Recuperação de Imagens na Web Baseada em Informações Textuais

André Ribeiro da Silva
Mário Celso Candian Lobato

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

{arsilva,mlobato}@dcc.ufmg.br

A Internet atual é um enorme repositório de imagens. Obter uma imagem desejada diante dessa enorme quantidade de dados espalhados por milhões de sítios Web, pode se tornar uma tarefa cansativa ou até mesmo impraticável se não tivermos uma ferramenta para nos apoiar. Daí a grande importância de se desenvolver técnicas para recuperar essas imagens com eficiência, ou seja, recuperar a imagem mais próxima da desejada em uma busca suficientemente rápida. A recuperação de imagens é normalmente baseada em duas abordagens principais: recuperação baseada no conteúdo da imagem e recuperação baseada em informações textuais do documento Web.

Na recuperação baseada em conteúdo são usadas, para classificar as imagens, as suas próprias características como a cor, a forma, a textura, etc. Como consulta, os usuários providenciam uma imagem, ou uma descrição das características da imagem, que será comparada com as imagens de um banco de dados, nesse caso com as imagens disponíveis na Internet.

O uso dessa abordagem tem algumas desvantagens. Primeiro, o custo computacional de extrair as características das imagens, como cor e forma, para uma grande coleção, será enorme. Segundo, gerar as consultas como descrito acima não é uma tarefa simples de ser feita.

Na recuperação baseada em informações textuais, alguma forma de descrição textual do conteúdo da imagem é assumido estar armazenada com a própria imagem nos documentos Web. O fato de uma imagem poder ser descrita utilizando textos para se referir ao seu conteúdo, faz com que esse tipo de classificação seja uma importante abordagem para recuperação de imagens na Web. As principais máquinas de busca de imagens (Google, Alta Vista, Lycos) [3] utilizam essa abordagem na classificação de imagens na Internet. A máquina de busca Google, por exemplo, analisa, nos documentos HTML, o texto em torno de uma imagem e a sua legenda, entre outras inúmeras partes do documento, procurando por informações que possam descrever o conteúdo dessa imagem [2].

Utilização de metadados HTML para recuperar imagens relevantes na Web

Esse artigo apresenta uma forma de se recuperar imagens baseada na análise do documento HTML levando em consideração partes do documento que podem apresentar alguma informação sobre o conteúdo das imagens presentes no documento analisado [5].

Imagens na Web normalmente se encontram em documentos HTML que possuem, além da informação textual, também uma estrutura hierárquica. Dessa forma, foram consideradas tanto as informações textuais quanto a estrutura hierárquica como metadados que foram analisados para obter imagens relevantes para uma determinada consulta.

Para que os experimentos fossem realizados sobre uma coleção de documentos da Web seria preciso obter tais documentos, para tanto se usou uma máquina de busca e os documentos

foram armazenados. Assim, uma ferramenta de busca de imagens foi criada, baseada em 4 módulos: *busca de texto*, *captura de documentos*, *análise de documentos* e *interface de resultados das buscas*.

Por meio do primeiro módulo é selecionado na Web um conjunto de páginas onde serão feitas as buscas por imagens. Esse módulo recebe uma consulta, ou seja, uma palavra-chave e passa a consulta para a máquina de busca Alta Vista, que retorna *links* para documentos Web que estão relacionados com a consulta.

Com base nos *links* retornados pelo Alta Vista o módulo de captura de documentos copia os documentos e realiza uma limpeza para tornar documentos mal estruturados em documentos bem formados no padrão XHTML. Esse processo é realizado com a ajuda do software *Tidy* [1].

O terceiro módulo realiza a análise do documento XHTML obtido do módulo anterior procurando por "pistas" que indiquem que a imagem associada ao documento satisfaz a consulta realizada. Algumas partes específicas (aqui chamadas recursos) dos documentos são utilizadas nessa análise, que considera uma consulta satisfeita se a palavra que forma a consulta é encontrada em um desses recursos. Foram considerados 8 recursos nos documentos obtidos:

1 - Nome do arquivo da imagem;
2 - Texto da tag <TITLE> do documento;
3 - Texto do atributo ALT na tag <IMG>;
4 - Texto de um *link* que aponta para uma imagem;
5 - Texto do atributo TITLE de um *link*. Tag <A></A>;
6 - Texto do parágrafo pai da imagem (imediatamente antes da imagem);
7 - Texto do parágrafo localizado dentro da mesma tag <CENTER> que uma imagem.
8 - Texto dos elementos que estão antes da imagem.

O último módulo gera uma página Web com enlaces para as imagens selecionadas no módulo anterior.

Experimentos de recuperação de imagens

A partir dos módulos criados foram realizados experimentos com o objetivo de responder a duas perguntas:

- Quais características de um documento HTML revelam mais informações sobre as imagens contidas nele?
- Os resultados obtidos na recuperação de imagens dependem do tipo de consulta realizada?

Foram feitas 12 consultas divididas em 5 categorias. São elas:

- Pessoas famosas;
- Pessoas não famosas;
- Lugares famosos;
- Lugares pouco famosos;
- Fenômenos Naturais;

Para cada consulta foram coletadas 30 das 200 páginas de resultados da máquina de busca Alta Vista. Assim, no total foram obtidas 360 páginas contendo 1.578 imagens excluindo as imagens pequenas que são usadas para decorar os documentos. Na análise foi associado a cada imagem o tipo de recurso do documento que possibilitou a seleção, ou seja, uma das 8 partes do texto que são analisadas. A classificação, de uma imagem retornada pela busca, como relevante ou não, foi realizada por um dos pesquisadores para todas as consultas realizadas.

Resultados e discussões

Durante a análise dos documentos, por meio da avaliação feita pela pesquisadora Tsymbalenko, foram calculadas as medidas de precisão e revocação. Está última normalmente é calculada sobre um conjunto de documentos conhecido. Nesse caso foram utilizados

documentos retornados pelo Alta Vista, o que faz com que os resultados tenham uma maior relevância, pois os documentos são selecionados de acordo com os critérios da máquina de busca. Como esse trabalho não tem por objetivo comparar os resultados obtidos com outros trabalhos, se atendo apenas em descobrir a influência das partes do documento analisado nos resultados da consulta, isso não representa um problema.

Os melhores resultados para a revocação foram os dos três primeiros recursos: *nome do arquivo* com 39,9%; *texto do título* (tag <TITLE>) com 79,5% e atributo *ALT* com 2,5%. Todos os três valores se referem à média para todas as consultas. A precisão média foi maior também nos três primeiros recursos considerados, sendo 83% para o recurso *nome do arquivo*; 55% para o recurso *texto do título* e 87,5 para o recurso *ALT*. Como nem todos os recursos obtiveram revocação para todas as consultas, não foi possível calcular a precisão para todos eles.

Os resultados obtidos mostram que os recursos 1, 2 e 3 apresentam grande utilidade na busca por imagens na Web enquanto os outros recursos apresentam muito pouca importância. Isso se deve aos projetistas de sítios Web preferirem nomes mnemônicos quando estão gerando imagens; e ao título da página normalmente representar um resumo sobre assunto principal da página. No caso do terceiro recurso, temos que o atributo *ALT* serve exatamente para descrever a imagem à qual está associado e, embora tenha apresentado um revocação menor que a dos 2 primeiros recursos (provavelmente porque o atributo não é muito utilizado), apresentou uma precisão excelente, na grande parte das vezes de 100%. Isso responde a primeira pergunta proposta.

Em relação à segunda pergunta concluímos que nomes de pessoas apresentam uma menor revocação que nome de lugares. Não foram observadas diferenças significativas em relação aos outros tipos de consultas propostos.

Trabalhos futuros

Esse trabalho apresentou um sistema para a realizar busca de imagens na Web, com a vantagem de não precisar salvar qualquer imagem localmente e sem precisar processá-las para realizar a classificação.

Entretanto os resultados estão condicionados à recuperação de documentos pelo Alta Vista, podendo apresentar resultados diferentes com outra máquina de busca. Além disso as consultas foram compostas de apenas uma palavra. Essa abordagem pode ser estendida para permitir consultas mais complexas.

Apesar de toda análise acontecer sobre informações textuais, nada impede que um módulo de análise de conteúdo de imagens seja adicionado ao sistema original para melhorar os resultados obtidos.

Recuperação de imagens na Web baseada em múltiplas evidências textuais

Os textos de documentos Web freqüentemente ou não possuem informações precisas ou enganam em relação ao conteúdo de suas imagens. Enquanto o texto pode incluir importantes informações sobre uma imagem, selecionar qual parte do texto melhor descreve a imagem não é uma tarefa fácil. Além disso, as imagens na Web são pobremente rotuladas (nome do arquivo, tag ALT). Conseqüentemente o uso individual das técnicas de recuperação de imagens não produz resultados adequados. O processo de achar imagens relevantes não depende exclusivamente de todo o texto de uma página, nem de informações textuais diretamente associadas com a imagem.

O trabalho *Recuperação de imagens na Web baseada em múltiplas evidências textuais* [4] propõe usar outras fontes de informações dentro de um documento da Web, para melhorar a classificação de uma imagem. O trabalho avalia quais partes do documento Web que podem ser usadas para complementar uma eficiente descrição das imagens e propõe um modelo de recuperação de imagens baseado em redes Bayesianas de Crenças. Redes Bayesianas de Crenças

são usadas, pois elas providenciam um flexível, eficiente e fortemente seguro método de combinar diferentes recursos de evidências em um simples modelo de recuperação de informações.

Fontes textuais de evidencias nos documentos Web

Documentos da Web incluem uma variedade de dados textuais que podem ser usados para recuperação de imagens. Entretanto, o conteúdo textual não necessariamente descreve a imagem na página. Uma possível solução é considerar cada parte do texto como um recurso independente de informações. Os recursos considerados nesse trabalho são:

- Tags de descrição: composto pelo nome do arquivo, pelo atributo ALT, e pelo texto entre as tags <A> e </A>. Esses termos são usados para descrever a imagem com a qual eles estão associados.
- Meta tags: composto pelos termos localizados entre as tags <TITLE> e </TITLE> e pela tag META do documento HTML. Esses termos são usados para descrever o conteúdo do documento.
- Texto completo: todas as palavras no documento da Web. É assumido que o conteúdo do documento esta relacionado com o conteúdo das imagens contidas nele.
- Passagem de texto: composta por palavras localizadas próximas à imagem. É esperado que o texto em volta da imagem relate o seu conteúdo.

Modelo de redes de crenças

O modelo proposto para recuperação de imagens é baseado no modelo de Redes de Crenças. Esse modelo adota uma visão epistemológica do problema de recuperação de informações e interpreta probabilidades com grau de crenças desprovido de experimentos. Por essa razão o modelo recebe esse nome.

As redes de crenças são modeladas usando redes Bayesianas. Essas redes providenciam um formalismo para representar independência entre variáveis aleatórias de uma distribuição de probabilidade conjunta. A distribuição de probabilidade conjunta é representada por um grafo acíclico direcionado onde os nodos representam as variáveis aleatórias de distribuição e onde o relacionamento entre essas variáveis é modelado por arestas direcionadas representando a dependência entre as variáveis ligadas. A força de uma dependência é expressa por uma probabilidade condicional.

Modelo de rede de crenças para recuperação de imagens

O modelo de rede de crenças para recuperação de informação tradicional modela um conjunto de k palavras chaves em k nodos. O modelo possui ainda nodos que representam a busca e nodos que representam os documentos. Um nodo de uma palavra chave possui aresta para um nodo de um documento ou para um nodo da busca se a palavra chave estiver contida no documento ou na busca, respectivamente. Assim, o cálculo da classificação de um documento é computado da seguinte forma:

$$P(d_j \mid q) = \eta \sum_k P(d_j \mid k) P(q \mid k) P(k)$$

A condição probabilística $P(d_j \mid q)$ computa a probabilidade de observar o documento d_j dado a busca q.

Dessa forma, para combinar distintos recursos de informações associados a imagens da Web em uma única classificação, foi feita uma extensão do modelo anterior, adicionando novas arestas, nodos e probabilidades à rede anterior.

A figura abaixo mostra o modelo de rede para imagens.

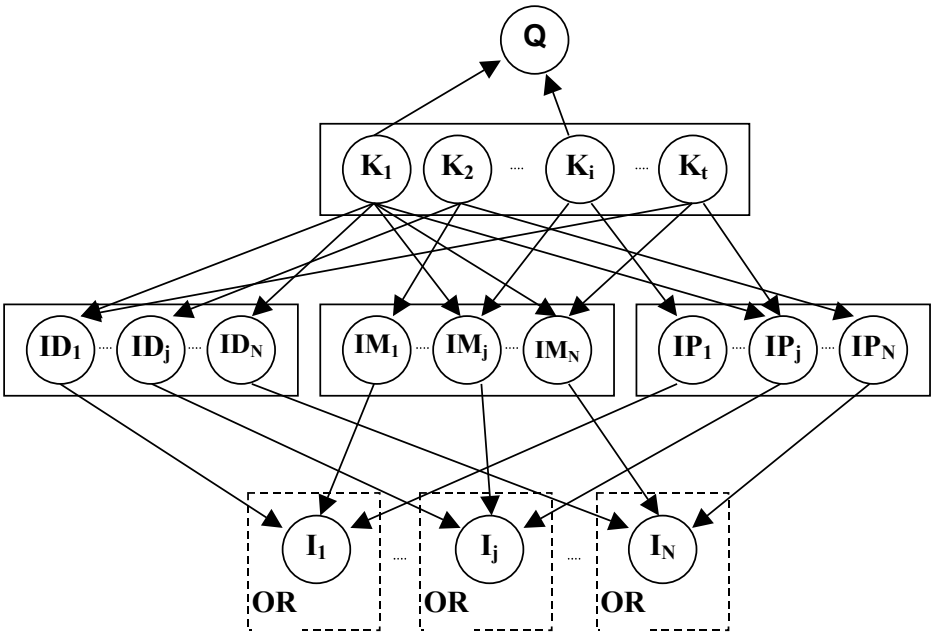

Figura 1 - Modelo de rede para combinação de múltiplas evidências extraídas de documentos Web

Cada novo nodo representa uma variável aleatória binária que modela um pedaço de informação. Os nodos k_i modelam os termos extraídos do documento Web usando todas as evidências. Os nodos ID_j modelam as evidências extraídas usando tags de descrição. Os nodos IM_j modelam as evidências extraídas usando meta tags. Os nodos IP_j modelam as evidências extraídas usando texto completo ou passagens de texto. Os nodos I_j modelam as imagens. As distintas evidências são combinadas através de um operador disjuntivo para produzir uma classificação para cada imagem.

A classificação de uma imagem I_j pode ser calculada da seguinte forma:

$$P(I_j \mid q) = \eta \sum_k [1-(1-P(ID_j \mid k))(1-P(IM_j \mid k))(1-P(IP_j \mid k))]P(q \mid k)P(k)$$

onde η é uma constante de normalização, introduzida para fazer a soma de todas as probabilidades iguais a 1, e k é o estado de todas as k_i variáveis. Para definir a classificação, temos que especificar os valores de $P(ID_j \mid k)$, $P(IM_j \mid k)$, $P(IP_j \mid k)$, $P(q \mid k)$ e $P(k)$. O valor de $P(k)$ é uma constante igual para todo k_i, para não dar preferência a nenhum conjunto de palavras chaves.

Para $P(q \mid k)$ temos:

$$P(q \mid k) = \begin{cases} 1 & se\ \forall i\ ki \in q \\ 0 & outro\ \mathrm{mod}\, o \end{cases}$$

Para $P(ID_j \mid k)$, $P(IM_j \mid k)$ e $P(IP_j \mid k)$ foi usado o conhecido modelo vetorial para recuperação de informação. Assim, temos para $P(ID_j \mid k)$ o seguinte valor:

$$P(ID_j \mid k) = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \times \sum_{i=1}^{t} w_{iq}^2}}$$

onde w_{ij} é o peso do termo k_i nas tags de descrição da imagem I_j, e w_{iq} é o peso do termo k_i na consulta do usuário. A definição dos pesos é: $w_{ij} = (1 + ln\ f_{ij})$, $w_{iq} = ln(1 + N/n_i)$ onde f_{ij} é a

freqüência do termo k_i nas tags de descrição da imagem I_j, N é o número total de imagens e n_i é o número de todas as tags de descrição na coleção que contém k_i. As equações são similares para $P(IM_j \mid k)$ e $P(IP_j \mid k)$.

Resultados

Os testes foram realizados utilizando uma coleção de 128.712 páginas, 54.571 imagens e 25 palavras chaves nas buscas. Para avaliar os resultados foram utilizadas as medidas *precisão* (*precision*) e *revocação* (*recall*), bastante conhecidas na área de recuperação de informações.

Com essas duas medidas é possível avaliar a média da precisão em relação a um dado nível de revocação. Nos testes foram realizadas medidas da média da precisão em 11 pontos: 0%, 10%, 20%, ..., 100% da revocação. Precisão em 0% de revocação é a precisão quando um documento relevante é encontrado no topo da classificação. Precisão em 10% de revocação é a precisão quando 10% dos documentos relevantes são encontrados nas primeiras posições da classificação. A média da precisão em 10% de revocação é a média da precisão para todos os testes de busca, obtida com 10% de revocação.

Os primeiros testes foram realizados para determinar qual o melhor tamanho da passagem de texto. Foram utilizadas passagens de seguintes tamanhos: 10 termos, 20 termos, 40 termos e o texto completo da página. Os testes mostraram que a passagem de texto com 40 termos obteve os melhores resultados. Nos experimentos restantes foram considerados somente passagem de texto com 40 termos.

Os experimentos seguintes foram realizados utilizando os recursos de evidências (tags de descrição, meta tags e passagem de texto) isoladamente. Os resultados obtidos mostraram que tags de descrição e passagem de texto apresentaram resultados similares na média da precisão sobre a maioria dos valores da revocação. Os resultados indicaram também que o uso dos meta tags para recuperação de imagens é pouco informativo.

O último experimento foi realizado combinando múltiplos recursos de evidências usando o modelo proposto. As combinações de evidências foram:

Combinação de Evidências	
A	Tags de descrição + meta tags
B	Tags de descrição + passagem de texto
C	Passagem de texto + meta tags
D	Tags de descrição + meta tags + passagem de texto.

Tabela 1 - Combinação de evidências utilizada nos experimentos.

Devido ao desempenho ruim dos meta tags, os testes utilizando as combinações A e C apresentaram resultados ruins. O teste com a combinação B produziu uma média total da precisão de 60,1% (obtida pela média da precisão em todos os níveis da revocação). Já a combinação de evidências D produziu uma média total de 59,0%. Apesar de ambos os desempenhos serem parecidos, o uso de B mostrou uma melhor precisão para os níveis de revocação de até 40%. Isso é um importante resultado para máquinas de busca na Web, onde precisão é mais importante entre os primeiros documentos da classificação.

Para níveis da revocação acima de 60%, a combinação D apresentou os melhores resultados. Isso mostra que apesar de os meta tags obterem resultados pobres quando usado como um simples recurso de evidência, eles podem ainda providenciar alguma importante informação.

O resultado obtido na abordagem B representa um melhoramento de 50,3% sobre o melhor resultado obtido quando um recurso isolado de evidência é usado.

Esses resultados indicam que o modelo proposto é uma estrutura adequada para combinar muitos recursos de evidências textuais em uma única e melhorada fórmula de classificação de imagens da Web. Uma vez que os resultados são baseados em recursos de

evidências que podem ser obtidos de modo totalmente automático, esse novo modelo providencia uma solução viável para o problema de classificação de imagens na Web.

Trabalhos futuros

Considerando a flexibilidade do modelo apresentado, é possível estender em número e forma as combinações das mais variadas evidências, é possível inclusive combinar as duas abordagens consideradas no início deste trabalho (recuperação baseada em texto e recuperação baseada em conteúdo), para tentar melhorar os resultados obtidos na recuperação de imagens.

Referências

[1]. D. Raggett. *Clean up Your Web Pages with HTML Tid*y.
URL: http://www.w3.org. Versão disponível em agosto de 2000.

[2]. Google FAQ: *Perguntas e respostas sobre a Busca de Imagens*. Acessado em maio de 2004.
URL: http://www.google.com/help/faq_images.html

[3]. R. Lempel, A. Soffer. *PicASHOW: Pictorial Authority Search by Hyperlinks On the Web*.

[4]. T. A. S. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, R. Muntz. *Image Retrieval Using Multiple Evidence Ranking*. IEEE Transactions on Knowledge and Data Engineering - April 2004.

[5]. Y. Tsymbalenko, E. V. Munson. *Using HTML Metadata to Find Relevant Images on the World Wide Web*. In Proceedings of Internet Computing 2001, Volume II, Las Vegas, pages 842–848. CSREA Press, June 2001.