Métricas de Análise de Links e Qualidade de Conteúdo: um estudo de caso na Wikipédia

Raíza Hanada*
Universidade de São Paulo
Av. do Trabalhador
São-Carlense, 400
São Carlos, Brasil
rhanada@icmc.usp.br

Marco Cristo†
Universidade Federal do Amazonas
Av. Rodrigo Otavio, 6200
Manaus, Brasil
marco.cristo@icomp.ufam.edu.br

Maria da Graça Campos Pimentel‡
Universidade de São Paulo
Av. do Trabalhador
São-Carlense, 400
São Carlos, Brasil
mgp@icmc.usp.br

ABSTRACT

Muitos apontamentos (links) entre páginas Web podem ser vistos como indicativos da qualidade e/ou importância da página apontada. Apoiando-se nessa ideia, diversas métricas baseadas em links foram propostas na literatura para identificar conteúdo de qualidade na Web. Vários métodos de avaliação demonstram que estas métricas são bem sucedidas na tarefa de ordenação(ranking) das páginas de resposta a consultas submetidas a máquinas de busca. Apesar disso, não é possível determinar qual a contribuição específica de fatores como qualidade, importância ou popularidade para o resultado obtido. Essa dificuldade se deve, em parte, ao fato de que tais informações não são de fácil obtenção para páginas da Web em geral. Diferentemente de páginas Web comuns, artigos da Wikipédia são avaliados por seres humanos segundo a sua qualidade, utilizando critérios estabelecidos previamente. As notas de qualidade dos artigos da Wikipédia estão disponíveis aos leitores. Com posse dessa informação, o objetivo deste trabalho é verificar a relação existente entre as métricas de análise de links e as notas de qualidade atribuídas aos artigos da Wikipédia. Para atingir este objetivo, a pesquisa foi dividida em etapas as quais serão relatadas.

Categories and Subject Descriptors

H.4 [**Sistemas Web**]: Recuperação de Informação, Análise de Links, Wikipédia

General Terms

Métricas de Análise de Links[qualidade de conteúdo]

Keywords

Qualidade, Wikipédia, Métricas de Análise de Links

*Estudante do programa de mestrado do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, com início em março de 2011 e defesa prevista para fevereiro de 2013

†Colaborador do Trabalho de Pesquisa

‡Orientadora do Trabalho de Pesquisa

1. INTRODUÇÃO

1.1 Contexto Teórico

A Web se expandiu ao longo das últimas décadas, se constituindo em uma enorme coleção de documentos, espalhados de forma descentralizada e desorganizada. Apesar disso, foram desenvolvidas ferramentas capazes de encontrar informações nesta coleção de forma bem sucedida. Tais ferramentas são as máquinas de busca. Elas permitem que um conjunto de páginas sejam buscadas de acordo com a sua relevância para um determinado conjunto de termos. O valor de relevância é atribuído por um algoritmo de *ranking*. Diversas evidências são combinadas pelos algoritmos de ranking para estipular valores de relevância. Dentre estas várias evidências estão as métricas que se baseiam em links. Neste caso, a Web pode ser vista como um grafo onde os nós são as páginas e as arestas são os hiperlinks. O estudo das propriedades e das relações nesse grafo chama-se Análise de Links.

Em análise de links, quando um autor de uma página cria um link para outra página, ele endossa o conteúdo da segunda, sugerindo que este é de qualidade e que a página apontada pode ser uma autoridade em um determinado assunto [9]. Outros fatores que contribuem para que uma página seja muito citada são a sua visibilidade e a sua popularidade [5], levando a ideia de que quanto mais uma página é apontada por outras maior é a sua relevância. Como consequência, páginas muito citadas obtêm posição de destaque em rankings de máquinas de busca.

Na literatura, são reportadas diversas métricas (e variações delas) para ordenar páginas aproveitando-se dos conceitos de análise de links [2]. O resultado desses métodos é avaliado por seres humanos que consideram o quão relevante são as páginas retornadas em relação aos termos dados. Uma vez que não se sabe determinar exatamente qual o grau de qualidade, de popularidade e de visibilidade de certa página,

não é possível verificar o impacto de cada um destes fatores para a relevância da mesma [1]. Porém, ao contrário de páginas da Web em geral, artigos da Wikipédia são explicitamente avaliados pela sua comunidade quanto à sua qualidade de conteúdo, o que fez com que muitos trabalhos sobre qualidade de páginas envolvendo a Wikipédia fossem publicados ([7], [13], [6]). Contudo, nenhum deles investigou, utilizando links internos e externos, em que grau métricas de análise de links são capazes de ordenar páginas da Wikipédia de acordo com a sua qualidade de conteúdo explicitamente indicada no processo de revisão humana.

1.2 Identificação do Problema

Existem na literatura vários trabalhos que consideram métricas baseadas em links para avaliar e ordenar páginas de acordo com a sua relevância [2]. Tais métricas foram elaboradas com base na hipótese de que links entre páginas podem indicar, entre outros fatores, maior ou menor qualidade de conteúdo [1]. Embora comum, até onde sabemos, esta hipótese não foi verificada adequadamente de forma a quantificar especificamente a importância do fator qualidade para os resultados obtidos pelas métricas de análise de links. Entre outros motivos, isso se deve ao fato de que páginas Web não possuem avaliação explícita de qualidade. Artigos da Wikipédia, entretanto, possuem avaliações de qualidade, o que nos permite determinar esta correlação. Contudo, artigos da Wikipédia podem possuir diferenças em relação a outras páginas Web, causando impacto no resultado de algumas métricas. A diferença existente entre o grafo de links da Web e da Wikipédia também deve ser estudada.

1.3 Objetivos

Este trabalho tem por objetivo investigar, com base na Wikipédia, qual é a relação existente entre a qualidade de conteúdo e os resultados obtidos pelas métricas de análise de links, utilizadas para avaliar e ordenar páginas de acordo com a sua relevância, e determinar como as conclusões observadas para Wikipédia podem ser estendidas para a Web em geral.

1.4 Contribuições Esperadas

Com este trabalho, esperamos fornecer as seguintes contribuições:

1. Verificar a correlação entre métricas de análise de links e qualidade de conteúdo através de um estudo em larga escala. Até onde sabemos, somente um trabalho analisou esta correlação [1]. Este, contudo, consistiu de um estudo limitado envolvendo poucas páginas, tópicos e avaliadores, o que motivou seus autores a afirmarem que um novo estudo, em maior escala, seria necessário para resultados mais conclusivos.

2. Possibilitar a criação de novas técnicas de inferência automática da qualidade de conteúdo. Esta é uma área de grande importância, dado seu impacto em várias outras aplicações como busca e recomendação. No caso particular da Wikipédia, a inferência automática de conteúdo também é desejável para a automatização de um processo essencialmente manual. Muitos trabalhos têm investigado esse problema, como por exemplo [16], [10] e [7].

3. Possibilitar a criação de novos algoritmos de ordenação em máquinas de busca através de uma melhor compreensão do impacto da qualidade em métricas de análise de links.

1.5 Organização

Este trabalho está organizado como descrito a seguir. A seção 2 aborda trabalhos relacionados a esta pesquisa e que deram conhecimento e suporte para a formulação dos procedimentos a serem seguidos e observados durante a resolução do problema; na Seção 3, são detalhadas as atividades que deverão ser realizadas para a execução do projeto; e, por fim, a Seção 4 apresenta quais atividades foram atingidas até o presente momento.

2. TRABALHOS RELACIONADOS

Durante o processo de revisão bibliográfica, foram encontrados trabalhos publicados na literatura que estão, de certa forma, relacionados ao trabalho proposto.

O trabalho considerado mais relevante para esta pesquisa é o realizado por Amento et. al (2000). Assim como desejamos fazer, os autores analisam a relação entre métricas de análise de links e qualidade de páginas. O principal problema encontrado neste trabalho é o pequeno número de especialistas, páginas e tópicos e as diferenças de opinião dos especialistas que dificultou a obtenção de resultados estatisticamente significativos. Ao usar a Wikipédia, poderemos confirmar seus resultados, pois possuiremos mais artigos revisados, cobrindo tópicos em um número maior de domínios e obtendo resultados estatísticamente significativos. Além disso, pretendemos considerar o impacto de outros fatores como popularidade e importância, que também podem ser obtidos de forma independente na Wikipédia.

Considerando que links na Wikipédia podem divergir de links na Web em geral, se assemelhando mais a citações em artigos cientificos, torna-se importante o trabalho de Smith et. al (2004), que aborda esta questão. Os autores concluem que citações e links mostraram-se semelhantes em apenas 20% da Web. Nesta mesma linha, Gleich et. al (2012), aplicaram o método PageRank sobre o grafo interno da Wikipédia, variando os valore do seu parâmetro de teletransporte (*coeficiente de dampenning*) e observaram que os rankings obtidos eram muito diferentes do esperado, levantando uma discussão sobre a diferença entre o grafo de links da Wikipédia e a Web. Finalmente, Kamps et. al (2009) realizaram um estudo e concluíram que a Wikipédia divergia da estrutura da Web em aspectos como a densidade de links, utilização de inlinks e outlinks para determinar importância da página, e utilização de evidências de links globais para realização de buscas. A discussão sobre as possíveis diferenças entre a Wikipédia e a Web são importantes, uma vez que desejamos saber o quanto é possível estender conclusões obtidas na Wikipédia para a Web em geral. Dessa forma, pretendemos realizar uma comparação da estrutura de links da Wikipédia com a nossa amostra da Web, similar à realizada por Kamps et. al (2009).

Outros trabalhos estudam o problema de estimativa automática de qualidade na Wikipédia, sugerindo o uso de informação de análise de links para esta tarefa (e.g. [6], [7] e [13]). Em todos os casos apresentados, os links usados

foram restritos à própria Wikipédia. Diferente destes trabalhos, pretendemos fazer uso de links internos e externos à Wikipédia, assim estaremos utilizando links independentes que não são usados com natureza enciclopédica e uma ampla vizinhança de páginas, necessárias para algumas métricas. Outra diferença é que nosso objetivo não é analisar métricas de links fora do contexto de previsão de qualidade na Wikipédia.

O trabalho apresentado por Berlt et. al (2010) faz adaptações de métricas agrupando páginas da Web por host e domínio e concluindo que todos os métodos apresentados tiveram desempenho melhor ou igual ao PageRank e ao Indegree tradicional com a vantagem adicional de serem menos sucestíveis a spam. Este trabalho está relacionado ao nosso estudo por apresentar adaptações dos métodos tradicionais que nós também pretendemos usar. Além disso, assim como faremos, Berlt et. al (2010) usam uma coleção derivada do domínio *br* como amostra da Web.

Finalmente, é importante verificar como fatores, tais como tópicos, influenciam no desempenho de métricas de análise de links no contexto da Wikipédia. Yamada et. al (2006) mostram que o comportamento dos links entre os artigos varia muito de acordo com a categoria ao qual eles pertencem. Assim, é interessante estudar como tais métricas se relacionam com qualidade quando artigos de diferentes áreas são considerados.

3. METODOLOGIA

O trabalho a ser realizado foi dividido nas etapas que estão descritas a seguir.

3.1 Estudo sobre a coleção e dados de interesse

O objetivo desta atividade é determinar quais dados são relevantes e devem ser capturados. Para a realização deste trabalho, é necessário uma amostra da Web com páginas que foram avaliadas quanto a sua qualidade (artigos da Wikipédia, neste trabalho). As bases obtidas foram as seguintes:

1. WBR10: base coletada no contexto do projeto INWEB[1], composta por documentos do domínio *br*. Essa coleção foi obtida em 2010 e contém 125 milhões de documentos HTML, com 6,8 bilhões de links entre eles, o que corresponde à maior amostra da Web brasileira disponível para estudo.

2. Ptwiki dump progress de 20100804: disponibilizada para o público pela Wikipédia no formato XML e diversos dumps SQL. A coleção corresponde ao mesmo período de coleta da base WBR10 e possui 2.363.341 páginas e 20.311.293 revisões.

Optamos por uma amostra da Wikipédia em português, visto que é provável que editores de páginas Web brasileiras citem mais artigos da Wikipédia em língua portuguesa do que em outros idiomas. Deve-se observar que os artigos da Wikipédia pertencem ao domínio *org* e, portanto, não fazem parte da coleção WBR10. Para a execução dos experimentos é necessária a união da coleção WBR10 com os artigos da Wikipédia.

Nesta etapa, desejamos verificar quais informações devem ser extraídas das coleções. Da amostra da Web WBR10, em princípio, estamos interessados no grafo para análise de links, podendo este ser enriquecido com informação de texto de âncora. Da amostra da Wikipédia, estamos interessados na qualidade do artigo, na popularidade (métricas de análise de links dão alta pontuação para páginas populares, devendo esse efeito ser considerado), e na categoria do artigo (útil para verificar o quanto a análise realizada vale para diferentes tópicos). Outras evidências podem ser consideradas de interesse para a pesquisa.

3.2 Construção das bases e extração dos dados

Esta etapa consiste na implementação dos extratores necessários para obter os dados de interesse, descritos anteriormente. Na Wikipédia, serão eliminados artigos que não possuem nota de qualidade e que não são citados nas páginas da WBR10 ou não citam páginas da WBR10. As informações necessárias serão extraídas por meio de consultas em SQL e depois combinadas às informações da coleção WBR10. Para obter o grafo de links final, serão utilizados ferramentas disponibilizadas com a própria coleção WBR10.

3.3 Comparação sistemática da Web e da amostra da Wikipédia

Como o objetivo deste trabalho é estabelecer a relação entre qualidade de conteúdo e métricas de análise de links na Web, é necessário determinar até que ponto a Wikipédia é representativa da Web. É importante observar que assumimos que a coleção WBR10 é representativa da Web. Baseados em trabalhos na literatura, sabemos que a Wikipédia possui diferenças em relação a amostras da Web, sendo importante determinar de forma sistemática como estas coleções (WBR10 e Wikipédia) se assemelham em termos da sua estrutura de ligações. Em particular, a comparação deve avaliar, pelo menos, características dos grafos de páginas, como a densidade de links e as distribuições de inlinks e outlinks.

3.4 Desenvolvimento e aplicação das métricas de ranking

Serão implementadas métricas encontradas na literatura baseadas em grafos (Indegree e Outdegree [4]) e baseadas em links (PageRank [11] e HITS [9]). Da mesma forma que foi feito por Berlt et. al (2010), as métricas serão usadas considerando páginas isoladas ou agrupadas em hosts e domínios. Será usada a mesma definição de hosts e domínios usada por Berlt et al. (2010). As métricas serão implementadas provavelmente utilizando as linguagens de programação C e/ou C++.

As métricas implementadas serão aplicadas sobre o grafo de páginas que integra as páginas da WBR10 e da Wikipédia. A aplicação de cada métrica implementada resultará em uma pontuação para cada artigo da Wikipédia que determinará a posição do artigo no ranking, permitindo a sua

[1] http://www.inWeb.org.br

comparação com a posição que teria no ranking baseado em sua qualidade.

3.5 Análise dos resultados e considerações finais

Para verificar a relação entre as métricas de análise de links e qualidade, usaremos uma estratégia de comparação de rankings. A ideia consiste em se obter rankings dos artigos da Wikipédia de acordo com cada métrica estudada e de acordo com a sua qualidade e aplicar métricas tradicionais de análise de ranking (Pearson [12] e Kendall Tau [8]). As métricas também serão comparadas entre si, de forma a permitir determinar qual a mais apropriada para o ambiente da Wikipédia. Além disso, as verificações deverão considerar os outros fatores que podem influenciar esta análise, como a popularidade e o tópico dos artigos, necessitando que outras comparações sejam realizadas.

Para o caso das métricas influenciadas pelo grafo interno da Wikipédia, as conclusões serão ponderadas de acordo com o estudo anterior sobre a similaridade entre a Wikipédia e a a Web. Em todas as análises, utilizaremos métodos estatísticos para garantir que conclusão serão tiradas apenas de resultados significativos em termos estatísticos.

4. CONTRIBUIÇÕES: ESTADO ATUAL DO TRABALHO

Inicialmente, foi feito um estudo preliminar no qual foram levantadas quais pesquisas estão sendo realizadas sobre a Wikipédia. A partir deste levantamento pudemos fundamentar esta pesquisa e encontrar trabalhos relacionados a ela.

O próximo passo foi obter as bases necessárias para a realização dessa pesquisa. Em parceria, com o grupo de pesquisa BDRI (Banco de Dados e Recuperacao de Informacao), da Universidade Federal do Amazonas, tivemos acesso à base WBR10 que contém informações sobre páginas com domínio *br*. Devido ao fato de a Wikipédia manter suas informações abertas, pudemos adquirir uma amostra da Wikipédia no mesmo período de coleta da WBR10 (Ptwiki dump progress de 20100804), sendo essa formada por arquivos XML e dumps SQL.

A amostra da Wikipédia possui arquivos em XML e dumps SQL. Foram realizadas operações para que pudéssemos possuir as informações da amostra em um banco de dados de fácil acesso. Na base WBR10, as informações estão armazenadas em imensos arquivos textuais. Está em andamento a implementação dos algoritmos pra manipulação dos dados existentes nas coleções escolhidas. Uma vez que todos os dados estierem em formatos fáceis de serem acessados, é possível fazer a combinação das duas bases para obter o novo grafo de links e ter acesso às informações importantes da Wikipédia.

5. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, 2000.

[2] R. Baeza-Yates, P. Boldi, and C. Castillo. Generalizing pagerank: damping functions for link-based ranking algorithms. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 308–315, New York, NY, USA, 2006. ACM.

[3] K. Berlt, E. S. de Moura, A. Carvalho, M. Cristo, N. Ziviani, and T. Couto. Modeling the web as a hypergraph to compute page reputation. *Inf. Syst.*, 35:530–543, July 2010.

[4] T. Bray. Measuring the web. *Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems*, page 993Ú1005, 1996.

[5] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 20–29, New York, NY, USA, 2004. ACM.

[6] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado. Automatic assessment of document quality in web collaborative digital libraries. *J. Data and Information Quality*, 2(3):14:1–14:30, Dec. 2011.

[7] P. Dondio and S. Barrett. Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipédia Project. *Informatica*, 31(2):151–160, 2007.

[8] M. Kendall. *Rank Correlation Methods*. Hafner Publishing Co, New York, 1955.

[9] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *JOURNAL OF THE ACM*, 46(5):604–632, 1999.

[10] E.-P. Lim, B.-Q. Vuong, H. W. Lauw, and A. Sun. Measuring qualities of articles contributed by online communities. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 81–87, Washington, DC, USA, 2006. IEEE Computer Society.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

[12] K. Pearson. The grammar o science. J. M. Dent and Company, 1892.

[13] L. Rassbach, T. Pincock, and B. Mingus. Exploring the Feasibility of Automatically Rating Online Article Quality, 2008.

[14] A. G. Smith. Web link as analogues of citations. *School of Information Management*, 2004.

[15] K. S. T. Yamada and K. Kazama. Network analyses to understand the structure of wikipédia. *In Proc. of AISB'06*, 3:195–198, 2006.

[16] B. S. M. B. Twidale. Assessing information quality of a community-based encyclopedia. In *In Proceedings of the International Conference on Information Quality*, pages 442–454, 2005.