UNIVERSIDADE METODISTA DE PIRACICABA

FACULDADE DE CIÊNCIAS EXATAS E DA NATUREZA

MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

RECUPERAÇÃO DE DOCUMENTOS TEXTO USANDO UM MODELO PROBABILÍSTICO ESTENDIDO

MARCELLO ERICK BONFIM

ORIENTADOR: PROFª. DRª. MARINA TERESA PIRES VIEIRA

PIRACICABA, SP
2006

UNIVERSIDADE METODISTA DE PIRACICABA

FACULDADE DE CIÊNCIAS EXATAS E DA NATUREZA

MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

RECUPERAÇÃO DE DOCUMENTOS TEXTO USANDO UM MODELO PROBABILÍSTICO ESTENDIDO

MARCELLO ERICK BONFIM

ORIENTADOR: PROFª. DRª. MARINA TERESA PIRES VIEIRA

Dissertação apresentada ao Mestrado em Ciência da Computação, da Faculdade de Ciências Exatas e da Natureza, da Universidade Metodista de Piracicaba – UNIMEP, como requisito para obtenção do Título de Mestre em Ciência da Computação.

PIRACICABA, SP
2006

Aos

Meus pais Vanderlei e Maria Rosa

AGRADECIMENTOS

A Deus, por mais esta etapa de caminhada, pois sei que não estaria aqui se não tivesse caminhado comigo.

Aos meus amados pais, que sempre me acompanharam, me ensinaram a superar obstáculos e lutar pelos meus sonhos, essa vitória também é de vocês.

À Deise e Antonio, que me acompanharam e incentivaram.

À Cristiane, pelo carinho, companheirismo, paciência e apoio.

Aos meus avós e familiares, pelo incentivo e apoio constantes.

À minha orientadora, Profa. Marina, por me propiciar a oportunidade de realizar este trabalho, pela indicação dos rumos e pelo constante incentivo.

À Claudia Mello pela disponibilidade e contribuição.

Aos meus colegas de curso, pela troca de conhecimentos, brincadeiras e amizade.

A todos, que de alguma forma contribuíram para a realização deste trabalho.

RESUMO

Neste trabalho são apresentadas estratégias utilizadas para a recuperação de informação, com base no modelo probabilistico de recuperação de informação. Nessas estratégias adotou-se os modelos probabilístico e probabilístico exponencial, que foram combinados com recursos do modelo vetorial, sendo denominados de modelo probabilístico estendido e modelo probabilístico exponencial estendido. A recuperação de informação considera os valores da probabilidade de relevância e de não-relevância durante a classificação dos documentos resultantes. São apresentados resultados de experimentos que comprovam que a combinação dos modelos probabilísticos com o modelo vetorial possibilita uma recuperação mais eficaz, trazendo como resposta documentos relevantes que não seriam recuperados utilizando somente um dos modelos.

PALAVRAS-CHAVE: Recuperação de Informação, Modelo Probabilístico Estendido e Modelo Probabilístico Exponencial Estendido.

ABSTRACT

Strategies are presented here which are used for information retrieval based on the probabilistic information retrieval model. These strategies involved the adoption of probabilistic and exponential probabilistic models, which were combined with resources from the vectorial model and are called extended probabilistic model and extended exponential vectorial model. Information retrieval considers the values of the probability of relevance and of non-relevance during the classification of the resulting documents. Results of experiments are presented which prove that the combination of these probabilistic models with the vectorial model leads to a more effective retrieval, bringing up as response relevant documents that would not otherwise be retrieved using only one of the models.

KEY WORDS: Information Retrieval, Extended Probabilistic Model and Extended Exponential Probabilistic Model

SUMÁRIO

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

IR	*Information Retrieval* (Recuperação de Informação).
RI	Recuperação de informação.
EI	Extração de Informação.

LISTAS DE TABELAS

1. 1. INTRODUÇÃO

1.1. CONSIDERAÇÕES INICIAIS

A ampla variedade e quantidade de informações armazenadas fazem com que a descoberta de informações implícitas e de grande importância na representação do conteúdo de um documento em um conjunto de dados seja alvo de pesquisas mais aprofundadas sobre recuperação de informação.

Sistemas de recuperação adotam palavras-chave (termo de indexação) para indexar e recuperar documentos. Um termo de indexação é uma palavra que aparece no texto de um documento em uma coleção. O sistema de recuperação apresenta os resultados à uma consulta do usuário, e cabe ao sistema identificar qual documento é relevante ou não-relevante à solicitação.

Pesquisas da área de Recuperação de Informação (Information Retrieval – IR) visam à descoberta de tecnologias de coleta, representação, indexação, recuperação e classificação de grandes coleções de informação (MACEDO, 2004).

1.2. MOTIVAÇÃO

Para a recuperação de informação existem três modelos clássicos: o modelo booleano, o modelo vetorial e o modelo probabilístico. Também existem variações desses modelos, dentre estes o modelo probabilístico exponencial.

Apesar da existência de vários modelos para recuperação de documentos, não existe um modelo ideal. Pesquisas atuais na área de recuperação de informação, relatadas na literatura, demonstram o interesse dos pesquisadores na busca de novas abordagens visando aprimorar as técnicas existentes. Neste trabalho foram adotados o modelo probabilistico tradicional e o modelo probabilístico exponencial adicionando-lhes recursos do

modelo vetorial, buscando propor uma estratégia de recuperação de documentos que apresente vantagens quando comparada com as existentes.

1.3. OBJETIVO DA PESQUISA

O objetivo deste trabalho foi desenvolver uma nova abordagem para recuperação de documentos, tomando-se como base o modelo probabilístico, no qual foram incorporados recursos do modelo vetorial. Duas versões dessa abordagem foram implementadas: uma que utiliza o modelo probabilístico clássico e outra que utiliza o modelo probabilístico exponencial, permitindo uma comparação dos dois modelos quanto à sua eficácia para a recuperação de documentos, quando comparados com outras abordagens.

1.4. ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada da seguinte forma: o capítulo 2 apresenta os conceitos que envolvem a recuperação de informação; no capítulo 3 são apresentados alguns dos modelos para a recuperação de informação; o Modelo Probabilístico e suas aplicações e variações estão no capítulo 4; no capítulo 5 são apresentados os recursos e técnicas utilizados no Sistema de Manipulação de Documentos desenvolvido nesta pesquisa, mostrando sua arquitetura, estrutura de dados, processo de armazenamento e recuperação e as estratégias utilizadas na recuperação da informação; o capítulo 6 apresenta os experimentos realizados para avaliação da abordagem proposta; no capítulo 7 são apresentadas as conclusões e propostas de trabalhos futuros. Os principais algoritmos utilizados são apresentados no Apêndice A.

2. RECUPERAÇÃO DE INFORMAÇÃO

2.1. CONSIDERAÇÕES INICIAIS

Conforme citado em Macedo (2004), no final da década de 60 surgiram os primeiros catálogos bibliográficos *on-line* que permitiam a recuperação de informação em alguns minutos. O usuário manuseava as informações através de um ambiente de consulta utilizando um conjunto controlado de operações e linguagens pré-definidas. Nas décadas seguintes, o tamanho das coleções de informações cresceu muito e, nos anos 90 surge a Web e populariza essa grande quantidade de informações.

Com o crescimento do volume de informação ocorre o desenvolvimento de computadores com maior capacidade de armazenamento e processamento, como também surgem pesquisas visando melhorar o desempenho da recuperação, integração e armazenamento dessas informações (GETOOR *et. al.*, 2002).

Neste capítulo são apresentados conceitos fundamentais de Recuperação de Informação.

2.2. CONCEITOS BÁSICOS

Um sistema de recuperação de informação pode ser representado por três componentes: entrada, processador e saída (Van RIJSBERGEN, 1979). Analisando as entradas (documentos e consultas), o principal desafio é obter uma representação de cada documento e consulta.

É possível ao usuário mudar sua consulta durante uma sessão de busca, melhorando a recuperação. Tal procedimento é chamado de realimentação. Em seguida, o processador inicia o processo de recuperação. Esse processo envolve a estruturação da informação através da classificação da informação recuperada. Na figura 2.1, os documentos foram colocados separados das consultas para enfatizar o fato que eles não são somente itens de entrada, mas podem ser usados durante o processo de recuperação de tal

modo que suas estruturas são vistas como parte do processo de recuperação. Como resultado desse processamento tem-se um conjunto de documentos.

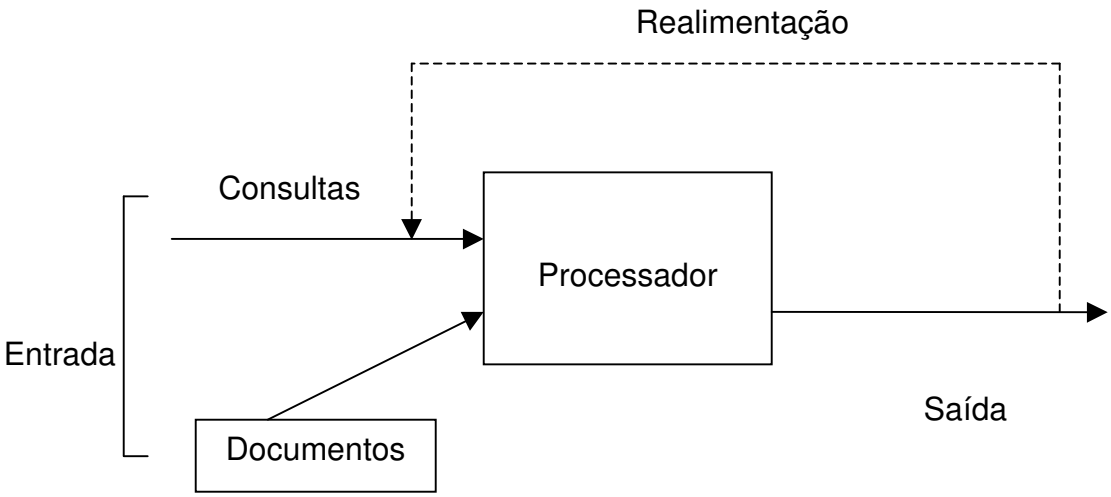

FIGURA 2.1 – SISTEMA TÍPICO DE RI (VAN RIJSBERGEN, 1979)

Em um processo de recuperação de informação podem ser realizadas algumas etapas que possibilitam o refinamento do texto por meio da aplicação de tarefas de mineração de textos. As seções seguintes descrevem características das tarefas de análise automática do texto, classificação automática, estruturação de arquivos e recuperação probabilística, dentre outras.

2.3. ANÁLISE AUTOMÁTICA DE TEXTO

Com o crescente volume de informações disponíveis torna-se necessário organizar e melhorar o armazenamento e a apresentação das informações visando facilitar uma pesquisa do usuário a um determinado tema.

Surgiram teorias e ferramentas computacionais para auxiliar a extração de informação, dando origem a uma área chamada de Descoberta de Conhecimento em Textos (Knowledge Discovery in Texts – KDT) (FELDMAN, *et al.,* 1995 *apud* CORREA, 2003).

A Descoberta de Conhecimento passa por várias etapas onde o usuário toma decisões que direcionam a busca. A mineração de textos visa explorar dados textuais desestruturados através de técnicas avançadas.

Na descoberta de conhecimento em documentos texto é necessário criar uma estrutura que possibilite a aplicação das técnicas de mineração. Para isso, é criada uma forma intermediária para uma coleção de documentos, composta por um conjunto de termos de indexação que representam esses documentos. O processo de mineração utiliza essa forma intermediária para obter as informações relevantes.

2.4. DESCOBERTA DE CONHECIMENTO EM TEXTOS (KDT)

A descoberta de conhecimento em textos também é conhecida como mineração de textos (*text mining*) (FELDMAN, *et al.,* 1995 *apud* CORREA, 2003), e surge da necessidade de organizar e padronizar automaticamente textos visando melhorar a análise dos mesmos.

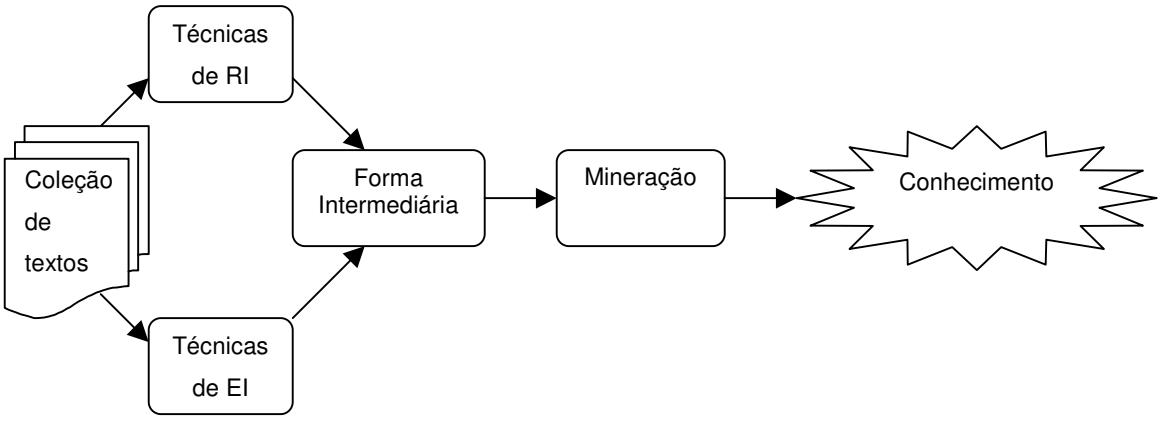

FIGURA 2.2 – PROCESSO DE MINERAÇÃO DE TEXTOS (CORREA, 2003)

Para conseguirmos estruturar os textos e obter as informações desejadas é necessário seguir um processo de mineração.

A mineração de textos possui etapas que envolvem técnicas de Recuperação de Informação (RI) e Extração de Informação (EI), onde são

aplicadas técnicas de Mineração da Informação. Essas etapas são descritas abaixo:

- **Recuperação de Informação**: localização e recuperação de documentos que podem ser relevantes a uma pesquisa. É necessário um sistema para filtrar esses documentos especificados pelo usuário e indexar as palavras-chave encontradas.

- **Extração de Informação**: os termos considerados relevantes nos documentos são extraídos e convertidos em dados afim de que possam ser utilizados durante o processo de mineração.

- **Mineração da Informação**: assim que a informação é armazenada de forma estruturada, a descoberta de informação é feita através da mineração sobre o banco de dados criado.

É através da interpretação dos padrões recuperados através da mineração, onde os resultados obtidos são interpretados, que se realiza a descoberta do conhecimento.

Os documentos devem ser pré-processados possibilitando a extração das palavras-chave, o que possibilitaria localizar um documento a partir da comparação entre um termo de consulta do usuário e os termos presentes em um documento.

2.5. INDEXAÇÃO

Os termos de indexação são utilizados para representar documentos e consultas. Os elementos do índice são os termos de indexação que são derivados do corpo do documento (Van RIJSBERGEN, 1979).

A classificação automática é realizada através do conceito de similaridade, como será apresentado nos capítulos posteriores, entre os termos de indexação. Os termos extraídos dos documentos ficam armazenados em vetores com referências para seus respectivos documentos. Assim, através do termo é possível encontrar o documento solicitado.

2.6. NORMALIZAÇÃO

A normalização ocorre em etapas possibilitando melhorar a análise e classificação do conjunto de documentos.

Em baixo nível, o documento é descrito por um conjunto de palavras. O primeiro passo da normalização é remover as palavras que aparecem em excesso no corpo do texto e que não possuem grande importância (preposições, artigos, conjunções etc). Assim, pode-se dizer que depois dessa primeira etapa teremos as palavras-chave. A próxima etapa é a classificação automática das classes de palavra-chave (Van RIJSBERGEN, 1979).

2.7. CLASSIFICAÇÃO AUTOMÁTICA

Na Recuperação de Informação em textos ocorre a descoberta do conhecimento através do refinamento dos documentos de uma coleção, transformando-os em uma estrutura intermediária armazenada em um banco de dados.

Os dados consistem em objetos e suas descrições correspondentes. Os objetos podem ser documentos, termos de indexação etc., e devem ser classificados para ser possível a recuperação da informação. A clusterização (agrupamento) trabalha os objetos visando possibilitar sua classificação e recuperação.

Segundo Salton (1983), alguns métodos da classificação são baseados em um relacionamento binário entre objetos. A base deste relacionamento pode ser um sistema de agrupamento (*cluster*). O relacionamento é descrito como similaridade, que é uma medida projetada para quantificar e relacionar objetos.

Existem várias tarefas de mineração, a seguir serão apresentadas as tarefas de sumarização, associação, classificação e

clusterização, que podem ser utilizadas em um sistema de recuperação de informação.

2.7.1. SUMARIZAÇÃO

A sumarização é utilizada para identificar palavras ou frases importantes em um documento ou num conjunto de documentos que trazem o conceito do documento.

A sumarização produz uma lista das sentenças presentes nos documentos resumindo o conteúdo dos mesmos (DIXON, 1997 *apud* CORREA, 2003).

2.7.2. ASSOCIAÇÃO

Na associação, as transações do banco de dados são do tipo implicação (X => Y) e significa que se um documento possui X tende a possuir Y. Essas associações são muito utilizadas na mineração de textos possibilitando descobrir associações entre termos e documentos. Para se realizar essas associações podem ser utilizados algoritmos específicos (CORREA, 2003).

2.7.3. CLASSIFICAÇÃO

A classificação automática dos documentos tem como base um conjunto pré-classificado. O algoritmo utilizado para extrair conhecimento analisa todos os exemplos de documentos, assimila as regras e armazena em uma base de conhecimento. Assim, os documentos passam pelo algoritmo de classificação, que é baseado em regras previamente definidas na base de conhecimento, e é classificado de acordo com a classe a que pertence (SALTON, 1983).

2.7.4. CLUSTERIZAÇÃO

A tarefa de clusterização agrupa os documentos similares com base nos termos de indexação do documento. Esse processo não requer um conjunto previamente definido e treinado. Os documentos similares podem ser agrupados de acordo com os termos de indexação, onde os termos similares são colocados na mesma classe.

Ao particionar um conjunto de dados em grupos (*clusters*), utiliza-se o conceito da clusterização, onde cada *cluster* é formado por objetos similares, sendo assim, os objetos de um mesmo *cluster* são mais similares entre si do que se comparados com objetos de outro *cluster.* A clusterização é utilizada na recuperação de informação automaticamente para organizar uma coleção de resultados recuperados, agrupando os documentos que pertencem ao mesmo tópico para facilitar a navegação pelos documentos resultantes (HEARST e PEDERSEN, 1996).

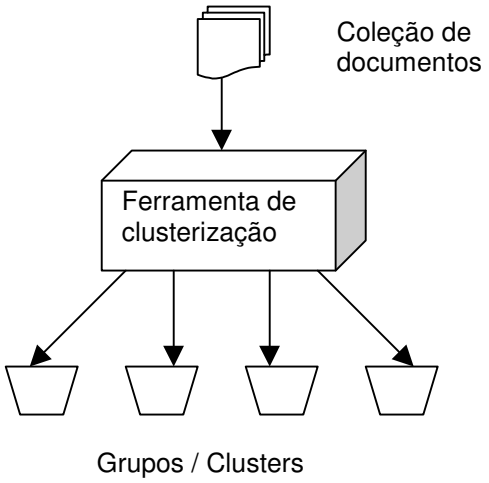

FIGURA2. 3 – PROCESSO DE CLUSTERIZAÇÃO

2.8. CONSIDERAÇÕES FINAIS

Este capítulo apresentou as etapas utilizadas para a Recuperação de Informação. Dentre as etapas mencionadas, destaque para a sumarização, onde as frases de um documento são extraídas formando um

sumário, e para a clusterização, que agrupa os documentos similares. As etapas de mineração são utilizadas com a finalidade de se obter uma forma intermediária e estruturada dos documentos, já que esses se encontram desestruturados.

3. OS MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO

3.1. CONSIDERAÇÕES INICIAIS

Segundo Baeza e Ribeiro (1999), os três modelos clássicos de recuperação de informação são: Booleano, Vetorial e Probabilístico; estes são responsáveis por recuperar os documentos relevantes utilizando um mecanismo de comparação entre a consulta e os documentos armazenados. O modelo Booleano é um modelo simples de recuperação de informação baseado na Álgebra Booleana. No modelo vetorial, documentos e consultas são representados como vetores em um espaço t-dimensional (algébrico). No modelo probabilístico, a estrutura para modelagem de documentos e consultas é baseada na teoria da probabilidade.

Há outros modelos de recuperação de informação na literatura tais como: booleano estendido (SALTON *et al*., 1983), redes bayesianas (TURTLE e CROFT, 1991) e probabilístico exponencial (TEEVAN e KARGER, 2003), entre outros. Esses modelos são apresentados nas seções seguintes.

3.2. MODELOS DE RECUPERAÇÃO DE INFORMAÇÃO

Os modelos de recuperação de informação consideram que cada documento é descrito por palavras-chave chamadas de termos de indexação. Um termo de indexação é uma palavra cuja semântica ajuda a localizar os temas principais de um documento. Adjetivos, advérbios, conjunções são menos úteis como termos de indexação.

Segundo Baeza e Ribeiro (1999), dado um conjunto de termos de indexação para um documento, nota-se que nem todos os termos podem ser usados para descrever o conteúdo do documento. Não é uma tarefa fácil determinar a importância de um termo de indexação em um documento. Considerando uma coleção com cem mil documentos, uma palavra que aparece em cada um dos cem mil documentos é completamente inútil como um termo de indexação porque ela não trás somente documentos de interesse do

usuário. Por outro lado, uma palavra que aparece em cinco documentos é completamente útil porque se estreita o espaço dos documentos que interessam na pesquisa.

Os modelos de recuperação de informação podem ser descritos de acordo com a seguinte terminologia (BAEZA e RIBEIRO, 1999):

- d_j representa um documento de uma coleção;
- q representa uma consulta;
- t representa a quantidade de termos de índice da coleção de documentos;
- k_i representa um termo de índice;
- $K = \{k_1, ..., k_t\}$ é o conjunto de todos os termos de indexação;
- $w_{i,j} > 0$ é o peso associado com cada termo de indexação k_i de um documento d_j. Quando um termo de indexação não está contido no corpo do documento seu peso $w_{i,j} = 0$;
- $sim(d_j,q)$ representa a função utilizada para comparar a consulta com os documentos da coleção, retornando um coeficiente de similaridade (relacionamento) entre a consulta q e o documento d_j.

O documento d_j é associado a um vetor de termos de indexação (d_j) representado por $d_j = (w_{1,j}, w_{2,j}, ... w_{t,j})$.

A seguir são apresentados alguns dos modelos utilizados na recuperação de informação. O modelo probabilístico e probabilístico exponencial, utilizados para a validação das técnicas de recuperação probabilística estendida, serão apresentados com mais detalhes no próximo capítulo.

3.2.1. MODELO BOOLEANO

O Modelo Booleano é um modelo de recuperação simples baseado na teoria da Álgebra Booleana (BAEZA e RIBEIRO, 1999). Como seu conceito é bastante intuitivo, o Modelo Booleano fornece uma estrutura de fácil compreensão para o usuário comum de um sistema de recuperação de informação (*IR – Information Retrieval)*. As consultas são estabelecidas como expressões booleanas com semânticas precisas. Dada a simplicidade e o formalismo puro, o modelo recebeu grande atenção e foi adotado por muitos dos primeiros sistemas bibliográficos comerciais.

O modelo booleano considera que termos de indexação estão presentes ou ausentes num documento. Como um resultado, assume-se que os pesos dos termos indexados são todos binários. Uma consulta (*query*) *q* é composta de termos unidos por três tipos de operadores lógicos: *NOT, AND* e *OR.* Assim, "a consulta é essencialmente uma expressão booleana convencional que pode ser representada como uma disjunção de vetores conjuntivos" (BAEZA e RIBEIRO, 1999).

Considerando uma expressão de busca $q = t_1$ *AND* t_2, são recuperados documentos indexados pelos termos t_1 e t_2. Essa operação equivale à intersecção do conjunto de documentos indexados pelo termo t_1 com o conjunto de documentos indexados pelo termo t_2. Utilizando o operador lógico *OR*, é realizada a união entre o conjunto de documentos indexados pelos termos da consulta. Com o operador *NOT*, são recuperados os documentos que não possuem o termo da consulta.

No modelo booleano um documento é considerado relevante ou não-relevante a uma consulta. Não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta. As principais vantagens do modelo booleano são o formalismo claro e a simplicidade do modelo.

O Modelo Booleano apresenta algumas desvantagens. Como principal desvantagem, a consulta pode trazer muito pouco ou muitos

documentos. Sua estratégia de recuperação é baseada no critério de decisão binária sem qualquer noção de balanceamento na classificação, que garanta um bom desempenho na recuperação. Expressões booleanas têm uma semântica precisa, porém, freqüentemente, não é simples de saber quão relevante é uma informação solicitada numa expressão booleana.

3.2.2. MODELO BOOLEANO ESTENDIDO

O modelo booleano estendido, proposto por Salton (1983), considera o peso dos termos nos documentos e permite que o usuário especifique as relevâncias dos termos para uma determinada consulta. Esse modelo se baseia na interpretação dos operadores de consulta conjuntivas e disjuntivas em termos de distâncias euclidianas em um espaço *t*-dimensional.

Nas expressões conjuntivas o ponto (1,1) é o mais desejável, significa que ambos os termos de uma expressão de busca estão no documento. Quanto menor a distância do documento em relação a este ponto maior é a similaridade em relação à busca. Nas expressões disjuntivas o ponto (0,0) representa que nenhum dos termos da expressão de busca está presente no documento.

Considerando a utilização de dois termos t_1 e t_2 para representar as consultas e documentos, é definido um espaço de busca bidimensional onde cada termo é associado a um eixo. Um documento é representado por um vetor com dois elementos contendo pesos dos respectivos termos. Esses pesos definem o posicionamento do documento no espaço euclidiano.

A similaridade entre um documento $d_i = (w_{1i}, w_{2i})$ e uma consulta $q = t_1 \ or \ t_2$ é calculada através da equação 3.1, onde w_{1i} e w_{2i} representam os pesos de cada um dos termos de indexação do documento.

$$sim(q, d_i) = \sqrt{\frac{w_{1i}^2 + w_{2i}^2}{2}} \tag{3.1}$$

3.2.3. MODELO VETORIAL

O modelo vetorial também é chamado de modelo espaço vetorial e representa cada documento como um vetor de termos e cada termo possui um valor associado que indica seu grau de importância (peso – *weight*) para o documento, ou seja, cada consulta possui um vetor resultado construído através do cálculo da similaridade baseado no ângulo (co-seno) entre o vetor que representa o documento e o vetor que representa a consulta. (BAEZA e RIBEIRO, 1999)

São acrescentados pesos aos termos das consultas e documentos. Os pesos especificam a relevância de cada termo para a consulta e para os documentos no espaço vetorial.

A consulta do usuário também é representada por um vetor. Desta forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado. Os documentos mais similares à consulta são considerados relevantes para o usuário e retornados como resposta. Os pesos são usados para computar a similaridade entre cada documento armazenado e uma consulta feita pelo usuário. Os métodos de cálculo se baseiam no número de ocorrências do termo no documento (freqüência).

O Modelo Vetorial é definido formalmente: no modelo vetorial, o peso $w_{i,j}$ associado com um par (k_i, d_j) é positivo e não-binário. Os termos de indexação nas consultas também possuem peso $w_{i,q}$ associado com um par $[k_{i,q}]$, onde $w_{i,q} \geq 0$. O vetor de consulta $q\rightarrow$ é definido como $q\rightarrow = (w_{1,q}, w_{2,q}, ...w_{t,q})$ onde t é o número total de termos de indexação. O vetor de documento d_j é representado por $d = (w_{1,j}, w_{2,j},..., w_{t,j})$. (BAEZA e RIBERO, 1999)

As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos diferentes que possuem os mesmos termos são colocados em uma mesma região do espaço e, em teoria, tratam de assuntos similares.

Um documento *d* e uma consulta *q* são representados como um vetor *t-dimensional*. O modelo vetorial propõe avaliar o grau de similaridade do documento com a consulta por meio de uma qualificação que pode ser feita através do cálculo do co-seno (*cosine vector similarity*) do ângulo entre estes dois vetores. Com os graus de similaridade calculados monta-se uma lista ordenada (*ranking*) de todos os documentos e seus respectivos graus de relevância à consulta, da maior para a menor relevância.

Quanto à freqüência de um termo num documento tem-se como definição que em um número total *N* de documentos são selecionados os n_i documentos em que o termo de indexação aparece; a freqüência é o número de vezes que o termo mencionado aparece no texto do documento selecionado . Se o termo não aparece no documento selecionado a freqüência é igual a zero ($f_{i,j} = 0$). Segundo Baeza e Ribeiro (1999), a melhor fórmula para calcular o peso do termo é dada por

$$w_{i,j} = f_{i,j} \times \log N/n_i \qquad (3.1)$$

O resultado da busca é um conjunto de documentos ordenados pelo grau de similaridade entre cada documento e a consulta. A expressão 3.2 de similaridade calcula a distância entre o vetor de documento e o vetor da consulta.

Fórmula da similaridade

$$sim(d_j, q) = \frac{\sum_{i=1}^{n} w_{iq} * w_{ij}}{\sqrt{\sum_{i=1}^{n} w_{iq}^2} * \sqrt{\sum_{i=1}^{n} w_{ij}^2}} \qquad (3.2)$$

Características:

- a atribuição de pesos aos termos melhora o desempenho da recuperação;

- sua estratégia de comparação (*matching*) parcial permite a recuperação de documentos que se aproximam das condições da consulta;

- a fórmula do co-seno classifica os documentos de acordo com seu grau de similaridade com a consulta;

A principal vantagem do modelo vetorial é a recuperação de documentos que satisfazem parcialmente a expressão de busca, trazendo também documentos similares como conjunto resposta.

Segundo Baeza e Ribeiro (1999), uma grande variedade de métodos de classificação alternativos vem sendo comparados ao modelo vetorial e concluiu-se que, em geral, o modelo clássico vetorial é superior ou quase tão bom quanto os métodos alternativos conhecidos. Além disso, é simples e rápido o que faz dele um modelo de recuperação popular.

3.2.4. O MODELO DE REDES BAYESIANAS

Uma outra área de pesquisa é a que utiliza a representação de redes para as dependências entre os documentos e termos (YANAI e IBA, 2005). Um formalismo probabilístico utilizado para se chegar a um bom resultado na recuperação de informação é o modelo Bayesiano. Em um modelo de rede Bayesiana, a independência entre as variáveis de uma distribuição de probabilidade conjunta é representada por meio de grafos dirigidos acíclicos. A cada variável aleatória da distribuição é associado um nodo no grafo. Estas variáveis podem representar eventos, estados, objetos, proposições ou outras entidades (GREIFF e PONTE, 2000). O relacionamento entre estas variáveis é modelado como arestas dirigidas. Estas arestas representam dependências entre as variáveis (ou nodos) ligadas (os). Uma interpretação para estas dependências é que representam influências causais cuja força é expressa por probabilidades condicionais (SILVA, 1999).

A principal vantagem das redes Bayesianas quando comparadas com outras representações de probabilidades é que elas

representam relacionamentos probabilísticos de uma forma concisa. O mecanismo é baseado no conceito de probabilidade condicional e no teorema de Bayes (SILVA, 1999). A estimativa de probabilidade é eficiente em várias situações práticas.

O primeiro modelo de rede Bayesiana para RI, chamado *inference network model*, foi proposto por Turtle e Croft (1991). Um modelo de rede Bayesiana mais genérico, chamado *belief network model*, foi proposto por Ribeiro e Muntz (1996). Esta variante introduz evidências de consultas passadas em uma rede Bayesiana com o objetivo de melhorar a qualidade da resposta.

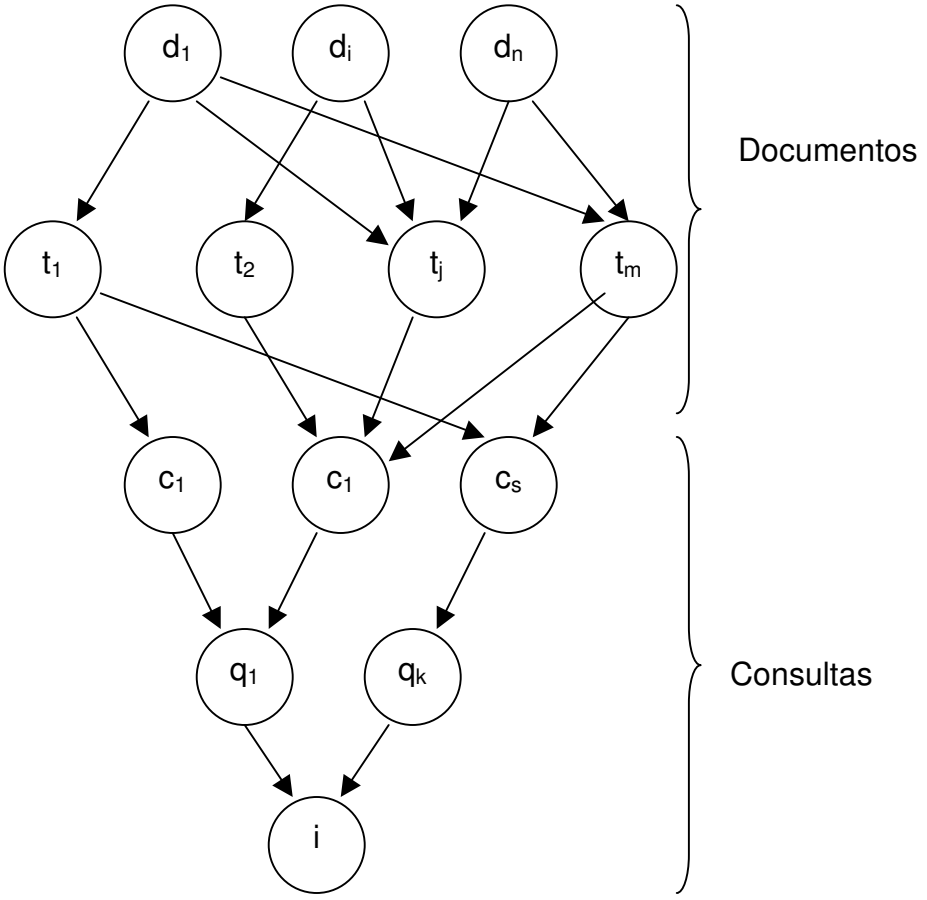

FIGURA 3.1: MODELO DE REDE BAYESIANA (CRESTANI ET.AL., 1998)

Segundo Pearl (1988), as duas escolas probabilísticas mais tradicionais são baseadas na visão freqüentista e na visão epistemológica. Na visão freqüentista a probabilidade é um conceito relacionado às leis de chance, obtidas através da repetição de experimentos. Na visão epistemológica, a probabilidade é tida como um grau de crença que pode ser especificado independente da experimentação.

A figura 3.1 ilustra um exemplo de Rede Bayesiana; de acordo com essa figura, os nodos representam entidades de IR como documentos, termos de indexação, conceitos, consultas, e necessidades de informação. Pode-se escolher o número e tipo dos nodos que se deseja utilizar, de acordo com a complexidade de representação de um documento numa coleção ou a necessidade de informação. Os arcos representam as dependências probabilísticas entre entidades. Eles representam probabilidades condicionais, quer dizer, a probabilidade de uma entidade ser verdadeira dada as probabilidades de seus pais serem verdadeiras.

A rede bayesiana normalmente é composta de duas redes: uma rede de documento e uma rede de consulta. A rede que representa a coleção de documentos é composta para uma determinada coleção e sua estrutura não muda. Uma rede de consulta é construída para cada necessidade de informação e pode ser modificada e estendida durante cada sessão, pelo usuário, de modo interativo e dinâmico. A rede de consulta é ligada à rede estática de documentos para processar uma consulta.

É possível implementar vários modelos tradicionais em IR nesta rede introduzindo nodos representando operadores booleanos ou fixando funções condicionais apropriadas de avaliação de probabilidade dentro de nodos. (CRESTANI, 1998)

Uma característica particular deste modelo é que múltiplas representações de documentos e consultas podem ser usados dentro de uma coleção particular de documentos. Além disso, dada uma única solicitação de informação, é possível combinar resultados de múltiplas consultas e de múltiplas estratégias de busca (ZHAI, 2002). A principal característica desse

modelo é que os nodos podem ser somente binários (presente ou não presente).

Segundo Silva (1998), considerando os documentos representados por termos de indexação, e que estes termos de indexação compõem o conjunto **U**, este conjunto é adotado como espaço amostral. Seja **t** o número de termos de indexação da coleção. Define-se:

- *ki*: um termo de indexação;
- U = $\{k_1, k_2,...,k_t\}$: *espaço amostral*. Cada k_i é interpretado como um conceito elementar. **U** é interpretado como um espaço de conceitos;
- u $\subseteq$ U: um conceito qualquer em U, formado por um conjunto de conceitos elementares.

Associada a cada termo de indexação k_i, é definida uma variável aleatória, também denotada por k_i. Tal variável recebe o valor 1 para indicar que o termo pertence a um conceito. Por exemplo, em uma coleção com **t** termos, um documento é representado como um conceito d = $\{k_1, k_2,...,k_t\}$ onde cada k_i é 1 para indicar que o termo ocorre no documento **d** e é 0 em caso contrário. De forma análoga, uma consulta é representada por um conceito q = $\{k'_1, k'_2,...,k'_t\}$. Seja g_i(u) uma função que retorna o valor da variável k_i de acordo com o conceito **u**, isto é, g_i(u) é uma função que define uma relação de pertinência de um termo k_i em um conceito **u**, onde g_i(u) = 0 se $k_i \notin$ u e g_i(u) = 1 se $k_i \in$ u.

Seja **P** uma distribuição de probabilidade definida sobre o espaço amostral **U**. A probabilidade P(c) associada a um conceito genérico **c** no espaço **U** é definida pela equação a seguir.

$$P(c) = \sum_u P(c \mid u) * P(u) \qquad (3.3)$$

P(c | u) define uma relação de cobertura entre os conceitos **c** e **u** do espaço **U**. A probabilidade P(c) define uma relação de cobertura entre o

conceito **c** e todo o espaço **U**. Tal interpretação permite interpretar a similaridade entre um documento e uma consulta como uma relação de cobertura.

Em princípio todos os conceitos u $\in$ U são igualmente prováveis e portanto a probabilidade *a priori* P(u) é dada por $P(u) = (1/2)^t$.

No modelo proposto por Silva (1998), consultas e documentos são modelados de forma idêntica. Ambos são modelados como conceitos do espaço **U**. Esta simetria induz naturalmente a rede bayesiana.

Os nodos d_i modelam documentos enquanto o nodo **q** modela a consulta do usuário. Uma variável aleatória binária **q** é associada ao nodo **q**. Esta variável é igual a 1 (um) para indicar que **q** cobre completamente o espaço amostral **U**. A interpretação semântica da probabilidade P(q) é que ela reflete nosso grau de crença na seguinte assertiva: *É verdade que **q** cobre completamente o espaço **U***? Um documento **d** é modelado de forma análoga e a probabilidade P(d) é interpretada como o grau de crença na seguinte assertiva: *É verdade que **d** cobre completamente o espaço **U***? Uma vez que **q** e **d** são conceitos no espaço amostral **U**, tem-se:

$$P(q) = \sum_u P(q \mid u) * P(u) \quad e \tag{3.4}$$

$$P(d) = \sum_u P(d \mid u) * P(u) \tag{3.5}$$

Para determinar um vetor resultado para uma consulta **q**, calcula-se P(d | q) para cada documento na coleção. A probabilidade P(d | q) reflete o grau de cobertura do conceito **d** dado o conceito **q**. De acordo com a lei de Bayes,

$$P(d \mid q) = P(d \wedge q) / P(q) \tag{3.6}$$

Uma vez que P(q) é constante para todos os documentos, basta obter

$$P(d \mid q) \approx P(d \wedge q) \tag{3.7}$$

onde

$$P(d \wedge q) = \sum_u P(d,q \mid u) * P(u) \quad (3.8)$$

Na estrutura da rede na figura 3.1, é possível observar que a instanciação dos termos de indexação t_i (o que gera um conceito **u**) separa **q** e **d**, tornando-os mutuamente independentes. Assim sendo,

$$P(d,q \mid u) = P(d \mid u) * P(q \mid u) \quad (3.9)$$

, e podemos escrever:

$$P(d, q) = \sum_u P(d \mid u) * P(q \mid u) * P(u) \quad (3.10)$$

Esta é a expressão genérica para obter um vetor resultado (SILVA, 1998). Tal expressão pode ser utilizada para representar qualquer um dos modelos clássicos.

3.2.5. Recuperação de Componentes Utilizando Técnicas de Agrupamento

Nessa abordagem é definido um repositório que manipule metadados de componentes de software proporcionando mecanismos eficazes para a sua localização e reuso (MELLO, 2005). Para a extração e armazenamento dos dados são utilizados conceitos de recuperação de informação e rede neural. A estratégia de busca explora a organização dos metadados de componentes no banco de dados para promover mecanismos eficazes para a sua localização e reuso. Essa abordagem parte do pressuposto que componentes construídos para reuso disponibilizam uma documentação com suas principais funcionalidades. Essa documentação é composta por termos, que são extraídos automaticamente, normalizados e armazenados em repositórios. Essas informações são utilizadas no agrupamento dos componentes. Após a normalização dos termos são obtidas as freqüências de ocorrência na documentação e calculados os pesos, de acordo com a equação 3.11 proposta por Salton e McGill (1983), onde f_i é a freqüência do termo *i*, *n* é a quantidade total de componentes e n_i é a quantidade de componentes que possuem o termo *i*.

$$w_i = f_i * \log \frac{n}{n_i} + 1 \tag{3.11}$$

O agrupamento é realizado através da identificação de conjuntos de componentes similares. Essa abordagem utiliza a arquitetura de rede neural artificial auto-organizável Art-2A (CARPENTER *et al.,* 1991*apud* MELLO, 2005).

Para a recuperação de informação são utilizadas duas formas alternativas, uma que utiliza o modelo vetorial e outra usando o modelo booleano, que, aplicados aos agrupamentos, trazem ao usuário o conjunto resposta.

3.3. CONSIDERAÇÕES FINAIS

Este capítulo apresentou conceitos relacionados à recuperação de informação e características dos modelos de recuperação de informação booleano, booleano estendido, rede bayesiana e vetorial.

No próximo capítulo é apresentado o modelo probabilístico e probabilístico exponencial de recuperação de informação, que são os modelos utilizados nesse trabalho para classificação e recuperação dos documentos.

4. MODELO PROBABILÍSTICO

4.1. A MODELAGEM PROBABILÍSTICA NA RECUPERAÇÃO DE INFORMAÇÃO

Na recuperação de informação, a modelagem probabilística é utilizada para classificar documentos em ordem decrescente de probabilidade de relevância de acordo com uma solicitação do usuário (CRESTANI, 1998). Pesquisas antigas e recentes usam a teoria probabilística e estatística para estimar as relevâncias, diferindo do modelo espaço vetorial (SALTON, 1968) em que cada documento é classificado de acordo com a sua similaridade para a consulta.

As primeiras tentativas para se desenvolver uma teoria probabilística de recuperação de informação são datadas de 1960 e desde então esta abordagem vem sendo desenvolvida (ALLAN, 2002). Existem diversos sistemas baseados em modelos probabilísticos e semi-probabilísticos, várias teorias e modelos que comprovam a eficácia do modelo probabilístico (ROBERTSON, 2000). O maior obstáculo para esses sistemas é encontrar métodos para estimar as probabilidades que serão usadas para avaliar a relevância e não-relevância dos documentos. Nos estágios iniciais de uma aplicação do modelo probabilístico, os documentos são tratados como independentes para facilitar a questão computacional (GILDEA, 2001). Um outro obstáculo segundo Pavlov e Smyth (2001), é o tempo gasto para a recuperação de uma informação solicitada, dado que se torna necessário estimar as probabilidades de relevância e não-relevância. Uma outra abordagem do modelo probabilístico é a que utiliza a freqüência dos termos nas estimativas das probabilidades; essa abordagem foi trabalhada por Amati e Van Rijsbergen (2002) e Greiff *et. al.* (2002). Em uma outra abordagem, sugerida por Gey (1994), para um termo *t* pertencente a um documento, pode-se utilizar a *freqüência inversa do documento*, através da razão entre o número de documentos da coleção (*N*) e o número de documentos com o termo *t* (n_t), para determinar a probabiliidade de relevância de um termo.

Também foram desenvolvidos modelos que utilizam hipóteses e teoria estatística. Seu principal inconveniente é a necessidade da heurística para a descrição e recuperação dos documentos, o que não é apropriado para estimar a probabilidade de relevância e não-relevância (COOPER, 1995).

Outro conceito para aplicação do modelo probabilístico é o modelo probabilístico exponencial, que considera a freqüência do termo no corpo do documento e o tamanho do documento para estimar as probabilidades de relevância e não-relevância numa consulta; essa abordagem é proposta por Teevan e Karger (2003).

Segundo Croft *et. al.* (2001) e FUHR (1986), o modelo probabilístico é muito eficaz para sistemas de recuperação de informação, mesmo tendo suas raízes na literatura há muitos anos atrás, devido ao sucesso na classificação de documentos.

4.2. A MODELAGEM PROBABILÍSTICA

Os modelos probabilísticos trabalham com um conjunto *Q* de consultas e um conjunto *D* de documentos de uma coleção (FUHR e PFEIFER, 1994). Na maioria dos modelos de recuperação de informação as consultas e os documentos são representados por palavras-chave (termos de indexação), freqüentemente extraídos manualmente ou automaticamente, como visto nas seções anteriores. Essas palavras-chave são representadas como um vetor onde cada elemento corresponde a um termo.

Uma consulta é uma expressão de uma solicitação de informação, sendo considerada um evento único. Se dois usuários solicitam uma mesma consulta ou se consultas semelhantes são solicitadas de dois usuários semelhantes em duas ocasiões diferentes, as duas consultas são consideradas diferentes. Uma consulta é submetida ao sistema que busca a informação relevante para a solicitação.

Entre os documentos recuperados, o usuário decide quais são relevantes ou não. O sistema usa essa informação para refinar a descrição do

conjunto ideal de respostas. O processo é repetido muitas vezes com a finalidade de melhorar a descrição do conjunto. O usuário sempre deve ter em mente a descrição da consulta ideal. Segundo Jin e Hauptmann (2002), a interação do usuário com o sistema possibilita uma recuperação mais eficaz.

Um documento é qualquer objeto que contém informação: um trecho de texto, uma imagem, um som, ou um vídeo. Porém a recuperação de informação concentra-se mais na recuperação de documentos texto. Algumas hipóteses são comuns aos modelos de recuperação:

- as informações solicitadas pelos usuários são submetidas a contínuos refinamentos.

- a recuperação é baseada somente nas representações dos documentos e consultas, e não nos próprios documentos e consultas.

- a representação de objetos é “incerta”, pois a extração de termos de indexação de um documento ou de uma consulta para representar documentos e consultas é um processo incerto.

Segundo Crestani *et. al.*(1998), a teoria probabilística é um caminho para tratar essa incerteza na recuperação. O modelo clássico probabilístico foi introduzido em 1976 por Roberston e Sparck Jones e mais tarde ficou conhecido como modelo de recuperação de independência binária (BIR) (BAEZA e RIBERO, 1999). Foram desenvolvidos modelos baseados no modelo clássico probabilístico que podem ser utilizados na recuperação de informação. Destaque para Cooper (1995), que porpôs uma variação do Modelo de Independência Binária, e para o Modelo Probabilístico de Indexação, proposto por Fuhr (1989) que considera o peso para os termos da consulta, dentre outros.

O modelo probabilístico tenta tratar o problema da recuperação de informação dentro da visão probabilística. Dada uma consulta de um usuário, há um conjunto de documentos que possui documentos relevantes e não-relevantes. Tendo a descrição desse conjunto não se teria problema em recuperar esses documentos. Considerando que as propriedades de um

conjunto ideal de respostas não são conhecidas de imediato na consulta, tenta-se adivinhar quais seriam essas propriedades. Nesta hipótese inicial pode-se gerar uma descrição probabilística preliminar do conjunto ideal de respostas, que é usado para recuperar o primeiro conjunto de documentos. Inicia-se uma interação com o usuário com a finalidade de melhorar a descrição probabilística do conjunto ideal de respostas.

Segundo Baeza e Ribeiro (1999), o modelo probabilístico é baseado na seguinte hipótese:

Hipótese (Princípio Probabilístico): Dados uma consulta *q* e um documento d_j numa coleção, o modelo probabilístico tenta calcular a probabilidade do documento d_j ser relevante para o usuário. O modelo assume que esta probabilidade de relevância depende somente das representações dos documentos e das consultas. O modelo assume que há um subconjunto de todos os documentos que satisfazem a solicitação do usuário como conjunto resposta para a consulta *q*. O conjunto ideal de respostas é chamado *R* que é a probabilidade global de relevância. Os documentos desse conjunto são os documentos relevantes à consulta e os que não estão nesse conjunto são os não-relevantes.

Esta hipótese não é a ideal, pois não declara como são computadas as probabilidades de relevância de maneira explicita, e nem o espaço de amostra, utilizado para definir as probabilidades, é determinado.

Baeza e Ribeiro (1999) definem o modelo probabilístico da seguinte maneira: para o modelo probabilístico, o peso do termo de indexação para uma consulta é representado por $w_{i,q}$ e o peso do termo para o documento é representado por $w_{i,j}$, esses são todos binários, $w_{i,q} \in \{0,1\}$, $w_{i,j} \in \{0,1\}$. A consulta, que é formada por um subconjunto de termos de indexação, é representada por *q*. $+R_q$ representa que o documento é relevante à consulta *q* e $-R_q$ representa que o documento não é relevante para a consulta *q*. $P(+R_q|d_j)$ é a probabilidade de que um documento d_j seja relevante para a consulta *q*, e $P(-R_q|d_j)$ é a probabilidade de que um documento d_j seja não-relevante para a consulta *q*.

Segundo Salton (1986), dado um documento d_j, é necessário calcular as probabilidades de relevância e não-relevância. Essas probabilidades dependem da relevância individual de um termo de indexação k_i no documento. Assume-se que o termo ocorre independentemente (princípio da independência dos termos) em cada um dos documentos, relevantes ou não-relevantes de uma coleção. O peso ideal do termo (w_i) pode ser calculado pela equação

$$wi = log \frac{P(k_i|+R_q)[1 - P(k_i|-R_q)]}{P(k_i|-R_q)[1- P(k_i|+R_q)]} \quad (4.1)$$

onde $P(k_i|+R_q)$ é a probabilidade de um documento que possui o termo de indexação ser relevante e, $P(k_i|-R_q)$ é a probabilidade de um documento que possui o termo de indexação ser não-relevante à consulta.

Assumindo este peso individual da relevância do termo, cada documento pode ser nomeado com um valor de relevância global igual a soma dos pesos wi para todos os termos de consulta presentes num documento (expressão 4.5).

Um documento d_j é relevante a uma consulta q quando: $P(+R_q |d_j) > P(–R_q |d_j)$. Assim, dada uma consulta q, o modelo probabilístico atribui a cada documento *d* (como medida de similaridade) um peso $W_{d/q}$, como sendo:

$$W_{d/q} = sim(d_j, q) = \frac{P(+R_q |d_j)}{P(–R_q |d_j)} \quad (4.2)$$

Aplicando a regra de Bayes (BAEZA e RIBEIRO, 1999) tem-se,

$$sim(d_j, q) = \frac{P(d_j |+R_q) \times P (+R_q)}{P(d_j |–R_q) \times P (-R_q)} \quad (4.3)$$

onde $P(d_j |+R_q)$ é a probabilidade de se selecionar aleatoriamente um documento d_j do conjunto de documentos relevantes $+R_q$ e $P(d_j |-R_q)$ é a probabilidade de se selecionar um documento d_j do conjunto de documentos não-relevantes $–R_q$. Tem-se que $P(+R_q)$ é a probabilidade que um documento

selecionado aleatoriamente de uma coleção inteira seja relevante, e $P(-R_q)$ é a probabilidade que um documento selecionado aleatoriamente de uma coleção inteira não seja relevante.

Considerando que $P(+R_q)$ e $P(-R_q)$ é inicialmente a mesma para todo documento da coleção tem-se:

$$sim(d_j, q) \approx \frac{P(d_j|+R_q)}{P(d_j|-R_q)} \tag{4.4}$$

Segundo Baeza e Ribeiro (1999), sabendo que $P(k_i|+R_q) + P(-k_i|+R_q) = 1$, após transformações algébricas pode-se escrever

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|+R_q)}{1 - P(k_i|+R_q)} + \log \frac{1 - P(k_i|-R_q))}{P(k_i|-R_q)}\right) \tag{4.5}$$

que é uma expressão chave para classificação computacional pelo modelo probabilístico.

Considerando que, a princípio, não conhecemos o conjunto R_q, é necessário criar um método para levantamento das probabilidades iniciais (ROBERTSON, *et. al.* 1980).

No início, logo depois da especificação da consulta, não existe nenhum documento recuperado. Assim, faz-se uma hipótese:

> 1 – $P(k_i|+R_q)$ é constante para todo termo de indexação k_i e igual a 0,5 (50% de possibilidade de ser ou não relevante);
>
> 2 – a distribuição dos termos de indexação entre os documentos não-relevantes pode ser aproximada da distribuição dos termos de indexação entre todos os documentos da coleção;

Assim temos:

$$P(k_i|+R_q) = 0,5 \tag{4.6}$$

$$P(k_i|-R_q) = n_i / N \quad (4.7)$$

Onde n_i é o número de documentos que contém o termo de indexação k_i e N é o número total de documentos da coleção. Dada essa hipótese, pode-se recuperar documentos que contém termos da consulta e promover uma classificação inicial probabilística.

Após a classificação inicial, é definido que tendo V como um subconjunto dos documentos inicialmente recuperados e classificados pelo modelo probabilístico, esse subconjunto pode ser definido como o topo *r* de documentos classificados onde *r* é um ponto inicial previamente definido, sendo V*i* um subconjunto de *V*, composto de documentos que contenham termos de indexação k_i. *V* e V_i também são utilizados para se referir ao número de elementos nos conjuntos. Para melhorar a classificação probabilística, é necessário melhorar as hipóteses para as probabilidades de relevância e de não-relevância. Isto pode ser feito da seguinte maneira: pode-se aproximar $P(k_i|+R_q)$ pela distribuição do termo de indexação k_i entre os documentos recuperados e pode-se aproximar $P(k_i|-R_q)$ pela consideração de que todos os documentos não-recuperados são documentos não-relevantes à consulta.

Assim, pode-se calcular as probabilidades de relevância e não-relevância:

$$P(k_i|+R_q) = Vi / V \quad (4.8)$$

$$P(k_i|-R_q) = (ni - Vi) / (N - V) \quad (4.9)$$

Esse processo pode ser repetido recursivamente. Assim, pode-se melhorar as hipóteses para as probabilidades $P(k_i|+R_q)$ e $P(k_i|-R_q)$ sem qualquer ajuda humana (KLUEV, 2000), diferente da idéia original. Porém, pode-se usar o auxílio do usuário para definir o subconjunto V.

As últimas fórmulas para $P(k_i|+R_q)$ e $P(k_i|-R_q)$ trazem problemas para valores pequenos de V e V*i*, pois sugerem na prática V = 1 e V*i* = 0. Para evitar esse problema é somado um fator de ajuste, resultando em:

$$P(k_i|+R_q) = (Vi+0,5) / (V+1) \quad (4.10)$$

$$P(k_i|-R_q) = (ni - Vi + 0,5) / (N - V + 1) \quad (4.11)$$

Segundo Baeza e Ribeiro (1999), definir um fator de ajuste constante e igual a 0,5 não é sempre satisfatório; uma alternativa é utilizar ni/N como fator de ajuste, como segue

$$P(k_i|+R_q) = (Vi + ni/N) / (V+1) \quad (4.12)$$

$$P(k_i|-R_q) = (ni - Vi + ni/N) / (N - V + 1) \quad (4.13)$$

Utilizando as expressões apresentadas é possível estimar as probabilidades de relevância e não-relevância para um conjunto de documentos. No capítulo 5 são apresentados mais detalhes sobre a utilização dessas expressões na classificação de um conjunto de documentos.

Vantagens do Modelo Probabilístico:

- Sua principal vantagem é que documentos são ordenados de forma decrescente de acordo com a probabilidade de relevância;
- Maior precisão na recuperação que os outros modelos clássicos;

Desvantagens do Modelo Probabilístico:

- Necessidade de descobrir a separação inicial de conjuntos relevantes e não-relevantes através de hipótese;
- O método clássico não explora a freqüência do termo de indexação no documento, utilizando pesos binários;

O desempenho do modelo depende da precisão da estimativa probabilística.

4.3. REALIMENTAÇÃO DE RELEVÂNCIA

A realimentação de relevância (*relevance feedback*) é a mais popular estratégia de reformulação de consulta. Em um ciclo de realimentação de relevância, o usuário é apresentado a uma lista de documentos recuperados e, depois de examiná-los, marca quais são relevantes. Segundo Salton e McGill (1983), na prática só os 10 documentos melhores classificados são examinados; a idéia principal consiste em selecionar termos importantes, ou expressões (termos compostos), dos documentos que são identificados como relevantes pelo usuário; esse processo aumenta a importância desses termos em uma nova formulação de consulta. Como resultado, numa nova consulta, esta será direcionada para os documentos relevantes e não serão verificados os não-relevantes.

A realimentação de relevância mostra uma boa melhoria de precisão para testes em pequenos conjuntos de documentos. Para essa melhoria podem ser usadas duas técnicas: consultas expandidas (adição de novos termos para consultas na coleção de documentos relevantes) e repesagem de termo (modificação do peso do termo baseado no julgamento de relevância feito pelo usuário). Neste trabalho, a realimentação de relevância é baseada na repesagem dos termos envolvidos nas consultas e nos documentos.

A realimentação de relevância apresenta outras estratégias de reformulação de consultas:

- O usuário interage com o sistema identificando documentos como relevantes ou não relevantes.
- Faz-se uma análise minuciosa dos resultados obtidos na consulta.
- Enfatiza a importância em alguns termos (relevantes) e não em outros (não-relevantes).

Para o modelo probabilístico há dois usos da realimentação de relevância, a repesagem de termos da consulta e a repesagem de termos da consulta através de uma variante do modelo probabilístico, conforme apresentados nas próximas seções.

4.3.1. REPESAGEM DE TERMOS PARA O MODELO PROBABILÍSTICO

O modelo probabilístico classifica dinamicamente documentos similares para uma consulta *q* de acordo com o princípio de classificação probabilística. Como definido anteriormente na expressão 4.5, a similaridade entre um documento d_j e uma consulta q é expressa como

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|+R_q)}{1 - P(k_i|+R_q)} + \log \frac{1 - P(k_i|-R_q))}{P(k_i|-R_q)}\right)$$

onde $P(k_i|+R_q)$ é a probabilidade do termo k_i estar no conjunto $+R_q$ (documentos relevantes) e $P(k_i|-R_q)$ é a probabilidade do termo k_i estar no conjunto $-R_q$ (documentos não-relevantes). Contudo, não se pode usar a equação acima quando as probabilidades de relevância e não-relevância não são conhecidas. Numa busca inicial, onde ainda não temos documentos recuperados, assume-se que a probabilidade de relevância $P(k_i|+R_q)$ é constante para todos os termos (0,5) e que a probabilidade de não-relevância $P(k_i|-R_q)$ pode ser aproximada da distribuição de toda a coleção. Assim temos as expressões 4.6 e 4.7, já apresentadas anteriormente,

$$P(k_i|+R_q) = 0,5$$
$$P(k_i|-R_q) = n_i / N$$

Onde, como já apresentado, n_i é o número de documentos na coleção que possuem o termo k_i. Substituindo na equação de similaridade obtém-se

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \log \frac{N - n_i}{n_i} \qquad (4.14)$$

Para buscas utilizando a realimentação, são utilizadas as estatísticas acumuladas sobre relevância e não-relevância em recuperações anteriores para estimar as probabilidades nas novas buscas. Tem-se então *Dr* como o conjunto de documentos relevantes de acordo com a seleção do usuário e Dr_i é um subconjunto de *Dr* composto de documentos que contém o termo k_i (BAEZA E RIBERO, 1999). Assim,

$$P(k_i|+R_q) = |\ Dr_i\ | / |\ Dr\ | \tag{4.15}$$

$$P(k_i|-R_q) = (n_i - |\ Dr_i\ |) / (N - |\ Dr\ |) \tag{4.16}$$

Assim, a expressão de cálculo de similaridade para a realimentação de relevância pode ser reescrita como:

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \log \left(\frac{|\ Dr_i\ |}{|\ Dr\ | - |\ Dr_i\ |} \Big/ \frac{n_i - |\ Dr_i\ |}{N - |\ Dr\ | - (n_i - |\ Dr_i\ |)} \right) \tag{4.17}$$

Os mesmos termos de consulta são repesados usando informação da realimentação. Quando os valores para |Dr| e $|Dr_i|$ são pequenos freqüentemente tendendo a |Dr| = 1 e $|Dr_i|$ = 0, utiliza-se um fator de ajuste para o cálculo das probabilidades de relevância e não-relevância,

$$P(k_i|+R_q) = (|Dr_i| + 0{,}5) / (|Dr| + 1) \tag{4.18}$$

$$P(k_i|-R_q) = (n_i - |Dr_i| + 0{,}5) / (N - |Dr| + 1) \tag{4.19}$$

A utilização do fator de ajuste não é satisfatória em alguns casos, surgindo uma outra proposta de ajuste:

$$P(k_i|+R_q) = (|Dr_i| + n_i/N) / (|Dr| + 1) \tag{4.20}$$

$$P(k_i|-R_q) = (n_i - |Dr_i| + n_i/N) / (N - |Dr| + 1) \tag{4.21}$$

As principais vantagens da realimentação de relevância são que o processo de realimentação é relacionado diretamente aos novos pesos

dos termos da consulta e que a repesagem do termo otimiza as hipóteses de independência do termo e indexação binária do documento, pois aproxima as probabilidades de relevância e não-relevância de 0 ou 1. Como desvantagens, os pesos dos termos no documento inicialmente calculados não são levados em conta durante o loop de realimentação e nenhuma consulta de expansão é usada (o mesmo conjunto de termos de indexação na consulta original é repesado várias vezes).

No modelo reportado neste trabalho será utilizado o modelo probabilístico combinado com o modelo vetorial durante a realimentação, visando obter um conjunto resposta mais efetivo que os conjuntos obtidos pelos modelos convencionais.

4.3.2. UMA VARIAÇÃO DA REPESAGEM DE TERMOS NO MODELO PROBABILÍSTICO

De acordo com Croft (1983) *apud* Baeza *et. al.* (1999), essa estratégia propõe a utilização de formulações distintas para a busca inicial e a realimentação. Surge uma adaptação à fórmula probabilística, utilizando a freqüência interna dos pesos dos documentos. Ela substitui na fórmula de similaridade as probabilidades de relevância e de não-relevância por um fator que depende da freqüência do termo no documento, assim temos

$$sim(d_j, q) = \sum_{i=1}^{t} w_{i,q} * w_{i,j} * F_{i,j,q} \qquad (4.22)$$

onde $F_{i,j,q}$ é interpretado como um fator que depende de uma tripla $[k_i,d_j,q]$ e é computada como uma função de $P(k_i|+R_q)$ e $P(k_i|-R_q)$.

A busca inicial é representada como

$$F_{i,j,q} = (C + idf_i)\, f'_{I,,j} \qquad (4.23)$$

$$f'_{i,,j} = K + (1 + K)\, (f_{i,,j} / \max (f_{i,,j}) \qquad (4.24)$$

onde $f'_{i,,j}$ é uma normalização da freqüência do termo no corpo do documento. C e K são constantes e podem ser ajustados de acordo com a coleção. Para coleções indexadas automaticamente utiliza-se C como 0 inicialmente. Assim temos

$$F_{i,j,q} = (C + \log \frac{P(k_i|+R_q)}{1 - P(k_i|+R_q)} + \log \frac{1 - P(k_i|-R_q))}{P(k_i|-R_q)} f'_{i,j} \quad (4.25)$$

4.4. O MODELO PROBABILÍSTICO EXPONENCIAL

O modelo probabilístico exponencial, proposto por Teevan e Karger (2003), considera a freqüência do termo no documento e o tamanho do documento, aplicados às expressões probabilísticas, para estimar as probabilidades de relevância e não-relevância, possibilitando uma melhor classificação dos termos e documentos envolvidos. Essa é a maior diferença entre o modelo probabilístico clássico e o modelo probabilístico exponencial.

A freqüência do termo no documento é o número de vezes *dt* que o termo *t* aparece em um documento, ℓ é o tamanho do documento representado pelo número total de termos do documento. A probabilidade de relevância de um termo no documento utiliza a freqüência deste no documento *dt* como função exponencial para obter o resultado. A probabilidade de não-relevância utiliza o tamanho ℓ do documento subtraído da freqüência do termo em questão como função exponencial para obter o resultado. Assim, a probabilidade inicial será

$$P(k_i|+R_q) = (0,5)^{dt} \quad (4.26)$$

$$P(k_i|-R_q) = (n_i / N)^{\ell - dt} \quad (4.27)$$

Após a classificação inicial, o modelo trabalha de maneira similar ao modelo probabilístico clássico, definindo V como um subconjunto dos documentos inicialmente recuperados e classificados pelo modelo probabilístico, sendo V*i* um subconjunto de *V*, composto de documentos que contenham termos de indexação *ki.* Para se melhorar as hipóteses probabilísticas é utilizado esse subconjunto para recalcular as probabilidades.

Assim, pode-se calcular as probabilidades de relevância e não-relevância:

$$P(k_i|+R_q) = (Vi / V)^{dt} \quad (4.28)$$

$$P(k_i|-R_q) = ((ni - Vi) / (N - V))^{\ell-dt} \quad (4.29)$$

Obtidos os valores das probabilidades de relevância e não-relevância de cada termo em um documento, aplica-se a expressão 4.6 para estimar a similaridade do documento em relação à consulta.

Esse modelo possibilita uma melhor classificação dos resultados, pois considera a freqüência do termo em cada documento para estimar as probabilidades. Como desvantagem, o modelo pode não ser tão eficaz se os documentos da coleção forem pequenos, possuindo poucos termos, assim os resultados seriam parecidos com os do modelo probabilístico clássico.

Após a classificação os documentos são apresentados em ordem decrescente de probabilidade de relevância e submetidos à realimentação de relevância de modo recursivo possibilitando aproximar a classificação do resultado ideal.

4.5. CONSIDERAÇÕES FINAIS

Este capítulo apresentou o modelo probabilístico clássico e suas variações. O sistema, que será apresentado no próximo capítulo, foi desenvolvido para possibilitar experimentos com o modelo probabilístico estendido e com o modelo probabilístico exponencial estendido, descritos neste capítulo, combinados com o modelo vetorial de recuperação de informação.

Na tabela 4.1 são apresentadas as vantagens e desvantagens dos modelos de recuperação de informação apresentados até aqui. Nos próximos capítulos são apresentados o sistema desenvolvido para a recuperação de informação utilizando o modelo probabilístico estendido, e os experimentos realizados.

TABELA 4.1: VANTAGENS E DESVANTAGENS DE CADA MODELO DE RECUPERAÇÃO DE INFORMAÇÃO

Modelo	Vantagens	Desvantagens
Booleano	- Formalismo claro - Simplicidade	- Consulta pode trazer poucos ou muitos documentos - Decisão binária - Sem balanceamento
Booleano Estendido	- Permite especificar as relevâncias dos termos	- Consulta pode trazer poucos ou muitos documentos - Decisão binária
Vetorial	- Recupera documentos similares - Utiliza a freqüência do termo para estimar os pesos	- Não realiza a realimentação de relevância
Probabilístico	- Ordena os documentos em ordem decrescente de probabilidade de relevância - Maior precisão na recuperação - Realimentação de relevância	- Utilização de hipótese - Não explora a freqüência do termo
Probabilístico Exponencial	- Ordena os documentos em ordem decrescente de probabilidade de relevância - Maior precisão na recuperação - Utiliza a freqüência do	- Utilização de hipótese - Por utilizar a freqüência dos termos no corpo do documento, não é o ideal para conjuntos com poucos

	termo para estimar as probabilidades - Realimentação de relevância	documentos.

5. MANIPULAÇÃO DE DOCUMENTOS USANDO UM MODELO PROBABILÍSTICO ESTENDIDO

5.1. CONSIDERAÇÕES INICIAIS

Para o tratamento da informação são utilizadas técnicas de Processamento e de Recuperação de Informação e de Descoberta de Conhecimento que realiza a aplicação de etapas de mineração de texto. Essas técnicas são utilizadas em Mello (2005), usando um algoritmo de mineração de dados para fazer os agrupamentos.

A pesquisa aqui reportada utiliza o modelo probabilístico e o probabilístico exponencial estendidos, combinados com recursos do modelo vetorial, como estratégia de recuperação de documentos texto. Essa estratégia toma como base a arquitetura do Sistema de Manipulação de Documentos proposta por Correa (2003), que realiza a extração automática de informação dos documentos texto, e as armazena em um banco de dados adaptado às novas especificações do sistema, possibilitando a recuperação com base no conteúdo desses documentos.

5.2. SISTEMA PARA MANIPULAÇÃO DE DOCUMENTOS

O Sistema para Manipulação de Documentos segue a arquitetura apresentada na figura 5.1. No pré-processamento e na recuperação de informação são utilizadas as técnicas do modelo probabilístico e do modelo probabilístico exponencial estendidos com recursos do modelo vetorial. A arquitetura do Sistema para Manipulação de Documentos utilizada na abordagem proposta é apresentada na figura 5.1.

De acordo com a proposta deste trabalho, as principais mudanças em relação ao modelo vetorial de recuperação de informação ocorrem no pré-processamento, no armazenamento dos dados e na recuperação da informação. O sistema extrai características que sirvam para identificar o conteúdo dos documentos, permitindo descobrir os

relacionamentos entre os termos dos documentos de uma coleção, através do uso combinado do modelo probabilístico com o modelo vetorial.

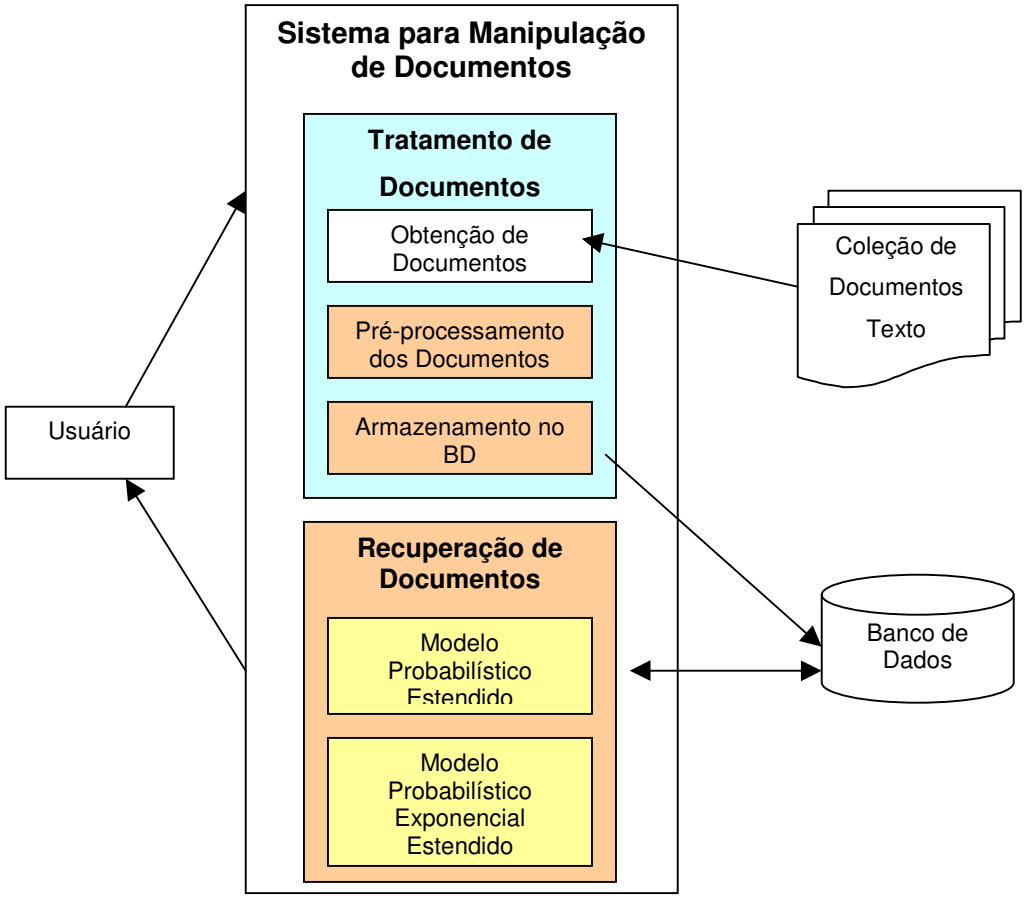

FIGURA 5.1 – ARQUITETURA DO SISTEMA DE MANIPULAÇÃO DE DOCUMENTOS

5.2.1. O MÓDULO DE TRATAMENTO DE DOCUMENTOS

No módulo de Tratamento de Documentos é realizada a obtenção dos documentos para que estes sejam submetidos à extração automática da informação, sua classificação e armazenamento.

Na obtenção dos documentos, o usuário pode escolher documentos armazenados em disco rígido, CD ou na Web. Para a extração da informação, o usuário seleciona os arquivos de seu interesse. A próxima etapa é o pré-processamento desses arquivos.

Durante o pré-processamento os documentos selecionados são analisados com o objetivo de identificar quais termos serão definidos como palavras-chave. Como resultado, é obtido um conjunto de palavras-chave (termos) que identificam o conteúdo do documento. Durante o pré-processamento são realizadas as seguintes etapas:

- Limpeza e Padronização do Texto;
- Remoção de *stop-words* (palavras que devem ser eliminadas do texto);
- *Stemming*: algoritmo que reduz as palavras na sua forma raiz;
- Determinação das probabilidades de relevância dos termos de acordo com o modelo probabilístico e com o modelo probabilístico exponencial, e os pesos de acordo com o modelo vetorial.

Quando o usuário elabora uma consulta o sistema busca todos os termos relacionados. As consultas podem ser elaboradas com um ou mais termos (composições).

Foi utilizada uma técnica de truncagem para selecionar os 50 termos mais relevantes com o objetivo de estabelecer um número máximo de características para representar um documento. Schitze e Silverstein (1997), indicam que, em geral, 50 termos são suficientes para representar um documento.

As informações obtidas dos textos são armazenadas em um banco de dados, o que permite que sejam reutilizadas por diversos usuários.

Para determinar as probabilidades iniciais de relevância e de não-relevância para cada termo são utilizadas as expressões a seguir, já apresentadas no capítulo anterior.

- Para o modelo probabilístico estendido:

$$P(k_i|+R_q) = 0,5 \quad (5.1)$$

$$P(k_i|-R_q) = n_i / N \quad (5.2)$$

- Para o modelo probabilístico exponencial estendido:

$$P(k_i|+R_q) = (0,5)^{dt} \quad (5.3)$$

$$P(k_i|-R_q) = (n_i / N)^{\ell - dt} \quad (5.4)$$

Essas probabilidades são armazenadas no banco de dados visando possibilitar o reuso dessas informações em futuras recuperações, enquanto não for alterado o conjunto de documentos.

Durante a recuperação é necessário refazer os cálculos de probabilidade para os termos envolvidos. São utilizadas as expressões 5.5 e 5.6 para o modelo probabilístico clássico e as expressões 5.7 e 5.8 para o modelo probabilístico exponencial, já apresentadas no capítulo 4.

$$P(k_i|+R_q) = Vi / V \quad (5.5)$$

$$P(k_i|-R_q) = (ni - Vi) / (N - V) \quad (5.6)$$

$$P(k_i|+R_q) = (Vi / V)^{dt} \quad (5.7)$$

$$P(k_i|-R_q) = ((ni - Vi) / (N - V))^{\ell - dt} \quad (5.8)$$

Essas expressões são utilizadas de maneira recursiva durante a realimentação de relevância possibilitando uma melhor classificação dos resultados.

O diagrama de classes, apresentado na figura 5.2, mostra como as informações estão organizadas no banco de dados e as relações entre essas informações.

Na classe *Term* constam até 50 termos extraídos do corpo de cada documento através dos atributos código do termo (*cod_term*) e nome do termo (*name*). Assim, um documento estará associado a *n* termos (n <=50) e

cada termo estará associado a *m* documentos. Na classe *Doc_Term_Relation* constam o código do documento (*cod_document*), o código do termo (*cod_term*), a freqüência do termo no documento (*frequency*), a probabilidade de relevância (*weight*) e o peso de acordo com o modelo vetorial (*weight_vet*), facilitando as novas buscas. A freqüência do termo no documento, a probabilidade de relevância e o peso pelo modelo vetorial são obtidos durante o pré-processamento.

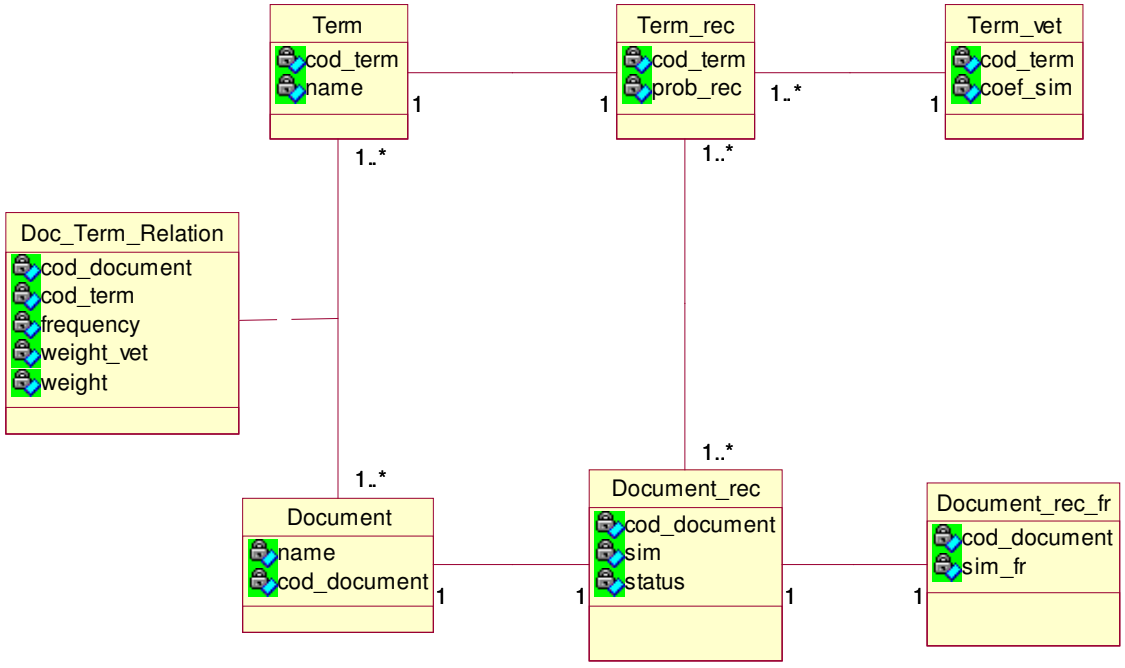

FIGURA 5.2 – ESTRUTURA DE CLASSES PARA INFORMAÇÕES ARMAZENADAS

A classe *Document* permite armazenar o código do documento (*cod_document*) e nome do documento (*name*). A classe *Document_rec* é utilizada para armazenar as informações dos documentos recuperados; nela constam o código do documento inicialmente recuperado (*cod_document*), a probabilidade de relevância do documento (*sim*) e o status (*status*), utilizado para indicar qual documento foi pré-selecionado como relevante pelo usuário na recuperação inicial. A classe *Document_rec_fr* é utilizada para armazenar os documentos submetidos à realimentação de relevância; nela constam os códigos dos documentos inicialmente recuperados (*cod_document*) e a nova probabilidade de relevância do documento (*sim_fr*).

Na classe *Term_rec* são armazenados os termos presentes nos documentos recuperados; nela constam os códigos dos termos de cada documento (*cod_term*) e a nova probabilidade de relevância do termo (*prob_rec*). A classe *Term_vet* é utilizada para armazenar os termos similares à consulta de acordo com o modelo vetorial de recuperação de informação; nela são armazenados o código do termo (*cod_term*) e o coeficiente de similaridade calculado (*coef_sim*). Os dados nas classes Term_rec, Term_vet, Document_rec e Document_rec_fr são temporários, auxiliando no processamento.

5.3. O MÓDULO RECUPERAÇÃO DE DOCUMENTOS

Este módulo é responsável por avaliar as consultas e retornar os documentos mais relevantes. Para validar a estratégia de busca proposta nesta pesquisa, são utilizados os modelos: probabilístico e probabilístico exponencial, ambos combinados com o modelo vetorial. Maiores detalhes são apresentados nas seções seguintes.

Nos modelos probabilísticos e exponencial as consultas são formuladas através de palavras-chave e a recuperação também é baseada nos termos fornecidos pelo usuário. Na interação com o modelo vetorial, o sistema encontra, através do cálculo de similaridade, os termos mais similares entre si, que pertencem aos documentos inicialmente considerados relevantes pelo usuário, busca os documentos que possuem esses termos similares, classifica-os pelo modelo probabilístico e apresenta ao usuário um conjunto resposta em ordem decrescente de probabilidade de relevância.

5.3.1. ESTRATÉGIA DE BUSCA PARA O MODELO PROBABILÍSTICO ESTENDIDO

A primeira abordagem adotada neste trabalho para a estratégia de busca probabilística estendida é apresentada no algoritmo 5.1.

ALGORITMO 5.1: ESTRATÉGIA DE BUSCA PROBABILÍSTICA ESTENDIDA

1: entrada: $q = \{t_1, t_2, ..., t_k\}$
2: saída: conjunto de documentos ordenados de acordo com a probabilidade de relevância
3: **para todo** termo t_k pertencente a q **faça**
4: submeter o termo t_k ao processo de normalização morfológica
5: **fim para**
6: **para** o conjunto q normalizado **faça**
7: *DocumentoRecuperado* = resultado da busca no banco de dados dos documentos que possuam o conjunto q entre seus termos
8: apresentar informações do(s) documentos(s) ao usuário
9: **fim para**
10: entrada: *DocumentoRelevante* = $\{dr_1, dr_2, ..., dr_n\}$
11: *RealimRelevancia* = resultado do cálculo de realimentação de relevância de cada documento
12: *TermoRec* = união ($\cup$) dos termos t de cada documento do conjunto *DocumentoRelevante*
13: **fim para**
14: **para todo** elemento de *TermoRec* **faça**
15: wt_k = resultado do cálculo do peso de acordo com o modelo vetorial
16: *MatrizSimTermo* = resultado do cálculo da similaridade entre o termo t_k e os demais termos (t_i) do conjunto de termos *TermoRec*
17: **fim se**
18: *qSim* = $\{t_1, t_2\}$ (2 termos mais similares)
19: *DocRecSim* = resultado da busca no banco de dados dos documentos que possuem os termos da busca *qSim*
20: **fim para**
21: *ConjuntoRelevante* = *DocumentoRelevante* $\cup$ *DocRecSim*
22: **para todo** elemento de *ConjuntoRelevante* **faça**
23: *RealimRelevancia* = resultado do cálculo de realimentação de relevância de cada documento
24: **fim para**
25: *DR* = conjunto de documentos de *RealimRelevancia*, ordenados pelo modelo probabilístico
26: **para todo** documento dj de DR **faça**
27: localizar no banco de dados as informações gerais estruturais
28: apresentar informações do componente para o usuário
29: **fim para**

Dada uma consulta $q = \{ t_1, t_2, t_3...t_k\}$, onde t representa um termo da consulta, a estratégia de busca aplica inicialmente cada termo da consulta ao processo de normalização morfológica. Para cada termo normalizado são localizados no banco de dados os documentos por ele indexados. Como resultado inicial temos um conjunto de documentos que possuem pelo menos um dos termos da consulta. Já neste primeiro momento

os documentos são apresentados em ordem decrescente de probabilidade de relevância. O cálculo da probabilidade de relevância de um documento é obtido através da expressão 4.5.

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|+R_q)}{1 - P(k_i|+R_q)} + \log \frac{1 - P(k_i|-R_q))}{P(k_i|-R_q)}\right)$$

Após ser determinada a probabilidade de relevância de cada documento inicialmente recuperado, estes são apresentados ao usuário que interage com o sistema selecionando alguns documentos que considerar relevantes para sua busca. Isso é necessário para que seja possível recalcular os pesos, realizando a realimentação de relevância. Durante a realimentação ocorre um processo recursivo no cálculo das probabilidades possibilitando uma melhor classificação dos documentos recuperados. Os termos dos documentos inicialmente selecionados pelo usuário são submetidos aos cálculos de similaridade pelo modelo vetorial. Cria-se uma matriz de similaridade entre esses termos. Os dois termos distintos melhores classificados são utilizados para uma nova consulta e recuperação de documentos; foi definido o total de dois termos para todos os casos com o objetivo de obter um conjunto resposta mais otimizado[1]. Esses novos documentos recuperados são submetidos aos cálculos de probabilidade e classificados pelo modelo probabilístico.

Para exemplificar a técnica de recuperação probabilística estendida, considere uma consulta $q = \{t_1, t_4\}$ e os documentos e termos de indexação apresentados na tabela 5.1.

Após os termos da consulta serem normalizados, é realizada uma busca no banco de dados pelos documentos indexados por pelo menos um desses termos. Em seguida é calculado o grau de similaridade do documento para com a consulta (expressão 5.9). No exemplo, seriam retornados cinco documentos (doc_1, doc_3, doc_5, doc_6 e doc_9).

[1] Nas experiências realizadas foi adotado esse limite de dois termos melhores classificados, que permitiu uma avaliação da técnica adotada. Devem ser realizadas experiências com um número maior de termos para uma comparação dos resultados.

TABELA 5.1: EXEMPLO DE UMA COLEÇÃO DE DOCUMENTOS ARMAZENADOS EM BANCO DE DADOS

Documentos	Termos de indexação									
	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
doc_1	0,45			0,35				0,56		
doc_2		0,70			0,89					
doc_3	0,45		0,78				0,70			
doc_4			0,78		0,89					0,40
doc_5		0,70		0,35					0,15	
doc_6				0,35				0,56		
doc_7							0,70			
doc_8			0,78							0,40
doc_9	0,45						0,70			
doc_{10}		0,70							0,15	

Nesse exemplo, a apresentação inicial dos documentos em ordem decrescente de probabilidade de relevância será:

TABELA 5.2: EXEMPLO DE DOCUMENTOS ORDENADOS NA BUSCA INICIAL

Classificação	Documentos
1	doc_1
2	doc_3
3	doc_9
4	doc_5
5	doc_6

Supondo que o usuário, interagindo com o sistema, considere os documentos doc_1 e doc_3 os mais relevantes para a primeira busca, é então realizada a realimentação de relevância, onde as probabilidades de relevância e não-relevância dos termos presentes nesses documentos são recalculadas. Nesse momento ocorre a combinação com o modelo vetorial com o objetivo de encontrar os dois termos mais similares entre os termos presentes nos

documentos doc_1 e doc_3. Considerando que os termos t_3 e t_8 tenham sido os dois termos distintos mais similares entre si, estes compõem uma nova expressão de busca. Nesse momento são recuperados também os documentos que possuem os termos t_3 e t_8.

Em seguida é realizada a reclassificação dos documentos envolvidos. Se o documento similar já estiver entre os documentos recuperados pelo modelo probabilístico que serão apresentados ao usuário, este será desconsiderado e não será apresentado, eliminando a duplicidade. Tendo sido formado o conjunto resposta, este é apresentado ao usuário (tabela 5.3).

TABELA 5.3: EXEMPLO DE CONJUNTO RESPOSTA APÓS A REALIMENTAÇÃO DE RELEVÂNCIA

Classificação	Documentos
1	doc_1
2	doc_3
3	doc_4

5.3.2. ESTRATÉGIA DE BUSCA PARA O MODELO PROBABILÍSTICO EXPONENCIAL ESTENDIDO

Esta abordagem utiliza conceitos da estratégia de recuperação probabilística exponencial apresentada por Teevan e Karger (2003) combinados ao modelo vetorial, seguindo a abordagem utilizada para o modelo probabilístico estendido apresentado anteriormente. A diferença entre as abordagens é em relação aos cálculos das probabilidades dos termos. Como o modelo probabilístico exponencial estendido utiliza a freqüência do termo no documento e o tamanho deste documento para estimar as probabilidades de relevância e não-relevância dos documentos, estas serão diferentes das probabilidades calculadas pelo modelo probabilístico estendido.

O algoritmo utilizado para a recuperação dos documentos é o mesmo utilizado para o modelo probabilístico estendido (algoritmo 5.1). A

diferença do modelo probabilístico exponencial estendido para o modelo probabilístico estendido está no momento de se estimar as probabilidades de relevância na recuperação inicial e na realimentação de relevância. Numa consulta $q = \{t_1, t_2, t_3...t_k\}$, onde t representa um termo da consulta, a estratégia de busca aplica cada termo da consulta ao processo de normalização morfológica e para cada termo normalizado são apresentados os documentos por ele indexados. A probabilidade de relevância do documento é obtida através da expressão 5.9 utilizando as probabilidades de relevância dos termos obtidas através da aplicação dos conceitos da recuperação probabilística exponencial estendida.

Determinada a probabilidade de relevância de cada documento inicialmente recuperado, estes são apresentados ao usuário, em ordem decrescente de probabilidade de relevância, que interage com o sistema selecionando alguns documentos que considerar relevantes para sua busca. A realimentação também ocorre como um processo recursivo no cálculo das probabilidades possibilitando uma melhor classificação dos documentos recuperados. Os termos dos documentos inicialmente selecionados pelo usuário são submetidos aos cálculos de similaridade pelo modelo vetorial. Cria-se uma matriz de similaridade entre esses termos. Os dois termos distintos melhores classificados são utilizados para uma nova consulta e recuperação de documentos. Esses novos documentos recuperados também são submetidos aos cálculos de probabilidade e classificados pelo modelo exponencial estendido. A última etapa é a apresentação final dos resultados ao usuário.

5.4 RECURSOS COMPUTACIONAIS

Os recursos computacionais utilizados na implementação da estratégia adotada para recuperação de documentos foram: Linguagem de programação: JAVA 2 SDK – Standard Edition, versão 1.5.0_01 e o Sistema Gerenciador de Banco de Dados: PostgreSQL 8.0.

5.5 CONSIDERAÇÕES FINAIS

Este capítulo apresentou os recursos desenvolvidos para um Sistema para Manipulação de Documentos, propondo uma forma de auxiliar a extração de informações relevantes de documentos, armazenando-as em um banco de dados.

A grande quantidade de variações deste modelo probabilístico sugere um aprofundamento maior das pesquisas afim de compará-los, visando obter resposta de qual modelo apresenta os melhores resultados para os diversos tipos de aplicação.

Foram utilizados o modelo probabilístico e o modelo probabilístico exponencial, combinados com o modelo vetorial, propondo, com isso, duas formas para a recuperação e classificação dos documentos. Isso possibilitou uma comparação entre essas duas abordagens, e algumas abordagens existentes. Os experimentos realizados para demonstrar a abordagem de classificação e comparações com outras abordagens são apresentados no próximo capítulo.

A seguir apresenta-se um resumo das expressões utilizadas na recuperação de informação:

Modelo Probabilístico Estendido

Probabilidade de relevância inicial

$$P(k_i|+R_q) = 0,5$$
$$P(k_i|-R_q) = n_i / N$$

Realimentação de relevância

$$P(k_i|+R_q) = Vi / V$$
$$P(k_i|-R_q) = (ni - Vi) / (N - V)$$

Modelo Probabilístico Exponencial Estendido

Probabilidade de relevância inicial

$$P(k_i|+R_q) = (0,5)^{dt}$$
$$P(k_i|-R_q) = (n_i / N)^{\ell-dt}$$

Realimentação de relevância

$$P(k_i|+R_q) = (Vi / V)^{dt}$$
$$P(k_i|-R_q) = ((ni - Vi) / (N - V))^{\ell-dt}$$

Expressão de similaridade (Modelo Probabilístico Estendido e Modelo Exponencial Estendido)

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|+R_q)}{1 - P(k_i|+R_q)} + \log \frac{1 - P(k_i|-R_q))}{P(k_i|-R_q)}\right)$$

6. EXPERIMENTOS

6.1. CONSIDERAÇÕES INICIAIS

Neste capítulo são apresentados os experimentos realizados com o objetivo de avaliar a estratégia proposta neste trabalho. Foram realizados experimentos com o conjunto de documentos *MEDLINE* (SHAW *et al.*, 1991). Os resultados foram submetidos às métricas de precisão (*precision*) e revocação (*recall*). A seguir são apresentados os detalhes sobre os experimentos e os resultados obtidos.

6.2. MÉTRICAS DE AVALIAÇÃO

As medidas de avaliação são utilizadas para analisar quão satisfatórios são os resultados obtidos num sistema de recuperação de informação. Para realizar essas avaliações são utilizadas as métricas de precisão (*precision*) e revocação (*recall*) sugeridas por Salton e McGill (1983).

A precisão (*precision*) representa a quantidade de documentos relevantes para o usuário dentre os itens que foram retornados como resposta a uma busca. Para estimar a precisão é necessário saber o total de itens relevantes na consulta (*tir*), e o total de itens recuperados do banco de dados (*tr*).

$$P = (tir / tr) \qquad (6.1)$$

A revocação (*recall*) representa a quantidade de itens relevantes recuperados dentre os itens relevantes existentes na base de dados. Para estimar a revocação é necessário saber o total de itens relevantes recuperados (*tirr*), e o total de itens relevantes armazenados no banco de dados (*ta*).

$$R = (tirr / ta) \qquad (6.2)$$

6.3. ABORDAGEM ADOTADA

O conjunto de documentos *MEDLINE* é composto por 1215 documentos publicados de 1974 a 1979 e são relacionados a documentos médicos. Não são documentos completos e sim resumos dos documentos originais. Esses documentos já encontram-se classificados, possibilitando comparar os resultados obtidos com os resultados fornecidos pelo *MEDLINE*. Foram utilizadas 30 consultas, baseadas nas 100 consultas sugeridas para este conjunto por Shaw *et al.* (1991), apresentadas na tabela 6.2.

Na primeira etapa, os documentos são submetidos ao módulo de extração de informação. Foram obtidos 6253 termos representativos, e esses termos foram armazenados no banco de dados. Para cada termo foram realizados os cálculos de peso de cada termo pelo modelo vetorial, da probabilidade de relevância e de não-relevância de acordo com o modelo probabilístico de recuperação de informação (BAEZA e RIBEIRO, 1999). Essas informações também foram armazenadas no banco de dados.

Quando o usuário realiza uma consulta, o módulo de recuperação de informação busca no banco de dados as informações referentes aos documentos que contém os termos envolvidos na consulta. O número de documentos relevantes, apresentados ao usuário como resultado da primeira busca, foi de 10 documentos seguindo proposta de Salton e McGill (1983). Em seguida, o usuário seleciona os documentos que são inicialmente considerados relevantes, com base nas informações fornecidas por Shaw *et al.* (1991), para que seja possível realizar os cálculos da realimentação de relevância. Os documentos inicialmente recuperados são submetidos à realimentação de relevância, o processo é repetido de maneira recursiva com o objetivo de possibilitar uma melhor classificação dos documentos que serão apresentados como conjunto resposta ao usuário. Após essa etapa, os termos dos documentos inicialmente recuperados e considerados relevantes são submetidos ao cálculo de similaridade de acordo com o modelo vetorial de recuperação de informação. É criado um vetor composto pelos 2 termos mais similares. Esses termos foram utilizados para recuperar outros documentos não

recuperados na busca inicial, que possuem termos similares aos termos dos documentos inicialmente recuperados e classificados. Esse número de termos foi assim definido para que os documentos recuperados sejam os mais similares; se esse número de termos for maior será recuperado um grande número de documentos, o que poderia comprometer a precisão da resposta. Foram aplicadas as mesmas consultas para o modelo probabilístico clássico e para o modelo probabilístico estendido, proposto neste trabalho, com o objetivo de comparar qual modelo trás como conjunto resposta os melhores resultados ao usuário.

Foi definido um limite de documentos para serem apresentados como resultado final para a busca com o objetivo de facilitar a visualização do conjunto de documentos pelo usuário. Tendo conhecimento do conjunto resposta ideal, o critério adotado para a apresentação dos resultados foi o de considerar como número de documentos recuperados a quantidade ideal de documentos considerados relevantes acrescidos de 50% (ex. Numa consulta onde o número ideal de documentos relevantes é 2, serão apresentados ao usuário 3 documentos como conjunto resposta (2 + (2 x 50%)) = 3)). O número de documentos recuperados através dos termos similares (modelo vetorial) apresentados como conjunto resposta é formado por um total de 50% do número ideal de documentos considerados relevantes (ex. 2 x 50% = 1). Tal procedimento é adotado visando obter um percentual de *precision* mais otimizado, tendo em vista que não estabelecendo o limite de documentos recuperados, a precisão para uma busca poderá ser muito baixa. Na tabela 6.1 são apresentados os documentos considerados relevantes para cada uma das 30 consultas, conforme disponibilizados em Shaw *et al.* (1991).

TABELA 6.1: CONJUNTO DE CONSULTAS ELABORADAS PARA UM CONJUNTO DE DOCUMENTOS MEDLINE

Consultas	Documentos Relevantes
1	74140, 74152, 74167, 75145, 76015, 76037, 76085, 76086, 76087, 76100, 76107, 76148, 76149, 76151, 76166, 76168, 76172, 76173, 76179, 77011, 77037, 77156, 77161, 78008, 78046, 78054, 78080, 78094, 78111, 78141, 78142, 79195, 79205, 79242
2	75002, 76080, 76100, 76144, 76145, 77010, 78094
3	76149, 76184, 76185, 76186, 76200, 77022, 77088, 77130, 78095
4	74023, 75002, 75156, 76015, 76031, 76081, 76085, 76109, 76142, 76143, 76144, 76145, 76149, 77023, 77102, 77129, 77169, 77190, 78086, 78094,

	78164, 79145
5	74166, 75007, 76007, 76015, 76059, 76089, 78013, 79012, 79060, 79135
6	74046, 75020, 75135, 76073, 76164, 77068, 77071, 77078, 77191, 78015, 78017, 78018, 78054, 78056, 78128, 78129, 79011, 79139, 79198, 79246
7	74130, 74154, 75039, 75060, 76009, 76015, 76059, 76069, 76089, 77060, 77075, 77076, 77077, 77079, 77144, 78035, 78155, 79013, 79015, 79036, 79182
8	74069, 74098, 74099, 75014, 75085, 76058, 76061, 76080, 76089, 77033, 77095, 77127, 77128, 77138, 77141, 77170, 77175, 77176, 78161, 79118, 79135, 79178, 79209
9	74099, 75012, 77017, 77175, 78077, 78125, 78132, 78133, 79020, 79103, 79116
10	74118, 75162, 75166, 76079, 76138, 76204, 77004, 77106, 77194, 78099, 78100, 78131, 78162, 79021, 79104, 79244, 79254
11	74032, 74121, 75044, 78030, 78044, 78108, 78133, 78135, 78140, 78146, 78168, 79125, 79129, 79139
12	74067, 75159, 76015, 76048, 77030, 77140, 77141, 77143, 79072
13	74046, 75130, 75135, 75156, 76015, 76037, 77021, 78161, 79018, 79126, 79127, 79128, 79135, 79204, 79210
14	74051, 74053, 76015, 76055, 76076, 76146, 76149, 76162, 76200, 76215, 77041, 77121, 77156, 78073, 78075, 78076, 79216, 79232
15	74044, 75093, 75143, 75181, 76081, 76086, 76141, 76151, 76189, 76217, 76222, 77019, 77052, 78048, 78067, 78173, 78181, 78193, 79078, 79176, 79177, 79179, 79180
16	76051, 76151
17	74043, 74062, 74149, 75188, 76015, 76118, 76119, 76165, 77143, 77183, 78061, 78130, 78134, 79254
18	75027, 75160, 75161, 76004, 76008, 76015, 76084, 76197, 77122, 77143, 78056, 78127, 78164, 79020, 79254
19	74011, 74057, 74149, 75161, 76015, 76194
20	74059, 75084, 75157, 75167, 77102, 77151, 77152, 78056, 78161, 79020, 79039, 79145
21	74149, 76015, 77143, 78130, 78134, 79020, 79113,
22	74008, 74161, 75167, 75180, 76029, 76202, 77008, 77197, 78056, 78132, 78142, 78172, 79006, 79106, 79132, 79199
23	76015, 76089, 76103, 77008, 77032, 77087, 77097, 77098, 77121, 78165, 79010, 79111, 79154, 79164,
24	74080, 74153, 75009, 75010, 75012, 75094, 75099, 76015, 76115, 76125, 77008, 77125, 77197, 78125, 78132, 79085, 79106
25	74012, 74030, 74043, 75026, 75091, 75156, 77021, 77024, 79117
26	74009, 74040, 74043, 74076, 76100, 76101, 76166, 76172, 76173, 77092, 78102, 79107
27	74052, 74069, 74136, 74141, 75024, 76037, 76061, 76184, 76185, 77170, 77176
28	74146, 75013, 75104, 75106, 76007, 76113, 77086, 77147, 78001, 78006, 79005, 79038, 79019, 79138
29	75102, 75158, 76095, 79012, 79211
30	74060, 75016, 76015, 76226, 78022, 78052, 79020, 79037, 79053, 79117, 79252

6.4. APLICAÇÃO DE ESTRATÉGIA DE BUSCA

Foi definida uma abordagem para a aplicação das estratégias de busca probabilística estendida. Essa abordagem considera os documentos

que o usuário classificou como relevantes, apresentados como resultado da primeira busca, para reclassificar os documentos através da realimentação de relevância (*feedbak relevance*), em seguida os termos dos documentos considerados relevantes são submetidos ao cálculo de similaridade pelo modelo vetorial (matriz de similaridade dos termos), os 2 termos mais relevantes são utilizados para buscar os documentos similares, estes são classificados e apresentados de acordo com o modelo probabilístico estendido.

A abordagem para a aplicação da estratégia de recuperação probabilística estendida utilizou como termos de busca os termos de indexação apresentados na tabela 6.2. Esses termos foram extraídos das expressões de consultas formuladas em linguagem natural, escolhidas entre as 100 consultas disponibilizadas pelo *MEDLINE.*

TABELA 6.2: CONSULTAS SUBMETIDAS PARA A AVALIAÇÃO DAS ESTRATÉGIAS DE BUSCA

Consulta
q1 = {effects, calcium, physical, mucus }
q2 = {mucus, hypersecretion, infection, submucosal glands, repiratory tract}
q3 = {lipid, respiratory, secretions}
q4 = {histochemical, respiratory, epithelia}
q5 = {liver, cirrhosis, vitamin A, metabilism}
q6 = {meconium, ileus, plug}
q7 = {dietary, supplementation, bile, salts}
q8 = {pancreatic, insufficiency, absorb, metabolize}
q9 = {concordance, biochemical, sibling}
q10 = {genetic, counseling, families, children}
q11 = {patient, normal, sweat, tests}
q12 = {concentration, potassium, sweat}
q13 = {vitamin D, metabilism, normal}
q14 = {proprieties, activity, galactosytransferase, enzymes}
q15 = {defects, systhesis, metabolism, cyclic, nucleotides}
q16 = {prolactin, patients}
q17 = {prognosis, episode, respiratory, failure}
q18 = {treat, pneumothorax}
q19 = {infants, wheezing, fibrosis}
q20 = {treatment, nasal, polyps}
q21 = {mechanical, ventilation, respiratory, failure}
q22 = {haemophilus, influenzae, pseudomonas, aeruginosa}
q23 = {viral, infection, lung}
q24 = {epidemiology, pseudomonas, aeruginosa}
q25 = {abnormalities, skeletal, muscle, functions, structure}

q26 = {incidence, dental, caries, periodontal}
q27 = {oxygen, transport, red, blood, cells, abnormal}
q28 = {effects, brain, central, nervous}
q29 = {abnormalities, taste}
q30 = {hypertrophic, osteoarthropathy}

6.4.1. AVALIAÇÃO DA ABORDAGEM UTILIZANDO O MODELO PROBABILÍSTICO ESTENDIDO

Para o Modelo Probabilístico Estendido esta abordagem trouxe como resultado documentos que possuíam os termos da consulta e também documentos relacionados aos termos similares encontrados através da matriz de similaridade no modelo vetorial.

Realizadas as consultas, os resultados foram submetidos às estimativas de precisão (*precision*) e revocação (*recall*) com base nas informações contidas na base de dados fornecida por Shaw *et al.* (1991). Nas tabelas a seguir, os campos *tir, tr, tirr* e *ta* significam, respectivamente, o total de documentos relevantes na consulta (*tir*), o total de documentos recuperados do banco de dados (*tr*), o total de documentos relevantes recuperados (*tirr*) e o total de documentos relevantes armazenados no banco de dados (*ta*).

Os campos *Pest* e *Rest* significam, respectivamente, a precisão e revocação da estratégia de busca probabilística estendida.

TABELA 6.3: PRECISION E RECALL PARA O MODELO PROBABILÍSTICO ESTENDIDO

Consulta	*tir*	*tr*	*tirr*	*ta*		Pest	Rest
1	8	68	8	34		11,76%	23,53%
2	2	14	2	7		14,29%	28,57%
3	3	18	3	9		16,67%	33,33%
4	8	44	8	22		18,18%	36,36%
5	5	20	5	10		25,00%	50,00%
6	14	42	14	21		33,33%	66,67%
7	13	42	13	21		30,95%	61,90%
8	5	50	5	25		10,00%	20,00%
9	3	22	3	11		13,64%	27,27%
10	8	34	8	17		23,53%	47,06%
11	8	28	8	14		28,57%	57,14%
12	4	18	4	9		22,22%	44,44%
13	5	30	5	15		16,67%	33,33%
14	4	36	4	18		11,11%	22,22%

15	6	50	6	25		12,00%	24,00%
16	2	3	2	2		66,67%	100,00%
17	6	28	6	14		21,43%	42,86%
18	8	30	8	15		26,67%	53,33%
19	2	12	2	6		16,67%	33,33%
20	5	26	5	13		19,23%	38,46%
21	4	14	4	7		28,57%	57,14%
22	6	32	6	16		18,75%	37,50%
23	3	28	3	14		10,71%	21,43%
24	7	34	7	17		20,59%	41,18%
25	2	18	2	9		11,11%	22,22%
26	3	24	3	12		12,50%	25,00%
27	2	22	2	11		9,09%	18,18%
28	4	30	4	15		13,33%	26,67%
29	3	10	3	5		30,00%	60,00%
30	4	22	4	11		18,18%	36,36%
				Média		**20,38%**	**39,65%**

Analisando os resultados da tabela 6.3 observa-se que a média percentual de precisão (precision) foi de 20,38%, e a revocação (*recall*) foi de 39,65%, valores que poderiam ser considerados baixos se não fosse a característica principal desse conjunto de documentos que é formado por resumos e não por documentos completos, o que pode comprometer a extração de termos de indexação representativos.

6.4.2. AVALIAÇÃO DA ABORDAGEM UTILIZANDO O MODELO PROBABILÍSTICO

Para o Modelo Probabilístico os resultados foram menos satisfatórios quando foram aplicados à realimentação de relevância e busca de documentos similares. Os campos *Ppro* e *Rpro* significam, respectivamente, a precisão e revocação da estratégia de busca probabilística exponencial.

TABELA 6.4: PRECISION E RECALL – MODELO PROBABILÍSTICO

Consulta	*tir*	*tr*	*tirr*	*ta*		Ppro	Rpro
1	7	68	7	34		10,29%	20,59%
2	2	14	2	7		14,29%	28,57%
3	3	18	3	9		16,67%	33,33%
4	6	44	6	22		13,64%	27,27%
5	4	20	4	10		20,00%	40,00%
6	14	42	14	21		33,33%	66,67%
7	11	42	11	21		26,19%	52,38%
8	5	50	5	25		10,00%	20,00%

9	2	22	2	11		9,09%	18,18%
10	8	34	8	17		23,53%	47,06%
11	4	28	4	14		14,29%	28,57%
12	4	18	4	9		22,22%	44,44%
13	4	30	4	15		13,33%	26,67%
14	3	36	3	18		8,33%	16,67%
15	5	50	5	25		10,00%	20,00%
16	2	3	2	2		66,67%	100,00%
17	4	28	4	14		14,29%	28,57%
18	6	30	6	15		20,00%	40,00%
19	1	12	1	6		8,33%	16,67%
20	4	26	4	13		15,38%	30,77%
21	3	14	3	7		21,43%	42,86%
22	6	32	6	16		18,75%	37,50%
23	2	28	2	14		7,14%	14,29%
24	6	34	6	17		17,65%	35,29%
25	2	18	2	9		11,11%	22,22%
26	3	24	3	12		12,50%	25,00%
27	1	22	1	11		4,55%	9,09%
28	3	30	3	15		10,00%	20,00%
29	3	10	3	5		30,00%	60,00%
30	3	22	3	11		13,64%	27,27%
				Média		**17,22%**	**33,33%**

Como resultado dessa aplicação observa-se que a média percentual de precisão (precision) foi de 17,22%, e a revocação (*recall*) foi de 33,33%. Observa-se que o modelo probabilístico não teve um desempenho muito bom, comprometendo os resultados. Porém, isso também se dá devido à característica do conjunto de documentos utilizado nos experimentos. A seguir são comparados os resultados dos modelos probabilístico e probabilístico estendido.

6.4.3. Comparação Entre os Modelos Probabilístico e Probabilístico Estendido

Analisando a tabela 6.5 observa-se que o modelo probabilístico estendido leva vantagem em relação ao modelo probabilístico. A diferença fundamental das duas aplicações é que para o modelo probabilístico estendido foram recuperados os documentos similares, o que melhorou a precisão e revocação.

TABELA 6.5: COMPARAÇÃO ENTRE OS MODELOS PROBABILÍSTICO E PROBABILÍSTICO ESTENDIDO

Consulta	Ppro	Pest		Rpro	Rest
1	10,29%	11,76%		20,59%	23,53%
2	14,29%	14,29%		28,57%	28,57%
3	16,67%	16,67%		33,33%	33,33%
4	13,64%	18,18%		27,27%	36,36%
5	20,00%	25,00%		40,00%	50,00%
6	33,33%	33,33%		66,67%	66,67%
7	26,19%	30,95%		52,38%	61,90%
8	10,00%	10,00%		20,00%	20,00%
9	9,09%	13,64%		18,18%	27,27%
10	23,53%	23,53%		47,06%	47,06%
11	14,29%	28,57%		28,57%	57,14%
12	22,22%	22,22%		44,44%	44,44%
13	13,33%	16,67%		26,67%	33,33%
14	8,33%	11,11%		16,67%	22,22%
15	10,00%	12,00%		20,00%	24,00%
16	66,67%	66,67%		100,00%	100,00%
17	14,29%	21,43%		28,57%	42,86%
18	20,00%	26,67%		40,00%	53,33%
19	8,33%	16,67%		16,67%	33,33%
20	15,38%	19,23%		30,77%	38,46%
21	21,43%	28,57%		42,86%	57,14%
22	18,75%	18,75%		37,50%	37,50%
23	7,14%	10,71%		14,29%	21,43%
24	17,65%	20,59%		35,29%	41,18%
25	11,11%	11,11%		22,22%	22,22%
26	12,50%	12,50%		25,00%	25,00%
27	4,55%	9,09%		9,09%	18,18%
28	10,00%	13,33%		20,00%	26,67%
29	30,00%	30,00%		60,00%	60,00%
30	13,64%	18,18%		27,27%	36,36%
	17,22%	**20,38%**		**33,33%**	**39,65%**

O conjunto de documentos possui algumas características que influenciaram as estimativas de precisão e revocação como:

- os documentos são compostos apenas por resumos dos documentos originais, impossibilitando uma melhor seleção de termos representativos;

- o conjunto de documentos é composto por muitos termos técnicos relacionados à medicina, o que impossibilita saber se

a busca deve ser composta pelos termos sugeridos ou se deve ser composta por termos técnicos (ex. *mucus*, *mucous* ou *mucin*?);

Com relação aos resultados obtidos foi observado que os documentos recuperados foram os documentos considerados mais relevantes por Shaw *et al.* (1991). Os documentos não recuperados não possuíam os termos envolvidos na busca.

TABELA 6.6: DOCUMENTOS RECUPERADOS PARA CADA CONSULTA

Consultas	Probabilístico	Probabilístico Estendido
1	76087, 76172, 76168, 76086, 76166, 78008, 76179	+ 75145
2	77010, 76144	
3	77022, 76185, 78095	
4	76085, 75156, 76031, 76142, 76145, 77190	+ 78086, 78094
5	75007, 76059, 74166, 78013	+ 76015
6	74046, 75020, 75135, 76073, 76164, 77068, 77071, 77078, 77191, 78015, 78017, 78018, 78054, 78056, 78128, 78129, 79011, 79139, 79198, 79246	
7	75039, 76069, 77075, 77076, 77077, 77079, 77144, 78035, 78155, 79013, 79182	+ 77060, 79036
8	79178, 74098, 77128, 77127, 79135	
9	78133, 78125	+ 77017
10	76079. 76204, 77004, 78099, 78131, 79021, 79104, 79244	
11	75044, 78030, 78133, 79125	+ 78044, 78135, 78140, 78146
12	74067, 75159, 77030, 77140	
13	79126, 79127, 79135, 79210	+ 79018
14	74051, 74053, 76215	+ 79232
15	75093, 76081, 76141, 78067, 78193	+ 76217
16	76051, 76151	
17	74149, 78130, 78134, 76119	+78061, 75188
18	75027, 75161, 76004, 76008, 76084, 78127	+ 77122, 76015
19	76194	+74149
20	75084, 77102, 77151, 78056	+ 77152
21	74149, 78130, 79113	+ 78134
22	74008, 74161, 75180, 76029, 78142, 79132	
23	77008, 79010	+ 77032

24	74080, 75009, 75010, 75099, 78125, 78132	+ 77008
25	74030, 75091	
26	76100, 76101, 77092	
27	77176	+ 77170
28	75013, 79038, 79138	+ 79039
29	75102, 75158, 76095	
30	75016, 76226, 79037	+ 79020

Na tabela 6.6 são apresentados os documentos recuperados para cada uma das consultas realizadas. Observa-se que em alguns casos não ocorreram melhoras no conjunto resposta, como nos casos das consultas 2, 3, 6 entre outras. Porém, na maioria das consultas foram recuperados documentos por meio de termos similares. Na consulta 11 por exemplo, o número de documentos considerados relevantes aumentou 100%, na consulta 17 a melhora foi de 50%.

6.5. COMPARAÇÃO COM OUTROS EXPERIMENTOS

Nesta seção são apresentados os experimentos realizados em classes Java API e os resultados apresentados de acordo com as métricas de precisão (*precision*) e revocação (*recall*), seguindo a mesma abordagem realizada por Mello (2005). A seguir são apresentados os detalhes sobre os experimentos e os resultados obtidos.

Foram realizados experimentos em um conjunto de classes da Java API tendo sido definidas 30 consultas em um conjunto de 100 componentes da biblioteca Java API, de acordo com a proposta de Mello (2005). As consultas estão representadas na tabela 6.11.

Na primeira etapa, os componentes são submetidos ao módulo de extração de informação. Foram obtidos 1553 termos representativos, e esses termos foram armazenados no banco de dados. Para cada termo foram realizados os cálculos de peso de cada termo pelo modelo vetorial, da probabilidade de relevância e de não-relevância de acordo com o modelo probabilístico de recuperação de informação (BAEZA e RIBEIRO, 1999) e

também a probabilidade de relevância e de não-relevância de acordo com o modelo probabilístico exponencial (TEEVAN e KARGER, 2003). Essas informações foram armazenadas no banco de dados.

O número de documentos relevantes apresentados ao usuário como resultado da primeira busca tem como base 5% do total de documentos da coleção (ex. 100 x 5% = 5), este percentual foi definido para limitar o primeiro subconjunto resposta, e por ser um conjunto relativamente pequeno (100 documentos). A configuração do ambiente e a estratégia de busca seguiram os conceitos mencionados nas seções anteriores. As tabelas 6.7, 6.8, 6.9 e 6.10 foram definidas por Mello (2005), e apresentam os documentos considerados relevantes para cada uma das 30 consultas.

TABELA 6.7: CONJUNTO DE CONSULTAS ELABORADAS PARA O PACOTE JAVA.NET

Consultas	Classes Relevantes
1	DatagramPacket e DatagramSocket
2	ServerSocket e Socket
10	InetAddress
17	Authenticator e PasswordAuthentication
18	ContentHandler, URL, URLConnection, URLStreamHandler e HttpURLConnection
28	HttpUrlConnection

TABELA 6.8: CONJUNTO DE CONSULTAS ELABORADAS PARA O PACOTE JAVA.ÚTIL

Consultas	Classes Relevantes
4	Stack, Vector, Treeset e LinkedList
6	StringTokenizer
9	Timer e TimerTask
11	Dictionary
16	Calendar, TimeZone e GregorianCalendar
19	Collections
20	TreeSet

TABELA 6.9: CONJUNTO DE CONSULTAS ELABORADAS PARA O PACOTE JAVA.IO

Consultas	Classes Relevantes
7	BufferedInputStream, BufferedOutputStream, ByteArrayInputStream, ByteArrayOutpuStream, DataInputStream, DataOutputStream, FileInputStream, FileOutputStream, InputStream, OutputStream, PipedInputStream, PipedOutputStream, PrintStream, PushbackInputStream, ObjectInputStream e ObjectOutputStream
8	File
15	BufferedReader, BufferedWriter, CharArrayReader, CharArrayWriter, FileReader, FileWriter, PipedReader, PipedWriter, PrintWriter, Writer, Reader, StringWriter, StringReader, PushbackReader, FilterReader e FilterWriter
23	CharArrayReader e CharArrayWriter
24	ObjectInputStream e ObjectOutputStream
26	PushbackReader e PushbackInputStream
27	StreamTokenizer

TABELA 6.10: CONJUNTO DE CONSULTAS ELABORADAS PARA O PACOTE JAVA.AWT

Consultas	Classes Relevantes
3	BorderLayout, Container, CardLayout, FlowLayout, GridBagLayout e GridLayout
5	Point, Polygon e Rectangle
12	Button, Canvas, Checkbox, CheckboxGroup, CheckboxMenuItem, Choice, Cursor, Dialog, Label, List, Menu, MenuBar, MenuItem, PopupMenu, Scrollbar, TextArea e TextField
13	JobAttributes e PageAttributes
14	Dialog e Frame
21	Color
22	TextArea, TextField e Label
25	PipedReader, PipedWriter, PipedInputStream, PipedOutputStream
29	FlowLayout
30	Menu, MenuItem e PopupMenu

Para avaliar a abordagem foram realizadas consultas, apresentadas na tabela 6.11. As consultas foram elaboradas com termos de indexação (k_i) presentes nos documentos d_j, de acordo com a proposta de Mello (2005).

TABELA 6.11: CONSULTAS SUBMETIDAS PARA A AVALIAÇÃO DAS ESTRATÉGIAS DE BUSCA (MELLO, 2005)

Consulta	Objetivo do usuário
q1 = {sends, receives, packets}	Obter classes que enviam e recebem pacotes pela rede
q2 = {port, host, socket}	Obter classes que disponibilizem serviços na rede
q3 = {interface, window, layout}	Obter classes para construção de interfaces gráficas
q4 = {vectors, arrays}	Obter classes que manipulam estruturas de dados
q5 = {polygon, rectangle}	Obter classes para criação de figuras geométricas
q6 = {break, string, tokens}	Obter classes que manipulem seqüência de caracteres
q7 = {read, write, file, bytes, streams}	Obter classes que permitam a leitura e escrita de arquivos
q8 = {represent, file, directory}	Obter classes que representam arquivos ou diretórios
q9 = {schedules, delay, task}	Obter classes que permitem o agendamento de execução de tarefas
q10 = {host, address}	Obter classes que representem endereço IP
q11 = {dictionary, key, value}	Obter classes que manipulem estruturas de dados do tipo dicionário
q12 = {widget, components, graphic,	Obter classes que representam objetos

event}	gráficos
q13 = {job, print, page, document}	Obter classes que controlem impressão de arquivos
q14 = {window, title, border}	Obter classes que permitem o gerenciamento de janelas gráficas
q15 = {read, write, character}	Obter classes de leitura e arquivo de caracteres em arquivos
q16 = {calendar, time, zone}	Obter classes que manipulam datas
q17 = {authentication, password}	Obter classes que realizam autenticações em rede
q18 = {connection, url}	Obter classes que estabelecem conexões através de urls
q19 = {binary, search}	Obter classes que realizam busca binária
q20 = {sorted, set, elements}	Obter classes que manipulam conjunto de dados ordenados
q21 = {rgb, color, red, green, blue)	Obter classes de gerenciamento de cores em interfaces gráficas
q22 = {graphics, text, component}	Obter classes que permitem manipulação de textos em interfaces gráficas
q23 = {reads, writes, characters, array}	Obter classes de leitura e escrita de caracteres em arrays
q24 = {storage, objects, file, stream}	Obter classes que permitem a persistência de objetos em arquivos
q25 = {thread, read, White, data}	Obter classes de leitura e escrita de dados em threads
q26 = {data, pushed, back, stream}	Obter classes que enviem dados para um buffer de escrita
q27 = {stream, tokenizes}	Obter classes que permitem ler partes de uma seqüência de dados
q28 = {request, http, server}	Obter classes que estabelecem conexões com protocolo http
q29 = {components, left, right, flow}	Obter classes que disponibilizem componentes gráficos da esquerda para a direita em uma interface gráfica
q30 = {menu, popup}	Obter classes para criação de menus gráficos

Para o Modelo Probabilístico Estendido esta abordagem trouxe como resultado documentos que pertencem ao conjunto ideal de respostas e também documentos relacionados aos termos similares encontrados através da matriz de similaridade no modelo vetorial.

Nas tabelas a seguir, os campos *tir, tr, tirr* e *ta* significam, respectivamente, o total de itens relevantes na consulta, o total de itens

recuperados do banco de dados, o total de itens relevantes recuperados e o total de itens relevantes armazenados no banco de dados.

Os campos *Ppro* e *Rpro* significam, respectivamente, a precisão e revocação da estratégia de busca probabilística clássica.

TABELA 6.12: PRECISION E RECALL – MODELO PROBABILÍSTICO ESTENDIDO

Consulta	*tir*	*tr*	*tirr*	*ta*		Ppro	Rpro
1	2	3	2	2		66,67%	100,00%
2	2	3	2	2		66,67%	100,00%
3	5	12	5	6		41,67%	83,33%
4	3	8	3	4		37,50%	75,00%
5	3	5	3	3		60,00%	100,00%
6	1	2	1	1		50,00%	100,00%
7	16	24	16	16		66,67%	100,00%
8	1	2	1	1		50,00%	100,00%
9	2	3	2	2		66,67%	100,00%
10	1	2	1	1		50,00%	100,00%
11	1	2	1	1		50,00%	100,00%
12	14	33	14	17		42,42%	82,35%
13	2	3	2	2		66,67%	100,00%
14	2	3	2	2		66,67%	100,00%
15	13	32	13	16		40,63%	81,25%
16	3	4	3	3		75,00%	100,00%
17	2	3	2	2		66,67%	100,00%
18	4	9	4	5		44,44%	80,00%
19	1	2	1	1		50,00%	100,00%
20	1	2	1	1		50,00%	100,00%
21	1	2	1	1		50,00%	100,00%
22	1	4	1	3		25,00%	33,33%
23	1	4	1	2		25,00%	50,00%
24	2	3	2	2		66,67%	100,00%
25	4	6	4	4		66,67%	100,00%
26	2	3	2	2		66,67%	100,00%
27	1	2	1	1		50,00%	100,00%
28	1	2	1	1		50,00%	100,00%
29	1	2	1	1		50,00%	100,00%
30	2	5	2	3		40,00%	66,67%
				Média		**53,28%**	**91,73%**

Analisando os resultados dessa aplicação observa-se que a média percentual de precisão (precision) foi de 53,28%, e a revocação (*recall*) foi de 91,73%. Para obtermos uma revocação melhor o grau de satisfação da precisão irá diminuir, porém se compararmos com os valores de revocação e

precisão apresentados por Mello (2005), veremos que a precisão praticamente dobrou no modelo probabilístico em comparação ao modelo vetorial e teve uma melhora considerável em relação ao modelo por agrupamentos (Tabela 6.14).

Em relação ao *recall,* as médias das abordagens deste trabalho só não foram superiores ao modelo vetorial, porém o modelo vetorial é o que tem a pior média de precisão por trazer muitos documentos não relevantes no conjunto de documentos recuperados.

Para o Modelo Probabilístico Exponencial os resultados foram mais satisfatórios quando foram aplicados à realimentação de relevância e busca de documentos similares; os resultados e a classificação foram mais precisos quando comparados aos outros modelos. Os documentos considerados similares melhores classificados pertenciam em sua maioria ao pacote Java relevante.

Os campos *Pexp* e *Rexp* significam, respectivamente, a precisão e revocação da estratégia de busca probabilística exponencial.

TABELA 6.13: PRECISION E RECALL – MODELO PROBABILÍSTICO ESPONENCIAL ESTENDIDO

Consulta	*Tir*	*tr*	*tirr*	*Ta*		Pexp	Rexp
1	2	3	2	2		66,67%	100,00%
2	2	3	2	2		66,67%	100,00%
3	6	9	6	6		66,67%	100,00%
4	3	8	3	4		37,50%	75,00%
5	3	5	3	3		60,00%	100,00%
6	1	2	1	1		50,00%	100,00%
7	16	30	16	16		53,33%	100,00%
8	1	2	1	1		50,00%	100,00%
9	2	3	2	2		66,67%	100,00%
10	1	2	1	1		50,00%	100,00%
11	1	2	1	1		50,00%	100,00%
12	14	33	14	17		42,42%	82,35%
13	2	3	2	2		66,67%	100,00%
14	2	3	2	2		66,67%	100,00%
15	13	32	13	16		40,63%	81,25%
16	3	4	3	3		75,00%	100,00%
17	2	3	2	2		66,67%	100,00%
18	4	9	4	5		44,44%	80,00%
19	1	2	1	1		50,00%	100,00%
20	1	2	1	1		50,00%	100,00%

21	1	2	1	1		50,00%	100,00%
22	1	5	1	3		20,00%	33,33%
23	1	4	1	2		25,00%	50,00%
24	2	3	2	2		66,67%	100,00%
25	4	6	4	4		66,67%	100,00%
26	2	3	2	2		66,67%	100,00%
27	1	2	1	1		50,00%	100,00%
28	1	2	1	1		50,00%	100,00%
29	1	2	1	1		50,00%	100,00%
30	3	5	3	3		60,00%	100,00%
				Média		**54,17%**	**93,40%**

Observou-se que a média percentual de precisão (precision) foi de 54,17%, e a revocação (*recall*) foi de 93,40%. Houve uma melhora em relação aos dados da recuperação probabilística clássica. Isso se dá justamente pela melhor classificação dos documentos recuperados. No modelo probabilístico exponencial o tamanho do documento (número de termos que este possui) e a freqüência de cada termo são de fundamental importância, pois são considerados durante os cálculos das probabilidades.

Analisando a tabela 6.14, assim como no modelo probabilístico estendido, o grau de revocação da abordagem utilizando o modelo probabilístico exponencial estendido tem média inferior à abordagem através do modelo vetorial. Isso se dá devido à melhora do grau de precisão, o que diminui o número de documentos relevantes apresentados. Outro ponto importante a ser analisado é que a precisão dos modelos probabilístico estendido e exponencial estendido foi de 100%; isso ocorre devido ao conjunto resposta sempre ser formado pelo número de documentos relevantes acrescidos de 50% desse número, por outros documentos que possuem os termos similares aos termos dos documentos considerados relevantes pelo usuário, visando contemplar o conjunto ideal de resposta.

Os campos Pvet e Rvet indicam, respectivamente, a precisão e revocação da estratégia de busca vetorial convencional. Os campos Pagr e Rarg indicam, respectivamente, a precisão e revocação da estratégia de busca vetorial utilizando agrupamentos. Os demais campos já foram descritos anteriormente.

TABELA 6.14: COMPARAÇÃO ENTRE OS MODELOS VETORIAL, POR AGRUPAMENTOS, PROBABILÍSTICO ESTENDIDO E PROBABILÍSTICO EXPONENCIAL ESTENDIDO

Consulta	Pvet	Pagr	Ppro	Pexp		Rvet	Ragr	Rpro	Rexp
1	25,00%	25,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
2	20,00%	25,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
3	28,57%	30,00%	41,67%	66,67%		100,00%	100,00%	83,33%	100,00%
4	6,90%	100,00%	37,50%	37,50%		100,00%	100,00%	75,00%	75,00%
5	20,00%	66,67%	60,00%	60,00%		100,00%	100,00%	100,00%	100,00%
6	1,03%	50,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
7	44,12%	61,11%	66,67%	53,33%		100,00%	73,33%	100,00%	100,00%
8	2,44%	100,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
9	100,00%	100,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
10	6,25%	10,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
11	5,56%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
12	26,32%	50,00%	42,42%	42,42%		29,41%	58,82%	82,35%	82,35%
13	50,00%	66,67%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
14	25,00%	5,00%	66,67%	66,67%		100,00%	50,00%	100,00%	100,00%
15	41,67%	62,50%	40,63%	40,63%		93,75%	62,50%	81,25%	81,25%
16	100,00%	75,00%	75,00%	75,00%		100,00%	100,00%	100,00%	100,00%
17	33,33%	100,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
18	22,73%	0,00%	44,44%	44,44%		100,00%	0,00%	80,00%	80,00%
19	11,11%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
20	11,11%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
21	33,33%	5,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
22	12,00%	15,00%	25,00%	20,00%		100,00%	100,00%	33,33%	33,33%
23	6,06%	0,00%	25,00%	25,00%		100,00%	0,00%	50,00%	50,00%
24	5,88%	11,11%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
25	17,39%	11,11%	66,67%	66,67%		100,00%	50,00%	100,00%	100,00%
26	6,25%	11,11%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
27	2,33%	50,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
28	3,03%	33,33%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
29	9,09%	5,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
30	15,79%	5,00%	40,00%	60,00%		100,00%	33,33%	66,67%	100,00%
	23,08%	**38,29%**	**53,28%**	**54,17%**		**97,44%**	**84,27%**	**91,73%**	**93,40%**

Para uma comparação entre os modelos considerando o grau de revocação máximo (100%) obtido para todas as consultas nos 4 modelos, foram observados os melhores graus de precisão e revocação dos modelos vetorial e de agrupamentos, comparando-os com o modelo probabilístico estendido e com o modelo probabilístico exponencial estendido. Os resultados são apresentados na tabela 6.15.

TABELA 6.15: COMPARAÇÃO ENTRE OS MODELOS VETORIAL, POR AGRUPAMENTOS, PROBABILÍSTICO E PROBABILÍSTICO EXPONENCIAL (RECALL MÁXIMO)

Consulta	Pvet	Pagr	Ppro	Pexp		Rvet	Ragr	Rpro	Rexp
1	25,00%	25,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
2	20,00%	25,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
3	28,57%	30,00%	42,86%	66,67%		100,00%	100,00%	100,00%	100,00%
4	6,90%	100,00%	16,00%	40,00%		100,00%	100,00%	100,00%	100,00%
5	20,00%	66,67%	60,00%	60,00%		100,00%	100,00%	100,00%	100,00%
6	1,03%	50,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
7	44,12%	61,11%	66,67%	53,33%		100,00%	73,33%	100,00%	100,00%
8	2,44%	100,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
9	100,00%	100,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
10	6,25%	10,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
11	5,56%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
12	26,32%	50,00%	26,15%	25,00%		29,41%	58,82%	100,00%	100,00%
13	50,00%	66,67%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
14	25,00%	5,00%	66,67%	66,67%		100,00%	50,00%	100,00%	100,00%
15	41,67%	62,50%	43,24%	38,10%		93,75%	62,50%	100,00%	100,00%
16	100,00%	75,00%	75,00%	75,00%		100,00%	100,00%	100,00%	100,00%
17	33,33%	100,00%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
18	22,73%	0,00%	26,32%	41,67%		100,00%	0,00%	100,00%	100,00%
19	11,11%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
20	11,11%	25,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
21	33,33%	5,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
22	12,00%	15,00%	23,08%	27,27%		100,00%	100,00%	100,00%	100,00%
23	6,06%	0,00%	18,18%	7,69%		100,00%	0,00%	100,00%	100,00%
24	5,88%	11,11%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
25	17,39%	11,11%	66,67%	66,67%		100,00%	50,00%	100,00%	100,00%
26	6,25%	11,11%	66,67%	66,67%		100,00%	100,00%	100,00%	100,00%
27	2,33%	50,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
28	3,03%	33,33%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
29	9,09%	5,00%	50,00%	50,00%		100,00%	100,00%	100,00%	100,00%
30	15,79%	5,00%	33,33%	60,00%		100,00%	33,33%	100,00%	100,00%
	23,08%	**38,29%**	**51,03%**	**53,16%**		**97,44%**	**84,27%**	**100,00%**	**100,00%**

Para obter a revocação a 100%, apresentando ao usuário todos os documentos relevantes armazenados no banco de dados, foi necessário expandir o número de documentos apresentados ao usuário; esse procedimento diminuiu um pouco o percentual de precisão, mas tornou possível uma melhor análise e comparação. A abordagem utilizada foi a mesma utilizada nas tabelas anteriores, a única diferença foi a realização da busca até que se completasse o conjunto resposta ideal, para algumas consultas como a 12 e a 15, que envolviam um maior conjunto ideal de resposta, as precisões foram menos satisfatórias se comparadas aos modelos

vetorial e por agrupamentos, porém, na grande maioria das consultas houve uma considerável melhora de desempenho, é o que podemos observar nas consultas 1, 2, 14, 23, 30 entre outras. A classificação probabilística, a realimentação de relevância e a utilização dessa realimentação de maneira recursiva possibilitaram uma melhor classificação de relevância para os documentos do conjunto utilizado neste trabalho.

6.7. CONSIDERAÇÕES FINAIS

Este capítulo apresentou os resultados obtidos em experimentos realizados para a estratégia baseada no modelo probabilístico estendido. Foram utilizados para os experimentos 2 conjuntos de documentos: resumos do *MEDLINE* e um conjunto de classes da biblioteca Java API. As consultas utilizadas foram as sugeridas por Shaw *et al.* (1991) para o conjunto *MEDLINE*, e por Mello (2005) para o conjunto de classes da biblioteca Java API, onde para cada consulta foram determinadas as classes consideradas relevantes como resposta.

A abordagem baseada no modelo probabilístico exponencial estendido, aplicada no conjunto de classes da biblioteca Java API, leva vantagem em relação aos resultados apresentados pelo modelo probabilístico estendido para este conjunto, pois também considera a freqüência dos termos envolvidos e o tamanho do documento para estimar as probabilidades de relevância e não-relevância. Em alguns casos o modelo vetorial e o modelo por agrupamentos levaram vantagem sobre os modelos utilizados neste trabalho, contudo, analisando as médias de performance observa-se que os modelos propostos neste trabalho possuem uma vantagem muito grande em relação aos outros modelos mencionados.

A modelagem probabilística estendida, proposta nesse trabalho, apresentou bom desempenho para os conjuntos de documentos utilizados, melhorando a precisão dos resultados.

7. CONCLUSÕES

7.1. CONSIDERAÇÕES INICIAIS

Este trabalho apresentou uma abordagem para recuperação de documentos de acordo com o modelo probabilístico de recuperação de informação, combinado com o modelo vetorial. Esse projeto teve como objetivo a pesquisa de técnicas e métodos, que visam a definição de estratégias para a recuperação de informação, e de contribuir para o desenvolvimento de um modelo probabilístico estendido, combinado com o modelo vetorial. É realizada a extração de termos de indexação que são armazenados em banco de dados e utilizados pelo sistema durante a recuperação dos documentos. Esses termos são submetidos aos cálculos de probabilidade de relevância e não relevância, de acordo com os modelos utilizados nos experimentos. Durante a realimentação de relevância ocorre a combinação com o modelo vetorial, que resulta na recuperação de documentos que possuem termos similares aos termos dos documentos considerados relevantes pelo usuário. Por fim, o sistema recupera e classifica os documentos relevantes, apresentando-os como conjunto resposta, em ordem decrescente de probabilidade de relevância.

7.2. CONTRIBUIÇÕES E RESULTADOS

A maior contribuição deste trabalho é a estratégia adotada para a recuperação de documentos. Para validar a idéia foi desenvolvido um protótipo do Sistema para Manipulação de Documentos, que possibilita ao usuário recuperar documentos com base nos termos de consulta.

A estratégia de recuperação leva em conta a probabilidade de relevância e de não-relevância dos termos para com as consultas, estimadas pelo modelo probabilístico estendido e pelo modelo probabilístico exponencial estendido. Foi proposto um conjunto de expressões para possibilitar a classificação dos documentos recuperados. Os resultados experimentais, apresentados no capítulo 6, comprovam a eficácia dessas estratégias.

7.3. TRABALHOS FUTUROS

Foram identificados alguns trabalhos futuros que seriam importantes para aperfeiçoar os recursos utilizados na recuperação de documentos. São eles:

- Definir uma interface gráfica para possibilitar ao usuário uma melhor análise dos resultados.
- Incorporar expressões de busca negativas no sistema, utilizando o operador NOT.
- Incorporar a possibilidade de realizar buscas com interpretação de termos técnicos.
- Comparar os resultados dos experimentos realizados neste trabalho com outros modelos baseados no modelo probabilístico.
- Verificar a viabilidade do uso combinado do modelo probabilístico estendido com outros modelos de recuperação de informação.
- Realização de testes com outras bases de dados.

APÊNDICE A

ALGORITMOS UTILIZADOS

1. CONSIDERAÇÕES INICIAIS

Neste apêndice são descritos os principais algoritmos utilizados para o tratamento dos documentos e para a recuperação destes.

2. ALGORITMO DO MÓDULO DE TRATAMENTO DE DOCUMENTOS

Nesta seção são apresentados os algoritmos responsáveis pela realização do tratamento dos documentos e termos de indexação.

O algoritmo 1 apresenta o primeiro algoritmo, que realiza o cálculo da probabilidade de relevância inicial do termo de acordo com o modelo probabilístico estendido.

ALGORITMO 1 – ALGORITMO PARA CÁLCULO DA PROBABILIDADE DE RELEVÂNCIA DO TERMO PELO MODELO PROBABILÍSTICO ESTENDIDO

```
1: N = número de documentos da coleção
2: para todo termo ki da coleção faça
3:     ni = número de documentos que possuem o termo ki
4:     para cada termo ki de cada documento dj faça
5:         P(ki|+Rq) = 0,5
6:         P(ki|-Rq) = ni / N
7:         weight = similaridade de acordo com a expressão 5.3
8:         armazena weight
9:     fim para
10:fim para
```

O algoritmo 2 apresenta o algoritmo que realiza o cálculo da probabilidade de relevância inicial do termo de acordo com o modelo probabilístico exponencial estendido.

ALGORITMO 2 – ALGORITMO PARA CÁLCULO DA PROBABILIDADE DE RELEVÂNCIA DO TERMO PELO MODELO PROBABILÍSTICO EXPONENCIAL ESTENDIDO

1: N = número de documentos da coleção
2: **para todo** termo k_i da coleção **faça**
3: n_i = número de documentos que possuem o termo k_i
4: **para cada** termo k_i de cada documento dj **faça**
5: *dt* = freqüência do termo no documento
6: ℓ = tamanho do documento
7: $P(k_i|+R_q) = (0,5)^{dt}$
8: $P(k_i|-R_q) = (n_i / N)^{\ell - dt}$
9: weight = similaridade de acordo com a expressão 5.3
10: armazena weight
11: f**im para**
12:**fim para**

1. ALGORITMO DO MÓDULO DE RECUPERAÇÃO DE DOCUMENTOS

Nesta seção são apresentados os algoritmos responsáveis pela realização da recuperação dos documentos e da realimentação de relevância (*feedback relevance*).

A recuperação dos documentos é baseada no algoritmo 3.

ALGORITMO 3 : ESTRATÉGIA DE BUSCA PROBABILÍSTICA COMBINADA COM O MODELO VETORIAL

1: entrada: $q = \{t_1, t_2, ..., t_k\}$
2: saída: conjunto de documentos ordenados de acordo com a probabilidade de relevância
3: **para todo** termo t_k pertencente a *q* **faça**
4: submeter o termo t_k ao processo de normalização morfológica
5: **fim para**
6: **para** o conjunto *q* normalizado **faça**
7: *DocumentoRecuperado* = resultado da busca no banco de dados dos documentos que possuam o conjunto *q* entre seus termos
8: apresentar informações do(s) documentos(s) ao usuário
9: **fim para**
10: entrada: *DocumentoRelevante* = $\{cr_1, cr_2, ..., cr_n\}$
11: *RealimRelevancia* = resultado do cálculo de realimentação de relevância de cada documento
12: *TermoRec* = união ($\cup$) dos termos *t* de cada documento do conjunto *DocumentoRelevante*
13: **fim para**
14: **para todo** elemento de *TermoRec* **faça**
15: wt_k = resultado do cálculo do peso de acordo com o modelo vetorial

16: *MatrizSimTermo* = resultado do cálculo da similaridade entre o termo t_k e os demais termos (t_i) do conjunto de termos *TermoRec*
17: **se** *sim* (t_k, t_i) > similaridade *n* **então**
18: *DocRecSim* = resultado da busca no banco de dados dos documentos que possuem o termo t_i
19: **fim se**
20: **fim para**
21: *ConjuntoRelevante* = *ComponenteRelevante* ∪ *CompRecSim*
22: **para todo** elemento de *ConjuntoRelevante* **faça**
23: *RealimRelevancia* = resultado do cálculo de realimentação de relevância de cada documento
24: **fim para**
25: *CR* = conjunto de componentes de *RealimRelevancia*, ordenados pelo modelo probabilístico
26: **para todo** componente cj de CR **faça**
27: localizar no banco de dados as informações gerais estruturais
28: apresentar informações do componente para o usuário
29: **fim para**

O próximo algoritmo é utilizado para calcular a probabilidade de relevância do documento em relação a uma consulta para o modelo probabilístico estendido.

ALGORITMO 4 – ALGORITMO PARA CÁLCULO DA PROBABILIDADE DE RELEVÂNCIA DE ACORDO COM O MODELO PROBABILISTICO ESTENDIDO

1: N = número de documentos da coleção
2: V = número de documentos inicialmente recuperados
3: **para todo** termo k_i da coleção **faça**
4: n_i = número de documentos que possuem o termo k_i
5: Vi = número de documentos inicialmente recuperados que possuem o termo k_i
6: **para cada** termo k_i de cada documento dj **faça**
7: $P(k_i|+R_q) = Vi / V$
8: $P(k_i|-R_q) = (ni - Vi) / (N - V)$
9: weight = similaridade de acordo com a expressão 5.3
10: armazena weight
11: **fim para**
12:**fim para**

Após a apresentação dos documentos inicialmente recuperados o usuário escolhe os documentos considerados relevantes (algoritmo 3, linha 10). Após essa primeira seleção, os documentos selecionados são submetidos à realimentação de relevância. De modo recursivo, os demais documentos

também são submetidos à realimentação de relevância, possibilitando uma melhor classificação dos documentos recuperados. Os algoritmos da realimentação de relevância para o modelo probabilístico estendido e para o modelo probabilístico exponencial estendido são apresentados abaixo.

ALGORITMO 5: ALGORITMO PARA CALCULAR A SIMILARIDADE DO DOCUMENTOS COM A CONSULTA

1: **para cada** documento da coleção **faça**
2: **se** o documento d_j *possui o* termo k_i **faça**
3: Similaridade do documento = Somatório das similaridades dos
4: termos k_i presentes no documento
5: **fim-se**
6: **fim-para**

ALGORITMO 6: REALIMENTAÇÃO DE RELEVÂNCIA PELO MODELO PROBABILÍSTICO EXPONENCIAL

1: N = número de documentos da coleção
2: V = número de documentos inicialmente recuperados
3: **para todo** termo k_i da coleção **faça**
4: n_i = número de documentos que possuem o termo k_i
5: Vi = número de documentos inicialmente recuperados que possuem o termo k_i
6: **para cada** termo k_i de cada documento dj **faça**
7: $P(k_i|+R_q) = (Vi / V)^{dt}$
8: $P(k_i|-R_q) = ((ni - Vi) / (N - V))^{\ell - dt}$
9: weight = similaridade de acordo com a expressão 5.3
10: armazena weight
11: **fim para**
12:**fim para**

Após a realimentação de relevância, os documentos são novamente apresentados ao usuário devidamente re-classificados em ordem decrescente de probabilidade de relevância.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLAN, J. **Challenges of Information Retrieval and Language Modeling**. Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, pp. 31-47, 2002.

AMATI, G.; VAN RIJSBERGEN, C. **Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness**. ACM Transactions on Information Systems, Vol. 20, No. 4, pp. 357-389, 2002.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval,** Addison- Wesley, 1999.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. URL: http://sunsite.dcc.uchile.cl/irbook/ - Consultado em 30/06/2004.

COOPER, W. S. **Some inconsistencies and misnomers in probabilistic Information retrieval**. Proceedings of ACM, Vol 13, No. 1, pp. 100-111, 1995.

CORREA, A. C. G. **Recuperação de Documentos baseada em Informação Semântica no Ambiente AMMO**. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, DC-UFSCar, São Carlos, Brasil, Agosto de 2003.

CRESTANI, F., LALMAS, M., VAN RIJSBERGEN, C. J., CAMPBELL I. **"Is This Document Relevant?… Probably": A Survey of Probabilistic Models in Information Retrieval**. ACM Computing Surveys, Vol. 30, No. 4, pp. 528-552, December 1998

CROFT, B. CALLAN, J. LAFFERTY, J. **Language Modeling and Information Retrieval**. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, pp. 98-112, 2001.

FUHR, N, PFEIFER, U. **Probabilistic Information Retrieval as a Combination of Abstraction, Inductive, Learning, and Probabilistic Assumptions**. ACM Transactions on Information Systems, Vol. 12, No, 1, pp. 92-115, 1994.

FUHR, N. **Two Models of Retrieval with Probabilistic Indexing**. ACM Conference on Research and Development in Information Retrieval, pp. 249-257, 1986.

GETOOR, L.; FRIEDMAN, N.; KOLLER, D.; TASKAR, B. **Learning Probabilistic Models of Link Structure**. Jornal of Machine Learning Research 3, pp. 679-707, 2002.

GEY, F. C. **Infering Probability of Relevance Using the Method of Logistic Regression**. UC Data Archive and Technical Assistance, University od California, Berkeley USA, pp. 222-231, 1994.

GILDEA, D. **Probabilistic Models of Verb-Argument Structure**. University of Pennsylvania, USA, pp. 1-7, 2001.

GREIFF, W. R.; PONTE, J. M. **The Maximum Entropy Approach and Probabilistic IR Models**. ACM Transactions on Information Systems, Vol. 18, No. 3, pp. 211-228, 2000.

GREIFF, W. R.; PONTE, J. M.; MORGAN, W. T. **The Rule of Variance in Term Weighting for Probabilistic Information Retrieval**. ACM CIKM´02, McLean, Virginia, USA, pp-252-259, 2002.

HEARST, M.A.; PEDERSEN, J.O. **Reexamining the cluster hypothesis**. In Proceedings of SIGIR 96, pp. 76-84, 1996.

JIN, R.; HAUPTMANN, A.G. **A New Probabilistic Model for Title Generation**. Carnegie Mellon University, Pittsburgh, USA, pp. 1-7, 2001.

KRUEV, V. **Compiling Document Collections from the Internet**. ACM, University of Aizu, Japan, pp. 9-14, 2000.

MACEDO, A. A. **Especificação, instanciação e experimentação de um arcabouço para criação automática de ligações hipertexto entre informações homogêneas**. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação – ICMC-USP, São Carlos, Brasil, Maio de 2004.

MEDLINE. **Cystic Fibrosys Database** . URL: http://sunsite.dcc.uchile.cl/irbook/cfc - Consultado em 15/10/2005.

MELLO, C. A. S. **Proposta de um Método para a Recuperação de Componentes utilizando Técnicas de Agrupamento**. Dissertação de Mestrado, Programa de Pós-Graduação em Ciência da Computação, DC-UFSCar, São Carlos, Brasil, Julho de 2005.

MICROSYSTEMS. S. (2000b). **Reference API specifications**. URL: http://java.sun.com/reference/api.index.html - Consultado em 13/08/2005.

PAVLOV, D.;SMYTH, P. **Probabilistic Query Models for Transaction Data**. ACM KDD´01 San Francisco, CA, USA, pp. 164-173, 2001

PEARL, J. **Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference**. Morgan Kaufmann Publishers, Inc., pp. 306-313, 1988.

POSTGRESQL. **PostgreSQL Core Distribution**. URL: http://www.postgresql.org/download/ - Consultado em 13/03/2005.

POSTGRESQL. **PostgreSQL Trac**. URL: http://www.postgresql.org.br/ - Consultado em 13/03/2005.

PRESSMANN, R. S. **Engenharia de Software**, Makron Books do Brasil, 1995.

RIBEIRO, B. A. N., MUNTZ, R. **A belief network models for IR**. Proc. Of the 19th ACM SIGIR Conference, Zurich, Switzerland, pp. 253-260, 1996.

ROBERTSON, S. **On Theoretical Argument in Information Retrieval**. ACM SIGIR, pp. 1-10, 2000.

ROBERTSON, S., VAN RIJSBERGEN, C. J., PORTER, M. F. **Probabilistic models of indexing and searching**. ACM, pp. 35-56, 1980

SALTON, G. **Automatic Information Organization and Retrieval**. Computer Science Series, USA: McGraw-Hill, 1968.

SALTON, G.; McGILL, M. J. **Introduction to Modern Information Retrieval**. Computer Science Series, USA: McGraw-Hill, 1983.

SALTON, G. **Recent Trends in Automatic Information Retrieval**. Proc. Of Conf. ACM, pp. 1-9, 1986.

SCHITZE, H.; SILVERSTEIN, C. **Projection for efficient document clustering**. Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 1-9, 1997.

SHAW, W.M. & WOOD, J.B. & WOOD, R.E. & TIBBO, H.R. **The Cystic Fibrosis Database: Content and Research Opportunities**. LISR 13, pp. 347-366, 1991.

SILVA, I. R. **Um Estudo de Desempenho em Recuperação de Informação: Modelos, Consultas e Índices**. SPG'98 - II Semana de Pós-Graduação em Ciência da Computação – DCC – UFMG, pp. 1-12, 1998.

SILVA, I. R. **Redes para Sistemas de Recuperação de Informação**. SPG'99 - III Semana de Pós-Graduação em Ciência da Computação – DCC – UFMG, pp. 1-13, 1999.

TEEVAN, J., KARGER, D. R. **Empirical Development of an Exponential Probabilistic Model for Text Retrieval**. Proc. Of Int. Conf. ACM SIGIR, Toronto, Canada, pp. 18-25, 2003

TURTLE, H.; CROFT, W. B. **Evaluation of an inference network-based retrieval model**. ACM Transactions on Information Systems, pp. 187-222, 1991.

VAN RIJSBERGEN, C. J. **Information Retrieval** (Second Ed.) Butterworths, London, 1979.

YANAI, K; IBA, H. **Probabilistic Distribution Models for EDA-based GP**. ACM GECCO´05, Washington, DC, USA, pp. 1775-1776, 2005.

ZHAI, C. X. **Risk Minimization and Language Modeling in Text Retrieval**. SIGIR Forum, Vol. 36, No. 2, pp. 100-111, 2002.