Iniciação Científica CESUMAR
Jul./Dez. 2009, v. 11, n. 2, p. 149-157

RECUPERAÇÃO DE DOCUMENTOS-TEXTO USANDO MODELOS PROBABILISTICOS ESTENDIDOS

Marcello Erick Bonfim*

RESUMO: Neste artigo são apresentadas estratégias utilizadas para a recuperação de informação, com base no modelo probabilístico de recuperação de informação. Nessas estratégias adotaram-se o modelo probabilístico e o modelo 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; Modelo Probabilístico Exponencial Estendido.

RECOVERY OF DOCUMENTS USING EXTENDED TEXT-PROBABILISTIC MODELS

ABSTRACT: This article presents strategies used for retrieval information, based on probabilistic model of information retrieval. These strategies have been adopted the probabilistic model and exponential probabilistic model and were combined with vector features, now are called probabilistic model of extended and stretched exponential probabilistic model. The information retrieval community values of the probability of relevance and non-relevance for the classification of the resulting documents. We present experiment results showing that the combination of probabilistic models with the vector model allows a more efficient recovery, bringing in response to relevant documents that would not be recovered using only one of the models.

KEYWORDS: Information Retrieval; Extended Probabilistic Model; Extended Exponential Probabilistic Model

INTRODUÇÃO

A ampla variedade e quantidade de informações recuperadas e 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.

Conforme citado em Macedo (2004), no final da década de 60 surgiram os primeiros catálogos bibliográficos *on-line* que permitiam a recuperação de informação armazenada em alguns minutos. O usuário manuseava as informações através de um ambiente de consulta utilizando um conjunto controlado de operações e

* Docente do Departamento de Informática do Centro Universitário de Maringá – CESUMAR. E-mail: mebonfim@yahoo.com.br

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.

O objetivo deste artigo é apresentar uma estratégia para a extração automática de informação de documentos-texto, visando sua recuperação com base no conteúdo de tais documentos. Essa recuperação utiliza uma estratégia de recuperação de documentos de acordo com o modelo probabilístico combinado com recursos do modelo vetorial de recuperação de informação, aqui denominado de modelo probabilístico estendido, e uma estratégia de recuperação de informação de documentos de acordo com o modelo probabilístico exponencial combinado com o modelo vetorial, aqui denominado de modelo probabilístico exponencial estendido. O sistema aqui proposto realiza a extração automática de informação dos documentos, armazenando-as em um repositório de dados. A recuperação de documentos é realizada através de expressões de consulta utilizando essas informações. A seguir são apresentados os trabalhos relacionados, os conceitos sobre recuperação de informação, o modelo probabilístico estendido, o sistema para manipulação de documentos, os experimentos realizados e as conclusões e trabalhos futuros.

2 TRABALHOS RELACIONADOS

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 semiprobabilísticos e 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). 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. Outra abordagem do modelo probabilístico é a que utiliza a frequência dos termos nas estimativas das probabilidades, essa abordagem foi trabalhada por Amati e Van Rijsbergen (2002) e Greiff, Ponte e Morgan (2002).

Outro conceito para aplicação do modelo probabilístico é o modelo probabilístico exponencial, que considera a frequência 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, Callan e Lafferty (2001) e Fuhr (1986), o modelo probabilístico e suas variações são muito eficazes para sistemas de recuperação de informação, mesmo tendo suas raízes na literatura há muitos anos, devido ao sucesso na classificação de documentos.

3 RECUPERAÇÃO DE INFORMAÇÃO

Os modelos de recuperação de informação consistem em 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. A seguir são apresentados os modelos utilizados nesse trabalho para a recuperação de informação.

3.1 MODELO VETORIAL

O modelo vetorial também é chamado de modelo espaço vetorial e representa cada documento como 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 como 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. Os métodos de cálculo se baseiam no número de ocorrências do termo no documento (frequência) (BAEZA-YATES; RIBEIRO-NETO, 1999). Quanto à frequência de um termo num documento tem-se como definido que em um número total N de documentos são selecionados os ni documentos em que o termo de indexação aparece; a frequência é o número de vezes que o termo mencionado aparece no texto do documento selecionado. Segundo Baeza-Yates e Ribeiro-Neto (1999), a melhor fórmula para calcular o peso do termo é dada por:

$$W_{i,j} = F_{i,j} \; X \log \frac{N}{n_i} \qquad (3.1)$$

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.

Expressão para cálculo da similaridade:

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

3.2 MODELO PROBABILÍSTICO

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-Yates e Ribeiro-Neto (1999) definem o modelo probabilístico da seguinte maneira: para o modelo probabilístico, o peso do termo de indexação para uma consulta é representado por wi,q e o peso do termo para o documento é representado por wi,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 dj seja relevante para a consulta q, e $P(–R_q | d_j)$ é a probabilidade de que um documento dj seja não-relevante para a consulta q.

Um documento dj 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 Wd/q, como sendo:

$$W_{d/q} = sim(d_j, q) = \frac{P(+R_q | d_j)}{P(–Rq | dj)} \quad (3.3)$$

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

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i|+R_q)}{1 - P(k_i|+Rq)} + \log \frac{1 - P(k_i|-R_q)}{P(k_i|-R_q)}\right) \quad (3.4)$$

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; VAN RIJSBERGEN; PORTER, 1980).

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

1 – P(ki|+Rq) é constante para todo termo de indexação ki 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:

$$P(k_i|+R_q) = 0,5 \quad (3.5)$$

$$P(k_i|-R_q) = n_i / N \quad (3.6)$$

Onde n_i é o número de documentos que contêm o termo de indexação k_i e N é o número total de documentos da coleção. Dada essa hipótese, podem-se recuperar documentos que contêm termos da consulta e promover 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 ki. V e Vi 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 ki 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, podem-se calcular as probabilidades de relevância e não-relevância:

$$P(k_i|+R_q) = 0,5 \quad (3.7)$$

$$P(k_i|-R_q) = n_i / N \quad (3.8)$$

Esse processo pode ser repetido recursivamente. Assim, podem-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 ideia 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 Vi = 0. Para evitar esse problema é somado um fator de ajuste, resultando em:

$$P(k_i|+R_q) = (V_i+0,5) / (V+1) \quad (3.9)$$

$$P(k_i|-R_q) = (n_i – V_i + 0,5) / (N – V + 1) \quad (3.10)$$

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

$$P(k_i|+R_q) = V_i / V \quad (3.11)$$

$$P(k_i|-R_q) = (n_i – V_i) / (N – V) \quad (3.12)$$

Utilizando as expressões apresentadas é possível estimar as probabilidades de relevância e não-relevância para um conjunto de documentos.

3.2.1 O Modelo Probabilístico Exponencial

O modelo probabilístico exponencial, proposto por Teevan e Karger (2003), considera a frequência 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 frequência 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 frequência deste no documento dt como função exponencial para obter o resultado. A probabilidade de não-relevância utiliza o tamanho ℓ do documento subtraído da frequência do termo em questão como função exponencial para obter o resultado. Assim, a probabilidade inicial será

$$P(k_i|+R_q) = (V_i+0,5) / (V+1) \quad (3.13)$$

$$P(k_i|-R_q) = (n_i - V_i + 0,5) / (N - V + 1) \quad (3.14)$$

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 Vi um subconjunto de V, composto de documentos que contenham termos de indexação k_i. Para se melhorar as hipóteses probabilísticas é utilizado esse subconjunto para recalcular as probabilidades.

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

$$P(k_i|+R_q) = (V_i+ n_i/N) / (V+1) \quad (3.15)$$

$$P(k_i|-R_q) = (n_i - V_i + n_i/N) / (N - V + 1) \quad (3.16)$$

Obtidos os valores das probabilidades de relevância e não-relevância de cada termo em um documento, aplica-se a expressão 3.4 para estimar a similaridade do documento em relação à consulta.

Esse modelo possibilita uma melhor classificação dos resultados, pois considera a frequência do termo em cada documento para estimar as probabilidades. Como desvantagem, o modelo pode não ser tão eficiente se os documentos da coleção foram 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.

3.2.2 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 examinálos, ele marca quais são relevantes. Segundo Salton e McGill (1983), na prática só os 10 documentos melhores classificados são examinados. A ideia principal consiste em selecionar termos importantes, ou expressões, 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 ESTRATÉGIAS PARA RECUPERAÇÃO DE INFORMAÇÃO

4.1 ESTRATÉGIA DE BUSCA PARA O MODELO PROBABILÍSTICO ESTENDIDO

Dada uma consulta q = {t1, t2, t3,..tk}, 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 tem-se 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 3.4.

Após 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. Esses novos documentos recuperados são submetidos aos cálculos de probabilidade e classificados pelo modelo probabilístico. 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.

4.1 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 frequência 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. 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.

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.

4.3 EXPERIMENTOS UTILIZANDO O MODELO PROBABILÍSTICO ESTENDIDO

Foram realizados experimentos com o objetivo de avaliar a estratégia proposta neste artigo. O conjunto de documentos utilizado nesses experimentos é o MEDLINE (SHAW et al., 1991). 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).

$$P = (tir / tr) \quad (4.1)$$

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).

$$R = (tirr / ta) \qquad (4.2)$$

O conjunto de documentos MEDLINE[1] 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. Foram utilizadas 30 consultas, baseadas nas 100 consultas sugeridas para este conjunto por Shaw e colaboradores (1991). Na primeira etapa, os componentes 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-YATES; RIBEIRO-NETO, 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, na prática é de 10 documentos (SALTON; McGILL, 1983). Em seguida, o usuário seleciona quais os documentos são inicialmente considerados relevantes 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 é assim definido para que os documentos recuperados sejam os mais similares. Se esse número de termos for maior, serão recuperados um grande número de documentos, o que poderá 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 traz 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 apresentados, a precisão para uma busca poderá ser muito baixa.

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 e 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, extraídos das expressões de consultas formuladas em linguagem natural, escolhidas entre as 100 consultas disponibilizadas pelo MEDLINE.

4.4 AVALIAÇÃO DA ABORDAGEM APLICADA

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

[1] Disponível em: <http://sunsite.dcc.uchile.cl/irbook/cfc>. Acesso em: 15 out. 2005.

nas informações contidas na base de dados fornecida por Shaw et al. (1991). Na tabela a seguir, os campos Pest e Rest significam, respectivamente, a precisão e revocação da estratégia de busca probabilística estendida. Os campos Ppro e Rpro significam, respectivamente, a precisão e revocação da estratégia de busca probabilística.

Analisando os resultados da tabela 1 observa-se que a média percentual de precisão (precision) foi de 20,38%, e a revocação (recall) foi de 39,65% para o modelo probabilístico estendido, 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 compromete a extração de termos de indexação representativos. Para o modelo probabilístico os resultados foram menos satisfatórios. 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 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.

O conjunto de documentos MEDLINE 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 a 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 e colaboradores (1991). Os documentos não recuperados não possuíam os termos envolvidos na busca.

Tabela 1. 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%	26,19%	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%**

4.4.1 Comparação com outros modelos

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 sugerida 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).

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-YATES; RIBEIRO-NETO, 1999) e também a probabilidade de relevância e de não-relevância de acordo com o modelo probabilístico exponencial (TEEVAN; 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.

Para avaliar a abordagem foram realizadas 30 consultas. As consultas foram elaboradas com termos de indexação (ki) presentes nos documentos dj, de acordo com a proposta de Mello (2005). Os campos Ppro e Rpro significam, respectivamente, a precisão e revocação da estratégia de busca probabilística estendida, os campos Pexp e Rexp significam a precisão e revocação da estratégia de busca probabilística exponencial estendida, os campos Pvet e Rvet indicam a precisão e revocação da estratégia de busca vetorial convencional e os campos Pagr e Rarg indicam a precisão e revocação da estratégia de busca vetorial utilizando agrupamentos.

Analisando os resultados da abordagem probabilística estendida observa-se que a mdia 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), observa-se 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 2).

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 estendido 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.

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 estendida. Isso se dá justamente pela melhor classificação dos documentos recuperados. No modelo probabilístico exponencial estendido o tamanho do documento (número de termos que este possui) e a frequência de cada termo são de fundamental importância, pois são considerados durante os cálculos das probabilidades.

Tabela 2. Comparação entre os modelos de recuperação de informação

Consulta	Pvet	Pagr	Pexp	Ppro	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%**

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem para recuperação de documentos de acordo com modelos probabilísticos de recuperação de informação, combinado com o modelo vetorial. Esse projeto teve como objetivo a pesquisa de técnicas, métodos e ferramentas que visam à 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.

A maior contribuição deste trabalho é a estratégia adotada para a recuperação de documentos. Para validar a ideia 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 comprovam a eficácia dessas estratégias.

Foram identificados alguns trabalhos futuros que seriam importantes para aperfeiçoar os recursos utilizados na recuperação de documentos tais como: definir uma interface gráfica para possibilitar ao usuário uma melhor análise dos resultados; incorporar expressões de busca negativas no sistema, utilizando o operador NOT; comparar os resultados dos experimentos realizados neste trabalho com outros modelos baseados no modelo probabilístico; verificar a viabilidade do uso do modelo probabilístico estendido com outros modelos de recuperação de informação.

REFERÊNCIAS

ALLAN, J. **Challenges of Information Retrieval and Language Modeling**. Massachusetts: University of Massachusetts Amherst, 2002.

AMATI, G.; VAN RIJSBERGEN, C. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. **ACM Transactions on Information Systems**, v. 20, n. 4, p. 357-389, 2002.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. [S. l.]: Addison- Wesley, 1999.

CROFT, B.; CALLAN, J.; LAFFERTY, J. **Language Modeling and Information Retrieval**. Pittsburgh, Pennsylvania, USA: Carnegie Mellon University, 2001.

FUHR, N. **Two Models of Retrieval with Probabilistic Indexing**. [S. l.]: ACM Conference on Research and Development in Information Retrieval, 1986.

GILDEA, D. **Probabilistic Models of Verb-Argument Structure**. [S. l.]: University of Pennsylvania, USA, 2001.

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

KRUEV, V. **Compiling Document Collections from the Internet**. Aizu, Japan: ACM, University of Aizu, Japan, 2000.

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

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

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

ROBERTSON, S. **On Theoretical Argument in Information Retrieval**. [S. l.]: ACM SIGIR, 2000.

ROBERTSON, S.; VAN RIJSBERGEN, C. J.; PORTER, M. F. **Probabilistic models of indexing and searching**. [S. l.]: ACM, 1980.

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

SHAW, W. M. et al. The Cystic Fibrosis Database: Content and Research Opportunities. **Library and Information Science Research**, v. 13, n. 4, p. 347-366, Oct./Dec. 1991.

TEEVAN, J.; KARGER, D. R. Empirical Development of an Exponential Probabilistic Model for Text Retrieval. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. 26, 2003, Toronto. **Anais...** Toronto, Canadá: ACM SIGIR, 2003. P. 18-25.

Recebido em: 30 Junho 2009
Aceito em: 09 Novembro 2008