

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

**Um Estudo Sobre
Técnicas de Recuperação de Informações
Com Ênfase em
Informações Textuais**

por

Leandro Krug Wives
T.I. nº 672 CPGCC-UFRGS

Trabalho Individual I

Prof. Dr. José Mauro Volkmer Castilho
Orientador

Porto Alegre, Dezembro de 1997

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitora: Profa. Dra. Wrana Panizzi

Pró-Reitor de Pós-Graduação: Prof. Dr. José Carlos Ferraz Hennmann

Diretor do Instituto de Informática: Prof. Dr. Roberto Tom Price

Coordenador do CPGCC: Prof. Dr. Flávio Rech Wagner

Bibliotecária-Chefe do Instituto de Informática: Dra. Zita Catarina Prates da Silveira

Sumário

| | |
|---|----|
| Lista de figuras | 5 |
| Resumo | 6 |
| Abstract | 7 |
| Introdução | 8 |
| 1. Conceitos | 11 |
| 1.1. Paradigma | 11 |
| 1.1.1. Abstração de informações | 12 |
| 1.1.2. Descrição da necessidade do usuário | 14 |
| 1.1.3. O processo de matching | 15 |
| 1.2. Medindo a eficiência de um SRI | 16 |
| 1.2.1. Abrangência (Recall) | 16 |
| 1.2.2. Precisão (Precision) | 17 |
| 2. Recuperação de informações textuais | 19 |
| 2.1. Indexação automática | 20 |
| 2.1.1. Identificação de palavras | 21 |
| 2.1.2. Remoção de stopwords | 22 |
| 2.1.3. Word stemming | 23 |
| 2.1.4. Identificação de termos compostos | 24 |
| 3. Modelos de recuperação de informações | 26 |
| 3.1. Modelo booleano | 26 |
| 3.1.1. Interseção | 26 |
| 3.1.2. União | 27 |
| 3.1.3. Subtração | 27 |
| 3.1.4. Combinação de operadores em expressões booleanas | 28 |
| 3.2. Modelo probabilístico | 30 |
| 3.2.1. Espaço de vetores (vector space model) | 30 |
| 3.3. Modelo difuso (fuzzy) | 35 |
| 3.4. Modelo contextual | 37 |
| 4. Ferramentas de auxílio | 42 |
| 4.1. Relevance feedback | 42 |
| 4.2. Thesaurus | 44 |
| 4.3. Dicionários | 46 |

| | |
|----------------------------------|----|
| 4.4. Hiperdicionários _____ | 47 |
| 4.4.1. Montagem manual _____ | 47 |
| 4.4.2. Montagem automática _____ | 48 |
| 5. Conclusões _____ | 50 |
| Referências bibliográficas _____ | 52 |

Lista de figuras

| | |
|--|----|
| Figura 1.1 - Paradigma da recuperação de informações _____ | 12 |
| Figura 1.2 - Processo de abstração _____ | 13 |
| Figura 1.3 - Descrição de uma consulta _____ | 15 |
| Figura 1.4 - Relações entre recall e precision _____ | 17 |
| Figura 2.1 - Função similaridade _____ | 20 |
| Figura 2.2 - Estrutura de uma lista invertida _____ | 21 |
| Figura 2.3 - Identificação de termos válidos _____ | 22 |
| Figura 2.4 - Identificação de stopwords _____ | 23 |
| Figura 2.5 - Processo de indexação textual _____ | 25 |
| Figura 3.1 - Resultado de uma consulta booleana com o operador AND _____ | 27 |
| Figura 3.2 - Resultado de uma consulta booleana com o operador OR _____ | 27 |
| Figura 3.3 - Resultado de uma consulta booleana com o operador NOT _____ | 28 |
| Figura 3.4 - Resultado da pesquisa booleana estendida _____ | 29 |
| Figura 3.5 - Modelo espaço de vetores _____ | 32 |
| Figura 3.6 - Representação de documentos em duas dimensões _____ | 32 |
| Figura 3.7 - Utilização de um contexto que converge _____ | 39 |
| Figura 3.8 - Utilização de um Contexto que não converge _____ | 39 |
| Figura 3.9 - Uma palavra em mais de um contexto _____ | 40 |
| Figura 4.1 - Figura abstrata de uma ave _____ | 43 |
| Figura 4.2 - Estrutura de um thesaurus _____ | 45 |
| Figura 4.3 - Estrutura de um hiperdicionário _____ | 47 |

Resumo

As técnicas de recuperação de informações têm sido desenvolvidas há muitos anos. Há um consenso na área de que estas técnicas estarão em constante desenvolvimento, devido a natureza dinâmica das informações e ao crescimento contínuo de documentos de todos os tipos, armazenados em meios magnéticos.

Este documento apresenta o esforço atual realizado na área, discutindo os problemas que surgem no processo de recuperação de informações. São apresentados as técnicas, os modelos e as ferramentas utilizadas na recuperação de informações textuais. Porém, apesar de focalizar as informações textuais, muitas das técnicas aqui descritas podem ser utilizadas, com algumas adaptações, no processo de recuperação de quaisquer outros tipos de informação.

Palavras-chave:

*Recuperação de Informações Textuais, Técnicas de Recuperação,
Modelos de Recuperação*

Abstract

For many years the techniques involved in the process of information retrieval have been developed. It is a consensus that these techniques will be constantly refashioned, reformed, reformed, due to the highly dynamic nature of information, and to the continuous increase in the volume of documents of all sorts, stored in electronic media.

This text presents the current research effort realized in the area, addressing the problems that arise in the process of textual information retrieval. The models and the tools used in this process are also presented, and despite focusing textual information, it is possible to use most of these techniques on any other kind of information, with some adaptations.

Keywords:

*Textual Information Retrieval, Retrieval Models,
Retrieval Techniques*

Introdução

Há muitos anos os homens preocupam-se em criar meios eficientes de localização de informações, para que possam recuperá-las assim que sejam necessárias. Desde as épocas mais remotas, quando surgiram as primeiras bibliotecas, pessoas eram encarregadas de estruturar e organizar as informações para que outros pudessem encontrá-las, sem ter que ler e compreender, documento por documento, até que os dados que estavam procurando fossem localizados.

Já naqueles tempos o que os encarregados pelas informações faziam era a indexação, a organização por assuntos ou tópicos, e a elaboração de hierarquias. Assim, não é necessário ler todos os documentos para que a informação desejada seja encontrada, mas sim, determinar seu contexto e estudar somente aqueles documentos que pertencem ao mesmo assunto.

Com o tempo estas técnicas foram sendo aprimoradas, e são utilizadas até hoje. Com a informatização do processo de armazenamento e manipulação das informações todas estas técnicas foram adotadas. Apesar, com as facilidades oferecidas pela tecnologia a quantidade de informações que passou a ser manipulada chega a níveis muito elevados. Com isso, é necessário refinar as técnicas antigas ou elaborar técnicas mais novas capazes de atender a nova demanda.

Segundo Calvin Moores [GUP97] o termo “Recuperação de Informações” identifica o ato intelectual, realizado pelo usuário, de especificação e de descrição da informação de que ele necessita. O termo abrange também os sistemas, as técnicas e as máquinas utilizadas neste processo de busca de informações relevantes para o usuário.

Na época, 1950, Calvin imaginava somente informações textuais (documentos), e durante muitos anos esta foi a abrangência da área: um pequeno mercado onde a maior parte das aplicações eram os bancos de dados bibliográficos. No entanto, a evolução tecnológica mais uma vez alterou os rumos da ciência, oferecendo uma série de facilidades para o manuseio dos dados, tais como meios de armazenamento mais rápidos com maior capacidade, e equipamentos de digitalização e obtenção de informações diversas.

Com isto, novos tipos de informação surgiram e a área passou a preocupar-se, também, com outras fontes de informação. Obviamente, recuperar informações já não significa mais buscar documentos, mas sim também sons, imagens, vídeos e outros tipos de dados que surgem a cada momento.

Com o surgimento de novos tipos de informação, novas técnicas de manipulação de dados precisaram ser criadas. Hoje em dia, facilidades antes consideradas impraticáveis, devido a capacidade dos sistemas, já não são mais novidade. É comum observar nos sistemas características como a utilização de linguagem natural, linguagens de consulta, resultados por ordem de relevância ao usuário e assistentes de formulação de consulta. Neste caso a área de passou preocupar-se, também, em otimizar a localização, recuperando a maior quantidade de itens relevantes no menor tempo possível, com o mínimo de esforço e com a máxima eficiência. Isso torna a área complexa, exigindo estudos em outras áreas tais como a Teoria da Informação, probabilidade, técnicas de otimização, reconhecimento de padrões e modelagem matemática.

Isso mostra que a área da recuperação de informações é uma área extremamente dependente da evolução tecnológica. Enquanto houver evolução, haverá a necessidade de novos estudos a fim de adaptar ou criar novos métodos. Métodos estes que atendam a demanda dos novos tipos de dados ou dos novos equipamentos que manipulam estes dados. Isso pode ser considerado um problema, e é um dos dos fatores que mais motivam os estudos na área, já que as técnicas anteriores precisam ser repensadas e refinadas de acordo com as novas tendências.

Atualmente a informação não só é dinâmica como também é volumosa, o que ocasiona um novo problema. A quantidade de informações com que o homem vem trabalhando já não é mais a mesma do que a de alguns anos atrás. O grande número de bancos de dados, disponíveis *on-line*, faz com que o volume de informações ao alcance de qualquer pessoa torne-se grandioso, e não há alguém capaz de assimilar tamanha quantidade.

É o que ocorre na *Internet* onde as fontes de informações são muito numerosas e dinâmicas. A *Internet* é uma grande rede de informações heterogêneas e distribuídas, que estão interligadas através dos chamados *links* (elos). Estas informações são mantidas e utilizadas por pessoas do mundo inteiro. Cada uma destas pessoas possui sua cultura, sua necessidade e seu modo de ver as coisas. Isso gera um ambiente totalmente desestruturado, o que dificulta a localização de informações tornando-a uma tarefa complicada. Segundo [CAT96], o usuário necessita localizar a informação de seu interesse navegando pelas ligações existentes entre estas inúmeras fontes de informação, realizando uma verdadeira garimpagem, o que pode gerar uma sobrecarga cognitiva e desorientação.

Em [KUO96] comenta-se que isso ocorre porque os métodos tradicionais são inadequados para este tipo de estrutura, havendo então uma necessidade por melhores técnicas de acesso. Segundo [CRO95], estas técnicas devem ser capazes de guiar o usuário por entre esse emaranhado de informações, mostrando-lhe somente àquelas que são realmente importantes, de acordo com sua necessidade.

Além dos problemas da dinâmica da informação é necessário solucionar também os problemas relacionados a sua integração. Não adianta criar métodos que solucionem somente os problemas relacionados a um determinado tipo de informação. Quando alguém está procurando informações sobre determinado assunto, é esperado que todas as informações relacionadas a este assunto, independente de sua forma, sejam retornadas. Esta não é uma tarefa trivial pois cada informação tem sua estrutura.

A integração de pessoas através da *rede* é um fato necessário e de grande utilidade para todos. A *Internet* é um dos únicos meios de comunicação que consegue tamanha diversidade, já que há uma certa liberdade de utilização, de expressão e de disponibilização das informações desta rede. É devido à toda esta diversidade que torna-se trabalhoso o processo de localização de informações. Chega a ser irônico o fato de um meio tão poderoso e irrestrito de comunicação não possuir meios de acesso rápido às informações de que o usuário necessita.

Este é mais um dos motivos que tornam as pesquisas na área tão importantes, já que as soluções ainda não são eficientes. É claro que a *Internet* não é a única preocupação da área. Há ainda as Bibliotecas Digitais, cujo papel também é prover informações. Este tipo de estrutura facilita muito a educação, e tem, portanto, muita utilidade.

As Bibliotecas Digitais são o futuro da informação eletrônica, onde o material básico é o texto, podendo conter gráficos, imagens, sons e vídeos. Neste caso os mesmos problemas citados anteriormente podem ser encontrados. Isso porque não há uma estrutura comum e existe muita redundância. Os fatores segurança e privacidade também não devem ser deixados de lado.

Estes são os motivos que geraram o estudo apresentado neste trabalho. Nas próximas seções buscar-se-á abordar as técnicas utilizadas no processo de recuperação de informações, já que tudo indica que a informação deverá tornar-se um dos maiores recursos da humanidade nas próximas décadas.

O objetivo deste trabalho é realizar um levantamento bibliográfico, a fim de demonstrar as principais técnicas e modelos envolvidos no processo de recuperação de informações textuais. Porém, sempre que possível, buscar-se-á abstrair as idéias básicas do processo possibilitando sua utilização em quaisquer outros tipos de informação, com pequenas adaptações.

No primeiro capítulo, *Conceitos*, descrever-se-á o paradigma envolvido no processo de Recuperação de Informações, incluindo o formalismo necessário para realizar o mapeamento das informações, possibilitando assim sua posterior localização. Além disto, serão apresentados alguns conceitos básicos necessários para uma melhor compreensão da área e dos capítulos seguintes.

O capítulo seguinte, *Recuperação de informações textuais*, desenvolve o paradigma, descrevendo como este é aplicado na busca de informações textuais. Lá pode ser encontrado também um pequeno *tutorial* sobre as técnicas envolvidas no processo de indexação de informações textuais.

No Capítulo 3, apresentar-se-á os modelos de recuperação de informações utilizados pela maioria dos sistemas de recuperação de informações. Já no Capítulo 4 serão discutidas algumas ferramentas de apoio à recuperação de informações encontradas na literatura, cujo objetivo é melhorar a eficiência dos sistemas e minimizar os problemas inerentes ao processo de recuperação de informações.

Finalmente são apresentadas algumas considerações e conclusões sobre o trabalho realizado, descritas no Capítulo 5.

1. Conceitos

1.1. Paradigma

Todos os Sistemas de Recuperação de Informações (SRI) possuem um único objetivo: fazer com que o usuário encontre a informação que está precisando rapidamente, de modo que este usuário não necessite analisar ele próprio todas as informações existentes na base de informações.

O SRI deve realizar esta análise e fornecer para o usuário aquelas informações que são mais *relevantes* para este usuário. O conceito de informação relevante é importante para um sistema de recuperação de informações. E diz respeito àquelas informações que devem ser retornadas como resposta a uma consulta do usuário. Em [SAL83] define-se relevância como sendo a correspondência contextual entre uma consulta e uma informação, ou seja, o grau de relevância indica o quanto a informação é apropriada para o que o usuário está requisitando; ou o quão importante para o usuário é determinada informação. É claro que determinar a relevância de determinada informação depende muito de como o usuário expressou sua necessidade, isto é, depende de como o usuário formulou sua consulta.

É através das características que o usuário fornece em uma consulta que o SRI vai determinar quais informações são mais relevantes para este usuário. Já que não há como obter informações diretamente do usuário, a não ser através de uma expressão formal de consulta, o usuário deve expressar corretamente sua necessidade. Do contrário o resultado não será satisfatório. Muitas vezes o sistema tem como saber se as informações recuperadas têm relação com a descrição feita pelo usuário. O que não é possível é saber se a consulta que o usuário elaborou descreve corretamente a sua necessidade.

Portanto, em muitos casos, o sistema busca informações relevantes para a descrição do usuário, mas irrelevantes para a informação realmente desejada pelo usuário, já que este usuário não descreveu corretamente sua necessidade.

Vários problemas podem surgir devido a isso, pois o paradigma utilizado na área de recuperação de informações é muito complexo. A Figura 1 apresenta o paradigma clássico, ou a estrutura geral de um SRI tradicional, que independe do tipo de informação utilizado.

Os elementos da Figura 1.1 representam os objetos e a interação entre estes objetos no paradigma. Observando esta figura, pode-se notar que existem três pontos-chave que devem ser trabalhados com atenção.

O primeiro é o processo de abstração de informações, determinado pela modelagem do sistema. O segundo é decorrente da abstração que o usuário faz ao descrever a informação de que necessita, em algum formalismo (linguagem de consulta). 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. Na literatura é comum encontrar este processo como sendo uma função de similaridade, que determina a semelhança entre os elementos da consulta e os elementos da base de informações.

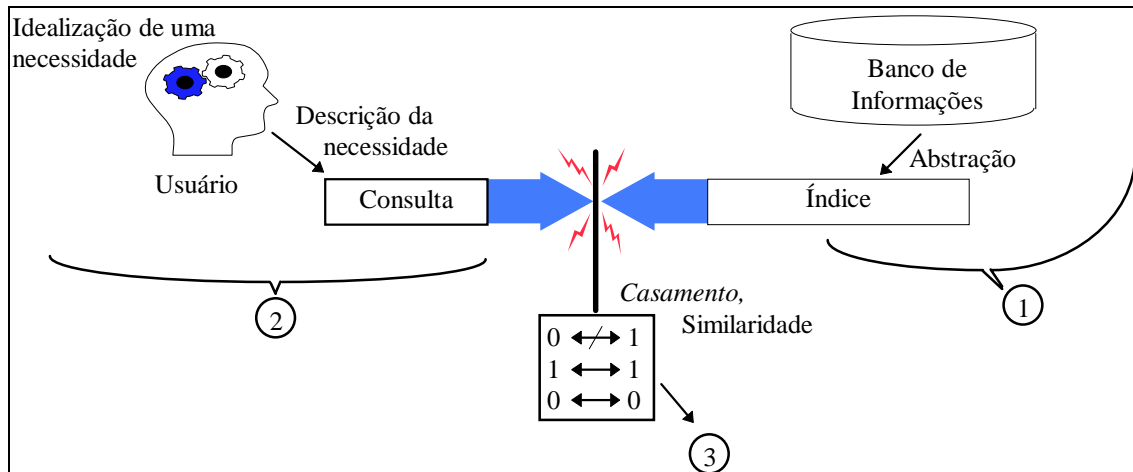


Figura 1.1 - Paradigma da recuperação de informações

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.

No texto a seguir estes problemas e suas origens serão apresentados. Serão discutidos também algumas soluções e cuidados que devem ser tomados na elaboração de um SRI e na sua utilização, a fim de minimizar estes problemas.

1.1.1. Abstração de informações

Como em todo processo de informatização a modelagem correta do objeto que está para ser informatizado é uma das fases mais importantes, e deve ser realizada com todo o cuidado. Em um Sistema de Recuperação de Informações modelar a informação que o sistema irá tratar é extremamente importante. É necessário identificar o que é relevante (importante) em determinado tipo de informação, isto é, o que caracteriza determinada informação e pode ser utilizado para distingui-la entre outras informações. Ou seja, é preciso identificar o conteúdo real da informação, o conteúdo que realmente consegue descreve-la. Para isto, pode ser importante descobrir como o usuário geralmente faz referência a esta informação.

Após identificado o conteúdo e as características de determinada informação, é necessário idealizar algum formalismo que possa descrevê-la e armazená-la. Em imagens, por exemplo, suas características são, entre outras coisas, as cores, as formas que a imagem contém e as texturas. Seu conteúdo é expresso através destas características. Uma forma de descrição textual para estas características pode não ser boa escolha, até mesmo porque quem vai fazer a descrição textual pode não ser capaz de captar todas as características de uma imagem. Cada um descreve uma imagem de acordo com que vê ou consegue ver.

Uma opção seria armazenar todos os *bits* da imagem e criar mecanismos ou funções capazes de detectar as características da imagem em cima destes *bits*. Assim, toda sua informação descritiva é armazenada. Quando o usuário requisitar uma consulta pela característica cor uma função ou algoritmo é aplicado na hora, vasculhando os *bits* a fim de encontrá-la. Do mesmo modo podem ser identificadas as formas e texturas, ou

qualquer outra característica que o usuário possa precisar, desde que existam algoritmos para isto. Caso estes não existam basta criá-los, e é possível criá-los pois o modelo contém todas as informações que qualquer algoritmo venha a precisar.

Um exemplo da complexidade da escolha correta de um modelo é apresentado em [GUP97], onde são apresentadas as necessidades de usuários que trabalham com sistemas de recuperação de informações visuais. No caso é discutido o problema de como o usuário poderia recuperar informações sobre um *videoclip*, com restrições tais como o tempo de duração máximo de dois segundos e surgimento de um carro vermelho, que velozmente deve desaparecer da pista ao passar por trás de uma montanha.

Logo, escolher um modelo 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 1.2 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.

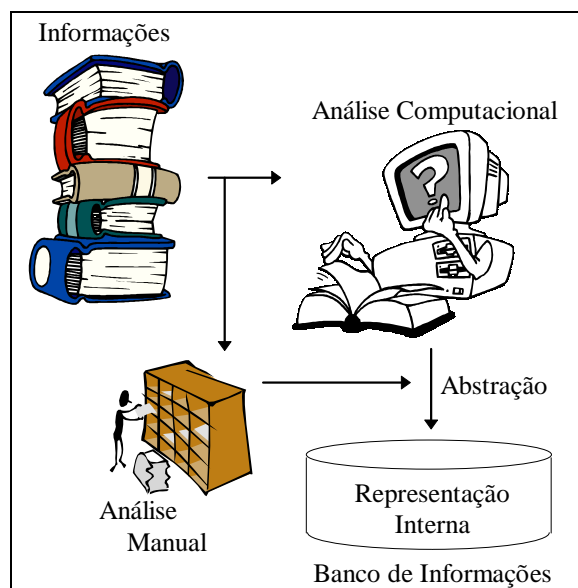


Figura 1.2 - Processo de abstração

A indexação é uma das técnicas de abstração muito utilizada na área de recuperação de informações. É através do índice que as informações são acessadas. Portanto o índice deve ser construído com muito cuidado. A não inclusão de uma característica importante de uma determinada informação no arquivo de índice fará com que o acesso a esta informação seja deficiente, ou até mesmo, dependendo da consulta, esta informação pode não ser recuperada. Logo, este processo de abstração tem seus problemas, como dito anteriormente, e deve ser tratado com cuidado.

No capítulo 2 encontra-se um pequeno tutorial do processo de indexação de informações textuais. As técnicas utilizadas são específicas para informações textuais, mas servem para mostrar a importância de uma boa modelagem, onde a identificação do

conteúdo que caracteriza a informação é importante. Descobrir o que é importante em uma informação nem sempre é uma tarefa fácil, mas uma vez definido, é possível aproveitar as técnicas textuais realizando algumas abstrações e modificações.

1.1.2. Descrição da necessidade do usuário

Do mesmo modo que o sistema precisa estar preparado para trabalhar com abstrações de informação, o usuário deve ser capaz de descrever a informação que ele necessita, identificando o maior número possível de características desta informação.

O processo de descrição da informação pelo usuário também é problemático por várias razões. A primeira em relação ao próprio usuário, pois devido à sua formação e seu conhecimento, pode não ser capaz de descrever a informação corretamente. A segunda, em relação ao usuário não estar familiarizado com o sistema, não sabendo utilizar a linguagem de consulta, o tipo de informação armazenado ou o modelo de abstração utilizado pelo sistema. A última, em relação ao próprio sistema, onde linguagem de consulta não permite que o usuário expresse corretamente suas intenções.

Alguns destes problemas podem estar diretamente relacionados com o modelo utilizado pelo sistema para modelar as informações. Segundo [GUP97], mesmo os usuários mais experientes costumam ter problemas quando o modelo de abstração utilizado não é bom para o tipo de informação. É o caso da utilização de descrição textual em sistemas de informação visuais, que tratam com imagens. Isso porque certas informações, como é o caso das imagens, são difíceis de serem descritas textualmente.

O usuário faz uma visualização mental da informação que quer. Esta mentalização geralmente é feita levando em conta o formato original da informação. Quando o usuário utiliza outra forma de descrição, que não utilize o formato original da informação, os problemas surgem. Decorrente disso pode ocorrer de várias pessoas descreverem uma mesma imagem de formas diferentes, mesmo que vejam ou imaginem a mesma imagem.

Este problema, onde vários usuários descrevem o mesmo objeto de formas diferentes, é descrito em [CHE94a, CHE94b, CHE96], e é conhecido por *Problema do Vocabulário* (Vocabulary Problem).

Pelo estudo de [CHE96] vários motivos levam as pessoas a utilizarem termos diferentes para o mesmo objeto. Estes motivos estão, muitas vezes, diretamente relacionados com a pessoa, pois são provenientes de sua cultura e de sua formação. Isso dificulta muito a solução do problema do vocabulário. Porém, algumas vezes, o problema surge pelo motivo do usuário não estar familiarizado com o tipo de informação que está procurando, e isto pode ser solucionado.

É o que ocorre quando pessoas de áreas diferentes utilizam o mesmo sistema para armazenar e consultar informações. Os estudos de [CHE94a] indicam que cada área possui termos científicos diferentes para situações e objetos similares. E quando uma pessoa consulta informações de uma área diferente da sua, não obtém bons resultados porque os termos utilizados não são os mesmos (já que os autores das informações pertencem à outras áreas ou outras culturas).

Em [CHE96] são realizados estudos que indicam que o problema do vocabulário pode ser solucionado com a utilização de dicionários de sinônimos e *Thesaurus* (ver página 44).

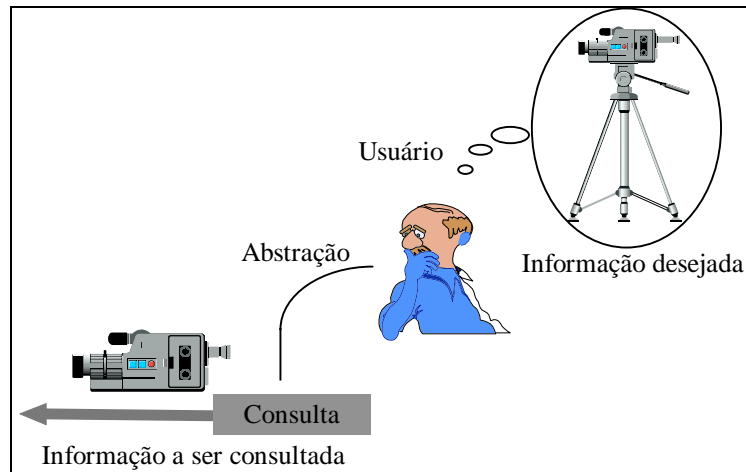


Figura 1.3 - Descrição de uma consulta

Em muitos casos ocorre o cenário descrito na Figura 1.3, onde o usuário deseja um tipo de informação mas, devido aos problemas citados, a descrição do que ele está procurando não é completa ou é errônea. Neste caso sua consulta irá retornar informações irrelevantes.

Cuidados especiais devem ser tomados a fim de evitar que isto ocorra. Os sistemas devem ser muito bem preparados, oferecendo recursos eficientes para que o usuário consiga descrever corretamente a informação de que necessita.

Porém, nem todos os usuários que utilizam o sistema são obrigados a conhecê-lo. É importante informar ao usuário como funciona a linguagem de consulta e quais são as características de descrição de informação, ou seja, o modelo adotado pelo sistema.

Portanto, procurar a informação relevante é um processo complexo que exige muito do usuário, principalmente se ele não está familiarizado com a informação que precisa ou com o sistema.

1.1.3. O processo de matching

O último, mas não menos importante, ponto-chave da recuperação de informações encontra-se justamente no mecanismo que faz o processo de identificação de quais informações são relevantes para a consulta do usuário. Este processo procura identificar a similaridade entre as informações armazenadas no sistema e a descrição de informação que o usuário deseja.

Muitos problemas surgem neste ponto. Primeiro, porque tanto o sistema quanto o usuário fazem abstrações da informação, o que pode acarretar numa perda de características importantes da informação. Além disso, muitos métodos utilizam um processo de comparação direta entre as características desejadas pelo usuário 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.

É o que ocorre muitas vezes em uma biblioteca, onde o usuário tem dificuldades para encontrar determinada informação porque o bibliotecário não a colocou na mesma categoria que o usuário pensa que ela deveria estar.

Decorrente disto duas pessoas diferentes, consultando, podem obter resultados diferentes, já que descrevem a informação de formas diferentes. Neste caso uma pode obter resultados melhores do que outra.

Um estudo realizado em [IIV95] indica que a utilização de um vocabulário controlado é uma solução que apresenta bons resultados, desde que os usuários sejam treinados para utilizar este vocabulário. Além disso, aqueles que procuram informações cujo assunto lhes é conhecido costumam obter melhores resultados, pois estão familiarizados com os termos utilizados. O estudo indica também que o usuário deve utilizar o maior número de características possíveis na descrição da informação que deseja, pois assim será maior a probabilidade do sistema localizar uma informação relevante (já que ele terá a sua disposição uma quantidade maior e características para comparar).

Existem várias técnicas que buscam minimizar estes problemas, realizando análises mais complexas que não utilizam somente a comparação direta entre características. Muitas destas técnicas variam de acordo com o modelo de recuperação utilizado pelo sistema. Estes modelos serão discutidos em um dos capítulos seguintes.

1.2. Medindo a eficiência de um SRI

O desenvolvimento de técnicas de recuperação de informações realmente eficientes tem sido o centro das atenções dos estudos na área de RI. Ao longo dos anos, várias métricas foram propostas a fim de testar e validar a eficiência destas técnicas. Em [SAL91] pode-se encontrar uma discussão sobre as principais medidas utilizadas na avaliação da eficiência de sistemas de recuperação de informações. A *Abrangência* (Recall) e a *Precisão* (Precision) são as medidas de eficiência mais conhecidas e utilizadas na área.

O objetivo por trás da utilização de medidas de eficiência, segundo [SAL91], é indicar se o sistema realmente consegue recuperar grandes quantidades de informações relevantes para o usuário (uma boa *abrangeência*), ao mesmo tempo em que consegue excluir os itens irrelevantes (uma boa *precisão*).

Com isso é possível avaliar se o usuário realmente conseguiu o que queria, pois toda a informação recuperada é útil para ele, ou se o usuário ainda não está satisfeito e deve realizar novas consultas, pois os itens recuperados não são relevantes.

1.2.1. Abrangência (Recall)

A Abrangência, como o próprio nome sugere, serve para indicar a proporção de itens relevantes, recuperados em resposta a uma consulta do usuário. É utilizada para medir a habilidade do sistema recuperar todos os itens relevantes.

A fórmula que mede a abrangência do resultado retornado por uma consulta é a seguinte:

$$\text{Recall} = \frac{\text{total de informações relevantes encontradas}}{\text{total de informações relevantes existentes}}$$

1.2.2. Precisão (Precision)

A Precisão mede a proporção de itens recuperados que são realmente relevantes. É a habilidade do sistema recuperar somente as informações que são relevantes para o usuário, e nada mais.

A fórmula que mede a precisão do resultado retornado por uma consulta é a seguinte:

$$\text{Precision} = \frac{\text{total de informações relevantes encontradas}}{\text{total de informações encontradas}}$$

A Figura 1.4, a seguir, apresenta as combinações possíveis entre abrangência e precisão, onde os documentos recuperados são aqueles que encontram-se no interior dos círculos. Nem sempre é possível conseguir o melhor caso, que é o de abrangência e precisão total. No entanto, é comum e aceitável que o método de localização não seja totalmente abrangente ou preciso (fique no meio-termo), desde que o usuário possa identificar de alguma forma quais são os documentos que mais aproximam-se do resultado esperado.

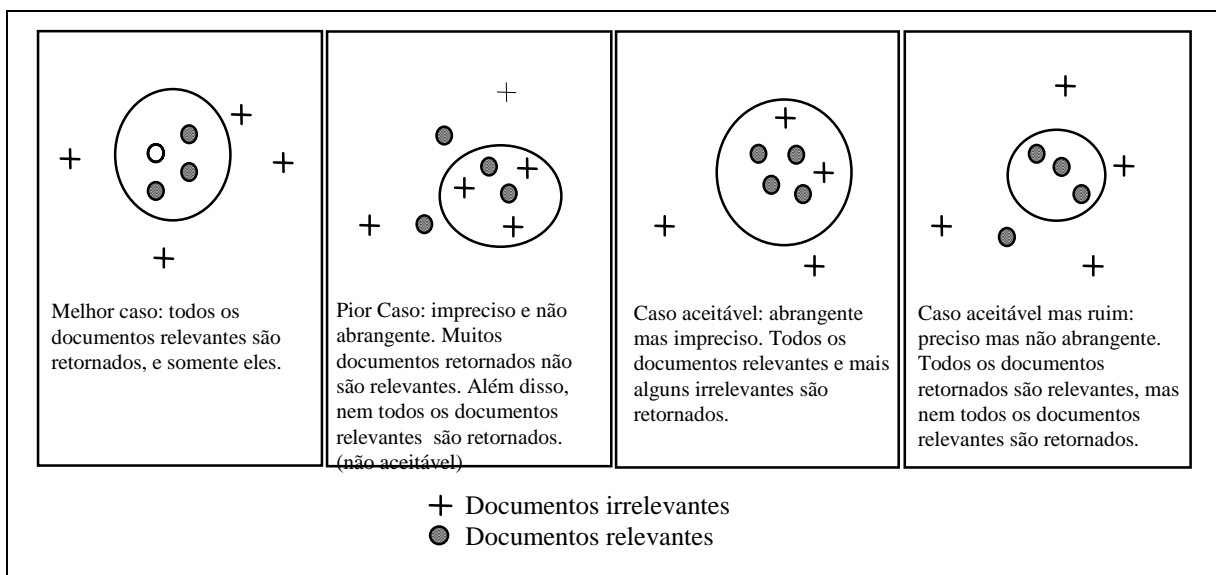


Figura 1.4 - Relações entre recall e precision

Apesar destas duas medidas serem muito utilizadas, há crítica sua utilidade. Isso porque a abrangência não garante que os documentos recuperados sejam realmente úteis para o usuário. Além disso, estas medidas não informam outros parâmetros também importantes para o usuário, tais como o número total de documentos e o número total de itens recuperados. E finalmente, não há como saber realmente quais são os documentos relevantes à determinada consulta. A não ser que se tenha conhecimento prévio de todas as informações presentes na base de dados, o que é impossível devido a grande quantidade de dados.

Geralmente a quantidade de documentos relevantes (utilizada nas fórmulas de *recall* e *precision*) é estimada *a priori* através de métodos estatísticos, onde somente alguns documentos são analisados e sua categoria identificada. Deste modo tem-se uma idéia aproximada de quantos documentos podem ser relevantes a determinado assunto.

Existem, é claro, outras formas de se medir a eficiência de um SRI, mas geralmente estas medidas são mais difíceis de serem interpretadas, além de exigirem informações que não são obtidas sem uma análise mais detalhada das informações, o que demanda um esforço computacional maior. Devido a estes fatores sugere-se, em [SAL91], que a abrangência e a precisão sejam utilizadas, além de serem mais facilmente interpretadas.

Porém, alguns fatores devem ser levados em consideração na construção de um sistema. Um estudo realizado em [CRO95] conclui que tão importante quanto avisar o usuário de que ele conseguiu ou não o que queria, é avisá-lo do porquê este resultado foi obtido. O usuário pode ter cometido um erro, expressando incorretamente sua necessidade. Geralmente os erros provocam pouco impacto na precisão e abrangência, mas muito impacto no usuário. Em [SAL83] complementa-se que é necessário analisar também o esforço intelectual ou físico realizado pelo usuário na elaboração de uma consulta, na condução da busca e na análise dos resultados.

Há também outros fatores, tais como o tempo de processamento de uma consulta e a quantidade de recursos computacionais utilizados. Conciliar todos estes fatores não é tarefa trivial, mas eles são necessários para obter-se um sistema mais eficiente e adequado às necessidades do usuário.

2. Recuperação de informações textuais

A busca de informações textuais difere da busca de informações tradicional que é realizada nos Bancos de Dados Tradicionais. Segundo [SAL83], os Bancos de Dados Tradicionais¹ preocupam-se com o armazenamento, manutenção e a recuperação de informações disponíveis explicitamente no sistema. Ao contrário dos Bancos de Dados Textuais onde a informação está implícita, muitas vezes escondida ou difícil de ser localizada, em forma de *Linguagem Natural*². Neste último, não há *Campos*, capazes de identificar os *Atributos* específicos de determinados *Registros*, ou seja, as informações não estão armazenadas em *Tabelas* como em Bancos de Dados Relacionais.

Por exemplo, para se buscar informações sobre determinada pessoa em um Banco de Dados (BD) tradicional basta percorrer no BD a Tabela que possui o atributo *Nome* e localizar o Registro (*Tupla*) que possui o nome da pessoa desejado (em [KOR94] podem ser obtidas maiores informações e exemplos sobre Bancos de Dados Tradicionais).

Caso o Banco de Dados citado fosse textual, os dados não estariam distribuídos de uma forma *tabular*. Até mesmo porque o texto é uma seqüência de caracteres, não existindo atributos. Não há como saber o que é um nome em um documento, a não ser que se faça uma análise de *Linguagem Natural* e se descubra o que pode vir a ser um nome - o que não é fácil de ser feito (maiores detalhes sobre as diferenças entre os vários tipos de Sistemas de Informação podem ser obtidos em [SAL83]).

Logo, para localizar as informações sobre determinada pessoa, em um Banco de Dados Textual, seria necessário analisar caracter-por-caracter do texto até que a seqüência de caracteres correspondente ao nome fosse localizada.

Este tipo de análise (caracter-a-caracter) não é conveniente. É necessário haver alguma forma mais eficiente de acesso aos documentos. Para isto, os documentos precisam de algo que os identifique entre os demais, permitindo a sua localização.

Sabe-se que os documentos textuais possuem um contexto, isto é, um assunto. Este assunto pode ser identificado pelas palavras (termos) que este documento contém. Portanto, o termo é o meio de acesso a um documento.

Decorrente disso, um Sistema de Recuperação de Informações Textuais tem como base a seguinte teoria, proposta por [SAL83]: perguntas (consultas) são submetidas pelo usuário. Perguntas estas baseadas em termos (palavras) que identificam a idéia desejada por este usuário. Os documentos são identificados pelos termos que eles contém, e, portanto, a localização de um documento desejado pelo usuário dá-se a partir da identificação da similaridade entre o(s) termo(s) fornecido(s) pelo usuário e os termos que identificam os documentos contidos na base de dados. A figura a seguir representa esquematicamente esta teoria:

¹ Relacional, Hierárquico, Redes.

² Linguagem Natural devido ao fato de ser a linguagem normalmente utilizada pelo homem para comunicar-se (exemplo: Português, Inglês, Alemão...).



Figura 2.1 - Função similaridade

Esta função *Similaridade* busca identificar uma relação entre os termos da *Query* (consulta) e os termos dos documentos. Teoricamente pode ser feita uma comparação direta entre estes termos, mas na prática é difícil estabelecer esta relação de similaridade entre estes termos devido a alguns problemas.

Um destes problemas é o já citado Problema do Vocabulário, onde as palavras utilizadas pelo sistema (palavras contidas nos documentos) podem ser diferentes das palavras utilizadas pelo usuário, mesmo que estas palavras (sinônimos) representem a mesma idéia.

Há ainda o problema da *Busca Incerta* (*Search Uncertainly*), onde os usuários não sabem quais são as melhores palavras que identificam o assunto que querem localizar. Por conseqüência, acabam não recuperando informações precisas. Este problema também é discutido por [CHE94c], [SAL83] e outros trabalhos.

Estes problemas fazem com que sejam recuperados muitos documentos ou documentos de assuntos variados (pois o termo é muito abrangente), ou ainda, podem recuperar informação alguma.

É buscando solucionar estes problemas (e alguns outros) que mecanismos de mapeamento entre os diferentes termos similares foram criados. Segundo [SAL83] há vários sistemas universitários e comerciais que se utilizam destes mecanismos: STAIRS (IBM), Dialog System (Lookhead Information Systems), BRS (State University of New York), MEDLARS (National Library of Medicine), SMART (Cornell University). Em [ACM96a] são citados mais alguns: WIN (West Publishing Company), DOWQUEST (Dow Jones Newswire), WAIS, e um muito conhecido, o INQUERY.

Nem sempre estes sistemas conseguem satisfazer o usuário, mas foram a base para as técnicas atuais e das que estão por vir. A metodologia básica destes sistemas é discutida a seguir.

Após estas definições e estudos iniciais percebe-se, portanto, que o meio de acesso aos documentos são as palavras que ele contém. Para tornar possível o acesso a estas palavras é preciso colocá-las em uma estrutura auxiliar - o índice, isso porque fica inviável pesquisar todos os textos toda a vez que for requisitada uma consulta.

A indexação é o processo de mapeamento citado anteriormente. Ela é o meio pelo qual a função de similaridade vai comparar os termos da Consulta com os termos presentes nos documentos, e, após, localizar os documentos relacionados com o assunto desejado pelo usuário.

2.1. Indexação automática

O método de *Salton* [SAL83] é um método bastante aceito, inclusive vários outros trabalhos utilizam este método (apesar de poderem variar em alguns aspectos).

Este método é conhecido por *Indexação Automática*, e constitui-se de várias etapas. Ao final das etapas, os termos resultantes são adicionados a um arquivo de índice cuja estrutura geralmente é baseada em *Arquivos Invertidos* (ou *Listas*

Invertidas). Segundo [SAL83], outros tipos de arquivos podem ser utilizados, mas a experiência mostra que este tipo de estrutura é uma das mais eficientes para a indexação de documentos. Na Figura 2.2 é apresentado um exemplo da estrutura de uma lista invertida.

Basicamente, a estrutura permite que um único termo aponte para vários documentos. Maiores detalhes sobre esta estrutura podem ser obtidos em [SAL83].

A Indexação Automática possui várias etapas, que variam dependendo do modelo utilizado. Porém, geralmente são encontradas as seguintes:

- a) **Identificação de palavras;**
- b) **Remoção de “stopwords”;**
- c) **“Word stemming”;**
- d) **Identificação de termos compostos.**

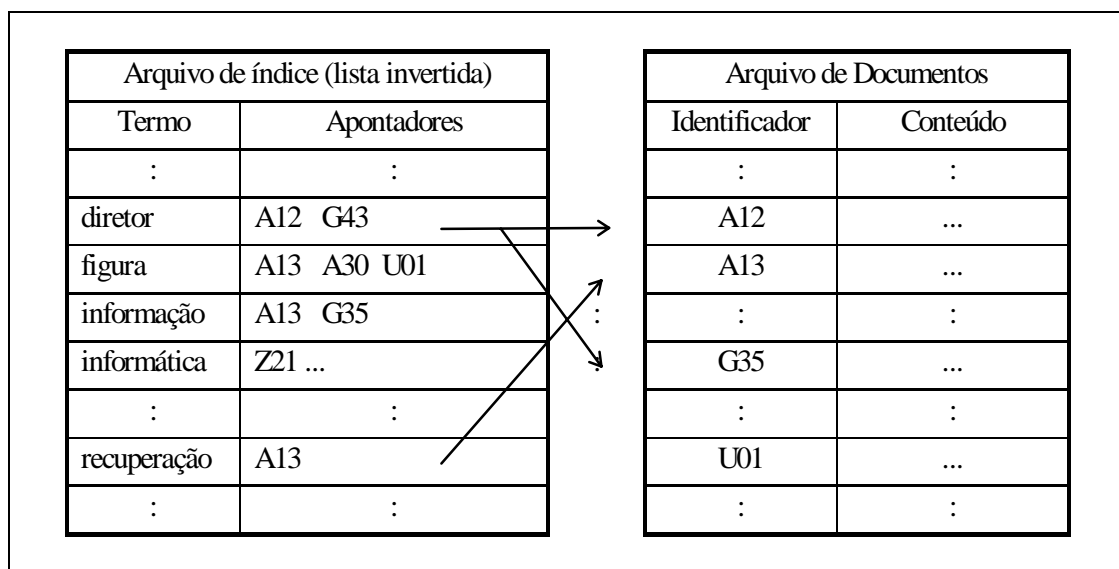


Figura 2.2 - Estrutura de uma lista invertida

2.1.1. Identificação de palavras

Identifica as palavras nos documentos a serem indexados. Nada mais é do que a identificação de palavras, analisando-se as seqüências de caracteres no texto. Salton [SAL83] aconselha fazer um *Dictionary lookup*, ou seja, comparar as seqüências de caracteres retiradas do texto com um dicionário (ver página 46) a fim de validar se estas palavras realmente existem.

Este processo de validação torna-se bastante útil, especialmente quando o documento apresenta muitos caracteres inválidos ou palavras com erros gramaticais. As seqüências de caracteres inválidas devem ser eliminadas e as palavras com erros corrigidas. Pode-se aplicar ainda um processo de filtragem naqueles arquivos que possuem formatos de texto específicos, a fim de eliminar as seqüências de controle e/ou formatação de texto.

O dicionário pode também auxiliar a identificação de termos específicos quando deseja-se utilizar palavras predefinidas no índice, evitando que palavras desconhecidas sejam identificadas (ou seja, evita a utilização de um vocabulário descontrolado).

Um simples *Analizador Léxico* que identifique seqüências de caracteres e monte palavras pode ser utilizado.

A figura a seguir apresenta o trecho de um documento com diversas seqüências de caracteres. As seqüências marcadas são seqüências inválidas, que não devem passar pela fase de identificação de palavras. As demais seqüências podem ser identificadas como termos válidos. Os termos sublinhados são termos identificados como incorretos pelo dicionário e devem ser corrigidos. Os caracteres de pontuação são desprezados.

... à;+• á`>`~` `þÿ Na maioria das vezes os documentos retornados pelas ferramentas de `→` recuperação de informacoes `→` envolvem um contexto mais amplo, fazendo com que o usuário tenha que garimpar, ou seja, especificar ou filtrar estes documentos (o que demanda tempo e conhecimento) a fim de obter a informação que ele realmente necessita `~` ...

Figura 2.3 - Identificação de termos válidos

2.1.2. Remoção de stopwords

Nem todas as palavras dos documentos podem ser adicionadas na estrutura de índice. As palavras que aparecem em todos os documentos, ou na maioria deles, são um exemplo. Isso porque a utilização de uma palavra com estas características não é capaz de selecionar documentos relativos a um assunto específico.

As preposições são um exemplo deste tipo de palavra, pois são termos que servem para fazer o encadeamento de idéias e palavras. Portanto, são termos inerentes a linguagem e não ao conteúdo dos documentos.

Logo, as palavras que aparecem em muitos documentos não devem ser indexadas, pois sua utilização compromete a precisão e a eficiência do sistema.

Nos sistemas já implementados foi construída uma estrutura (uma lista) contendo todas as palavras que não devem ser indexadas. A esta estrutura foi atribuído o nome de *Stoplist*, e as palavras presentes nesta lista são conhecidas como *Stopwords*.

O processo de obtenção das *stopwords* pode ser manual, onde o projetista do sistema avalia quais palavras devem ou não ser indexadas (o que varia de língua para língua, ou até mesmo entre sistemas). Há ainda a possibilidade de montar-se esta lista automaticamente, verificando-se quais são as palavras com maior freqüência (que aparecem em mais documentos), e selecionando-as como *stopwords*.

Então, após uma palavra ser reconhecida no processo de indexação, sua presença na *Stoplist* é verificada. Caso exista na lista de palavras negativas ela não é adicionada ao índice.

A figura abaixo apresenta o documento resultante da etapa anterior após ser validado por uma *stoplist*. Neste caso a lista de *stopwords* contém artigos, preposições, conjunções e algumas seqüências de caracteres que não devem ser adicionadas ao índice, por possuírem freqüência elevada.

... Na maioria das vezes os documentos retornados pelas ferramentas de recuperação de informações evoluem um contexto mais amplo fazendo com que o usuário tenha que garimpar ou seja especificar ou filtrar estes documentos o que demanda tempo e conhecimento a fim de obter a informação que ele realmente necessita ...

Figura 2.4 - Identificação de stopwords

Geralmente com estas etapas já é possível criar-se índices que auxiliem a localização. Esta localização de documentos é feita a partir da comparação direta entre os termos da consulta do usuário e os termos presentes nos documentos. Este é um método ainda ineficiente, e algumas técnicas adicionais podem ser utilizadas a fim de melhorá-lo.

As técnicas a seguir permitem ao sistema melhorar sua eficiência, mas possuem alguns inconvenientes. São técnicas recomendadas por *Salton*, mas podem ser feitas à parte. Alguns sistemas não as utilizam, como é o caso de [WIV96a] e [WIV96b], mas conseguem recuperar documentos de forma razoável. Há ainda quem diga que a utilização destas técnicas não compensa. É o caso de [RIL95] que realizou alguns testes, e chegou a conclusão de que em alguns casos a eficiência do sistema pode ser pior se o *Stemming* (ver adiante) for utilizado. Comenta ainda que até a fase de remoção de *Stopwords* pode comprometer o sistema porque estas palavras têm papel importante *em alguns domínios*.

Os estudos realizados até o momento sobre a eficiência destes métodos de remoção de palavras não indicam ainda uma conclusão. Utilizá-los ou não depende muito do sistema e dos próprios documentos. *Church* [CHU95] apresenta algumas comparações de eficiência entre a utilização ou não destes métodos. De qualquer modo eles são apresentados a seguir.

2.1.3. Word stemming

“Word stemming” corresponde à identificação de radicais (agrupamento de palavras similares), a fim de melhorar a eficiência e solucionar o problema do vocabulário. É uma técnica que procura reduzir a variância morfológica de um termo (uma normalização), e portanto depende muito da linguagem utilizada nos documentos (técnicas elaboradas para uma língua não podem ser utilizadas em outra). Vários experimentos para a língua Inglesa foram realizados, e funcionam de maneira eficiente. Um dos mais recentes experimentos sobre *stemming* pode ser encontrado em [KRA96].

A técnica consiste em identificar os radicais das palavras e adicioná-las no arquivo de índice desta forma. Uma maneira de identificar os radicais das palavras é remover seus sufixos e prefixos. Outro exemplo é a eliminação dos plurais das palavras.

Assim, todas as palavras que possuem o mesmo radical, e portanto com significados similares (mas categorias diferentes de linguagem: adjetivo, verbo, advérbio...) são reconhecidas pelo mesmo identificador. As palavras são armazenadas de uma só forma - o radical - facilitando a consulta. A desvantagem deste método é que ele pode tornar as palavras muito abrangentes, pois os termos específicos são eliminados. Neste caso os documentos específicos não são recuperados. Em outras palavras, pode melhorar a *abrangência*, mas piora a precisão da resposta a uma consulta.

Porém, [CRO95] sugere que esta técnica seja utilizada somente na normalização de termos da consulta, se o usuário assim desejar. Desta forma, não há o problema de um mesmo termo do índice apontar para vários documentos, que possuem o mesmo radical, podendo tornar a consulta mais abrangente. Em muitos casos o usuário pode querer ser mais específico, e a normalização de termos dificulta este processo.

2.1.4. Identificação de termos compostos

Também conhecida por *Identificação de frases-termo*. Junta as palavras adjacentes para formar novos termos, buscando solucionar o problema dos termos abrangentes. Isso porque as idéias estão agrupadas em contextos, e palavras compostas geralmente categorizam melhor os assuntos (os termos passam a ser mais específicos).

A utilização de palavras mais específicas consegue fazer com que o sistema recupere documentos de forma mais precisa, justamente pelo fato destas palavras aparecerem em um número menor de documentos (geralmente os documentos de contextos específicos utilizam termos específicos).

Para exemplificar, pode-se imaginar uma pessoa buscando informações sobre *programas de computador*. Esta pessoa poderia formular uma consulta utilizando a palavra *Programa*, o que poderia ocasionar a recuperação de muitos documentos, que contém a palavra *programa*, mas que não pertencem ao contexto *computador*.

Uma solução para este problema seria utilizar o termo composto “programa de computador”, ou simplesmente “programa computador” (pela eliminação da preposição). Esta *frase* contextualiza melhor a palavra *programa* tornando-a menos abrangente e mais específica. Deste modo os documentos retornados por esta *frase-termo* fariam parte somente do contexto *programa de computador*.

Deve-se tomar o cuidado para não confundir o conceito de *frase-termo* com a utilização das duas palavras de forma independente. Ou seja, caso o usuário não tenha de alguma forma especificado que as duas palavras devem aparecer juntas, ou o sistema não possua alguma técnica que unifique as duas palavras, a consulta pode tornar-se ainda mais abrangente. Isso significa que seriam retornados tanto documentos que tratam do assunto *computador* quanto documentos que tratam do assunto *programa*.

Em geral não é necessário armazenar as palavras de forma composta pois este processo de unificação das palavras exige tempo. *Salton*, em seus estudos, e *Croft* [CRO82] recomendam que ela não seja utilizada, pois não aumenta de forma considerável a eficiência do sistema. O que pode ser feito é o armazenamento da informação sobre as distâncias entre as palavras de um mesmo documento e deixar com que a técnica de consulta avalie se as palavras são ou não adjacentes (no livro de *Salton* [SAL83], é descrita uma técnica de consulta que utiliza a distância entre as palavras).

A figura seguinte resume o processo total de Indexação. Pode-se ver que os documentos são fornecidos à ferramenta de indexação e ao final é produzido um arquivo de índices que consegue localizar os documentos apresentados.

É importante salientar que este é um dos tipos de indexação automática mais simples, e pode ser chamada de *FullText*, pois analisa todo o documento. Esta técnica não considera a semântica do documento e nem a posição sintática das palavras nas orações. Baseando-se nestas duas últimas considerações surgiram outras formas de indexação mais complexas: a indexação *Sintática* e a indexação *Semântica*.

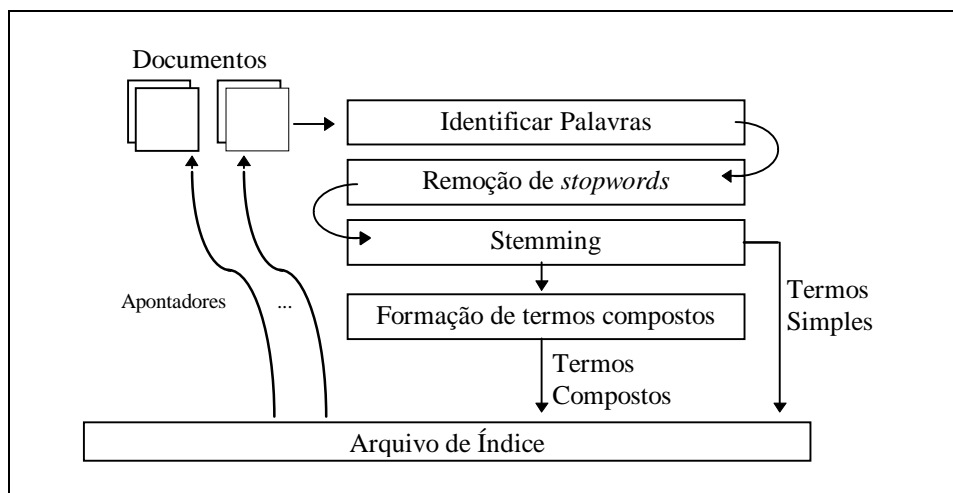


Figura 2.5 - Processo de indexação textual

Na indexação sintática utiliza-se de uma análise sintática para descobrir quais são as palavras mais importantes de uma oração. A linguagem do documento permite que este tipo de análise seja feito, já que as orações (a exemplo do português) possuem posições sintáticas predefinidas para os termos (sujeito, predicado, local do verbo...) e alguns destes termos são mais importantes do que os outros (seus auxiliares).

Somente os termos *importantes* são adicionados a estrutura de índice. Esta técnica exige uma *Base de Conhecimento* que contenha todas as combinações sintáticas possíveis, além de exigir mais poder computacional e tempo. Portanto, geralmente não é utilizada. Um estudo sobre o assunto pode ser encontrado em [SAL88] e [SAL83].

A indexação semântica baseia-se no princípio de que o documento já possui estruturas de formatação que indicam a semântica dos termos. Por exemplo, em HTML existem marcações (*Tags*) que indicam onde se encontram os títulos, as palavras-chave e algumas outras estruturas importantes ao documento.

O processo de indexação deve identificar estas marcações e indexar os termos presentes entre estas marcações com maior importância. Podem surgir alguns problemas, como o da *indexação incerta*, onde a pessoa encarregada de demarcar o documento não utiliza palavras que identificam corretamente o documento. Há alguns trabalhos que se destacam neste assunto: [YAT96] e [MOU92].

3. Modelos de recuperação de informações

3.1. Modelo booleano

O termo *booleano* é derivado do nome do matemático britânico *George Boole*, que no século XIX realizou diversos trabalhos integrando a lógica com a álgebra matemática.

O modelo booleano baseia-se na teoria de conjuntos, onde cada documento é representado por um conjunto de palavras (termos). A união de todos os documentos constitui o universo de documentos. A consulta é expressa por um conjunto de termos, combinados com o sentido de restringir o universo de documentos. Somente os documentos que contêm os termos da consulta são considerados relevantes. Os demais documentos são considerados irrelevantes, o que caracteriza o modelo como sendo restritivo.

Portanto, somente os documentos correspondentes às combinações de termos logicamente verdadeiras, isto é, que atendem todas as especificações da consulta, são retornados. Segundo [CRO94] não há meios de expressar o quanto o documento é relevante para a consulta, e nem há como o usuário determinar a importância dos termos da consulta, especificando quais devem ter maior consideração na análise de relevância de documentos.

Supondo que um usuário deseje localizar documentos sobre *programas de computador*, é necessário que este usuário formule uma consulta ao sistema e indique que os documentos que ele deseja devem conter ambas as palavras.

Há uma série de combinações possíveis entre as palavras “programa” e “computador”. O usuário deve especificar, através dos chamados *operadores booleanos*, quais são as combinações que ele deseja.

Os operadores booleanos disponíveis em qualquer sistema de recuperação de informações que utilize o modelo booleano são o *AND* (e), o *OR* (ou) e o *NOT* (não/negação). Um conjunto de palavras ligadas por operadores booleanos constitui uma *Expressão Booleana*.

Estas *expressões booleanas* oferecem ao usuário três possibilidades básicas, que podem ser combinadas em consultas mais complexas:

- a) **Interseção;**
- b) **União;**
- c) **Subtração.**

3.1.1. Interseção

O operador *AND* permite que o usuário indique que os documentos retornados devem conter ambas as palavras. A maneira utilizada pelo sistema para localizar documentos que atendam estas necessidades é localizar o conjunto de documentos que contém a palavra *X*, localizar o conjunto de documentos que contém a palavra *Y*, e retornar os documentos correspondentes à interseção destes dois conjuntos.

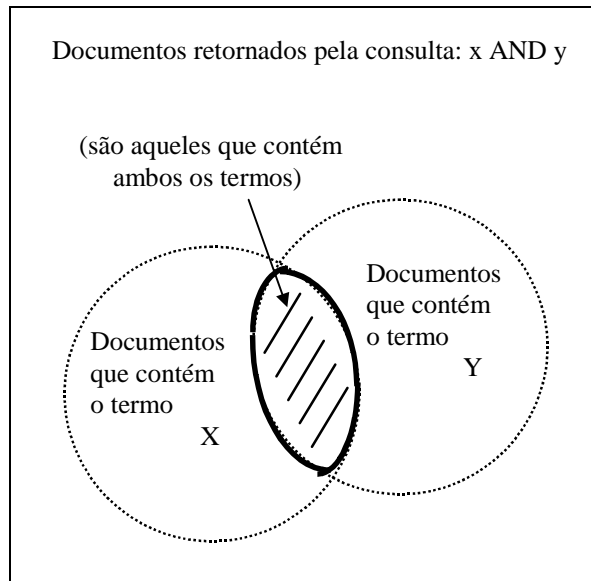


Figura 3.1 - Resultado de uma consulta booleana com o operador AND

3.1.2. União

A união corresponde a localização dos documentos que contém ambas as palavras, independente do fato delas ocorrerem no mesmo documento. O operador de união é o *OR*, e é utilizado quando o usuário deseja documentos que pertençam a mais de um contexto ou quando o usuário deseja abranger um número de documentos maior. Neste caso todos os documentos que possuem o termo *X* ou o termo *Y*, ou ambos, são retornados.

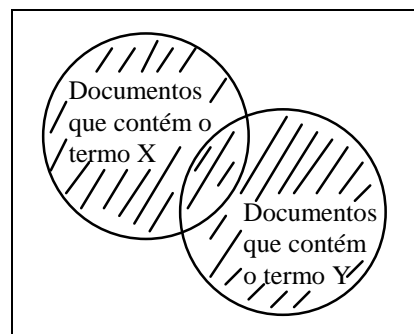


Figura 3.2 - Resultado de uma consulta booleana com o operador OR

3.1.3. Subtração

Corresponde ao operador *NOT*. Uma consulta do tipo *X Not Y* corresponde a localização de todos os documentos que contém *X* com exceção àqueles que também contém *Y*.

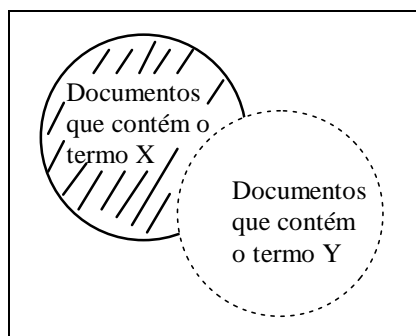


Figura 3.3 - Resultado de uma consulta booleana com o operador NOT

3.1.4. Combinação de operadores em expressões booleanas

Os operadores podem ser utilizados em conjunto, isto é, combinados a fim de restringir ainda mais o conjunto de documentos que deve ser retornado. É possível construir consultas do tipo: *Programa AND Computador OR Micro*. Porém, é necessário estabelecer a ordem de execução dos operadores, pois a ordem de execução influencia no resultado.

Uma opção é deixar o usuário estabelecer a ordem através da utilização de *parênteses* - os indicadores de precedência universais. Caso o usuário não utilize parênteses o sistema deve adotar uma ordem *default*, como por exemplo a execução dos NOT, depois AND e em seguida os OR.

Exemplificando, utilizando-se a consulta: (programa AND computador) OR micro poderiam ser recuperados documentos que tratam do assunto “programa de computador” em conjunto com documentos que possuam apenas a palavra “micro”. Poderiam ainda ser recuperados outros conjuntos, mudando-se a ordem dos parênteses.

Aparentemente a *pesquisa booleana* apresenta resultados satisfatórios, pois permite que consultas complexas sejam realizadas. Apesar disso os usuários (mesmo os mais experientes) têm algumas dificuldades em utilizar este tipo de consulta, especialmente as mais complexas. Por consequência acabam não utilizando toda as facilidades que o método oferece.

Existem muitas ferramentas que utilizam este modelo. Algumas destas ferramentas podem ser encontrados na *Internet*, como é o caso das ferramentas *Altavista* (www.altavista.digital.com), *Excite* (www.excite.com), *Yahoo* (www.yahoo.com) e *Lycos* (www.lycos.com).

Porém na prática, como é o caso das ferramentas citadas, o modelo é aperfeiçoado a fim de proporcionar um melhor desempenho. O modelo aperfeiçoado chama-se *modelo booleano estendido* [SAL83], e neste caso os documentos retornados não precisam pertencer completamente ao conjunto especificado pela consulta.

Nestas ferramentas é possível que o usuário selecione quais termos são mais importantes, quais termos devem aparecer nos documentos e quais termos não devem aparecer nos documentos. Isso é feito especificando-se um sinal de mais (+) na frente dos termos que devem ser localizados e especificando-se um sinal de menos (-) na frente dos termos que *não devem* ser localizados.

Para localizar documentos que satisfaçam a estas requisições o sistema localiza todos os documentos que possuem os termos requisitados pelo usuário - uma operação de união que leva em conta todos os termos. Após são excluídos deste conjunto os documentos que possuem os termos negativos, indicados pelo usuário através do sinal de - (esta é uma operação de subtração). Finalmente, o sistema retorna primeiramente os documentos localizados no interior da interseção de todos os conjuntos de termos que possuem o sinal + (identificados como relevantes pelo usuário); e após são retornados os documentos que fazem parte da união de todos os conjuntos de termos que possuem o sinal + (a periferia, que não satisfaz completamente a consulta).

Deste modo, além dos documentos que satisfazem completamente a consulta, são retornados também os documentos que possuem algumas das palavras requisitadas. Neste caso estes documentos são listados em local posterior aos que satisfazem completamente a consulta.

Pode-se dizer que a ordem de um documento é determinada pela soma aritmética dos termos corretos que ele possui, após terem sido eliminados os termos identificados como negativos (ou exclusivos). Para isto, atribui-se o valor 1 para os documentos que possuem determinado termo e zero para os documentos que não contém este termo. Os documentos são listados em ordem decrescente de pontos obtidos.

Exemplificando, supondo-se que os termos que indexam os documentos A, B e C sejam respectivamente: (recuperação, informação, textual), (recuperação, informação) e (recuperação, informação, visual). E supondo-se que a consulta desejada seja: **+recuperação +informação +textual -visual**. O seguinte resultado seria apresentado: *documento A* (3 pontos) e *documento B* (2 pontos). O documento C não seria recuperado, já que possui a palavra *visual*.

A figura seguinte facilita a identificação dos documentos relevantes no exemplo citado.

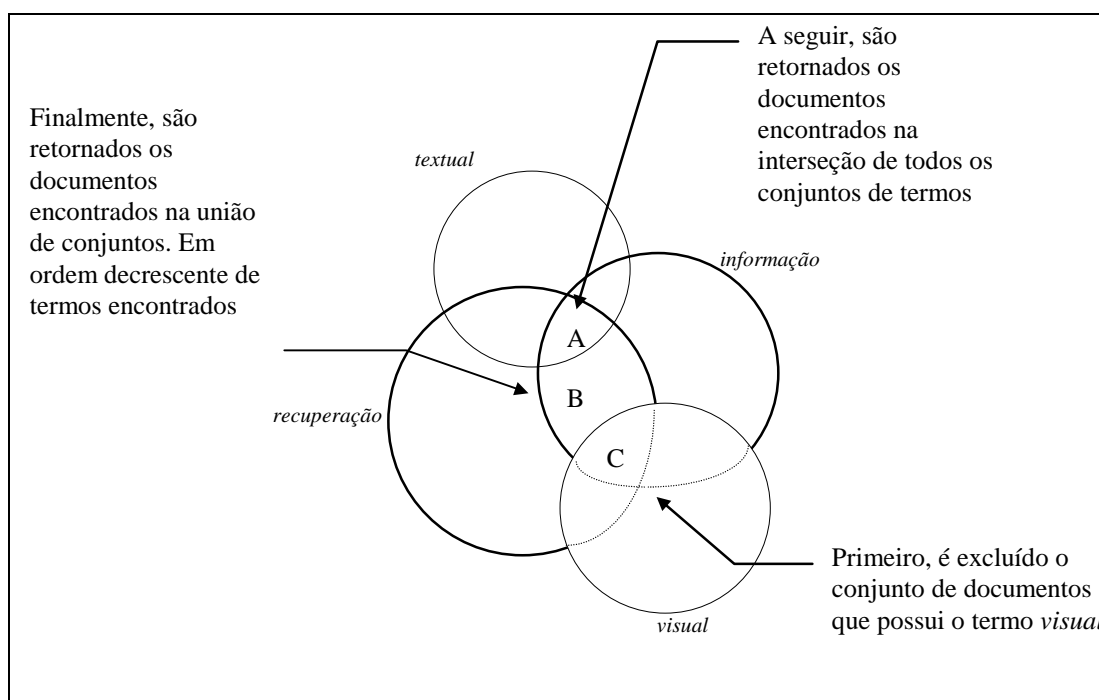


Figura 3.4 - Resultado da pesquisa booleana estendida

3.2. Modelo probabilístico

Apesar da estrutura de índice ser capaz de indicar quais são os documentos que possuem determinado termo, e a consulta *Booleana* conseguir recuperar estes documentos, encontrar a informação que realmente satisfaz a necessidade do usuário não é uma tarefa simples.

São necessários métodos mais eficientes, que consigam de alguma forma distinguir um documento de outro. Não basta recuperar os documentos que possuem determinada palavra, é necessário também especificar o quão importante são estes documentos para esta palavra. Havendo esta distinção alguns documentos podem ser considerados mais importantes do que outros, mesmo que possuam os mesmos termos de consulta.

É justamente este aspecto que o modelo probabilístico busca tratar, ou seja, o objetivo de um sistema de recuperação de informações probabilístico é indicar quais documentos são mais relevantes para o usuário. Segundo [COO92] o sistema deve ordenar os elementos do universo de informações em ordem decendente em relação à sua probabilidade estimada de utilidade para o usuário.

Isso facilita o trabalho do usuário pois ele não precisa analisar todos os documentos retornados para saber se servem ou não à sua necessidade. O resultado de uma consulta é apresentado em uma lista ordenada de acordo com a *relevância*, onde os documentos que aparecem no início desta lista possuem um grau de relação maior com o assunto. O usuário pode simplesmente analisar estes documentos porque *estatisticamente* são mais importantes para ele.

No entanto, estimar a probabilidade de uma informação ser ou não útil para o usuário é tarefa que exige muito poder computacional. Várias técnicas foram desenvolvidas ao longo dos anos. Segundo [VIL95] a maioria destas técnicas tentam identificar o grau de relação entre um termo e um documento.

Estas técnicas atribuem *pesos* ou *graus* de relação entre uma palavra e os documentos em que ela aparece. Elas partem do princípio de que havendo distinção entre os documentos é possível obter uma melhor performance, já que os itens relevantes podem ser recuperados isoladamente sem que os seus *vizinhos* de menor importância sejam recuperados.

Dentre estas técnicas, a que mais se destaca é a análise do *espaço de vetores* (Vector Space Model), descrita a seguir.

3.2.1. Espaço de vetores (vector space model)

A similaridade entre dois objetos quaisquer, documentos ou não, geralmente é identificada pelo número de propriedades (ou características) que estes objetos possuem em comum. É possível também adicionar ao cálculo de similaridade as características que estão ausentes em ambos os objetos.

Neste modelo, desenvolvido por *Salton* [SAL83], os documentos são representados por vetores de termos (características). Estes vetores são representados pela forma $D_i = (t_1, t_2, \dots, t_n)$, onde D_i é o *i-ésimo* documento da base de dados, e t_n o *n-ésimo* termo do documento.

Cada um dos termos do vetor possui um valor associado, que indica o grau de importância deste termo no documento - geralmente chamado de *peso*. Logo, cada documento possui um vetor com pares de elementos na forma: {(palavra1, peso1), (palavra2, peso2), ... , (palavra n, peso n)}.

Os pesos, quando normalizados, possuem valores que variam de zero (0) a um (1). Onde os pesos mais próximos de um (1) indicam termos mais importantes, e pesos menores caracterizam termos menos importantes. Os termos que os documentos não possuem são considerados nos vetores, mas possuem o valor 0.

Segundo [SAL83] o cálculo do peso de uma palavra de um documento pode ser realizado de várias formas. Porém, geralmente este cálculo está baseado no número de ocorrências deste termo no documento e o número de ocorrências deste termo em outros documentos.

Ou seja, a técnica baseia-se na teoria de que as palavras que aparecem com maior frequência em um documento têm uma forte relação com seu conteúdo. Em [SAL87a] diz-se que estes termos de maior frequência são capazes de aumentar a abrangência.

Porém, esta relação tende a diminuir quando este termo aparece em outros documentos. Os termos que aparecem em muitos documentos tendem a recuperar toda a coleção, o que afeta a precisão do sistema, conforme [SAL87a].

Logo, existem dois fatores utilizados no cálculo da relação entre um termo e um documento. Um deles é a *frequência do termo* (term frequency), que mede quantidade de vezes em que determinado termo aparece em um documento. O outro é a *frequência inversa* (inverse document frequency), que indica o número de documentos em que determinado termo aparece.

Com estas informações é possível atribuir um valor de relação entre esta palavra e o documento, e este valor é dado pela fórmula abaixo:

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

Onde $Peso_{td}$ é o grau de relação entre o termo t e o documento d ; a $Freq_{td}$ é a *Frequência do Termo*, o número de vezes que o termo t aparece no documento d ; e a $DocFreq_t$ é a *Frequência Inversa*, que representa o número de documentos em que o termo t aparece.

No entanto, segundo [SAL87a], esta fórmula é criticada devido ao fato do número de documentos relevantes e o número de documentos irrelevantes não serem utilizados no cálculo. Este tipo de informação não é facilmente obtido sem uma análise profunda das informações nos documentos.

Conforme [SAL83] uma série de fórmulas que estimam o peso de um termo, sem a necessidade de informações completas de relevância, foram desenvolvidas. Porém elas são diretamente relacionadas ao modelo específico utilizado pela ferramenta.

Um aspecto importante é a normalização do valor do peso, o que independe da fórmula utilizada. Sem a normalização, segundo [SAL87a], os documentos pequenos

são representados por vetores pequenos; e os documentos maiores são representados por vetores maiores.

Com isso os documentos maiores possuem melhores chances de serem recuperados, já que no cálculo de similaridades receberão valores maiores. O que não acontece na normalização, onde todos os vetores de pesos dos documentos são transformados e possuem um tamanho único.

Este processo de cálculo do peso deve ser realizado em cada documento, para cada termo. De preferência durante o processo de indexação, onde estes pesos são armazenados nos arquivos de índice da coleção de documentos.

Quando uma consulta é colocada pelo usuário, ela também é transformada em um vetor de características, com formato idêntico ao dos vetores dos documentos.

Todo os documentos e a consulta, por serem tratados como vetores, podem ser dispostos em um espaço euclidiano de n dimensões (onde n é o número de termos). Ver figura seguinte.

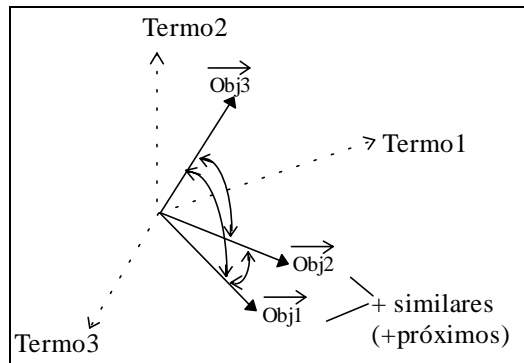


Figura 3.5 - Modelo espaço de vetores

Cada dimensão corresponde a um termo, e o valor do documento em cada dimensão varia entre 0 (irrelevante ou não presente) e 1 (totalmente relevante). A figura seguinte apresenta um espaço com duas dimensões (2 termos) representando os documentos. Neste caso só dois termos são considerados importantes na descrição dos documentos - todos os outros termos são desconsiderados na representação e indexação.

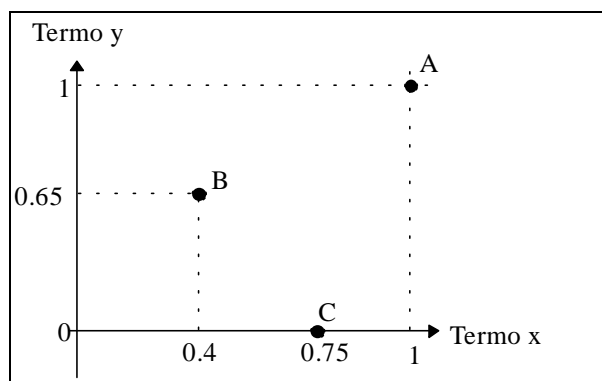


Figura 3.6 - Representação de documentos em duas dimensões

Na figura anterior o documento *A* possui o vetor $\{(x, 1), (y, 1)\}$ pois possui todos os termos e estes são considerados completamente relevantes na descrição e identificação deste documento. O documento *B* possui o vetor $\{(x, 0.4), (y, 0.65)\}$ e o documento *C* possui o vetor $\{(x, 0.75), (y, 0)\}$, pois não possui o termo *y* em sua descrição.

De posse destes vetores é possível calcular a similaridade (ou distância) entre os documentos. A recuperação é realizada fazendo-se uma comparação entre os vetores, onde os vetores de documentos mais próximos do vetor da consulta são considerados mais similares, e, portanto, mais relevantes. Nota-se, deste modo, a complexidade crescente ao aumentar-se o número de termos que descrevem os documentos.

Logo, a recuperação é uma operação neste espaço euclidiano, que busca identificar a similaridade entre as consultas e os documentos. É devido a isso que o modelo é conhecido por espaço de vetores (*Vector Space Model*). A figura seguinte descreve o processo.

Uma das formas de calcular-se a proximidade entre os vetores é testar o ângulo entre estes vetores. Quanto menor for este valor, mais próximos estão os vetores um do outro, ou seja, maior a sua similaridade.

A função de similaridade por cosenos (*cosine vector similarity*), detalhada em [SAL87a], calcula o produto dos vetores normalizados, utilizando a seguinte fórmula:

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

Onde: *Q* é 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.

Com isso é possível calcular o grau de similaridade de todos os documentos da coleção com a consulta e montar uma lista dos mais relevantes - um *rank*.

Além desta, segundo [SAL83], existem várias outras funções capazes de indicar a similaridade entre dois objetos quaisquer. Porém, a similaridade *cosine* é uma das mais utilizadas e menos complexa. Além disso os estudos de *Salton*, no sistema SMART, comprovam sua eficiência, já que muitos dos resultados de consultas realizadas neste sistema encontram-se dentro dos padrões de abrangência e precisão considerados aceitáveis (ver capítulo 2 para obter informações sobre os padrões aceitáveis).

Atualmente existem variações que buscam aperfeiçoar o modelo a fim de obter melhores resultados, como é o caso de [SIG96]. Porém, os estudos indicam que os resultados não são muito significativos.

Alguns consideram que o modelo de espaço de vetores apresenta os resultados de uma forma difícil de ser compreendida para o usuário. Segundo [COO92] o ideal é

informar para o usuário a probabilidade do documento ser relevante para a consulta realizada. Assim fica mais fácil decidir se determinado documento é ou não importante, pois, a princípio, os usuários estão mais acostumados com porcentagem do que com valores entre zero (0) e um (1). No caso da função *cosine*, para obter a porcentagem de relevância bastaria que os valores resultantes fossem multiplicados por 100, mas isto não é feito nos experimentos de *Salton*.

Vários experimentos com este modelo foram realizados. Um dos experimentos de maior sucesso é o sistema SMART [SAL83]. SMART é um sistema desenvolvido durante algum tempo na universidade de Cornell em conjunto com a universidade Harvard, cujo autor foi Gerard Salton.

Uma consulta típica em um modelo como este é apresentada a seguir, como exemplo.

Supondo-se que o usuário deseje recuperar documentos que tratem do assunto recuperação de informações. Este usuário poderia utilizar na consulta os seguintes termos: *Recuperação*, *Informações* e *Documentos*.

Considerando que o arquivo de índice possui a estrutura a seguir:

| <u>Palavra</u> | <u>Documentos</u> | | |
|--------------------|-------------------|------|------|
| <i>Documentos</i> | U(8) | | |
| <i>Informações</i> | A(5) | H(2) | X(7) |
| <i>Recuperação</i> | A(3) | U(4) | X(6) |

Os resultados seriam os seguintes:

Espaço de vetores:

A: {(Recuperação, 3), (Informações, 5), (Documentos, 0)}

H: {(Recuperação, 0), (Informações, 2), (Documentos, 0)}

U: {(Recuperação, 4), (Informações, 0), (Documentos, 8)}

X: {(Recuperação, 6), (Informações, 7), (Documentos, 0)}

Consulta: {(Recuperação, 1), (Informações, 1), (Documentos, 1)}

$$\text{Similaridade(A,Consulta)}: \frac{\{(3 \times 1) + (5 \times 1) + (0 \times 1)\}}{\sqrt{(3^2 + 5^2 + 0^2) \times (1^2 + 1^2 + 1^2)}} = \frac{8}{\sqrt{34 \times 3}} = \frac{8}{10.1} = 0.79$$

$$\text{Similaridade(H,Consulta)}: \frac{\{(0 \times 1) + (2 \times 1) + (0 \times 1)\}}{\sqrt{(0^2 + 2^2 + 0^2) \times (1^2 + 1^2 + 1^2)}} = \frac{2}{\sqrt{4 \times 3}} = \frac{2}{3.5} = 0.57$$

$$\text{Similaridade(U,Consulta)}: \frac{\{(4 \times 1) + (0 \times 1) + (8 \times 1)\}}{\sqrt{(4^2 + 0^2 + 8^2) \times (1^2 + 1^2 + 1^2)}} = \frac{12}{\sqrt{80 \times 3}} = \frac{12}{15.5} = 0.77$$

$$\text{Similaridade(X,Consulta)}: \frac{\{(6 \times 1) + (7 \times 1) + (0 \times 1)\}}{\sqrt{(6^2 + 7^2 + 0^2) \times (1^2 + 1^2 + 1^2)}} = \frac{13}{\sqrt{85 \times 3}} = \frac{13}{15.9} = 0.82$$

Resultado da Consulta:

| <u>Documentos</u> | <u>Graus de Relevância</u> |
|-------------------|----------------------------|
| X | (0 . 82) |
| A | (0 . 79) |
| U | (0 . 77) |
| H | (0 . 57) |

3.3. Modelo difuso (fuzzy)

Segundo [OLI96] o termo *fuzzy* foi introduzido por volta dos anos sessenta em um estudo realizado por *Latfi A. Zadeh*. Neste estudo são apresentados os conjuntos difusos, onde pode-se utilizar a lógica difusa (fuzzy). Pode-se dizer que a lógica fuzzy está para o raciocínio aproximado assim como a lógica tradicional está para a o raciocínio preciso, conforme [OLI96].

Analisando-se os componentes básicos do paradigma da Recuperação de Informações (pontos-chave, descritos no Capítulo 1), verifica-se que a *imprecisão* é uma das características mais marcantes do processo de recuperação de informações.

Na modelagem de documentos há possibilidade de representá-los de forma imparcial ou imprecisa - devido a utilização de uma Linguagem Natural. Há imprecisão, também, na avaliação da relevância dos documentos em relação à consulta do usuário. E finalmente, há imprecisão na formulação de consulta, pois muitos usuários possuem pouco conhecimento sobre o que estão procurando.

Já que a recuperação de informações é caracterizada pela imprecisão, é possível utilizar a teoria *Difusa* (Fuzzy) no processo de recuperação, já que esta está, segundo [CRO92], preparada para manipular com facilidade as incertezas.

Tradicionalmente um objeto pertence ou não pertence a um conjunto através de uma relação exclusiva e total, isto é, não há possibilidade do objeto pertencer parcialmente ao conjunto - ou ele pertence ou não pertence. 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 “mais ou menos”, “muito”, “pouco”, “talvez”, etc.

Na teoria fuzzy, descrita em [ZAD73], há a possibilidade de trabalhar-se com estes valores intermediários, indicando-se o quanto determinado objeto pertence e o quando o objeto não pertence ao conjunto.

É como que se a teoria fuzzy trabalhasse com valores entre 0 e 1. Quanto maior o grau de pertinência do objeto ao conjunto, mais próximo de 1 é o valor correspondente, e quanto mais próximo de 0 é o valor menor é o grau de pertinência. 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 .

Esta é a maior vantagem deste modelo em relação aos outros, já que, ao contrário do modelo probabilístico, é possível determinar se os documentos são ou não são relevantes a uma consulta, além de ser possível também indicar o grau de relevância destes documentos.

O modelo fuzzy é considerado por [CRO94], que apresenta um *survey* sobre o assunto, uma formalização do modelo booleano estendido, pois continuam existindo conjuntos e operadores lógicos que atuam nestes conjuntos, com restrições. Porém estes operadores são aperfeiçoados para que a incerteza possa ser considerada.

De modo geral, a recuperação de informações fuzzy utiliza os conjuntos para representar as informações. Atribui graus de pertinência que determinam a relevância de determinado documento a uma consulta, e utiliza-se de operadores lógicos fuzzy para a definição de consultas.

Os conjuntos que representam os documentos são similares aos conjuntos utilizados na recuperação de informações probabilística, onde há pares {termo, peso} para cada documento. O *peso* é o valor fuzzy (entre 0 e 1) que indica a importância do termo no documento. Uma descrição mais formal, descrita em [CRO94], utiliza a seguinte notação:

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

Onde 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 $\mu_{FI}(d, w)$.

Segundo [CRO94] os graus de importância de um termo podem ser determinados por uma das funções utilizadas recuperação probabilística, pois não há uma função *fuzzy* específica para isto. Pode ser utilizada a função *Peso*, descrita na seção anterior, que utiliza a frequência absoluta do peso no documento e sua frequência inversa. Um outro exemplo, citado por [CRO94], é a frequência relativa, que indica o total de vezes que determinado termo aparece em um documento dividido pelo número total de termos deste mesmo documento.

As consultas também podem ser expressas por conjuntos fuzzy, outra vantagem do modelo, permitindo que o usuário especifique o grau de importância (relevância) de cada um dos termos da descrição da consulta. Termos com menor importância não afetam o resultado. Já os termos com maior importância devem ser priorizados pelo sistema.

Para descrever a consulta o usuário pode utilizar os operadores AND, OR e NOT, restringindo o universo de documentos que devem ser retornados. Para cada um destes operadores há uma ou mais funções fuzzy que podem ser utilizadas. Geralmente, os operadores lógicos são processados pelas seguintes funções, conforme [BAR82]:

- a) **AND** - o operador de interseção booleano é substituído pela função máximo, que retorna o maior valor entre dois valores.
- b) **OR** - o operador de união booleano é substituído pela função mínimo. A função mínimo de dois valores retorna o menor valor.
- c) **NOT** - utiliza-se a função *complemento de um* quando utiliza-se o operador NOT em frente a um termo.

A consulta é manipulada por estas funções e um único conjunto fuzzy é gerado. Este conjunto resultante é o conjunto que representa a intenção do usuário.

A identificação de quais documentos são relevantes à consulta dá-se a partir da identificação do grau de pertinência dos conjuntos de termos que representam os documentos ao conjunto de consulta. Quanto maior o grau de pertinência maior é o grau de relevância.

Conforme [CRO94] um dos maiores problemas do modelo fuzzy diz respeito à complexidade do cálculo de consultas que utilizam muitos operadores. Isso ocorre pelo fato de algumas funções fuzzy serem insensíveis, não levando em conta todos os componentes de uma consulta.

É o caso de uma consulta do tipo “A(0.9) AND B(0.1)”, onde o termo A é determinado como 0.9 pontos relevante para a consulta e o termo B 0.1 pontos. Pelo fato do AND ser substituído pela função mínimo, a fim de combinar o conjunto de A com o conjunto de B, o termo A será considerado irrelevante para a consulta “A AND B”, já que o valor de relevância final tornar-se-á 0.1.

Este problema pode ser solucionado, segundo [CRO94], utilizando-se outros operadores ou funções fuzzy, capazes combinar ou agregar dois ou mais conjuntos fuzzy.

Porém, existem muitos destes operadores para cada operador lógico, não havendo um padrão que dite qual deles deve sempre ser utilizado. Este problema é difícil de resolver. Para cada caso deve ser feita uma análise de qual função obtém melhor os resultados esperados, já que técnicas diferentes obtém resultados diferentes.

Estas funções de agregação de conjuntos fuzzy são chamadas de *agregadores compensatórios*, e buscam compensar os valores de cada termo quando é feita a agregação de conjuntos.

Em [OLI96] podem ser encontrados diversos estudos que demonstram 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.

3.4. Modelo contextual

Nenhum dos modelos apresentados até o momento soluciona completamente os problemas da busca incerta, onde o usuário não utiliza termos apropriados para a descrição da informação de que ele necessita.

Os problemas relacionados à busca incerta já foram discutidos no primeiro capítulo, e ocorrem porque pessoas diferentes podem utilizar termos similares para idéias diferentes ou então utilizar termos diferentes para as mesmas idéias.

Esse tipo de problema não é resolvido facilmente pois a causa de tais imprecisões e ambigüidades pode estar diretamente relacionada com o processo de comunicação, sendo impossível transmitir significados. Conforme [STE72] e [DAV72] só os signos, que são as marcas ou figuras que representam os significados, podem ser transmitidos. E estes dependem muito de cada pessoa, variando de indivíduo para indivíduo.

Porém os signos não andam soltos, há todo um ambiente que os certa - o *Contexto*. Dentro de um contexto os signos são mais facilmente identificados. O contexto é tudo aquilo que envolve um processo de comunicação, seja ou não transmitido explicitamente. Devem ser analisados não só os elementos físicos

envolvidos, mas também os aspectos sociais, humanos e emotivos que acompanham o processo de comunicação. Em documentos, segundo [KON73], o contexto é o resultado da análise do texto pela pessoa.

Portanto, a análise do contexto é indispensável para o bom entendimento dos termos. De outro modo, a busca poderia retornar documentos não relevantes, ou deixar de retornar documentos interessantes para o usuário.

Sabendo disso, muitos dos trabalhos mais recentes sugerem que melhorias significativas na recuperação de informações são obtidas por técnicas que, de alguma forma, compreendam o conteúdo dos documentos e o conteúdo das consultas. A compreensão dos conteúdos depende do contexto considerado. Compreendendo o conteúdo da informação identifica-se o significado dos termos.

A *busca contextual* foi desenvolvida tendo em vista o conteúdo das informações. Através dela espera-se ser possível amenizar o problema do vocabulário. Além disso, espera-se também determinar graus de relação mais adequados que identifiquem melhor quais termos são realmente importantes em uma coleção de documentos, descartando-se os termos menos cruciais - um aspecto muito importante e necessário, segundo [SAL87a].

A abordagem mais usual, definida em [CHE96], define o contexto como sendo um conjunto de palavras que representam o assunto ou a área do conhecimento.

Discute-se em [WIV96a], e em [SAL87a], que normalmente as palavras que co-ocorrem com certa frequência em uma coleção de documentos de um mesmo contexto (assunto) possuem uma forte relação entre si.

Decorrente disto, o modelo contextual busca formar conjuntos de palavras correlacionadas que descrevem um contexto. Deste modo, os termos utilizados nas consultas tendem a recuperar informações mais relevantes, já que mais de uma palavra é utilizada na consulta, maximizando a abrangência e a precisão.

Além disso, estas palavras não são escolhidas arbitrariamente pelo usuário. O sistema de alguma forma identifica o melhor conjunto de palavras que descreve a intenção do usuário e parte para uma busca mais elaborada, com palavras que possuam uma relação mais forte e correta.

A identificação do conjunto de palavras que melhor descrevem um contexto é uma tarefa complicada. Caso este conjunto não seja bem definido a busca contextual pode tornar-se ineficiente. Isso porque a inclusão de palavras incorretas na consulta torna a recuperação de documentos imprecisa ou mais abrangente, recuperando documentos fora do escopo desejado pelo usuário.

Pode ser feita uma analogia matemática, onde se diz que um método é eficiente ou correto quando ele converge, ou seja, *caminha* em direção ao resultado esperado. Uma consulta contextual que utilize palavras corretamente relacionadas leva a ferramenta de recuperação de informações a um conjunto de documentos que satisfaz o usuário, ou seja, o método converge. Caso as palavras que definem o contexto não sejam escolhidas corretamente a localização pode não convergir, não satisfazendo o usuário.

Na figura seguinte tem-se um exemplo de palavras que definem um contexto corretamente. Cada palavra está associada a um subconjunto de documentos pertencentes a um conjunto maior que é o *Universo de Documentos*. Este, contém todos os documentos da base de documentos. Cada palavra utilizada na consulta vai

restringindo o universo, refinando o conjunto de documentos a ser retornado. É lógico que quanto mais palavras forem utilizadas, menor e mais preciso será o conjunto de documentos. O conjunto resultante da intercessão dos conjuntos associados a cada palavra, ou seja, o conjunto de documentos que possuem todas as palavras do contexto, será o conjunto mais relevante para o usuário.

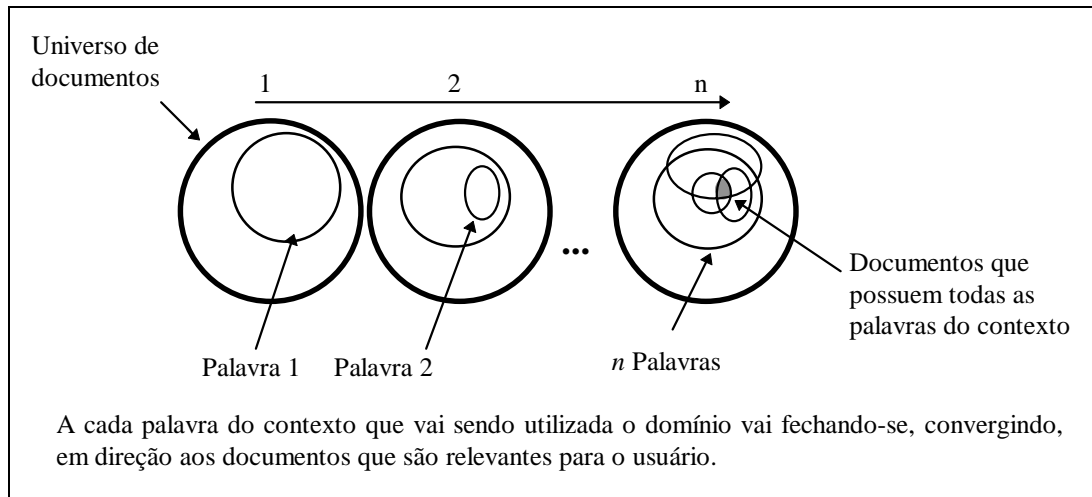


Figura 3.7 - Utilização de um contexto que converge

A figura seguinte descreve o processo errôneo de busca de documentos, que ocorre caso o conjunto de palavras escolhido para definir o contexto esteja incorreto. Por conseqüência, não é possível identificar um conjunto de documentos relevantes para o usuário.

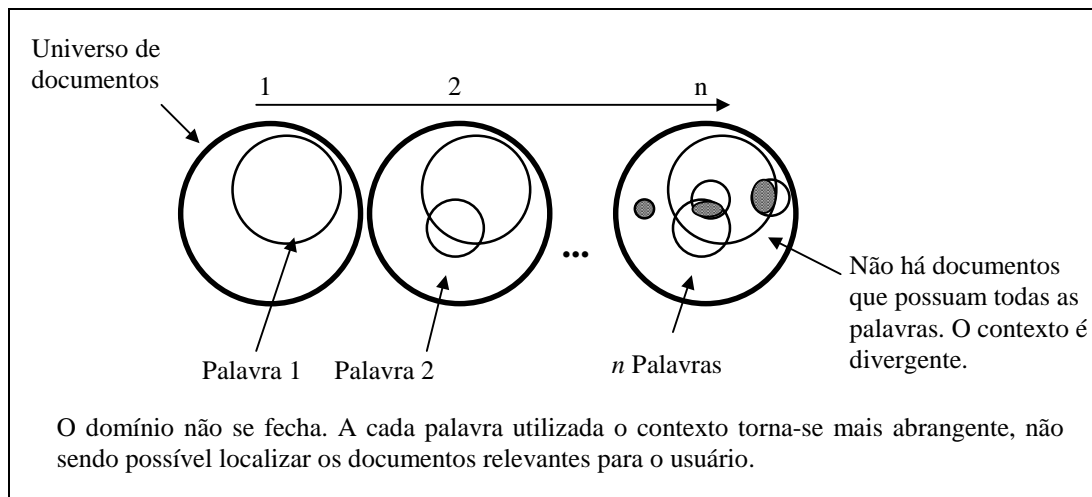


Figura 3.8 - Utilização de um Contexto que não converge

Decorrente destes dois aspectos, é necessário definir o melhor conjunto de palavras que descrevem o contexto. No caso, o melhor conjunto é aquele que abrange todos os documentos do contexto, e somente estes documentos (há precisão).

Na literatura podem ser encontrados muitos métodos de agrupamento de palavras. Métodos que montam classes de Thesaurus [CHE96], dicionários de sinônimos, redes semânticas e clusters [SAL83]. A utilização de palavras compostas (frases-termos) e de termos normalizados também não são novidades (ver capítulo 2). O modelo contextual

busca unificar as melhores características destas técnicas, utilizando-as na busca contextual de informações.

A figura seguinte apresenta uma palavra (signo *Jogador*) que de acordo com o contexto possui um significado diferente. A figura mostra também que as palavras que se relacionam com ela também mudam com o contexto. O número de relacionamentos também pode variar com o contexto.

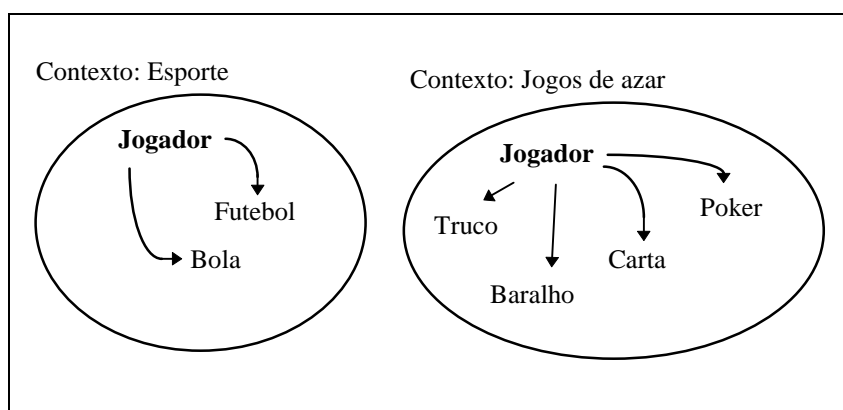


Figura 3.9 - Uma palavra em mais de um contexto

Uma ferramenta de recuperação de informações, que utilize o modelo contextual, deve analisar a palavra fornecida pelo usuário e identificar o contexto em que ela enquadra-se. Os contextos podem ser armazenados em estruturas especiais, um *hiperdicionário* (pág. 47) por exemplo, capazes de recuperar facilmente todos os relacionamentos de determinada palavra em um determinado contexto.

Após são utilizadas na consulta todas as palavras que estão relacionadas com ela no mesmo contexto.

O processo de identificação do contexto de uma palavra torna-se complicado no caso de haver mais de um contexto. É o caso da palavra *Jogador* demonstrado na figura anterior. Neste caso há o contexto *Esporte* e o contexto *Jogos de Azar*. Existem várias soluções para este problema. Uma delas é mostrar os contextos para o usuário, que então deve indicar qual é o contexto por ele desejado.

Outras formas incluem a descoberta do perfil do usuário, através da análise de consultas anteriores, que pode indicar sua tendência. Caso o usuário tenha realizado consultas anteriores na área do esporte talvez este contexto seja o mais adequado para a nova consulta. Mas nem sempre isto é válido. O usuário pode ter mudado de idéia e estar realizando consultas em outras áreas. A análise de perfil e histórico do usuário exige técnicas que ainda estão em estudo. Alguns resultados sobre este tipo de pesquisa podem ser obtidos em [LAM96].

Há ainda a opção de recuperar-se todos os documentos de todos os contextos em que a palavra encontra-se. Após, utilizando técnicas de *feedback* (ver pág. 42), é possível restringir os documentos através de refinamentos sucessivos, até que o contexto adequado seja identificado.

Como nos outros modelos, os documentos são retornados em ordem de relevância, onde aqueles que obtêm uma pontuação maior são considerados mais relevantes.

A pontuação de um documento é determinada pela presença dos termos de busca. Quanto mais termos do contexto forem encontrados no documento maior será a sua pontuação. Poderiam existir técnicas mais sofisticadas, que utilizassem a lógica *fuzzy* para determinar o quão importante é determinada palavra para o contexto. Isso porque algumas palavras podem ser mais importantes do que outras na identificação do contexto de um documento. Porém, não há resultados científicos que avaliem a utilização destas técnicas.

4. Ferramentas de auxílio

Nem sempre as técnicas empregadas nos modelos de recuperação de informações são suficientes para a resolução dos problemas inerentes ao paradigma, citados no segundo capítulo.

Dentre os problemas existentes o problema do vocabulário é o mais comum de ser encontrado, e um dos mais difíceis de serem solucionados. Devido à isso, talvez este seja o problema mais discutido nos artigos da área.

Relembrando, o problema do vocabulário diz respeito à utilização de um termo (característica) que não está presente no índice ou na coleção de documentos. Nestes casos é provável que o contexto (assunto) que envolve o termo esteja presente, mas este termo específico não é utilizado para descrever a informação.

Fazendo-se uma análise mais detalhada, pode ser possível identificar vários documentos enquadrados em um mesmo assunto, dos quais somente alguns possuem o termo requisitado pelo usuário e outros não. Esse tipo de problema resulta em uma abrangência muito baixa.

É possível também que um termo seja utilizado em vários documentos de assuntos variados. Neste caso a precisão é baixa, devido à recuperação de documentos variados.

Todos estes problemas sugerem que as palavras, quando utilizadas individualmente, não conseguem captar a semântica do documento. Há a necessidade de dicionários de sinônimos, categorias de palavras, generalizações, especializações, entre outros.

É buscando solucionar estes problemas que são desenvolvidas as ferramentas de auxílio, que podem ou não estar integradas à ferramenta de recuperação de informações.

Estas ferramentas têm o objetivo de auxiliar o usuário na escolha do melhor termo (da melhor característica) que descreve o documento (o objeto) que ele está necessitando.

A seguir são citadas algumas destas ferramentas.

4.1. Relevance feedback

Segundo [SAL87b], *Relevance Feedback* é uma técnica desenvolvida com o objetivo de facilitar o processo de recuperação de informações, utilizada como ferramenta de auxílio à formulação de consultas mais elaboradas.

Esta técnica originou-se a partir da observação de que a maioria dos usuários possui dificuldades em formular consultas. Esta dificuldade advém de problemas já discutidos anteriormente, que são a falta de conhecimento sobre o assunto ou sobre o sistema utilizado.

Geralmente o usuário faz uma tentativa inicial na sua primeira consulta. No resultado desta consulta inicial o usuário é capaz de obter uma visão geral sobre o assunto, melhorando seus conhecimentos e descobrindo os termos utilizados nos documentos. Decorrente disto, o usuário vai sucessivamente refinando sua consulta

baseando-se no resultado obtido, excluindo ou adicionando novos termos, guiando a consulta em direção ao melhor conjunto de informações.

Visto que este processo é manual procurou-se automatizá-lo, fazendo com que o sistema refine automaticamente a consulta inicial do usuário. É claro que este processo depende muito do usuário, pois ele deve comentar a qualidade do resultado apresentado pelo sistema.

Segundo [DUN97] o processo funciona da seguinte forma: o usuário especifica quais dos documentos retornados são mais relevantes e quais não são. A partir destas informações, o sistema adiciona à consulta alguns termos presentes nos documentos considerados relevantes pelo usuário e exclui os termos presentes nos documentos irrelevantes. Deste modo acredita-se que, sucessivamente e iterativamente, o sistema vá adequando corretamente a consulta, obtendo uma melhor performance.

Em [GUP97] há um exemplo de utilização desta técnica em uma ferramenta de recuperação de informações visuais (imagens, vídeos). Inicialmente o usuário descreve as características da imagem que ele deseja, desenhando-as em um editor gráfico simples que acompanha a ferramenta. Supondo que ele deseje uma *ave*, basta que ele desenhe algo similar ao que se encontra na figura seguinte.

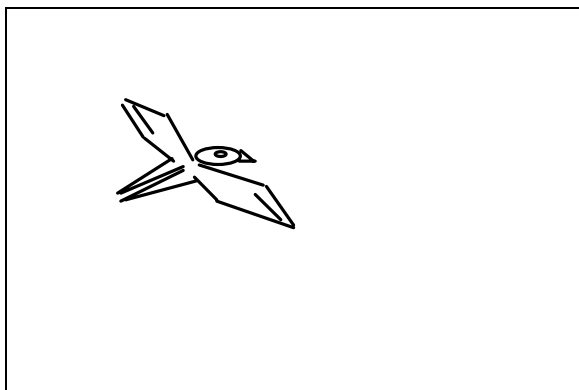


Figura 4.1 - Figura abstrata de uma ave

A seguir, o sistema analisa a imagem e seu conteúdo tentando descobrir sua semântica. Então, o sistema busca as figuras que possuem objetos similares e apresenta-as para o usuário. Após, o usuário visualiza as figuras retornadas pelo sistema e altera a consulta inicial (o desenho) retirando ou adicionando novas características, refinando-a até que o sistema encontre a figura correta.

Por outro lado, o usuário pode selecionar algumas das imagens recuperadas indicando aquelas que mais se assemelham com o que ele está procurando. Deste modo o sistema analisa estas imagens descobrindo o que elas têm em comum e, automaticamente, refaz a consulta buscando novas imagens.

Em [SAL87b] há uma análise sobre as várias técnicas desenvolvidas durante as décadas de setenta (70) e oitenta (80). Neste trabalho o objetivo não é discutir qual das técnicas é a melhor, até mesmo porque muitas continuam surgindo a todo o momento. Porém, avalia-se cada técnica em si, verificando se, de um modo geral, ela apresenta as seguintes vantagens:

- a) Ela auxilia o usuário no processo de formulação de consultas, permitindo a construção de requisições mais úteis sem que haja a necessidade do usuário conhecer o sistema ou o assunto;
- b) A operação de consulta é quebrada em várias etapas. Deste modo o usuário vai gradualmente aproximando-se do assunto desejado, recuperando documentos mais relevantes;
- c) O processo de refinamento das consultas é controlado, já que a ferramenta é que o faz;
- d) O processo de feedback é facilmente implementado, não necessitando muitas alterações na interface do sistema;
- e) A interface do sistema pode tornar-se bem mais agradável e amigável. Como exemplo, ver ferramenta Altavista (www.altavista.digital.com);
- f) Segundo [BUK95] a técnica chega a apresentar resultados de 10 a 15% melhores do que quando não é utilizada.

Há também experimentos mais recentes ([HAR92], [AAL92] e [ALL96]) que buscam aperfeiçoar a técnica. Porém, os resultados não apresentam mudanças significativas.

4.2. Thesaurus

O *Thesaurus* é uma estrutura hierárquica de palavras, permitindo que o usuário descubra os relacionamentos entre estas palavras. Estas palavras são geralmente agrupadas em classes onde cada classe possui um termo-chave, que a identifica.

O usuário pode navegar nesta estrutura hierárquica e localizar todas as palavras que pertencem à uma mesma classe ou categoria. Com isso é possível saber quais termos são mais abrangentes e quais são mais precisos, já que a utilização de uma classe contendo várias subclasses indica que existem muitas subcategorias (e portanto é muito abrangente). Ao aproximar-se dos ramos folha os termos vão tornando-se mais específicos.

Através da estrutura de classes *Thesaurus* o usuário pode facilmente combinar os termos a fim de conseguir uma abrangência ou precisão maior, conforme sua necessidade. Isso porque a hierarquia facilita a identificação de palavras relacionadas, e dá uma idéia geral sobre o contexto dos assuntos.

O *Thesaurus* pode também ser utilizado para fornecer a tradução de termos entre domínios. O usuário fornece um termo e ele fornece os sinônimos ou os termos que devem ser utilizados no sistema.

Em [CHE96] são apresentados os resultados de uma experiência onde biólogos de especialidades diferentes são convidados a utilizar ferramentas de recuperação de informações. O primeiro grupo possui especialidade em *minhocas*, e o segundo em *moscas*. Muitos conceitos utilizados em um grupo são similares aos conceitos utilizados no outro, porém os termos empregados são diferentes. Neste caso o *Thesaurus* mostrou-se uma boa opção pois facilitou a utilização de um sistema de uma área específica por especialistas de outra área. O especialista, especificando o termo desejado, conseguia obter o sinônimo correspondente em outra área, facilitando a consulta.

A seguir é apresentado um exemplo de uma estrutura Thesaurus que apresenta o conhecimento relacionado em classes. A figura representa a expansão de uma destas classes, a classe *esporte*, que possui como subclasses *futebol* e *voleibol*. Esta estrutura apresenta o relacionamento entre diversas palavras, incluindo sua abrangência. O usuário *navega* nesta estrutura procurando a palavra desejada.

No caso a utilização da palavra *esporte* deve retornar informações (documentos) que tratam de todos os assuntos (tais como esporte, jogadores, goleiros...) contidos nas subclasses. Porém a palavra *goleiro* está dentro de um sub-contexto de *Esporte*, e deve retornar um numero menor e mais preciso de informações (informações que tratam somente de goleiros).

Neste thesaurus exemplo mostra-se uma vantagem da estrutura, que pode conter definições, sinônimos, termos similares e traduções da palavra selecionada.

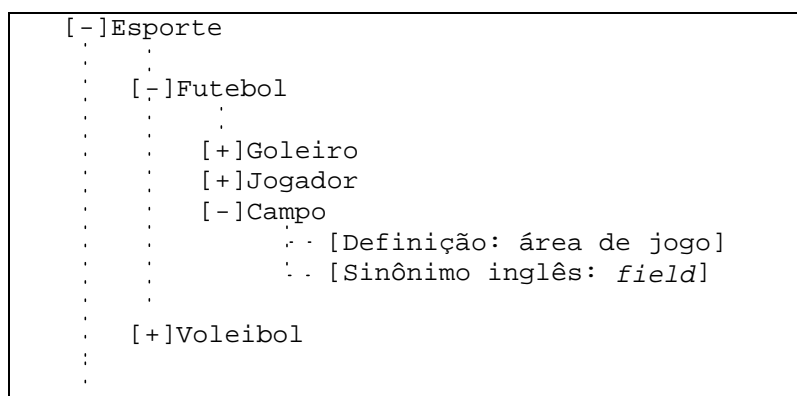


Figura 4.2 - Estrutura de um thesaurus

A utilização de termos abrangentes significa na recuperação de grandes proporções de documentos. Apesar, é possível torna-los menos gerais combinando-os com outros. Já a utilização de termos mais precisos faz com que poucos documentos sejam retornados.

Nos trabalhos [CHE96, CHE94b] são realizados vários experimentos sobre a utilização de estruturas deste tipo, além de discutidas as utilidades de sua utilização. Foram feitos vários testes e foi apresentada uma possível solução para o problema do vocabulário, com um algoritmo de construção automática.

Segundo [SAL83] as vantagens de utilizar-se esta estrutura são:

- a) Ajudar o usuário na formulação de consultas, indicando termos mais adequados que possivelmente recuperem um número de informações maior;
- b) Familiarizar o usuário com o vocabulário do sistema.

A estrutura também pode ser utilizada para normalizar os termos utilizados em um sistema, pois o vocabulário descontrolado pode ser substituído por identificadores de categorias.

Podem existir estruturas *Thesaurus* em forma de grafos, onde cada nodo representa um termo e as ligações entre estes nodos representam o tipo de relação entre eles (sinônimos, antônimos, etc). As ligações podem possuir valores associados, que indicam o grau da relação.

O Thesaurus pode ser construído manualmente, por um especialista que identifica os relacionamentos entre as classes de palavras, ou automaticamente, através de técnicas estatísticas de correlação que identificam relações entre palavras e colocam-nas em classes. Estas técnicas de construção podem ser encontradas em [CHE96], [SAL83] e [CRO92].

A maior dificuldade do Thesaurus está em mantê-lo atualizado, principalmente se a quantidade de informações incluídas ou atualizadas no sistema for muito alta. Isso impossibilita a atualização em tempo real desta estrutura, pois as técnicas de atualização e construção exigem uma análise complexa das informações.

4.3. Dicionários

Os dicionários têm um papel importante na área de recuperação de informações. Seu uso advém das técnicas de linguagem natural que necessitam de uma estrutura capaz de prover informações sobre a morfologia das palavras, informações sintáticas e sinônimos.

Na literatura costuma-se chamar estes dicionários de *lexicons*. O *lexicon*, conforme [BRI91], é uma estrutura similar a um dicionário que possui o papel de oferecer um vocabulário adequado a uma aplicação. Em [BRI91] comenta-se o fato do termo *dicionário* ser atribuído ao convencional objeto impresso, ideal para uso humano. Já o termo *lexicon* é uma formalização que diz respeito a um componente lingüístico, possivelmente implementado e estruturado com o objetivo de suprir informações lingüísticas a um outro software.

Utilidades:

- a) Dicionários de sinônimos para auxílio a formulação de consultas;
- b) Auxílio na seleção de termos mais adequados quando o usuário necessitar de um contexto mais amplo (indicando palavras mais abrangentes) ou contexto menos genérico (indicando palavras mais precisas);
- c) Fornecer significados de combinações de palavras, como por exemplo os termos compostos ou frases-termo;
- d) Indicação de vocábulos mais adequados ao sistema, inclusive identificando o significado de jargões locais e termos de outras línguas;
- e) Dicionários morfológicos e sintáticos para auxílio ao processamento de linguagem natural. Estes podem ser utilizados por uma ferramenta de indexação de informações textuais, permitindo uma análise mais detalhada sobre as sentenças e facilitando a fase de análise léxica;
- f) Auxílio a construção de corretores ortográficos para que os usuários possam desenvolver melhor os documentos que serão adicionados na base de informações. Há a possibilidade de padronizar os termos utilizados, e neste caso o dicionário pode sugerir os termos que devem ser empregados.

4.4. Hiperdicionários

O Hiperdicionário é uma estrutura que armazena relacionamentos entre palavras de um mesmo contexto. Esta estrutura é amplamente utilizada como base de conhecimento para ferramentas de busca de informações contextuais.

A estrutura em si não apresenta maiores dificuldades, pois ela nada mais é do que uma estrutura de *nodos*, que representam as palavras, interligados por *elos*, que representam as relações entre estas palavras. Os elos possuem informações sobre o tipo de relacionamento, o grau deste relacionamento e o contexto deste relacionamento. Ver a figura seguinte.

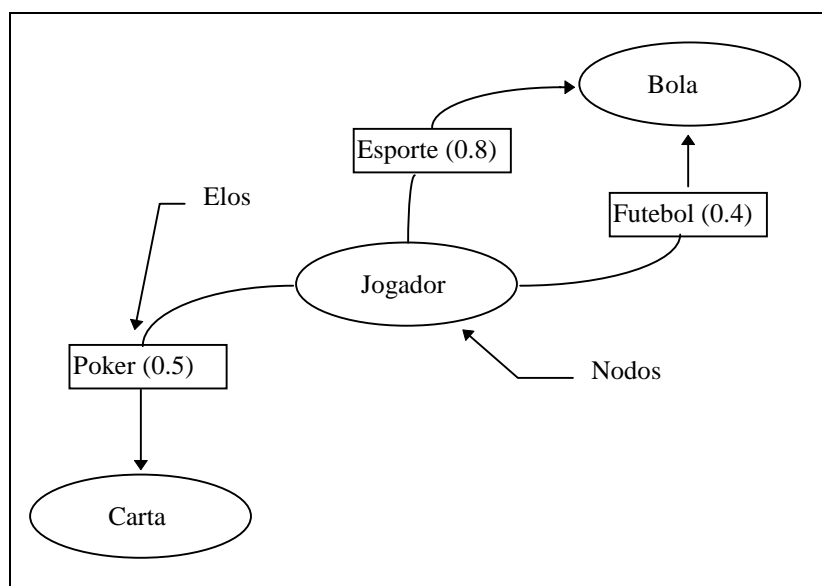


Figura 4.3 - Estrutura de um hiperdicionário

É possível percorrer esta estrutura e identificar quais são as palavras que pertencem a um determinado contexto. É possível também saber o quanto uma palavra está relacionada com o contexto ou também o quanto uma palavra está relacionada com outra palavra em determinado contexto.

Esta estrutura pode ser construída manualmente ou automaticamente.

4.4.1. Montagem manual

Para a montagem (construção) manual de contextos é necessário um especialista capaz de identificar os termos e os relacionamentos entre as diversas palavras nos diversos contextos.

É claro que não existem especialistas especializados em todas as áreas, mas não é obrigatório que somente um especialista participe do processo. Além disso, muitas vezes a estrutura é utilizada em uma área específica, como a Medicina, e neste caso é possível a utilização de um único especialista.

No caso da Medicina um médico pode ser o especialista. Este médico deve ser uma pessoa já acostumada com a área, e que sabe quais são os termos empregados por seus colegas e todas as outras pessoas que trabalham na mesma área. Deste modo este especialista é capaz de aumentar a abrangência do contexto, onde serão recuperados

todos os documentos relevantes ao assunto, e a precisão, recuperando somente estes documentos.

Uma boa técnica, no caso da medicina, é elaborar contextos sobre doenças onde são descritos os sintomas e os procedimentos envolvidos em seu tratamento.

Os maiores problemas da montagem manual de contextos estão relacionados com o tempo, disposição e conhecimentos do especialista. Muitas vezes os especialistas não possuem tempo suficiente para esta tarefa, que é minuciosa; além disso, o especialista pode não ser capaz de identificar todas as palavras relevantes do contexto.

4.4.2. Montagem automática

Alguns dos problemas relacionados com a montagem manual podem ser amenizados com a montagem automática. A princípio a montagem automática de contextos foi desenvolvida visando substituir o especialista. Porém, estudos sugerem que as duas abordagens sejam utilizadas em conjunto. Em [CHE94b] a montagem automática de relacionamentos é utilizada como um complemento à manual, facilitando o trabalho do especialista.

A montagem automática baseia-se em técnicas fuzzy ou estatísticas de análise de co-ocorrências, já citadas anteriormente, e são similares à técnica utilizada na montagem automática de *Thesaurus* utilizada por [CHE96]. São métodos estatísticos que baseiam-se na análise de ocorrência das palavras nos documentos. Estas técnicas são aplicadas sobre a base de informações, já que é nos próprios documentos que os contextos estão descritos.

Geralmente as técnicas de identificação de contextos possuem as seguintes etapas: a *identificação de palavras nos documentos*, a *determinação do grau de relação entre as palavras e o documento que as contém*, e a *análise das relações entre as palavras*.

Para identificar as palavras nos documentos podem ser utilizadas as técnicas de indexação descritas no capítulo 2. Já as outras duas etapas baseiam-se na premissa de que as palavras que aparecem repetidamente em um único documento e as palavras que aparecem em muitos documentos são boas candidatas, conforme [CHE96].

Após selecionadas as palavras que devem fazer parte do processo, é realizada uma análise de co-ocorrência das palavras nos documentos. É através desta análise que é possível definir o grau de relação de cada palavra com o contexto em questão.

Duas fórmulas são utilizadas para esta análise: a fórmula que analisa o grau de relação entre uma palavra e um documento, e a fórmula que analisa as relações entre palavras (definindo assim os contextos).

A fórmula abaixo define a relação entre uma palavra e o documento em que ela aparece, onde d_{ij} é o valor combinado da palavra j no documento i :

$$d_{ij} = tf_{ij} \times \log \left(\frac{N}{df_j} \right)$$

N representa o número total de documentos na BD, tf_{if} é a frequência da palavra j no documento i e df_j é a frequência inversa de documentos (número de documentos em que a palavra j aparece).

A segunda fórmula avalia os resultados gerados pela fórmula anterior, detectando as relações entre as palavras:

$$\text{Valor combinado} = \frac{\sum_{i=1}^n d_{ijk}}{\sum_{i=1}^n d_{ij}}, \text{ onde } d_{ijk} = tf_{ijk} \times \log \left(\frac{N}{df_{jk}} \right),$$

sendo que tf_{ijk} representa o número de ocorrências de ambas as palavras j e k no documento i (o menor número de ocorrências entre as palavras deve ser escolhido), df_{jk} representa o número de documentos (em uma coleção de N) no qual as palavras j e k ocorrem ao mesmo tempo.

Desta forma, é possível identificar as relações entre as palavras em diversos contextos, criando a base de Contextos - Hiperdicionário. É importante lembrar que duas palavras podem estar relacionadas entre si em mais de um contexto e para cada contexto pode existir um grau de relação diferente.

Em [WIV96b] são relatados experimentos realizados com uma ferramenta Hiperdicionário. Esta ferramenta utiliza informações sobre os sintomas de doenças que podem ser causadas pela utilização de drogas. O contexto de cada droga é montado, e após utilizado em prontuários médicos a fim de localizar pacientes com sintomas similares.

Nos trabalhos de [WIV96a] e [WIV96b] sugere-se que os documentos que vão ser analisados na montagem automática pertençam todos a um mesmo contexto. Portanto deve ser feita uma análise prévia de um especialista, que separa os contextos. Desta forma os contextos vão sendo analisados separadamente, um após o outro, tornando a análise mais rápida e eficiente. Os resultados são satisfatórios mas indicam que ainda são necessários aperfeiçoamentos.

Algumas considerações sobre a construção automática de hiperdicionários:

- a) O processo de construção de uma base de conhecimento estilo *hiperdicionário* é trabalhoso pois as técnicas de identificação de relações entre palavras exigem muito poder computacional;
- b) Quando o critério de construção adotado é muito rígido o número de relações encontradas torna-se muito pequeno e o ganho de performance do sistema não é significativo;
- c) Quando o critério de construção é relaxado, utilizando discriminação mais fraca, um número maior de relações significantes é encontrado, mas também muitas relações sem utilidade são construídas, e neste caso é necessário realizar uma filtragem muito rígida;
- d) Na maioria dos casos, segundo alguns estudos realizados em [SAL87a], os relacionamentos encontrados são válidos somente na coleção onde foram descobertos (onde o ganho de desempenho pode chegar a 20%). Porém o estudo é baseado em técnicas mais antigas. As técnicas mais atuais, [CHE96] por exemplo, apresentam resultados mais animadores.

5. Conclusões

Neste trabalho foram apresentados os problemas envolvidos no paradigma clássico de recuperação utilizado na maioria dos sistemas de recuperação desenvolvidos até o momento.

Foram apresentados também os modelos tradicionais de recuperação de informações, utilizados por estes sistemas. Estes modelos buscam solucionar alguns dos problemas inerentes ao paradigma.

Foram abordadas também algumas ferramentas ou técnicas que podem ser utilizadas em conjunto com os modelos, e que, quando combinados, obtêm uma melhor performance e minimizam os erros.

Porém o assunto discutido neste trabalho é muito amplo, sendo difícil realizar um *survey* (resumo abrangente e preciso) da área em tão pouco tempo. Além disso, como já salientado, a área de recuperação de informações está em constante desenvolvimento devido a própria natureza da informação. Portanto novas técnicas surgem a cada momento.

Uma das maiores preocupações atuais é a possibilidade de integração dos diversos tipos de informação existentes, devido à WEB³ e às *Bibliotecas Digitais*⁴.

Nestes dois ambientes, segundo [WIE96] e [DAO96], há a necessidade de **novos modelos e técnicas** que consigam manipular informações heterogêneas neles contidas (textos, vídeos, imagens, sons, multimídia). As técnicas atuais não estão capacitadas à manipulação dos tipos de informação heterogêneos já existentes ou que estão por vir.

Espera-se que este trabalho possa servir como introdução ao assunto para aqueles que desejem aperfeiçoar as ferramentas de recuperação de informações existentes ou que queiram desenvolver outras novas.

A seguir são ressaltados alguns aspectos e conclusões importantes do estudo realizado neste trabalho, que podem ser considerados como direções interessantes para pesquisas futuras:

³ Segundo [DAO96] a WEB, ou WWW, é um ambiente que se desenvolve rapidamente, além de ser uma rede de informações em larga escala, que possui informações interligadas mantidas por autônomos. É um grande e poderoso meio para compartilhar e distribuir múltiplos meios de informação. Apesar disso, um dos grandes problemas deste ambiente é a falta de abstrações conceituais (recursos) capazes de manipular o tamanho e a diversidade das informações que possui.

⁴ As Bibliotecas Digitais são o futuro da informação eletrônica, segundo [WIE96]. Nelas, o material básico é o texto, mas este pode conter gráficos, imagens, sons e vídeos. Esse tipo de recurso gera um novo desafio para a área de recuperação de informações, pois não há uma estrutura comum (os dados são heterogêneos) e existe muita redundância nos dados. Além dos fatores segurança e privacidade, que não devem ser deixados de lado.

- a) O *problema do vocabulário* (vocabulary problem) ainda não apresenta solução definitiva. Porém a utilização de dicionários de sinônimos ou *thesaurus* oferece resultados bem melhores, solucionando em parte este problema. O modelo *contextual* também é uma técnica que minimiza as diferenças de vocabulário, mas a identificação destas diferenças através de técnicas automáticas (ou não) necessita de algumas melhorias.
- b) O *modelo booleano*, utilizado por grande parte das ferramentas de recuperação de informações *on-line* - WEB, apresenta muitas deficiências. A abrangência de tal modelo é muito grande, pois os usuários têm dificuldades na utilização de termos corretos que descrevam sua necessidade (conforme já analisado no capítulo 1). O modelo estendido é uma solução para o modelo booleano tradicional, porém os estudos realizados (capítulo 1 e 3) indicam que descrever uma consulta desta forma é tarefa trabalhosa para a maioria dos usuários. A técnica de *feedback* é uma solução interessante para este problema, pois o usuário vai refinando gradualmente sua consulta, através de análise dos resultados intermediários.
- c) Estudos anteriores [WIV96b] indicam que a *busca contextual* apresenta resultados melhores, porém mais demorados. Cabe salientar, também, que o sucesso desta abordagem depende em muito de como a base de contextos é criada. Uma boa base permite melhores interpretações dos interesses do usuário, enquanto que uma base pobre ou mal-definida ocasiona o retorno de documentos não desejados ou não importantes.
- d) A análise literária realizada neste trabalho sugere que ferramentas mais inteligentes, capazes de compreender o usuário (através da análise de perfil e assistentes de consulta) e o conteúdo das informações da base de dados (através da análise contextual), podem ser a solução para muitos problemas inerentes ao paradigma, discutidos no primeiro capítulo. Porém não há estudos que comprovem esta afirmação.
- e) O modelo *fuzzy* está sendo muito estudado ultimamente. Sugere-se que estudos utilizando a lógica *fuzzy* em conjunto com o modelo *contextual* sejam realizados, pois são dois modelos que apresentam, individualmente, resultados animadores. Ambos podem ser a solução para muitos dos problemas que estão surgindo com os novos tipos de informação, já que o contexto envolve muito mais coisas do que o simples texto - conhecimento heterogêneo, e a teoria *fuzzy* trabalha muito bem com as incertezas inerentes ao homem.

Referências bibliográficas

- [AAL92] AALBERSBERG, Ijsbrand J. Incremental Relevance Feedback. In: ACM SIGIR'92. **Proceedings...** Copenhagen: ACM PRESS, 1992. p.11-22.
- [ALL95] ALLAN, James. Relevance Feedback with Too Much Data. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 337-343.
- [BAR82] BARTSCHI, M; FREI, H. P. Adapting a Data Organization to the Structure of Stored Information. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. **Proceedings...** Berlin: ACM PRESS, 1982.
- [BRI91] BRISCOE, Ted. Lexical Issues in Natural Language Processing. In: Natural Language and Speech. **Proceedings...** Springer-Verlag. 1991. p.39-68.
- [BUK95] BUCKLEY, Chris; SALTON, Gerard. Optimization of Relevance Feedback Weights. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 351-357.
- [CAT96] CATARCI, Tiziana. Databases and the Web: New Requirements for Easy Access. **ACM Computing Surveys**. Nº28, Dezembro. 1996. Disponível por WWW em <http://www.acm.org/pubs/citations/journals/surveys/1996-28-4es/a131-catarci/> (em Junho de 1997).
- [CHE96] CHEN, H. at alli. **A concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System.** Disponível por WWW em <http://ai.bpa.arizona.edu/papers> (2 de Julho de 1996).
- [CHE94a] CHEN, H. **The Vocabulary Problem in Collaboration.** Disponível por WWW em <http://ai.bpa.arizona.edu/papers> (5 de Julho de 1996).
- [CHE94b] CHEN, H. at alli. **Generating a Domain-specific Thesaurus Automatically: An Experiment on flyBase.** Disponível por WWW em <http://ai.bpa.arizona.edu/papers> (5 de Julho de 1996).
- [CHE94c] CHEN, Hsinchun. **A textual database/knowledge-base coupling approach to creating computer-supported organizational memory.** Disponível por WWW em <http://ai.bpa.arizona.edu/papers> (5 de Julho de 1996).

- [CHU95] CHURCH, Kenneth W. One Term or Two?. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 310-318.
- [COO92] COOPER, Willian et alli. Probabilistic Retrieval Based on Staged Logistic Regression. In: ACM SIGIR'92. **Proceedings...** Copenhagen: ACM PRESS, 1992. p. 198-210.
- [CRO95] CROFT, Bruce. **What Do People Want from Information Retrieval?**. Disponível por WWW em <http://cnri.dlib/november95-croft> (1 de Julho de 1997).
- [CRO94] CROSS, Valerie. Fuzzy Information Retrieval. **Journal Of Intelligent Information Systems**, Boston, v.3, n.1, p. 23-56, Fevereiro. 1994.
- [CRO92] CROUCH, Carolyn J; YANG Bakyung. Experiments in Automatic Statistical Thesaurus Construction. In: ACM SIGIR'92. **Proceedings...** Copenhagen: ACM PRESS, 1992. p.77-88.
- [CRO82] CROFT, W. B; RUGGLES, L. The Implementation of a Document Retrieval System. In: CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. **Proceedings...** Berlim: ACM PRESS, 1982.
- [DAO96] DAO, Son; PERRY, Brad. Information mediation in cyberspace: scalable methods for declarative information networks. **Journal Of Intelligent Information Systems**. Boston. v6, n2/3, p. 131-150, Junho.1996.
- [DAV72] DAVIS, K. **Human behavior at work: human relations and organizational behavior**. 4a.ed. McGraw-Hill, 1972.
- [DUN97] DUNLOP, Mark D. The Effect of Acessing Nonmatching Documents on Relevance Feedback. **ACM Transactions On Information Systems**. v.15, n.2, pp. 137-153, Abril. 1997.
- [GUP97] GUPTA, Amarnath; JAIN, RAMESH. Visual Information Retrieval. **Communications of the ACM**. v.40, n.5, Maio. 1997.
- [HAR92] HARMAN, Danna. Relevance Feedback Revised. In: ACM SIGIR'92. **Proceedings...** Copenhagen: ACM PRESS, 1992. p. 1-10.
- [IIV95] IIVNEN, Mirja. Searches and Searches: Differences Between the Most and Least Consistent Searches. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 149-157.

- [KON73] KONITZ, Ruth et alli. O signo e sua tipologia. IN: AZEVEDO, M. C. (coordenador). **Atenção - signos - graus de informação**. Série Cadernos Universitários, n.4. Ed. UFRGS, 1973.
- [KRA96] KRAAIJ, Wessel. Viewing Stemming as Recall Enhancement. In: ACM SIGIR'96. **Proceedings...** Zurich: ACM PRESS, 1996. p. 40-48.
- [KOR94] KORTH, H. F.; SILBERSCHATZ, A. **Sistemas de Bancos de Dados**. 2ª edição. Makron Books, Macgraw-Hill. 1994.
- [KUO96] KUOKKA, Daniel; HARADA, Larry. Integrating Information via Matchmaking. **Journal Of Intelligent Information Systems**. Boston, v.6, n.2,3, p 261-279, Junho. 1996.
- [LAM96] LAM, W. Et alli. Detection of User Interests for Personalized Information filtering. In: ACM SIGIR'96. **Proceedings...** Zurich: ACM PRESS, 1996. p. 317-325.
- [MOU92] MOULIN, Bernard; ROUSSEAU, Daniel. Automated knowledge acquisition from regulatory texts. **IEEE Expert**. Outubro. 1992.
- [OLI96] OLIVEIRA, Henry M. **Seleção de Entes Complexos Usando Lógica Difusa**. Instituto de Informática da PUC-RS, Porto Alegre, Julho de 1996. (dissertação de mestrado).
- [RIL95] RILOFF, Ellen. Little Words Can Make a Big Difference for Text Classifications. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 130-136.
- [SAL91] SALTON, Gerard. **The State of Retrieval System Evaluation**. Technical Report 91-1206. New York: Department of Computer Science, Cornell University. 1991.
- [SAL88] SALTON, Gerard; SMITH, Maria. **On the Application of Syntactic Methodologies in Automatic Text Analysis**. Technical Report. New York: Department of Computer Science, Cornell University. 1988.
- [SAL87a] SALTON, Gerard; BUCKLEY, Chris. **Term Weighting Approaches in automatic Text Retrieval**. Technical Report. New York: Department of Computer Science, Cornell University. 1987.
- [SAL87b] SALTON, Gerard; BUCKLEY, Chris. **Improving Retrieval Performance by Relevance Feedback**. Technical Report. New York: Department of Computer Science, Cornell University. 1987.

- [SAL83] SALTON, Gerard. **Introduction to Modern Information Retrieval**. MCGRAW-HILL, 1983.
- [SIN96] SINGHAL, A. Et all. Pivoted Document Length Normalization. In: ACM SIGIR'96. **Proceedings...** Zurich: ACM PRESS, 1996. p. 21-29.
- [SOL97] SOLOWAY, Elliot; Wallace Raven. Does The Internet Support Student Inquiry? Don't Ask. **Communications of The ACM**. v40, n5. Maio de 1997.
- [STE72] STEWART, D. K. **A psicologia da comunicação**. Ed. Forense. 1972.
- [VIL95] VILES, Charles L; FRENCH, James C. Dissemination of Collection Wide Information in a Distributed Information Retrieval System. In: ACM SIGIR'95. **Proceedings...** Washington: ACM PRESS, 1995. p. 12-20.
- [WIE96] WIEDERHOLD, Gio. Foreword: Intelligent Integration Of Information. **Journal Of Intelligent Information Systems**. Boston. v6, n2/3, p. 93-98, Junho. 1996.
- [WIV96a] WIVES, Leandro K.; SARDI, Filipe L. M.; LOH, Stanley. Definição de uma ferramenta de busca em bases de dados textuais usando um Hiperdicionário. In: JORNADAS DE INFORMATICA E INVESTIGACION OPERATIVA, VI ENCUESTRO DEL LABORATORIO DE CIENCIA DE LA COMPUTACION. MONTEVIDEO, 3. **Anais...** Uruguai: 1996.
- [WIV96b] WIVES, Leandro K. **Um Modelo de Hiperdicionário: Estudo de Caso em Prontuários Médicos**. Curso de Graduação em Ciência da Computação - UCPEL. Dezembro de 1996. (Trabalho de Conclusão de Curso).
- [YAT96] YATES, R. B. An extended model for full text databases. **Journal of the Brazilian Computer Society**. v.2, n.3, Abril. 1996.
- [ZAD73] ZADEH, Cotf. A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. **IEEE Transactions on Systems, Man and Cybernetics**, V. SMC-3, n.1. Janeiro. 1973.