MO815 – Análise Forense de Documentos Digitais – Fichamento #1

Local Rank: Ranking Web Pages Considering Geographical Locality by Integrating Web and Databases

by Jianwei Zhang et al.

10 de Agosto de 2010

Revisor: Fulano Beltrano — RA XXYYZZ

1 Visão global

Os autores propõem um método (*LocalRank*) para ordenar informações na Web relativas a uma determinada localidade geográfica. Para isso, o método integra um banco de dados local, que expressa entidades relativas a uma região de interesse, e a relação semântica de entradas deste banco de dados e páginas Web relacionadas.

2 Resumo

Os autores apresentam uma abordagem para consultas na Web com restrições de localidade que considera *rankings* dos relacionamentos semânticos entre entidades locais presentes em um banco de dados e seus respectivos sítios na Web.

Segundo os autores, consultas na Web com restrições de localidade constituem um problema difícil dado que páginas com popularidade (*ranking*) alto podem não servir para uma busca com restrição de localidade. Por outro lado, uma página importante para uma determinada localidade geográfica pode não estar bem ordenada por um sistema de busca que considera reputação e popularidade.

Ao invés de proceder consultas com restrição de localidade utilizando apenas a Web (e.g. [5, 4, 3]), os autores estendem a abordagem apresentada em [1] para criar e avaliar relacionamentos entre instâncias considerando o banco de dados, a Web e relacionamentos entre a Web e instâncias do banco de dados.

Os autores criam um banco de dados (BD) estendido que associa informações do BD local e a Web. Para isso, cria-se um grafo de transferência de conhecimentos chamado ATG que contém os pesos considerados importantes no novo sistema de ordenação.

Para cada entidade presente no BD local, acessa-se o sítio desta entidade e procura-se páginas pelas quais este sítio aponta e páginas que apontam para esse sítio considerando os critérios de localidade. Para cada nova página localizada, extrai-se suas informações de localidade e avalia-se se estas são relevantes ou não à pesquisa do usuário. Caso sejam, recebem um peso e são incluídas no ATG, caso não, são eliminadas. Ao final, a ordenação (*ranking*) das páginas é computada sobre os *links* gerados pelas entidades (páginas) relevantes coletadas e presentes no grafo de conhecimentos.

Os autores validaram o artigo com um pequeno banco de dados com 54 restaurantes na cidade japonesa de Tsukuba.

3 Contribuições

A principal contribuição do artigo diz respeito à associação de um banco de dados local representando entidades locais e a Web para resolver o problema de consultas na Web com restrições de localidade.

Pode-se entender o trabalho como uma extensão de técnicas que ordenam páginas por popularidade e reputação [2], técnicas que permitem consultas com restrição de localidade usando apenas a Web [5, 4, 3], e técnicas que consideram relacionamentos entre entradas em um banco de dados como *links* e utilizam uma análise destes *links* para a recuperação semântica de informações [1].

A contribuição dos autores é de ordem mais prática que teórica. No entanto, o trabalho é relevante pois a abordagem apresentada, embora validada apenas em um pequeno banco de dados, permite que páginas importantes no contexto local sejam avaliadas, o que pode não acontecer de forma eficaz utilizando-se apenas as técnicas anteriores. Além disso, pode-se visualizar muitas extensões para o trabalho (vide Sec. 6, neste documento).

4 Defeitos/Desvantagens

O artigo parece correto tecnicamente e suas alegações são comprovadas por uma validação pequena, mas razoável. De forma geral, o trabalho apresentado é claro.

No entanto, os autores pecam em alguns pontos específicos exigindo confiabilidade por parte do leitor ou um esforço adicional de entendimento. Abaixo, estão alguns problemas presentes no artigo:

1. Os autores citam que o projeto HITS (Sec. 2) gera alguns *hubs* e *scores* de autoridade mas não explica estes termos.

2. Os pesos dos *links* presentes no grafo de transferência de conhecimento (ATG) (Sec. 3.2) não são explicados e os autores apenas dizem que é um procedimento empírico e difícil.

3. A fórmula apresentada para cálculo do *score/ranking* parece ser a fórmula geral utilizada para ordenação de páginas (e.g. [2, 1]) mas os autores não referenciam de onde a mesma proveio ou mesmo a explicam com detalhes. Da mesma forma, o parâmetro d desta fórmula, representa um fator de *damping* a ser fornecido pelo usuário. No entanto, o seu significado não é apresentado.

4. A validação do artigo foi feita em um banco de dados relativamente pequeno (54 entradas) e isto resultou em uma análise de quase 13.000 páginas relacionadas. Os autores poderiam validar a abordagem em um banco de dados maior e fazer comentários sobre a performance geral do sistema. Seria interessante, colocar na validação o resultado de uma busca com restrição de localidade considerando um buscador comum na Web e mostrar as diferenças entre as 20 páginas mais relevantes segundo o buscador e segundo a abordagem de *LocalRank* apresentada.

5. Não foram fornecidos detalhes do cálculo de distâncias feito pelo algoritmo. Utiliza-se para isso, um serviço típico para cálculos de distâncias presente no Japão. Por exemplo, não se sabe se este cálculo é meramente Euclidiano desprezando-se a distribuição das ruas ou leva em conta esta informação.

5 Trabalhos correlatos

5.1 ObjectRank: Authority-Based Keyword Search in Databases – Andrey Balmin et al. [1]

Relação com o artigo avaliado: *LocalRank* utiliza os conceitos do cálculo de (conhecimento) entre instâncias para criar o grafo de relacionamentos entre instâncias do Banco de Dados e instâncias Web semanticamente relacionadas. Está citado corretamento no texto.

Descrição: Neste artigo, Andrey Balmin et al. apresentam uma abordagem para consultas em Bancos de Dados levando em consideração *rankings* de autoridade sobre uma determinada palavra-chave. Um *ranking* de autoridade denota o grau de conhecimento (autoridade) associado a uma palavra-chave.

Conceitualmente, este conhecimento é modelado através de um grafo de autoridade que denota o quanto um conceito se relaciona com outros conceitos. Inicialmente, uma autoridade se origina nos pontos em que a palavra-chave consultada aparece, depois nos pontos em que há ligações entre essas palavras-chaves e outras palavras com a mesma semântica e assim sucessivamente até um um certo critério de parada.

Um ponto fraco desta abordagem que se reflete em trabalhos que a estendem tal como o artigo sendo avaliado, é o custo computacional para o cálculo do grafo de autoridade. Este grafo tem que ser computado previamente pois seu cálculo é inviável durante a execução. Para isso, computa-se um índice invertido para cada possível palavra-chave de consulta o que, por sua vez, pode ser extremamente complexo em Bancos de Dados de médio ou grande porte. Esta informação não foi considerada na abordagem de *LocalRank*.

5.2 1 – Web Information Retrieval Based on the Localness Degree 2 – A Localness-Filter for Searched Web Pages – Qiang Ma et al. [5, 4]

Relação com o artigo avaliado: Apresentam uma abordagem para localização de informações regionais a partir da Web. O primeiro é uma extensão do segundo e os autores citam apenas o segundo, mas corretamente. *LocalRank* utiliza esta idéia em uma etapa de localização de relações entre páginas Web associadas ao tópico consultado e propõe utilizar alguns conceitos como extensão do trabalho realizado até o momento.

Descrição: Nestes artigos, Qiang Ma et al. apresentam uma abordagem para avaliar o grau de localidade de uma consulta Web. O grau de localidade (*localness degree*) proposto pelos autores consiste em uma estimativa de dependência local e da natureza ubíqua (onipresente) de páginas Web. O grau de localidade de uma determinada consulta é calculado em duas etapas. (1) Analisa-se o conteúdo de páginas Web relacionadas com uma consulta considerando-se a ordem das que são mais citadas (importantes). Para estas páginas, determina-se a freqüência de ocorrência de certas palavras com contexto geográfico e sua relação com a área geográfica (latitude/longitude) coberta pelo conteúdo descrito na página analisada. Em seguida, (2) Faz-se uma comparação das páginas analisadas com outras páginas Web considerando informações relacionadas.

A primeira etapa resulta em um grau de localidade para páginas próximas ao local em que o usuário tem interesse. A segunda etapa permite uma consulta de comunidades associadas ao tema consultado. Por exemplo, a primeira etapa permitiria achar restaurantes localizados no centro de uma cidade e a segunda permitiria localizar uma área da cidade com maior concentração de restaurantes.

Um ponto fraco desta abordagem que se reflete em trabalhos que a estendem é que a coleta de informações de uma página é complicada pelo próprio caráter ambíguo das descrições geográficas que esta página possa conter.

6 Extensões

A pesquisa realizada pelos autores poderia ser utilizada para fazer a categorização dos serviços presentes em uma determinada cidade. Através de um banco local da prefeitura ou de um portal da cidade, contendo informações sobre farmácias, médicos, hospitais, bombeiros, clubes, restaurantes entre outros serviços, poderia-se ordenar buscas relativas a regiões específicas desta cidade. Desta forma, o sistema desenvolvido funcionaria como um conselheiro. Por exemplo, na existência de cinco hospitais na cidade, uma busca neste sistema resultaria aquele que tem maior relevância segundo o banco de dados local e a sua reputação segundo descrições na Web.

Outra abordagem igualmente interessante seria a extensão da abordagem de associar um banco local a serviços Web utilizando os grafos de relacionamentos semânticos propostos no trabalho. Desta forma, uma central de polícia poderia ter informações atualizadas sobre a situação do trânsito, tempo, acidentes entre outras para melhor orientar seu efetivo policial em direção às chamadas realizadas. Outra possibilidade seria a associação de um banco de dados de contextos que faz a categorização das páginas por seu conteúdo separando-as em classes diferentes. Uma consulta por "pintores São Paulo" resultaria nas categorias "Artistas em SP" e "Serviços de Pintura em SP", por exemplo.

7 Notas

1. Relevância: 8.0
2. Originalidade: 6.5
3. Qualidade científica: 7.5
4. Apresentação: 7.0
5. Nota final: 7.5

Referências

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: Authority-based keyword search in databases. In *VLDB*, pages 564–575, 2004.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] R. Lee et al. Map-based range query processing for geographic web search systems. In *Digital Cities III: Information Technologies for Social Capital*, 2005.

[4] Qiang Ma, Chiyako Matsumoto, and Katsumi Tanaka. A localness-filter for searched web pages. In *Web Technologies and Applications: 5th Asia-Pacific Web Conf.*, pages 525–536, 2003.

[5] Chiyako Matsumoto, Qiang Ma, and Katsumi Tanaka. Web information retrieval based on the localness degree. In *13th Intl. Conf., DEXA*, 2002.