

Análise quantitativa e temporal do Wikigrafo-PT

Marcelo Zambiasi^{1 2}, Thiago A. Presa^{1 2}, Luciana S. Buriol¹, Viviane M. Orengo¹

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

²Programa de Educação Tutorial em Computação

{mzambiasi,tapresa,buriol,vmorengo}@inf.ufrgs.br

Abstract. *The Wikipedia is an online encyclopedia available in about 200 languages. Its Portuguese version currently contains over 200 thousand articles. If we consider each Wikipedia article as a vertex and each link as an arc, we have what we call a “Wikigraph”. This graph differs from other Web mainly graphs because it has temporal information associated to its nodes.*

The aim of this paper is to do a quantitative analysis of the Portuguese Wikigraph’s (Wikigraph-PT) temporal data. The analysis of this graph revealed that it presents features that are commonly found on other Webgraphs, which confirms results of prior studies on the English Wikigraph (Wikigraph-EN).

Resumo. *A Wikipédia é uma enciclopédia online disponível em cerca de 200 línguas que possui mais de 200 mil artigos na sua versão em português. Se considerarmos cada artigo da Wikipédia como um vértice e cada hiperlink como um aresta (direcionada) temos o que chamamos de um Wikigrafo. Além da estrutura topológica, os wikigrafos contêm informações temporais que permitem reproduzir um histórico da evolução dos dados.*

O objetivo deste trabalho é analisar quantitativamente os dados temporais do Wikigrafo da língua portuguesa (Wikigrafo-PT). A análise deste grafo revelou que o mesmo apresenta características comumente encontradas em Webgrafos, confirmando resultados anteriormente encontrados no Wikigrafo da língua inglesa (Wikigraph-EN).

1. Introdução

O surgimento da Web alterou a vida da sociedade como um todo, tornando-se em poucos anos, um dos principais meios de comunicação utilizados pela humanidade. Um dos fatores que certamente propiciam o seu uso e sua difusão é a interatividade, característica não encontrada em outros meios de comunicação como rádio e televisão. Por ter se tornado um meio popular para busca e divulgação de informações, a quantidade de dados disponibilizada online tem crescido de forma acentuada. Estudos realizados em 2005 [Gulli and Signorini 2005] estimam que o tamanho da Web “indexável” excede 11,5 bilhões de páginas. *Sites* especializados que contabilizam a dimensão da Web¹ confirmam este resultado e apresentam estimativas de um aumento de pelo menos 1 bilhão de páginas no ano de 2006.

É inviável organizar manualmente esta grande quantidade de informação de forma a atender a uma consulta de maneira rápida e satisfatória. O estudo de modelos para a

¹<http://www.worldwidewebsize.com>

Web torna-se importante para organizar este volume de informações sem a intervenção humana. Um dos modelos mais empregados para a Web são os grafos, uma vez que a natureza das páginas (conteúdo e referências) é naturalmente mapeada para esta estrutura. O estudo de Webgrafos (grafos formados a partir da estrutura de *links* das páginas Web) contribui principalmente para o aprimoramento de algoritmos para ferramentas de busca. Uma outra aplicação importante é o desenvolvimento de modelos estocásticos para a geração de grafos que capturam as propriedades da Web [Laura et al. 2003].

Em um estudo recente [Ntoulas et al. 2004] observou-se que após um ano apenas 24% dos *hiperlinks* continuam presentes na Web, ao passo que 50% do conteúdo permaneceu inalterado. Uma vez que muitos algoritmos de classificação de páginas baseiam-se na estrutura de *hiperlinks*, como PageRank descrito em [Brin and Page 1998, Pandurangan et al. 2002], podemos observar a relevância da informação temporal associada às referências.

Nossa abordagem para contribuir com este estudo faz uso da Wikipédia. A Wikipédia² é uma enciclopédia online com mais de 5 milhões de artigos em 250 línguas. Sua principal característica, que a difere de outras enciclopédias, é a autoria colaborativa, i.e. qualquer usuário pode editar seu conteúdo. Foi criada em 15 de Janeiro de 2001, com a publicação de cerca de 25 artigos em inglês. Sua versão em português foi criada na metade daquele mesmo ano e conta atualmente com mais de 200 mil artigos, correspondendo à oitava maior base de dados. A maior base de dados é a da língua inglesa que possui mais de 1,5 milhão de artigos e foi recentemente analisada [Buriol et al. 2006].

Uma coleta de dados da Web (*Web crawling*) faz a busca e o armazenamento do conteúdo de páginas Web. O confronto de coletas realizadas em diferentes épocas fornece uma base para estimativas sobre o futuro da Web [Modesto et al. 2005]. Esta é uma das grandes facilidades da Wikipédia, já que toda a informação temporal está associada a cada vértice. Com isso podemos saber o estado exato do grafo em um determinado instante, o que não é possível em coletas na Web [Baeza-Yates and Ribeiro-Neto 1999].

O restante deste artigo é organizado da seguinte maneira. A Seção 2 apresenta as características da base de dados. A Seção 3 descreve o processo de geração do grafo da Wikipédia (*Wikigrafo*). A Seção 4 relata os principais resultados obtidos com a análise quantitativa e temporal do Wikigrafo. Por fim, a Seção 5 traz as conclusões deste estudo.

2. Características da Base de Dados

Existem diversas razões que nos levam a considerar a Wikipédia como uma boa base de dados para a realização de experimentos do tipo Webgrafo. Alguns aspectos importantes foram apresentados em [Buriol et al. 2006], dentre os quais podemos destacar:

- **Diversidade de participantes:** por se tratar de uma enciclopédia em que qualquer usuário (cadastrado ou não) pode editar seu conteúdo, seus editores constituem um grupo altamente heterogêneo. Além disso, cada usuário tem a liberdade de colocar o conteúdo que desejar em um artigo, apenas devendo respeitar o modelo pré-definido. Sabe-se que autonomia é uma característica bastante comum no contexto de criação de conteúdo na Web, isto não é diferente na Wikipédia. A comunidade

²<http://www.wikipedia.org>

que a edita é composta basicamente por pessoas que têm o intuito de disponibilizar conteúdo sobre tópicos de conhecimento que considerem relevantes.

- **Independência de *links* externos:** os artigos da Wikipédia apontam, em sua maioria, apenas para outros artigos da Wikipédia. Quando existem referências externas, estas podem ser simplesmente deletadas sem descaracterizar o grafo gerado. Logo não é difícil isolar um Wikigrafo do resto da Web.
- **Informação temporal:** cada artigo contém todo seu histórico de modificações desde a sua criação. Logo, é possível resgatar o exato estado em que um determinado artigo - ou toda base de dados - se encontrava em um determinado instante.

Por ser uma enciclopédia de conteúdo livre, as bases de dados de todos os idiomas estão disponíveis para download ³. A base de dados é formada, principalmente, por um gigantesco arquivo XML compactado, sendo que a versão atual deste arquivo ocupa 25,5 GB quando descompactado. Neste arquivo estão contidas todas as informações de páginas atuais - sejam elas artigos, páginas de discussão, páginas de usuários, além de todo o histórico de atualizações. Um outro importante arquivo de cada base de dados é um *script* SQL, denominado “page.sql”⁴, para gerar uma tabela com diversas informações sobre cada página (metadados), excetuando-se seu conteúdo. Abaixo descreveremos as principais características do arquivo XML.

O arquivo “pages-meta-history.xml” contém todo o texto de cada artigo para cada atualização, desde a sua criação. No início de cada arquivo é descrito a qual *namespace* a página pertence. Um *namespace* define se a página é um artigo propriamente dito, se é uma página administrativa da Wikipédia, se é uma página de conteúdo de um respectivo usuário, se é uma página de discussão sugerindo possíveis mudanças em um artigo, entre outros. Os principais campos deste arquivo são:

- **<page>**: marcador de início de artigo;
- **<title>**: título da página e seu *namespace*. Se a página é um artigo, o *namespace* não existe;
- **<restrictions>**: restrições de atualização da página, como por exemplo: somente usuários administradores podem editar;
- **<revision>**: inicia uma revisão, ou seja, alguma atualização ou até mesmo a criação de um artigo;
- **<id>**: identificador único da página.
- **<timestamp>**: o tempo exato em que uma revisão foi criada, no formato Ano-Mês-DiaTHora:Minuto:SegundoZ, por exemplo: 2006-11-07T12:00:00Z;
- **<comment>**: comentário sobre as alterações feitas na página. Através deste campo pode-se saber, por exemplo, se uma página foi revertida para alguma versão anterior;
- **<contributor>**: grava o nome do usuário, através da tag <username>, que realizou a edição da página. Neste campo também pode ser gravado o nome do programa (robô) que fez uma edição automática, sendo que estas são geralmente utilizadas para adaptar artigos a novos padrões, criar *hiperlinks* ou corrigir o duplo-redirecionamento entre páginas. Além disso, se o usuário que realiza a alteração não está cadastrado, a tag <IP> armazena o IP do editor;

³<http://download.wikipedia.org>

⁴Maiores detalhes sobre o arquivo podem ser obtidos em http://meta.wikimedia.org/wiki/Page_table

- **<text>**: conteúdo textual da página, juntamente com seus *hiperlinks*;
- **<minor>**: indica que o usuário marcou a caixa *minor edit* durante a edição de uma página. Isto significa que sua edição não foi substancial ao conteúdo do artigo. Esta opção é usada, geralmente, em atualizações para corrigir a grafia de palavras, formatação de um texto ou remoção de vandalismo.

3. Gerando o Wikigrafo

Um Wikigrafo é um grafo direcionado em que cada artigo é considerado um vértice e cada referência entre artigos (*hiperlink*) é representada por um arco. Existe a possibilidade de um artigo apontar para alguma página da Web fora da enciclopédia ou para algum artigo de um outro idioma da Wikipédia. A fim de isolar o grafo, todos estes *links* externos são desconsiderados para a geração do Wikigrafo. Somente foram considerados os artigos de enciclopédia propriamente ditos, as páginas pertencentes aos demais *namespaces* foram excluídas. Além disso, os artigos redirecionados, ou seja, aqueles cuja única função é apontar para um outro artigo, como o artigo “UFRGS” que é redirecionado para o artigo “Universidade Federal do Rio Grande do Sul”, também foram excluídos. Por exemplo, se a página X é redirecionada para a página Y, então a página X é deletada do grafo e todos os arcos que chegam a X são redirecionadas para Y. Logo, os *hiperlinks* de artigos redirecionados não foram perdidos na geração do Wikigrafo. Na Wikipédia em português, aproximadamente 30% de todos os artigos são redirecionados.

Com o objetivo de analisar o crescimento da Wikipédia em português, foram geradas 12 imagens diferentes (*snapshots*) do Wikigrafo-PT com intervalos de três meses. Estas imagens contêm todas as informações consideradas relevantes da Wikipédia na determinada data. A primeira imagem do grafo é de 7 de Fevereiro de 2004 que continha 1.594 artigos. A última é de 7 de Novembro de 2006 na qual o número de artigos era 199.642. Para gerarmos os grafos a partir do arquivo “pages-meta-history.xml” foi utilizada uma série de *scripts* em PERL desenvolvida por [Buriol et al. 2006]. Foram feitas algumas alterações sobre os *scripts* para adequá-los ao formato de dados da Wikipédia que evoluiu desde a realização do trabalho citado. Além de arquivos de texto, também foi utilizado um banco de dados relacional MySQL para o armazenamento das informações do Wikigrafo e a biblioteca COSIN⁵, para a análise de grafos de grande dimensão.

4. Resultados Obtidos

Esta seção descreve os principais resultados obtidos através da análise quantitativa e temporal do Wikigrafo-PT. A análise de dados dá-se em três etapas. Inicialmente uma análise estatística da dinâmica das páginas é apresentada e discutida. Em seguida, algumas propriedades básicas do grafo gerado são estudadas e, finalmente, a estrutura topológica deste é apresentada.

4.1. Crescimento da Wikipédia

Notadamente, a Wikipédia tem aumentado sua popularidade nos últimos anos, tornando-se um dos principais *sites* de conteúdo de toda Web. Na Fig.1 podemos perceber que o crescimento desta enciclopédia, na sua versão em língua portuguesa, aconteceu de forma exponencial considerando-se o número de artigos, atualizações e editores.

⁵<http://www.dis.uniroma1.it/~cosin/html/pages/COSIN-Tools.htm>

O número de artigos (Fig.1-A) da Wikipédia teve um alto crescimento no período analisado de 33 meses, aumentando de 1.594 para 199.642, sendo que a taxa de crescimento atual está em 16,78% (entre as duas últimas imagens do grafo). Por sua vez, o número de atualizações (Fig.1-B) ocorridas, entre cada imagem do grafo, passou de 2.040 para 558.239 neste mesmo período. Considerando as duas últimas imagens do grafo, a taxa de crescimento atual do número de atualizações é de 13,70%. Esse aumento do número de atualizações mostra que, com o passar do tempo, os usuários estão dedicando mais atenção a cada artigo, ou seja, adicionando maior quantidade de conteúdo ou fazendo mais correções. Quanto ao número de editores distintos (Fig.1-C) da Wikipédia, por imagem do grafo, o aumento foi de 692 para 80.822 nos 33 meses analisados, sendo que a taxa de crescimento atual está em 36,49% por trimestre. Isto demonstra que um maior número de pessoas está interessado em adicionar conteúdo à enciclopédia livre.

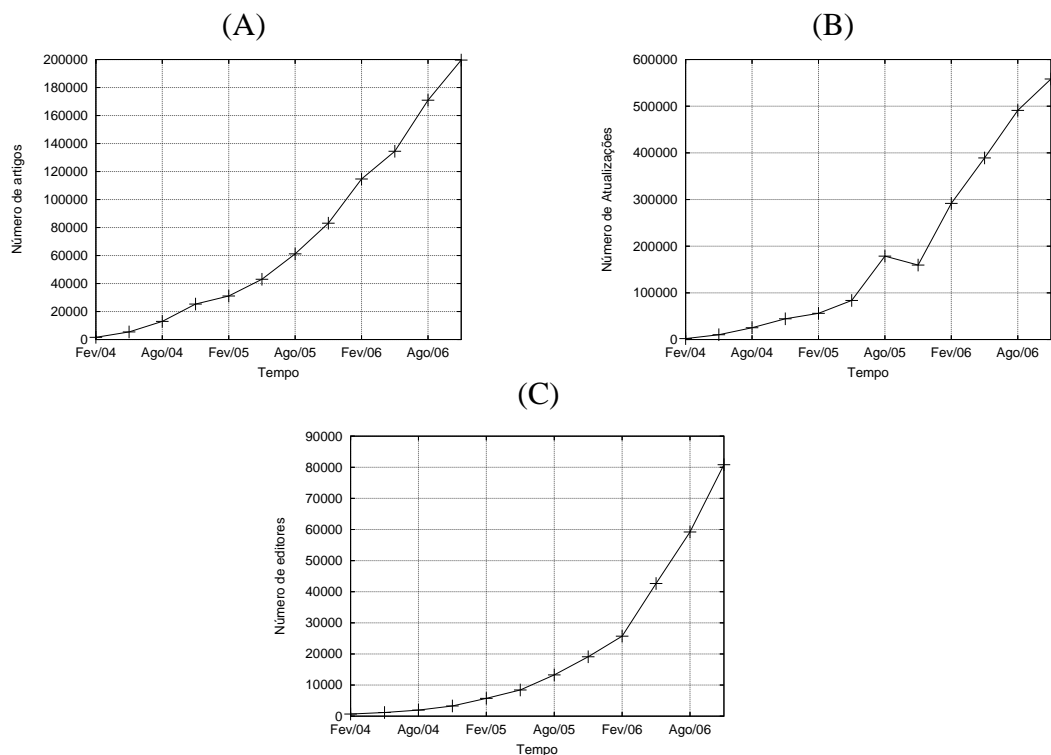


Figura 1. Crescimento da Wikipédia. A) número de artigos; B) número de atualizações; C) número de editores distintos.

Além do aumento no número de artigos, o tamanho médio dos artigos da Wikipédia (Fig.2-A) também cresceu. Na primeira imagem do grafo o tamanho médio era de 1,6 KB contra 2,18 KB da última imagem, sendo que o crescimento atual está em 5,5% por trimestre. Isso mostra que os artigos da Wikipédia estão ficando cada vez mais completos, tendo em vista o aumento da quantidade de seus conteúdos. O tamanho médio das páginas Web do domínio .br também têm aumentado, conforme relatado por [Modesto et al. 2005]. Entretanto, o número médio de atualizações por usuário dos artigos da base de dados da Wikipédia (Fig.2-B) está decrescendo a uma taxa de 16,77%. Isto se deve, provavelmente, ao grande crescimento do número de editores.

Na Fig.3-A verificamos como se distribui o número de modificações por artigo. Podemos perceber que aproximadamente dois terços do total de artigos receberam menos

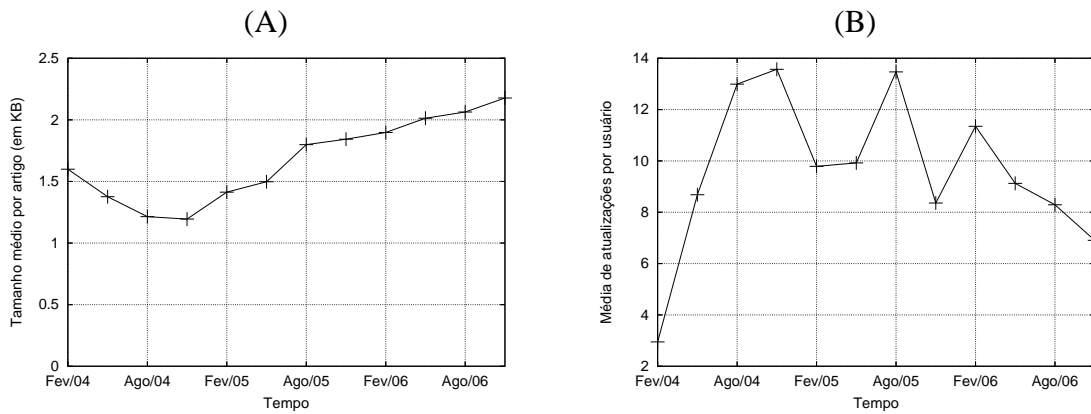


Figura 2. A) tamanho médio em KB por artigo; B) média de atualizações por usuário.

de 10 atualizações, enquanto que apenas 1,13% dos artigos foram atualizados mais que 100 vezes. Já a Fig.3-B apresenta a distribuição do número de editores distintos que atualizaram cada artigo. Nota-se que aproximadamente 84% dos artigos têm menos de 10 editores, entretanto o percentual de artigos com editores únicos é bem menor: 12,30%.

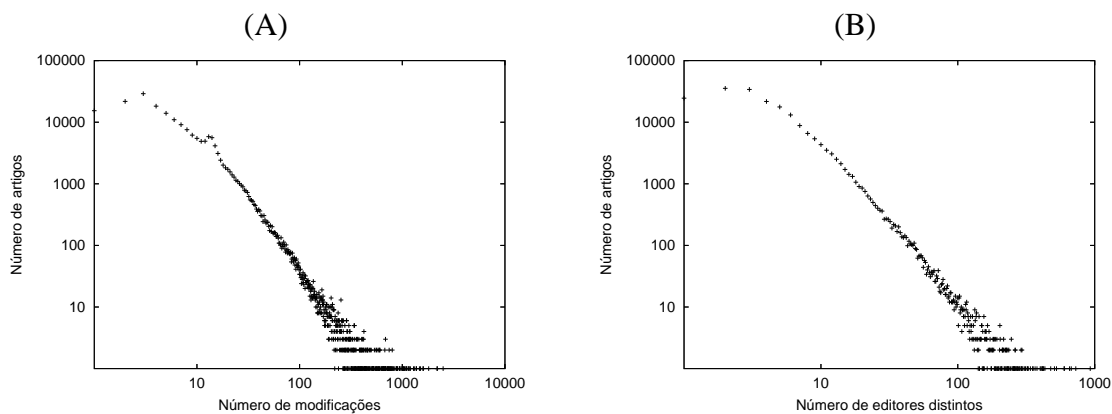


Figura 3. Distribuição das atualizações por artigo. A) número de atualizações; B) número de editores distintos envolvidos com cada artigo.

Assim como para o Wikigrafo-EN [Buriol et al. 2006], a distribuição do número de modificações e a distribuição do número de editores distintos pelo número de artigos caracteriza uma distribuição de lei de potência (*power law*). Numa distribuição de lei de potência, a relação entre duas variáveis x e y é igual a $y = ax^{-k}$, onde a e k representam a constante de proporcionalidade e o expoente da lei de potência [Faloutsos et al. 1999] respectivamente. Esta distribuição é facilmente identificada em gráficos em escala logarítmica. Como valores de y variam inversamente com a potência de x , o gráfico da distribuição se apresenta como uma reta de inclinação k .

Uma característica da Wikipédia é a alta colaboração entre usuários na edição de artigos, visando uma maior qualidade. Na Fig.4-A é mostrado como isto acontece de forma rápida, tendo em vista que em média 14% de todas atualizações, durante o período analisado, ocorreram em um intervalo inferior a 24h entre duas atualizações.

O vandalismo certamente é um dos maiores problemas encontrados na Wikipédia.

A fim de corrigí-lo é possível reverter um artigo para qualquer uma das suas versões anteriores. Isto é possível porque todo histórico de atualizações de cada artigo é mantido. Reversões podem ser feitas por qualquer usuário, cadastrado ou não, e são detectadas a partir de um comentário inserido pelo editor (*rv* ou *revert* na edição de uma página). Foi observado que o número de reversões duplas de uma página é extremamente baixo (não ultrapassando 0,1% em nenhum dos trimestres do período analisado). Logo, podemos concluir que não existe uma guerra de reversões entre usuários disputando o conteúdo de um artigo, ao contrário do observado na Wikipédia em inglês [Buriol et al. 2006]. Desta maneira, o fato do número de reversões por atualização estar aumentando (Fig.4-B) pode ser considerado fortemente correlacionado com o crescimento do vandalismo, já que o principal objetivo das reversões é justamente combatê-lo. Além disso, um outro mecanismo de combate ao vandalismo é o bloqueio de páginas. Uma alternativa comumente adotada é permitir que somente os administradores (ou a usuários cadastrados há um certo tempo) possam editar o conteúdo de um determinado artigo⁶.

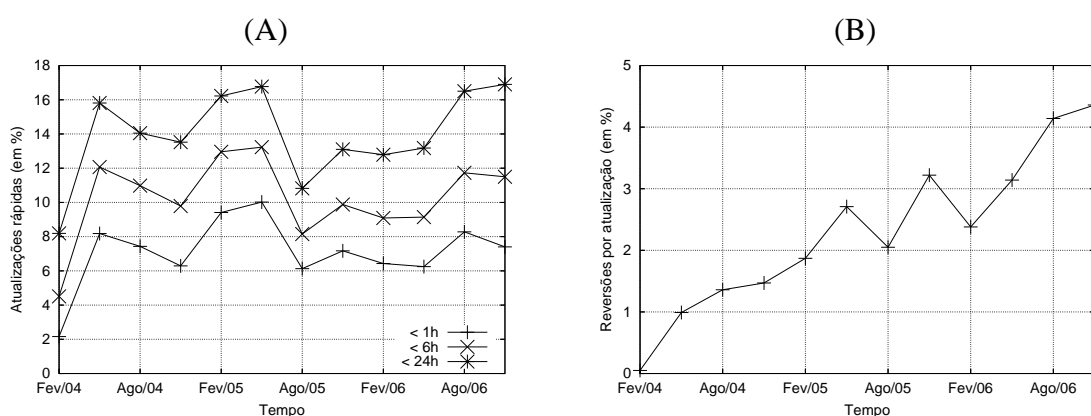


Figura 4. A) percentual de atualizações rápidas (em menos de 24h) por atualização; B) percentual do número de reversões por atualização.

4.2. Análise dos hiperlinks

Esta seção tem por objetivo analisar algumas propriedades básicas de grafos. O

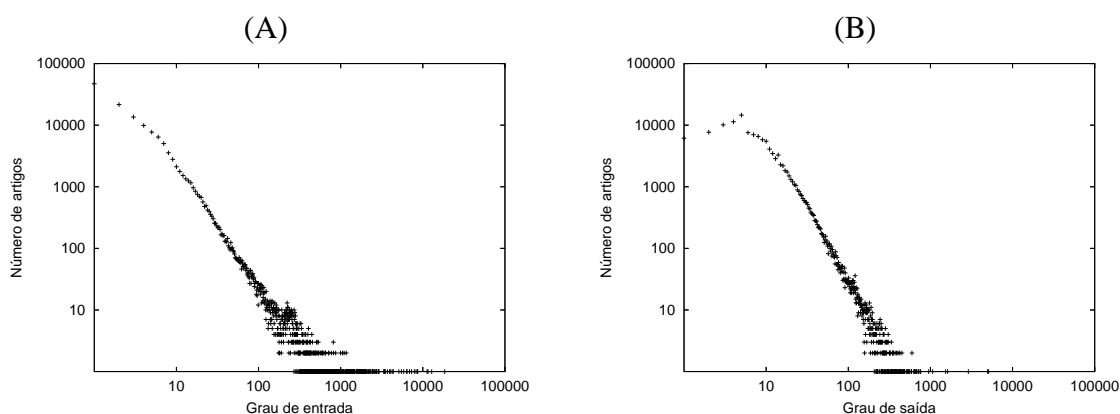


Figura 5. Distribuição do grau de entrada e saída da Wikigrafo-PT atual.

⁶Maiores detalhes podem ser obtidos em http://en.wikipedia.org/wiki/Wikipedia:Protection_policy

Wikigrafo-PT é composto por 170.201 nós e 1.749.830 arcos. A distribuição dos graus de entrada e saída dos nós deste grafo é apresentada na Fig.5. A análise do grau de entrada (Fig.5-A) assim como do grau de saída (Fig.5-B) dos nós deste grafo caracteriza uma distribuição de lei de potência. Esta distribuição ocorre de forma muito similar em Webgrafos, como observado em [Broder et al. 2000, Laura et al. 2003, Modesto et al. 2005].

A partir da Fig.6, que mostra o grau de saída médio, podemos perceber que o Wikigrafo-PT está ficando cada vez mais denso, à medida que o número médio de *hyperlinks* entre os artigos está aumentando linearmente, com crescimento atual de 12,15%.

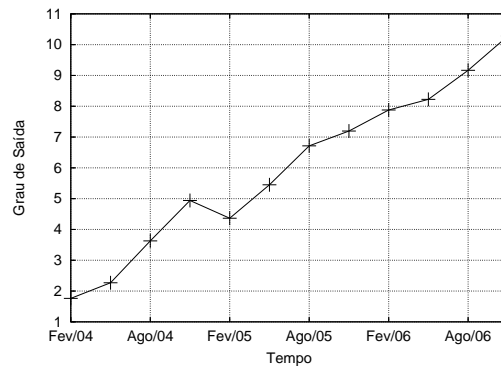


Figura 6. Grau de saída médio do Wikigrafo-PT.

O PageRank [Brin and Page 1998] é um algoritmo que atribui pesos numéricos às páginas da Web. Ele é a base do motor de busca Google e serve para quantificar a importância (autoridade) de uma página. O PageRank de uma determinada página reflete a probabilidade de que uma *random walk* na Web passe pela página em questão. Vale ressaltar que não só o número de *hyperlinks* para uma determinada página é levado em consideração, como também a importância da página que possui o *hyperlink*.

A Fig.7-A apresenta a distribuição dos valores de PageRank, uma análise típica de Webgrafos. Como este grafo também representa uma rede social relacionada com o

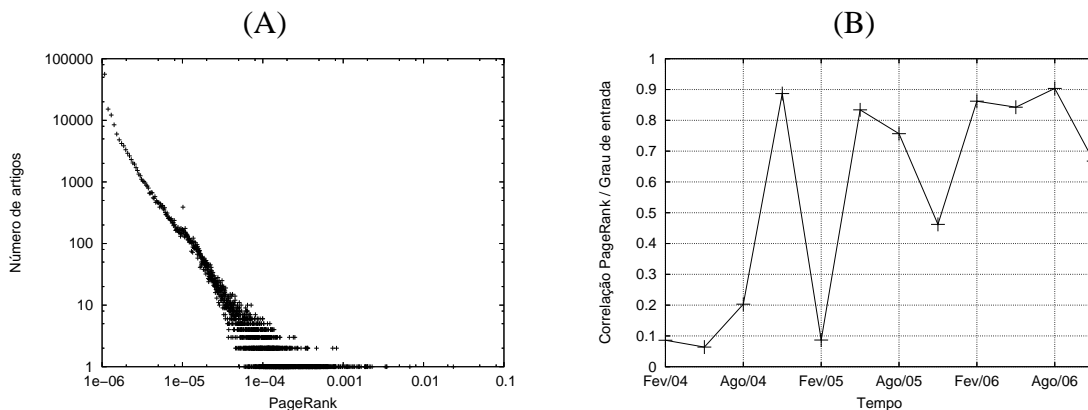


Figura 7. A) Distribuição dos valores do PageRank para o Wikigrafo-PT atual. B) Correlação entre o grau de entrada e o PageRank.

Webgrafo, torna-se importante realizar este experimento. Verifica-se que, assim como em Webgrafos, a distribuição do PageRank também caracteriza uma lei de potência no

Wikigrafo-PT. Por sua vez, a Fig.7-B tem por objetivo analisar a correlação entre o grau de entrada e o PageRank do grafo, sendo que esta correlação é pequena nos experimentos para o Webgrafo de [Donato et al. 2004]. No entanto, foi observado que para o Wikigrafo-PT, assim como no Wikigrafo-EN (observado por [Buriol et al. 2006]), esta correlação não é pequena. Diferente da Web, na Wikipédia não existe interesse sobre os valores de PageRank das páginas e os links são sempre inseridos sem objetivo de burlar o PageRank. Talvez esta seja a justificativa da diferença de correlação previamente analisada.

4.3. Estrutura macroscópica do Wikigrafo-PT

O último experimento desta seção tem por objetivo estudar a evolução dos componentes da estrutura topológica do Wikigrafo-PT. Os componentes da estrutura topológica de Webgrafos em geral é subdividida em cinco elementos: núcleo, entrada, saída, tentáculos e ilhas [Broder et al. 2000]. O componente *núcleo* é formado pelo conjunto de nós que possuem um caminho entre cada par de nós deste conjunto. Os nós do conjunto *entrada* são aqueles que possuem um caminho até qualquer nó do conjunto *núcleo*, mas nenhum nó do conjunto *núcleo* possui um caminho para qualquer nó do conjunto *entrada*. O conjunto *saída* possui conceito similar, somente invertendo a lógica aplicada ao conjunto *entrada*. Os *tentáculos* são grupos de nós que possuem caminhos para nós do conjunto *saída* ou de nós do conjunto *entrada*, sem que este caminho passe pelo *núcleo*. Em medições realizadas em Webgrafos, o maior componente conexo possui cerca de 90% dos nós do grafo [Broder et al. 2000, Donato et al. 2004]. No Wikigrafo-PT, este componente possui cerca de 98% dos nós. As *ilhas* são todos os demais componentes do grafo. A Fig.8 apresenta a evolução da dimensão dos três principais componentes do Wikigrafo-PT.

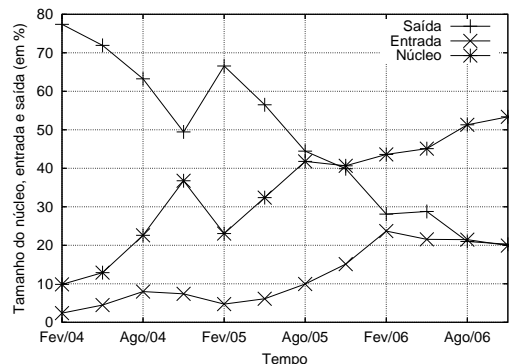


Figura 8. Evolução do tamanho dos componentes Entrada, Saída e Núcleo.

5. Conclusões e trabalhos futuros

O objetivo deste trabalho foi apresentar uma análise quantitativa e temporal do grafo gerado a partir da versão em língua portuguesa da Wikipédia. O enfoque principal foi mostrar como está acontecendo o crescimento Wikipédia. Além disso, com o intuito de comparar as características deste Wikigrafo com os demais Webgrafos, fizemos uma análise dos *hyperlinks*. A estrutura macroscópica do Wikigrafo-PT também foi estudada. Podemos concluir que existem sinais de um regime de crescimento misturados com um regime de maturidade no Wikigrafo-PT, confirmando a semelhança com o Wikigrafo-EN.

Dentre os sinais de regime de transição (crescimento), podemos citar: (i) o número de artigos, editores e atualizações está crescendo exponencialmente; (ii) o tamanho médio

de cada artigo está crescendo linearmente, assim como o número médio de *hiperlinks* entre artigos; e (iii) o número de reversões também está crescendo linearmente, o que indica fortemente que o vandalismo está aumentando.

Sinais de regime permanente (maturidade) encontrados são: (i) existe uma lei de distribuição de potência do grau de entrada e saída do nodos do Wikigrafo-PT, assim como na distribuição do PageRank; e (ii) mais de 50% dos artigos pertencem ao componente fortemente conexo do grafo (núcleo) e cerca de 98% pertencem ao componente conexo.

Como trabalho futuro, pretendemos analisar outros Wikigrafos a fim de comparar suas características com os resultados deste trabalho e de outros Webgrafos já estudados. Além disso, pretendemos estudar outros tipos de redes de relacionamento na Web e verificar se suas características se comparam as encontradas em wikigrafos e webgrafos.

Agradecimentos

Agradecemos a Carlos Castillo e Debora Donato por gentilmente cederem os scripts utilizados na geração dos grafos. Os dois primeiros autores agradecem à SESu e ao MEC pelo incentivo à realização deste trabalho.

Referências

- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press, Harlow, England.
- Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, S., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33:309–320.
- Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., and Millozzi, S. (2006). Temporal analysis of the wikigraph. In *Proceedings of the Web Intelligence Conference (AWIC)*, Hong Kong. IEEE CS Press.
- Donato, D., Laura, L., Leonardi, S., and Millozzi, S. (2004). Large scale properties of the web-graph. *The European Physical Journal B*, 38:239–243.
- Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA. ACM Press.
- Gulli, A. and Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *International World Wide Web Conference*, pages 902–903, Chiba, Japan. ACM.
- Laura, L., Leonardi, S., Millozzi, S., Meyer, U., and Sylbeyn, J. F. (2003). Algorithms and experiments for the webgraph. In *11th Annual European Symposium on Algorithms (ESA)*, volume 2832 of LNCS, pages 703–714. Springer.
- Modesto, M., Pereira Jr, I. R., Ziviani, N., Castillo, C., and Baeza-Yates, R. (2005). Um novo retrato da web brasileira. In *XXXII Seminário Integrado de Software e Hardware (SEMISH)*, São Leopoldo (RS).
- Ntoulas, A., Cho, J., and Olston, C. (2004). What's new on the web?: The evolution of the web from a search engine perspective. In *WWW 2004*, New York. ACM.
- Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using pagerank to characterize web structure.