p.533-543. 1935.
- [TER 97] TERVEEN, Loren et al. PHOAKS: A System for Sharing Recommendations. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.59-62, 1997.

- [WAR 92] WARTIK, S. et al. Hashing Algorithms. In: FRAKES, William B.; BAEZA-Yates, Ricardo A. **Information Retrieval: Data Structures & Algorithms**. Upper Saddle River, New Jersey: Prentice Hall PTR, 1992. p.293-363.
- [WEB 98] WEBBER, Alam. O que queremos dizer com conhecimento. In: DAVENPORT, T.; PRUSAK, L. **Conhecimento Empresarial**. 1998. p.1-28.
- [WIE 95] WIENER, Erik D. et al. A Neural Network Approach to Topic Spotting. In: FOURTH ANNUAL SYMPOSIUM ON DOCUMENT ANALYSIS AND INFORMATION RETRIEVAL (SDAIR'95), 1995, Las Vegas. **Proceedings...** [s.l.:s.n.], 1995. p.317-332. Disponível em: <<http://www.stern.nyu.edu/~aweigend/Research/Papers/TextCategorization>>. Acesso em: 23 Mai. 2000.
- [WIL 92] WILBUR, J. W.; SIROTKIN, K. The Automatic Identification of Stop Words. **Journal of Information Society**, v.18, , p.45-55. 1992.
- [WIV 97] WIVES, Leandro Krug. **Um Estudo sobre Técnicas de Recuperação de Informações com ênfase em Informações Textuais**. Porto Alegre, 1997. 55p. Trabalho Individual (TI-672) – CPGCC, UFRGS.
- [WIV 98a] WIVES, Leandro Krug. **Um Estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering"**. Porto Alegre, 1998. 102p. Dissertação (Mestrado em Ciência da Computação) – PPGC, UFRGS.
- [WIV 98b] WIVES, Leandro Krug; LOH, Stanley. Hyperdictionary: a knowledge discovery tool to help information retrieval. In: STRING PROCESSING AND INFORMATION RETRIEVAL: A SOUTH AMERICAN SYMPOSIUM (SPIRE'98), 1998, Santa Cruz de la Sierra, Bolívia. **Proceedings...** Los Alamitos: IEEE Computer Society, 1998. p.103-109.
- [YAN 97] YANG, Yiming; PEDERSEN, Jan O. **A comparative study on features selection in text categorization**. School of Computer Science, Carnegie Mellon University, 1997.
- [YAN 99] YANG, Yiming; LIU, Xin. An evaluation of statistical approaches to text categorization. **Journal of Information Retrieval**, v.1, n.1/2, p.67-88. 1999.
- [YEO 97] YEO, Boon-Lock; YEUNG, Minerva. Retrieving and Visualizing Video. **Communications of the ACM**, New York: ACM Press. v.40, n.12, p.43-52, 1997.
- [ZAD 65] ZADEH, Laft A. Fuzzy Sets. **Information and Control**, v.8, n.1, p.338-353. 1965.
- [ZAD 73] ZADEH, Lofti A. Outline of a new approach to the analysis of complex systems and decision processes. **IEEE Transactions on Systems, Man and Cybernetics**, Los Alamitos: IEEE Computer Society. v.SMC-3, n.1, p.28-44. 1973.
- [ZAN 98] ZANASI, Alessandro. Competitive Intelligence though datamining public sources. **Competitive Intelligence Review**, Alexandria, Virginia: SCIP. v.9, n.1, 1998.