OPIS: Um método para identificação e busca de páginas-objeto apoiado por realimentação de relevância e classificação de páginas web*

Miriam Pizzatto Colpo[1], Edimar Manica (Colaborador)[1], Renata Galante (Orientadora)[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{mpcolpo,edimar.manica,galante}@inf.ufrgs.br

Nível: Mestrado

Programa: Programa de Pós-graduação em Computação (PPGC) da Universidade Federal do Rio Grande do Sul (UFRGS)

Ingresso: Março/2012

Época esperada para conclusão: Março/2014

Etapas concluídas: Créditos (2012), Proposta (Outubro/2012) e Seminário de Andamento (Maio/2013)

Etapas futuras: Submissão de Artigos (Julho/2013 – Março/2014) e Defesa da Dissertação (Março/2014)

Abstract. *This paper proposes a new method for identifying and searching object pages named OPIS (acronyms to* ***O****bject* ***P****age* ***I****dentifying and* ***S****earching). Object pages are pages that represent exactly one inherent real-world object on the web. The purpose of OPIS is to address the search for these real-world objects pages, since the General Search Engines (GSEs) cannot answer satisfactorily this type of search today. The kernel of our method is to adopt feedback relevance and machine learning techniques in the task of content-based pages classification. OPIS, when integrated into a GSE, enables the filtering of object pages, in which only pages classified as object pages are retrieved by user keyword queries instead of all pages that contain those words. Preliminary experiments show that OPIS improved on average 37% of the precision in 20 (p@20) of the results retrieved when compared with a GSE.*

Palavras-chave: páginas-objeto, busca-objeto, classificação de páginas *web.*

* Este trabalho é parcialmente financiado pelo Instituto Nacional de Pesquisa da Web, pelo CNPq e pela CAPES.

1. Introdução

Os motores de busca convencional da web (do inglês, *General Search Engines* – GSEs) são programas que visam recuperar informações da web e apresentá-las, de forma organizada e eficiente, aos usuários [Baeza-Yates e Ribeiro-Neto 2011]. Um GSE, basicamente, recebe um conjunto de palavras-chave e, analisando apenas o texto não estruturado, gera uma lista de páginas que contêm essas palavras.

`Objetos da web` são unidades de dados sobre as quais informações da web são coletadas, indexadas e ordenadas. Esses objetos são conceitos usualmente reconhecidos (como autores ou conferências), relevantes a um domínio de aplicação e que podem ser representados por um conjunto de atributos, os quais dependem do domínio do objeto [Nie et al. 2007]. `Páginas-objeto` são páginas que descrevem um único objeto inerente na web. Isso significa que páginas que listam diversos objetos não são consideradas páginas-objeto por não representarem um objeto em particular. A busca por páginas-objeto é feita através de consultas restringidas por atributos de domínio e pode ser chamada de `busca-objeto` [Pham et al. 2010]. Um exemplo desse tipo de consulta é "*professor de banco de dados da UFRGS*", que restringe a área e a instituição de atuação de um objeto professor e tem como objetivo recuperar páginas (pessoais, institucionais, de currículo, etc.) que descrevam esse objeto.

Embora os GSEs consigam atender à maioria das consultas realizadas atualmente, eles se mostram inadequados para recuperar páginas-objeto [Pham et al. 2010]. Dentre as limitações do processo de busca convencional que podem estar relacionadas a esse problema, encontra-se a ambiguidade das palavras-chave. Por exemplo, mesmo que o objetivo de uma busca com a palavra-chave "*Paris*" seja encontrar páginas relacionadas à cidade capital da França, muitas páginas relacionadas ao primeiro nome de *"Paris Hilton"* serão retornadas [Miklós 2010].

Este artigo propõe um novo método para a identificação e a busca de páginas-objeto, denominado OPIS (acrônimo para ***O**bject **P**age **I**dentifying and **S**earching*), que adota realimentação de relevância e técnicas de pré-processamento de texto e de aprendizagem de máquina na classificação baseada em conteúdo de páginas web. O OPIS envolve a integração de um classificador a um motor de busca, de modo que os resultados recuperados pelo motor de busca sejam filtrados, permitindo que somente as páginas identificadas (classificadas) como páginas-objeto sejam apresentadas aos usuários. A principal contribuição deste método é a melhoria na precisão de buscas-objeto, permitindo que usuários finais encontrem resultados que melhor atendam a suas necessidades de informação.

O OPIS foi avaliado através de experimentos preliminares no domínio real de pesquisadores. Os resultados mostram que o OPIS superou o motor de busca convencional com aumentos de precisão de 112% (0,144 vs. 0,068) nos primeiros cinco resultados e de 37% (0.157 vs. 0.115) nos primeiros 20.

O restante desse artigo está organizado da seguinte forma. Na Seção 2 são apresentados trabalhos relacionados. Na Seção 3, o OPIS é especificado detalhadamente. Na Seção 4 é apresentada a implementação e a experimentação do OPIS no domínio de pesquisadores. Na Seção 5, o artigo é concluído e direções futuras são apontadas.

2. Trabalhos Relacionados

Muitos esforços têm sido feitos para melhorar os resultados recuperados pelos GSEs. A maioria deles considera que os motores de busca são independentes de domínio e, em geral, usam a mesma função de *ranking* para todas as páginas, ou seja, os GSEs não consideram as particularidades de cada domínio. O OPIS propõe uma nova solução nesse sentido. A seguir, são apresentados e comparados os principais trabalhos relacionados à busca-objeto e tópicos relacionados.

Motores de busca vertical [Ji et al. 2009][Lee et al. 2011][Luo 2009], que são abordagens para criar motores de busca específicos a determinados domínios, relacionam-se à busca-objeto por também restringirem a busca à um domínio específico. Além disso, outros trabalhos [Bennett et al. 2010][Geng et al. 2009][Pham et al. 2010] usam funções de *ranking* específicas para recuperar páginas relacionadas a domínios específicos. Em geral, esses trabalhos usam técnicas de processamento de texto e aprendizagem de máquina para extrair o conteúdo das páginas e, com base nisso: aprender um determinado domínio e filtrar apenas páginas consideradas pertencentes a esse domínio no processo de coleta; ou determinar as características do domínio a serem usadas em uma função de *ranking* específica. O OPIS se difere desses trabalhos à medida que não deseja aprender apenas o tópico das páginas, mas também seu tipo funcional (se a página é ou não uma página-objeto).

Blanco et al. (2008) propõem um método para coletar automaticamente páginas da web que publicam dados relacionados às instâncias de entidades conceituais com um esquema implícito. Esse método assume que o usuário fornece exemplos de páginas de entidades a partir de sites distintos e percorre cada um desses sites procurando por páginas que apresentam *templates* e caminhos similares aos respectivos exemplos. Esse trabalho difere do OPIS por focar na coleta de páginas de entidades com *templates* similares, enquanto o OPIS busca identificar páginas-objeto, sem considerar *templates* específicos, para melhorar a busca-objeto. Também com relação à coleta de páginas, Assis (2008) propõe um coletor focado para tópicos de interesse que possam ser representados por características de gênero e de conteúdo. Quando o usuário deseja buscar por páginas de planos de ensino de disciplinas de banco de dados, por exemplo, um conjunto de características (termos) que descreva o gênero (planos de ensino) e outro que descreva o conteúdo (banco de dados) devem ser informados por um usuário especializado, de modo que o coletor possa analisar cada página através da sua similaridade com os termos de ambos os aspectos. No OPIS, o conteúdo e o gênero das páginas não são considerados separadamente, o que reduz o nível de especialidade do usuário, uma vez que ele não precisa discernir entre esses dois aspectos e nem selecionar termos manualmente para caracterizá-los.

Para Pham et al. (2010), cuja proposta esta mais próxima do OPIS, o problema de busca-objeto se assemelha ao de aprendizagem de *ranking*, em que o principal objetivo é aprender uma função de *ranking* através de uma função de aprendizagem, com base em um conjunto de características relevantes. A solução proposta consiste em desenvolver diversos motores de busca vertical para suportar a busca por páginas-objeto em diferentes domínios. Para isso, uma função de *ranking* deve ser aprendida para cada

domínio específico. O desenvolvedor[1] deve submeter consultas por palavras-chave e anotar um conjunto de treinamento a partir das páginas recuperadas. Esse conjunto de treinamento tem suas características extraídas automaticamente e usadas em uma função de aprendizagem. O OPIS também adota o conteúdo das páginas para melhorar a busca-objeto através da classificação funcional (páginas-objeto ou não) das páginas. Porém, ele não considera a busca-objeto como um problema de aprendizagem de *ranking*. Ao invés disso, o processo de *ranking* fica a cargo do motor de busca ao qual acoplamos o método. Isso torna desnecessária a análise das informações estruturadas embutidas nas páginas, durante o processo de busca, para casá-las com as características que integram a função de *ranking* aprendida (por exemplo, "a palavra professor aparece no título").

3. OPIS: *Object Page Identifying and Searching*

O OPIS (acrônimo para ***O**bject **P**age **I**dentifying and **S**earching*) é um método que visa a identificação e a busca de páginas-objeto. OPIS permite que somente páginas identificadas como páginas-objeto, para um determinado domínio, sejam recuperadas pelas consultas dos usuários, o que torna os resultados de buscas-objeto mais precisos e, dessa forma, mais adequados às necessidades dos usuários. O método caracteriza-se por adotar realimentação de relevância e técnicas de pré-processamento de texto e de aprendizagem de máquina na construção, baseada no conteúdo de páginas web, de um classificador, que será responsável pela identificação das páginas-objeto. Esse classificador é, então, integrado a um motor de busca, adicionando uma etapa de filtragem (classificação) ao processo de busca convencional.

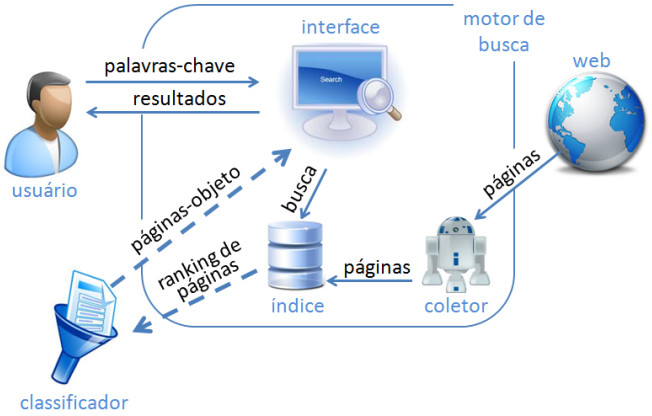

Figura 1. Visão geral do OPIS.

Na Figura 1 é apresentada uma visão geral do OPIS, mostrando a integração das atividades de identificação e busca. Note que o usuário submete, através da interface, uma consulta por palavras-chave. Essa consulta é executada sobre o índice e obtém como resposta o *ranking* das páginas nas quais os termos da consulta foram encontrados, seguindo, até então, o processo de busca convencional. A diferença introduzida pelo OPIS (ilustrada com seguimentos tracejados) esta no fato de que esse *ranking* de páginas não é apresentado diretamente ao usuário, passando antes por uma

[1] Pessoa responsável por guiar o treinamento da função de *ranking* para um domínio específico, permitindo que usuários possam, então, submeter consultas relacionadas a esse domínio.

atividade de filtragem adicional, na qual essas páginas passam por um processo de classificação e somente as classificadas como páginas-objeto são apresentadas para o usuário.

O classificador é o cerne do OPIS, pois é o responsável pela tarefa de identificação das páginas-objeto, da qual depende a filtragem e, dessa forma, os resultados da busca. A construção de um classificador depende, além da escolha e parametrização de um algoritmo de classificação, de um conjunto de páginas de treinamento do domínio de interesse, que é a base do processo de aprendizagem do algoritmo. Para obter esse conjunto, o OPIS faz uso de Realimentação de Relevância [Baeza-Yates e Ribeiro-Neto 2011], na qual o usuário avalia a relevância de um conjunto de páginas e este passa a ser considerado como base do treinamento. O classificador produzido está intrinsecamente relacionado com o usuário em questão, uma vez que será este quem definirá a base de treinamento a ser usada na geração do modelo de classificação. Dessa forma, a corretude da coleção de treinamento e, consequentemente, do classificador, está condicionada à correta avaliação do usuário quanto ao que é ou não uma página-objeto, devendo, assim, esse usuário ter conhecimento prévio sobre o domínio a ser treinado.

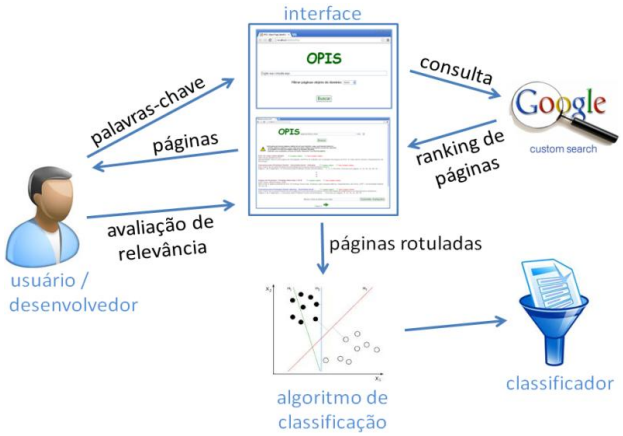

Figura 2. Construção do classificador, apoiada pela realimentação de relevância.

Inicialmente, o usuário deve estabelecer quais tipos de páginas-objeto podem existir no domínio a ser considerado (no exemplo do domínio de pesquisadores: páginas institucionais, pessoais, de currículo, etc. que descrevam um objeto pesquisador). Com base nisso, uma ou mais consultas por palavras-chave, visando recuperar essas páginas, devem ser criadas (como "*homepage professor doutor*"). O processo de realimentação de relevância é ilustrado na Figura 2. Após submeter uma dessas consultas, o usuário pode avaliar a relevância dos resultados recuperados, indicando páginas que tenham sido ou não consideradas páginas-objeto para o domínio em treinamento. Esse procedimento pode ser repetido com mais consultas, atualizando a coleção de treinamento e, consequentemente, o classificador, até que o usuário julgue necessário. Isso significa que o processo continua até que o usuário considere ter rotulado páginas suficientes para exemplificar tanto páginas não objeto quanto os principais tipos de páginas-objeto estabelecidos inicialmente.

4. Implementação e Experimentos

Uma interface web foi desenvolvida para suportar a tarefa de realimentação de relevância, permitindo que o usuário submeta suas consultas e possa avaliar a relevância das páginas resultantes, por meio de caixas de seleção apresentadas ao lado de cada resultado. A API *Google Custom Search* [Google 2012], sem nenhuma customização, foi usada para recuperar e ordenar os resultados das consultas submetidas pelo usuário. Mesmo não tendo sido mostrado na Figura 2, antes de serem usadas no treinamento do algoritmo de classificação, as páginas rotuladas passam por atividades de pré-processamento (como remoção de *tags* HTML, tradução de termos estrangeiros e remoção de *stopwords*) e são representadas através do Modelo de Espaço Vetorial, considerando TF-IDF [Baeza-Yates e Ribeiro-Neto 2011] como forma de ponderação. Na classificação, foi utilizado o algoritmo *LibSVM*, através da biblioteca de mineração *Weka* [Universidade de Waikato 2013], com núcleo linear (por ser de rápida execução).

Para a realização de experimentos, foi considerado um classificador construído, de acordo com o processo explicado na Seção 3, para o domínio de pesquisadores. Foram rotulados 10 exemplos de páginas-objeto (incluindo páginas institucionais, pessoais ou de currículo, que representassem um objeto pesquisador) e 10 exemplos negativos (incluindo páginas relacionadas a concursos, corpo docente e notícias), que serviram como base para o aprendizado do algoritmo. O classificador foi integrado à API *Google Custom Search* (sem nenhuma customização), usada como motor de busca convencional. Os experimentos contaram com a participação de 10 usuários, que criaram cinco consultas, para o domínio de pesquisadores, cada. Os usuários também especificaram objetivos e critérios de relevância para cada consulta, que guiaram a avaliação de relevância dos resultados recuperados.

Tabela 1. Resultados com a média das 50 consultas criadas pelos usuários.

	p@5	p@10	p@15	p@20
Google	0.068	0.094	0.108	0.115
OPIS	0.144	0.166	0.155	0.157

Na Tabela 1 são apresentados os resultados obtidos a partir da submissão e avaliação das 50 consultas criadas pelos usuários, considerando o Google (sem filtragem) e o OPIS (com a filtragem dos resultados recuperados pelo Google). A métrica de precisão em *n* (p@*n*), que considera somente os primeiros *n* resultados recuperados pelo sistema, foi usada com *n* de 5, 10, 15, e 20 para medir a precisão dos resultados apresentados e ter um indicativo das variações de precisão em relação ao posicionamento no *ranking*. Pode-se notar que o OPIS obteve melhores resultados em todas as precisões, considerando a média das consultas realizadas, tendo essa melhoria variado de 112% (p@5) a 37% (p@20). Isso significa que a filtragem realizada pelo OPIS permitiu que páginas-objeto relevantes para a consulta do usuário substituíssem páginas classificadas como não objeto no *ranking* das páginas recuperadas, apresentando ao usuário resultados que melhor atendam a sua necessidade de informação para buscas-objeto.

5. Considerações Finais e Trabalhos Futuros

Neste artigo foi introduzido o problema da busca-objeto em motores de busca convencional e proposto um novo método, chamado OPIS, para a identificação e a busca

de páginas-objeto. O OPIS usa realimentação de relevância e técnicas de pré-processamento de texto e de aprendizagem de máquina na construção, baseada no conteúdo de páginas web, de um classificador. A integração desse classificador a um motor de busca convencional permite a filtragem dos resultados de busca-objeto, fazendo com que apenas páginas classificadas como páginas-objeto sejam apresentadas aos usuários. Experimentos preliminares mostraram que o OPIS proporcionou um ganho de 37% de precisão, considerando as 20 primeiras páginas recuperadas pelas buscas-objeto, em relação a um motor de busca convencional.

Como trabalhos futuros, pretende-se: (i) incorporar expansão de consultas ao OPIS, para melhorar a precisão dos resultados; (ii) testar o método em um domínio adicional, a fim de atribuir maior confiabilidade à validação; e (iii) usar como *baseline* a abordagem de Pham et al.(2010), apresentada nos trabalhos relacionados.

Referências

Assis, G. T. (2008) Uma Abordagem Baseada em Gênero para Coleta Temática de Páginas da Web. Tese (Doutorado em Ciência da Computação) - Instituto de Ciências Exatas, Universidade Federal de Minas Gerais.

Baeza-Yates, R. e Ribeiro-Neto, B. (2011), Modern Information Retrieval: The Concepts and Technology behind Search. Addison Wesley, 2ª edição.

Bennett, P. N., Syore, K. e Dumais, S. T. (2010) Classification-Enhanced Ranking. In: 19th International Conference on World Wide Web, p. 111-120.

Blanco, L., Crescenzi, V., Merialdo, P. e Papotti, P. (2008) Supporting the Automatic Construction of Entity Aware Search Engines. In: 10th ACM Workshop on Web Information and Data Management, p. 149-156.

Geng, B., Yang, L., Xu, C. e Hua, X. (2009) Ranking Model Adaptation for Domain-Specific Search. In: 18th Conference on Information and Knowledge Management, p. 197-206.

Google (2013) "Google Custom Search", https://developers.google.com/custom-search, Junho.

Ji, L., Yan, J., Liu, N., Zhang, W., Fan, W. e Chen, Z. (2009) ExSearch: A Novel Vertical Search Engine for Online Barter Business. In: 18th Conference on Information and Knowledge Management, p. 1357-1366.

Lee, H., Nazareno, F., Jung, S. e Cho, W. (2011) A Vertical Search Engine for School Information Based on Heritrix and Lucene. In: 5th International Conference on Convergence and Hybrid Information Technology, p. 344-351.

Luo, G. (2009) Design and Evaluation of the iMed Intelligent Medical Search Engine. In: IEEE International Conference on Data Engineering, p.1379-1390.

Miklós, Z. (2010) From Web Data to Entities and Back. In: 22nd International Conference on Advanced Information Systems Engineering, p. 302-316.

Nie, Z., Ma, Y., Shi, S., Wen, J. e Ma, W. (2007) Web Object Retrieval. In: 16th International Conference on World Wide Web, p. 81-90.

Pham, K. C., Rizzolo, N., Small, K., Chang, K. C. e Roth, D. (2010) Object Search: Supporting Structured Queries in Web Search Engines. In: NAACL HTL 2010 Workshop on Semantic Search, p. 44-52.

Universidade de Waikato (2013), "Weka Data Mining Software API", http://www.cs.waikato.ac.nz/ml/weka, Junho.