ISBN 978-85-61091-05-7

VI EPCC
Encontro Internacional de Produção Científica Cesumar
27 a 30 de outubro de 2009

RECUPERAÇÃO DE DOCUMENTOS TEXTO USANDO MODELOS PROBABILISTICOS ESTENDIDOS

Marcello Erick Bonfim[1]

RESUMO: Neste trabalho são apresentadas estratégias utilizadas para a recuperação de informação, com base no modelo probabilístico 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, Modelo Probabilístico Exponencial Estendido.

1 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 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.

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

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

[1] Docente do CESUMAR. Departamento de Informática do Centro Universitário de Maringá – CESUMAR, Maringá-PR. mebonfim@yahoo.com.br

consulta. Os métodos de cálculo se baseiam no número de ocorrências do termo no documento (freqüência). (BAEZA e RIBEIRO, 1999). Quanto à freqüência de um termo num documento tem-se como definição que em um número total *N* de documentos são selecionados os *ni* 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. O resultado da busca é um conjunto de documentos ordenados pelo grau de similaridade entre cada documento e a consulta.

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 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*. $P(+R_q \mid d_j)$ é a probabilidade de que um documento d_j seja relevante para a consulta *q*, e $P(-R_q \mid d_j)$ é a probabilidade de que um documento d_j seja não-relevante para a consulta *q*.

Um documento d_j é relevante a uma consulta q quando: $P(+R_q \mid d_j) > P(-R_q \mid d_j)$. Assim, dada uma consulta *q*, o modelo probabilístico atribui a cada documento *d* (como medida de similaridade) um peso $W_{d/q}$. Segundo Baeza e Ribeiro (1999), sabendo que $P(k_i|+R_q) + P(-k_i|+R_q) = 1$, após transformações algébricas pode-se escrever

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

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

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.

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.

Após a classificação inicial, o modelo trabalha de maneira similar ao modelo probabilístico clássico. Obtidos os valores das probabilidades de relevância e não-relevância de cada termo em um documento, aplica-se a expressão 1 para estimar a similaridade do documento em relação à consulta. 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.

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, 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.

2 MATERIAL E MÉTODOS

Pela estratégia de busca utilizando o modelo probabilístico estendido, 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 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.

A Estratégia de busca para o modelo probabilístico exponencial estendido 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. 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. A última etapa é a apresentação final dos resultados ao usuário.

3 RESULTADOS E DISCUSSÃO

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). Os resultados foram submetidos às métricas de precisão (*precision*) e revocação (*recall*). 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*).

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*).

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. Foram utilizadas 30 consultas, baseadas nas 100 consultas sugeridas para este conjunto por Shaw *et al.* (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 e RIBEIRO, 1999). Essas informações também foram armazenadas no banco de dados.

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.*

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). Analisando os resultados 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.

Também foram realizados experimentos 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). Foram definidas 30 consultas em um conjunto de 100 componentes da biblioteca Java API. Analisando os resultados da abordagem probabilística estendida 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.

4 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. 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 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 comprovam a eficácia dessas estratégias.

REFERÊNCIAS

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

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.

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.

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

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.

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