

A Webometria e as restrições dos mecanismos de busca: crise ou oportunidade?

Eixo temático: Informetria e Webometria

Modalidade: Apresentação oral

1 INTRODUÇÃO

Criada a cerca de 15 anos em um campo fértil da informetria que surge no desenvolvimento das TICs, a webometria vivencia hoje um cenário de crise e novos desafios diante das restrições de acesso a informação impostos pelos mecanismos de busca, colocando em cheque tradicionais análises de hiperlinks e iniciando um debate crucial entre os pesquisadores: como trabalhar e obter daqui para frente dados webométricos? Quais os caminhos possíveis e os mais indicados para obtenção de métricas na web?

Historicamente, o principal objeto de estudo da webometria é o *link*, considerado unidade central de informação e a ligação direta entre os atores (páginas) que compõem essa vasta rede virtual. Em sua origem, os estudos webométricos se basearam na obtenção desta rede virtual por intermédio de *web crawlers* especializados que levantavam todo o conteúdo interligado dos websites estudados e faziam o mapeamento dos *links* existentes. Mudanças tecnológicas que tornavam os *links* inacessíveis (ex: conteúdos em flash, etc), bem como o acesso restrito apenas a web visível sempre foram desafios. Além disso, a própria instabilidade da manutenção dos conteúdos na Web, parte da dinâmica natural de seu desenvolvimento, já foram motivo de debate e questionamento (BAR-ILAN, 1999; ROUSSEAU, 1999). O próprio significado do *link* e sua motivação, bem como a analogia com a citação motivou ao longo dos anos estudos voltados para o melhor entendimento da dinâmica de interligação entre páginas na web (PRIME *et al.*, 2002; ZUCCALA, 2006).

Como um campo de pesquisa ainda recente e diretamente relacionado ao avanço das tecnologias e da ciência da computação, a webometria tem se debruçado, ao longo dos anos, sobre estudos de cunho exploratório que visam testar ferramentas e novas metodologias, e investigar possibilidades de aplicação na cientometria (LANG *et al.*, 2010; HERRERO-SOLANA & MORALES-DEL-CASTILLO, 2004; ORTEGA *et al.* 2008; VAUGHAN, 2006), e para além dela, como na área de competitividade comercial (VAUGHAN & YOU, 2006) e de Museus e Centros de Ciência (GOUVEIA & KURTENBACH, 2009).

Com o passar do tempo, estabeleceram-se como fonte frequente de informação para os estudos webométricos os mecanismos de busca comerciais, mas suas funcionalidades variavam e a escolha por determinado mecanismo dependia, prioritariamente, do tipo de busca pretendida. Nos últimos anos, o Yahoo! era o único que ainda permitia o uso de buscas combinadas (*co-links*) e automatizadas por meio de APIs (*Application Programming Interface*). A atual crise tem início em 2009, quando o Yahoo! é adquirido pela Microsoft, que já havia interrompido a possibilidade de buscas avançadas no Bing e desejava migrar a base de dados do Yahoo!. Em abril de 2011, o uso de APIs foi totalmente bloqueado pelo Yahoo!, impossibilitando buscas automatizadas em larga escala (YAHOO!, 2012). No início de 2012 já não era mais possível realizar qualquer tipo de busca por hiperlinks.

Se por um lado o momento atual representa uma crise, por outro, gera oportunidades de desenvolvimento de novas metodologias para o campo. Diante disso, pretendemos a seguir aportar algumas reflexões e explorar possibilidades que vem sendo discutidas entre os pesquisadores, como retorno ao uso de *web crawlers* e a citação da URL e da palavra em mecanismos de busca comerciais, além de trazer uma abordagem ainda pouco explorada, por meio do uso de um operador do Google, conhecido como *allinanchor*.

2 DESAFIOS AO LONGO DO TEMPO

A webometria – termo cunhado por Almind & Ingwersen (1997) – trata dos aspectos quantitativos tanto da construção, quanto do uso da Web, compreendendo quatro áreas de pesquisa no campo: análise de conteúdo das páginas web; análise da estrutura dos *links*; análise do uso da Web; e análise de tecnologias da Web (THELWALL *et al.*, 2003).

No final da década de 1990, os principais mecanismos de busca utilizados eram o AltaVista e o Infoseek, em razão das funcionalidades de busca oferecidas como o uso de operadores booleanos. No entanto, à época, ainda havia poucos estudos sobre o processo de indexação por esses mecanismos, o que levantou questionamentos sobre a confiabilidade dos dados e a opção de pesquisadores pelo uso de *web crawlers* (THELWALL, 2001). De fato, deve-se considerar que a indexação destes mecanismos não é completa. Gulli & Signorini (2005), em um estudo de cobertura dos mesmos verificaram que da web visível o Google indexava 79%, o Yahoo/AltaVista, 69%, o MSN, 62% e o Ask/Teoma, 58% sendo que por

terem conjuntos distintos, apenas 28,85% das páginas eram indexadas simultaneamente pelos quatro mecanismos de busca.

Vanti (2002) alertava para a natureza dinâmica e em tempo real da Internet, de sua enorme massa de dados não padronizados, bem como das inconsistências entre os mecanismos de busca. A preocupação com a estabilidade dos dados coletados por meio destes foi debatida entre os pesquisadores em busca de possíveis soluções (BAR-ILAN, 1999; ROUSSEAU, 1999; METTROP & NIEUWENHUYSEN, 2001). Dez anos após a criação formal da webometria, Payne e Thelwall (2007) realizam um estudo longitudinal (2000 a 2005) sobre universidades apontando para uma tendência de estabilização da rede acadêmica.

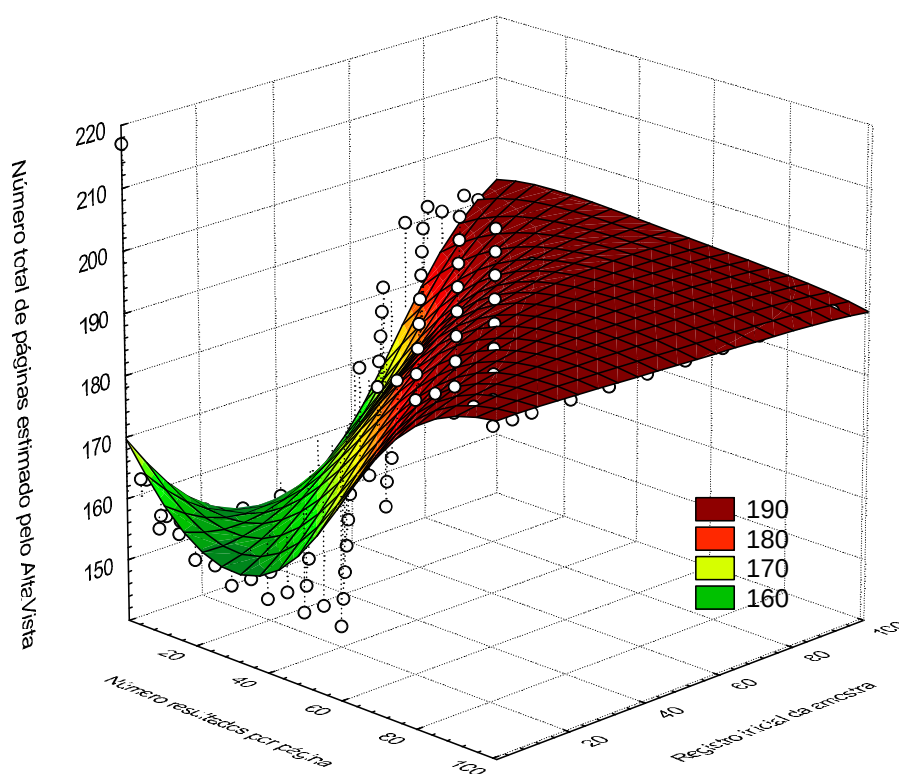


Figura 1 - Variação da estimativa de resultados para um mesmo link partindo de diferentes registros iniciais e números totais de resultados apresentados por página (GOUVEIA, 2007)

Com relação aos mecanismos de busca, a estimativa do número de páginas que satisfazem uma determinada consulta também podia variar. Com o intuito de padronizar a

metodologia aplicada à pesquisa no AltaVista, foi conduzido um estudo (GOUVEIA, 2007) para verificar qual a melhor abordagem para se obter valores mais estáveis. Variando os registros iniciais entre 1 e 101 e o número de resultados por página de 1 a 100 foi obtido uma matriz de resultados que demonstravam uma tendência a estabilização quando o número de resultados por página era ampliado para 100 (figura 1).

Em 2005, no entanto, o AltaVista foi comprado pelo Yahoo! e houve uma fusão das bases de dados de ambos os mecanismos. Os pesquisadores passaram então a optar pelo uso do Yahoo! para condução de seus estudos e quatro anos mais tarde, a compra do mesmo pela Microsoft anunciou uma crise para grande parte dos estudos webométricos.

3 ALTERNATIVAS

As duas principais abordagens que vem sendo discutidas se referem à citação da URL e citação da palavra (THELWALL & SUD, 2011; VAUGHAN & YOU, 2011). Estas exploram as possibilidades de se continuar utilizando mecanismos de busca comerciais a partir de uma nova perspectiva, que não tem mais o *link* como unidade central da informação.

A citação da URL se refere à busca textual pelo endereço de uma dada página na web (ex. www.brasil.gov.br). Já a citação da palavra corresponde à menção textual de um nome ou termo em uma dada página, sem necessariamente estar associado a um *link*. Este é um tipo de busca que sempre existiu, mas que até então não era objeto de estudos webométricos.

No caso da contagem de *links*, embora reconheça que todo indicador tem suas limitações, incluindo os *inlinks*, e não deve utilizado isoladamente como medida de impacto ou qualidade, um estudo conduzido por Thelwall & Sud (2011) comparou os métodos de citação da URL e da palavra, utilizando dois mecanismos de busca (Yahoo! e Bing). Os resultados obtidos para universidades britânicas validaram ambos os métodos como medida de impacto e apontaram para uma correlação positiva com a contagem de *inlinks*.

Em outro estudo, Vaughan & You (2009) compararam os resultados de *co-inlink* de uma pesquisa prévia com a busca textual pelos nomes das instituições estudadas, no intuito de investigar a validade do método como medida de similaridade e proximidade na web. A similaridade observada entre os dois diagramas de MDS gerados a partir dos *co-inlinks* e da co-ocorrência de palavras apontou para uma correlação positiva entre as metodologias.

No entanto, apesar das correlações encontradas, a citação da palavra não estabelece uma relação direta entre as páginas e não representa uma intencionalidade do produtor de conteúdo na indicação do outro ente virtual. Perde-se, portanto, a premissa de legitimidade conferida por intermédio do *link*. Outra desvantagem reside no fato de um mesmo termo pode ter significados distintos ou ter variações, complicando a análise dos dados.

Já a citação da URL, embora se refira ao *link* e, nesse sentido, mantenha uma relação direta entre as páginas, não segue à lógica da construção de conteúdo na web. Não é comum que a URL seja textualmente escrita nas páginas web, mas sim inserida no texto âncora, não deixando claro se ela poderia ser considerada, de fato, uma boa fonte ou evidência de impacto na internet. Neste sentido, a principal desvantagem seria conceitual.

No contexto das alternativas frente ao desafio colocado pelos mecanismos de busca, duas metodologias que ainda não têm ocupado um papel central nesse debate, merecem destaque por manter a relação de análise com foco no *link*. A primeira delas é a volta do uso de *web crawlers*. A principal crítica a este método está no fato de estes encontram novas páginas na web seguindo os links existentes nelas. Por isso, é possível que páginas ou coleções de páginas que estejam isoladas, com links quebrados ou ainda em JavaScript ou Flash podem não ser conseguir ser rastreadas.

Em uma abordagem mais recente e inovadora, Friedrich *et al.* (2009) efetuaram um levantamento de websites atuantes em uma determinada temática fazendo uso do texto âncora dos links levantados no Google. Esta metodologia se baseia no operador “allinanchor” que fornece como resultado a lista de páginas que recebem links com um determinado texto âncora. Com isto, passa a ser possível o levantamento de websites que abordam questões temáticas que podem estar vinculadas a termos específicos presentes nos links distribuídos pela web. A vinculação efetuada pela delimitação da busca ao termo âncora traz uma vantagem em relação aos métodos tradicionais de busca por palavra-chave dentro dos mecanismos de busca, pois a necessidade de tratamento dos dados diminui e a questão conceitual relativa a validação do vínculo entre as páginas que o link se resolve.

4 CONSIDERAÇÕES

Diante de um cenário de crise, novas oportunidades sempre se apresentam. O campo de estudo da webometria sempre esteve diante das mudanças que o desenvolvimento das TICs

apresentam a cada dia. Mensurar, avaliar e mapear as redes de relações na web e o impacto de seus conteúdos estará sempre entre as necessidades deste campo. Assim sendo, o constante desenvolvimento de metodologias e a validação das mesmas fazem parte da dinâmica deste campo de estudo, e devemos considera-los como oportunidades de propor novas saídas criativas para as barreiras impostas.

AGRADECIMENTOS

O autor gostaria de agradecer o apoio da FAPERJ.

REFERÊNCIAS

- ALMIND, T.C.; INGWERSEN, P. (1997) Informetric analyses on the world wide web : methodological approaches to 'Webometrics'. *Journal of Documentation*, v. 53, n. 4, p. 404-426.
- BAR-ILAN, J.(1999) Search engine results over time – a case study on search engine stability. *Cybermetrics*, Madri. v. 2, a. 1. Disponível em: <http://cybermetrics.cindoc.csic.es/articles/v2i1p1.html> Acesso em: Mai. 2012.
- FRIEDRICH, M. P.; GOUVEIA, F. C.; LETA, J. (2009) How is the term Environmental Education linked on the Internet? A study of its representation on the web environment. *In: International Conference of the International Society in Scientometrics and Informetrics (ISSI)*, 2009, Rio de Janeiro, v. 2, p. 930-931.
- GOUVEIA, F. (2007) *Estudos webométricos de associações de museus e centros de ciência*. Rio de Janeiro, Tese (Doutorado em Química Biológica) – Universidade Federal do Rio de Janeiro, Rio de Janeiro.
- GOUVEIA, F.; KURTENBACH, E. (2009) Mapping the web relations of science centres and museums from Latin America. *Scientometrics*, v. 79, n. 3. p. 491-505.
- GULLI, A.; SIGNORINI, A. (2005) The Indexable Web is More than 11.5 billion pages. *International World Wide Web Conference*, p. 902-903. Disponível em: <http://www.divms.uiowa.edu/~asignori/papers/the-indexable-web-is-more-than-11.5-billion-pages/size-indexable-web.pdf> Acesso: Mai. 2012.
- HERRERO-SOLANA, V.; MORALES-DEL-CASTILLO, J. (2004) "Geopolitical" maps of the Internet: application of new information representation methods. *Ciência da Informação*, v. 33, n. 3, p. 69-75.
- LANG, P.; GOUVEIA, F. C.; LETA, J. (2010). Site co-link analysis applied to small networks: A new methodological approach. *Scientometrics*, v. 83, n. 1, p. 157-166.
- METTROP, W.; NIEUWENHUYSEN, P. (2001) Internet search engines – fluctuations in document accessibility, *Journal of Documentation*, v. 57, n. 5, p. 623-651.

ORTEGA, J. L.; AGUILLO, I.; COTHEY, V.; SCHARNHORST, A. (2008). Maps of the academic web in the European Higher Education Area: an exploration of visual web indicators. *Scientometrics*, v. 74, n. 2, p. 295-308.

PAYNE, N.; THELWALL, M. (2007). A longitudinal study of academic webs: Growth and stabilisation, *Scientometrics*, v. 71, n. 3, p. 523-539

PRIME, C.; BASSECOULARD, E.; ZITT, M. (2002) Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics*, v. 54, n. 2, p. 291-308.

ROUSSEAU, R. Daily time series of common single word searches in AltaVista and Northern Light, *Cybermetrics*, Madri, 1999. Disponível em:
<http://cybermetrics.cindoc.csic.es/articles/v2i1p2.html> Acesso: mai. 2012.

THELWALL, M. (2001) A Web crawler design for data mining, *Journal of Information Science*, v. 27, n. 5, p. 319-325.

THELWALL, M. (2003) What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, v. 8, n. 3, a. 151. Disponível em: <http://informationr.net/ir/8-3/paper151.html> Acesso: Mai. 2012.

THELWALL, M.; SUD, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology*, v. 62, n. 8, p. 1488–1497.

VANTI, N.A.P. (2002) Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ciência da Informação*, v. 31, n. 2, p.152-162.

VAUGHAN, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology*, v. 57, n. 9, p. 1178-1193.

VAUGHAN, L.; YOU, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, v. 68, n. 3, p. 611-628.

VAUGHAN, L.; YOU, J. (2010). Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept. *Journal of Informetrics*, v. 4, n. 4, p. 483-491.

YAHOOO! (2012). Web Search APIs from Yahoo! Search. Disponível em:
<http://developer.yahoo.com/search/web/webSearch.html> Acesso: Mai. 2012.

ZUCCALA, A. (2006). Author Cocitation Analysis is to Intellectual Structure As Web Co-link Analysis is to... ? *Journal of the American Society for Information Science and Technology*, v. 57, n. 11, p. 1487-1502.