Reordenação de resultados de busca na Web conforme critério de relevância definido pelo usuário

João C. Medau Maria Cristina R. Belderrain Fabrício J. Barth
Faculdades Tancredo Neves
Av. Divino Salvador 876, Moema, São Paulo - SP
{*jmedau,mbelderrain,fbarth*}*@tancredo.br*

Abstract

This paper shows how the results produced by a search engine like Google can be ordered according to a user profile. The profile is built from any number of documents provided by the user. All unique words are then extracted from the documents for further statistic comparison with the search results. The availability of the user profile allows the results to be shown in such a way that the user will see the most relevant ones first.

1 Introdução e Objetivo

As ferramentas de busca disponíveis atualmente permitem que o usuário obtenha informações com relativa rapidez e facilidade. No entanto, alguns problemas ainda devem ser solucionados para que ferramentas mais eficientes venham a produzir informações mais relevantes para o usuário [1].

Entre os principais problemas estão os seguintes: interfaces inadequadas acabam induzindo o usuário, que normalmente não domina as linguagens de consulta, a gastar um tempo considerável apenas para formular a consulta que deseja fazer; em geral, verifica-se um desvio significativo entre a ordenação dos resultados (*ranking*) produzida pelas ferramentas de busca e aquela que o usuário desejaria, e; frequentemente, centenas de resultados são retornados pelas ferramentas, mas apenas alguns são relevantes para o usuário.

A solução destes problemas implica na obtenção de resultados ordenados de acordo com as necessidades do usuário, de modo que os documentos mais relevantes para ele apareçam no topo da lista [1]. Uma vez que as abordagens convencionais de recuperação da informação não têm esse tipo de personalização como meta, outras linhas de ação devem ser exploradas. Uma alternativa promissora é aquela que pressupõe a existência de um perfil do usuário. Assim, se as preferências do usuário, representadas no perfil, passam a ser conhecidas, então os resultados poderão ser ordenados de acordo com elas.

Neste contexto, o objetivo do trabalho é propor uma metodologia e implementar, a partir dela, uma ferramenta que permita uma reordenação dos resultados consistente com as necessidades do usuário.

Este trabalho está estruturado da seguinte forma: na seção 2 são descritos os conceitos e algoritmos utilizados na solução do problema; na seção 3 é descrita a implementação de um software que utiliza os conceitos e algoritmos descritos na seção 2; na seção 4 são apresentados os resultados alcançados, e; na seção 5 são apresentadas as considerações finais e trabalhos futuros.

2 Conceitos e Algoritmos

Formalmente, dado um conjunto de documentos $D = (d_1, \cdots, dn)$ recuperados através de uma consulta X, e um conjunto de usuários S, deve-se ordenar os elementos em D levando-se em consideração uma função de utilidade: $\forall d_i \in D \ \mu(d_i, s)$, onde $\mu(d, s)$ é uma função que mede a utilidade de um item $d \in D$ para um usuário $s \in S$.

A implementação da função de utilidade $\mu(d, s)$ pode ser realizada de diversas maneiras. Entre elas, a que se adapta melhor ao domínio de recuperação da informação é a abordagem baseada em conteúdo. Esta abordagem utiliza como entrada uma descrição do conteúdo dos itens e do perfil do usuário. O conteúdo de cada item é então comparado com as preferências do usuário, descritas em seu perfil [2]. Por exemplo, a utilidade $\mu(d, s)$ do item d para o usuário s é estimada com base na utilidade $\mu(d_i, s)$ atribuída pelo usuário s aos itens d_i pertencente a D similares ao item d.

Neste trabalho, o perfil do usuário é representado por um vetor $\overrightarrow{s_j}$ com as probabilidades ($p(w_k \mid$

$Relevante$)) que refletem o quanto uma determinada palavra w_k é relevante para o usuário s_j.

Dado um conjunto de documentos DR onde os itens são considerados relevantes pelo usuário s_j, deve-se criar o vetor $\overrightarrow{s_j}$ através do algoritmo apresentado na figura 1[1].

CRIA_PERFIL_INICIAL(DR)

Vocabulário ← conjunto de todas as palavras distintas que ocorrem em qualquer documento em DR
$Text_j$ ← um único documento criado pela concatenação de *DR*
n ← número de palavras do documento $Text_j$
for cada palavra w_k em Vocabulário **do**
  n_k ← número de vezes que a palavra w_k aparece em $Text_j$
  $P(w_k \mid Relevante) \leftarrow \frac{n_k+1}{n+|Vocubalario|}$
end for

Figura 1: Algoritmo para criação do perfil inicial

Dado o perfil do usuário s_j, representado pelo vetor $\overrightarrow{s_j}$, e dado o conjunto de documentos D recuperado por um mecanismo de busca tradicional[2], o algoritmo apresentado na figura 2 terá que reordenar o conjunto D.

FILTRAGEM($\overrightarrow{s_j}$,D):D'

Entrada: perfil do usuário ($\overrightarrow{s_j}$)
Entrada: conjunto de documentos (D)
Saída: conjunto de documentos reordenado (D')

for cada documento d_i em D **do**
  w_i ← conjunção de todas as palavras distintas que ocorrem em d_i ($w_{1,i} \wedge w_{2,i} \wedge \cdots \wedge w_{n,i}$)
  $P(Relevante \mid w_i) = \prod_{j=1}^{n} P(w_{j,i} \mid Relevante)$
end for
D' = conjunto ordenado segundo $P(Relevante \mid w_i)$

Figura 2: Algoritmo para reordenação dos documentos

O cálculo da relevância de um determinado documento d_i para um usuário s usa a fórmula de *naive bayes* (bayes ingênuo). A fórmula é assim chamada porque assume que a ocorrência dos eventos (neste caso, as palavras) são independentes. Apesar desta simplificação, a fórmula de *naive bayes* funciona de maneira adequada para o problema de classificação de documentos [4].

Refletindo a metodologia aqui proposta, a ferramenta implementada recalcula a relevância de cada documento retornado pelo serviço de buscas Google em resposta a uma consulta. Para tanto, é considerado o valor da relevância de cada palavra contida no documento. Tal valor foi previamente estabelecido quando o usuário criou seu perfil. Uma vez recalculada a relevância de cada resultado em função do perfil do usuário, este recebe uma lista com os resultados apresentados em ordem decrescente de relevância.

[1]Para calcular $P(w_k \mid Relevante)$ é utilizado a equação *m-estimate* [3]

[2]Neste caso, o *Google* - www.google.com

3 Implementação

Com o intuíto de validar a proposta aqui apresentada, foi implementada uma ferramenta de busca que faz uso dos algoritmos descritos na seção 2. A ferramenta, denominada *Faro Fino*, é um filtro que reordena os resultados de buscas realizadas por meio da Google Web APIs[3]. Em sua primeira versão, a ferramenta oferece as seguintes funcionalidades básicas: configuração de certas preferências do usuário e de aspectos dependentes de plataforma; criação de um perfil personalizado a partir de documentos selecionados pelo usuário; busca de documentos via Google a partir de uma consulta baseada em termos ou palavras-chave; recuperação e reordenação dos resultados mediante comparação estatística com o perfil do usuário, e; apresentação dos resultados reordenados.

O usuário poderá acrescentar qualquer documento ao seu perfil, desde que o arquivo a ser processado esteja em um dos formatos suportados pelo sistema (PS, PDF, DOC, XML, HTML ou texto). Sempre que um documento for acrescentado, a probabilidade de relevância de cada palavra nele contida será atualizada no perfil.

É importante notar que qualquer documento correspondente a um resultado poderá ser acrescentado ao perfil, bastando, para isso, que o usuário marque os resultados a serem acrescentados usando o *checkbox* apropriado. Esta possibilidade torna o perfil dinâmico e adaptável, em tempo real, à relevância que o usuário vai atribuindo aos resultados da busca.

Para evitar distorções decorrentes do acréscimo, ao perfil, de palavras irrelevantes (artigos, pronomes, preposições e outras), tais palavras são mantidas em um arquivo texto. Assim, ao ser acrescentado ao perfil, o documento é filtrado para evitar que qualquer uma das palavras registradas no arquivo venha a fazer parte do perfil. Estratégias mais refinadas, como a técnica conhecida por TF-IDF (*Term Frequency / Inverse Document Frequency*) [5] poderão ser utilizadas em futuras versões do sistema.

Para realizar a busca, basta que o usuário forneça um ou mais termos isolados, uma frase entre aspas, ou mesmo qualquer combinação de termos e operadores de busca reconhecida pelo Google. As URLs dos resulta-

[3]http://www.google.com/apis/

dos serão recuperadas, assim como outras informações pertinentes. O conteúdo de cada URL será, então, baixado e gravado no diretório previamente configurado pelo usuário.

Finalmente, cada um dos documentos é confrontado com o perfil de acordo com o algoritmo descrito na figura 2. Este confronto vai produzir um valor que representa a relevância de cada documento relativa àquele perfil. Tal valor é utilizado para reordenar os resultados da busca, apresentados ao usuário em ordem decrescente de relevância.

4 Avaliação Qualitativa

A avaliação objetiva deste tipo de ferramenta costuma fazer uso de dois conceitos probabilísticos: precisão e *recall* [5]. Ambos dependem de julgamentos puramente subjetivos emitidos por mais de um especialista na área de conhecimento.

Neste caso, a avaliação será conduzida mediante a aplicação de um questionário a pelo menos três usuários-especialistas em uma área de conhecimento previamente selecionada. Cada usuário será solicitado a atribuir uma nota a um número fixo de resultados, obtidos em 10 consultas sucessivas: quanto maior a nota, maior a relevância do resultado para o usuário.

A partir das notas e de alguns critérios simples que permitam particionar o conjunto de resultados em relevantes/não-relevantes e recuperados/não-recuperados, os valores de precisão e *recall* poderão ser calculados para cada um dos usuários. Naturalmente, quanto maior o número de usuários envolvidos, mais significativos serão os resultados da avaliação.

Seja qual for o resultado final da avaliação qualitativa, cabe destacar que a ferramenta foi projetada para favorecer o aprimoramento contínuo do perfil do usuário. Assim, se a escolha inicial dos documentos que compõem o perfil for criteriosa, é provável que os resultados da primeira busca se aproximem das expectativas do usuário; se, a cada nova busca, este acrescentar ao seu perfil os resultados mais relevantes, a qualidade do perfil será sempre reforçada, o que vai determinar, por sua vez, a produção de resultados cada vez melhores.

5 Considerações Finais

Este trabalho apresenta dois aspectos que o distinguem das abordagens usualmente adotadas na recuperação personalizada de informações: seu ponto de partida não foi a criação de mais um mecanismo de busca, e sim a utilização, via API, de um serviço de busca existente, e; não são os resultados já obtidos, simplesmente, que vão influenciar a apresentação dos próximos resultados, mas a relevância dos mesmos em relação ao perfil do usuário.

O modelo adotado na construção do perfil do usuário é muito simples, baseando-se, exclusivamente, na ocorrência estatística de palavras em textos. Em consequência, as limitações inerentes às abordagens baseadas em conteúdo [6, 7] aplicam-se ao modelo: embora adapte-se bem ao processamento de textos, não é um modelo facilmente transferível a outros domínios (áudio, vídeo, imagens); dois itens distintos representados pelo mesmo conjunto de atributos serão considerados iguais, ou seja, artigos bem escritos e mal escritos serão indistinguíveis, para todos os efeitos, e; os itens recomendados sempre estarão entre aqueles já vistos pelo usuário.

As próximas etapas deste trabalho incluem a aplicação do questionário, o processamento e a análise dos resultados da avaliação qualitativa, assim como a implementação de mudanças focalizadas no aumento do desempenho da ferramenta. Dependendo dos resultados da avaliação, outras modificações e extensões, ligadas à filtragem mais refinada dos documentos acrescentados ao perfil, poderão ser consideradas.

Referências

[1] C. Danilowicz and H. C. Nguyen, "Using user profiles in intelligent information retrieval," in *Foundations of Intelligent Systems. 13th International Symposium*, M.-S. Hacid, Z. W. Rás, D. A. Zighed, and Y. Kodratoff, Eds., no. LNAI 2366. Lyon, France: Springer-Verlag, June 2002, pp. 223–231.

[2] I. Koychev and I. Schwab, "Adaptation to drifting user's interests," in *Proceedings of ECML 2000 Workshop: Machine Learning in New Information Age*, 2000.

[3] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[4] S. J. Russel and P. Norvig, *Artificial intelligence: a modern approach*, 2nd ed. Prentice-Hall, 2003.

[5] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 2003.

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, June 2005.

[7] M. Montaner, B. López, and J. L. de la Rosa, "A taxonomy of recommender agents on the internet," *Artificial Intelligence Review*, vol. 19, no. 4, pp. 285–330, June 2003.