Análise de Diferentes Técnicas de Amostragem Para Coleta de Dados na Web 2.0

Thalisson Luiz Vidal de Oliveira, Fabrício Benevenuto
PPGCC - Programa de Pós-Graduação em Ciência da Computação
UFOP - Universidade Federal de Ouro Preto
Ouro Preto, Minas Gerais, Brasil
email: thalissonvidal@hotmail.com, benevenuto@gmail.com

***Resumo*—Diversos estudos recentemente realizados tem como objetivo analisar os dados coletados na Web por meio das estratégias de amostragem. Muitos desses resultados podem ser tendenciosos devido à estratégia de amostragem utilizada para coletar tais dados. Com este projeto, pretendemos avaliar a complexidade de diferentes algoritmos de amostragem e quantificar o quanto cada algoritmo pode levar a obtenção resultados tendenciosos. Como forma de realizar os estudos e observações pertinentes aos resultados a serem obtidos, utilizaremos 4 bases *Peer-to-Peer network* de grafos direcionados que juntas possuem 31.978 usuários, comparando a quantidade de vértices e arestas encontradas com o *BFS (breadth-frist search)* e com o *Snowball* em diferentes níveis com a quantidade relatada na biblioteca SNAP fonte das bases utilizada neste estudo e ainda apresentar o caminho percorrido, por meio dos métodos de busca em largura implementados para este estudo.**

***Keywords*-Twitter, análise, complexidade, grafos.**

I. Introdução

É visível o grande crescimento do uso dos serviços e atrativos oferecidos pela Internet na atualidade, principalmente quando comparados com os anos anteriores. Este crescente volume de informações e dados disponíveis na Web e Web 2.0 tem motivado diversos estudos e pesquisas que tem como objetivo a coletada de dados utilizando estratégias de amostragem. Porém diversas destas estratégias estudadas geram resultados que podem ser tendenciosos devido à estrutura adotada para amostragem durante o processo de implementação e coleta dos dados.

Com o constante crescimento do volume de informação na Internet surgiram também formas de se compartilhar essas informações, como por exemplo os serviços de redes sociais (Orkut, Facebook, Twitter), protocolos de comunicação (SMTP utilizado para envio de e-mail), mensageiros instantâneos (Gtalk, MSN) e aplicativos para compartilhamento de arquivos.

Essas ferramentas que a sociedade usufrui para se comunicar, enviar e receber dados são caracterizadas como redes de P2P (*Peer-to-Peer ou Ponto a Ponto*) que possui arquitetura distribuída onde cada nó é capaz de realizar funções de servidor e cliente dependendo da necessidade de uso do serviço. A interação que ocorrer entre os usuários das redes P2P são trabalhadas como grafos, onde os usuários são considerados vértices e as ligações entre estes usuário consideras arestas [1]. Juntamente com este constante crescimento existe a necessidade de se coletar tais informações, para que seja possível assim conhecer o que é disseminado na Internet, gerando estudos e análises para melhorar a qualidade destas informações e serviços.

Por meio deste trabalho pretendemos implementar alguns algoritmos de amostragem, avaliando a complexidade de cada algoritmo, comparando os resultados obtidos pelos mesmos e investigando se realmente é possível obter resultados tendenciosos a partir de um método ou abordagem escolhida durante o processo de coleta dos dados. Como forma de comparar os resultados obtidos vamos utilizar 4 bases *Peer-to-Peer* que juntas possuem 31.978 usuários (nós) e 110.154 ligações (arestas) conforme apresentados na tabela I.

O artigo está organizado da seguinte forma. Na Seção II apresentamos a relação que este trabalho possui com outros estudos, na Seção III, descrevemos os métodos abordados para o desenvolvimento das comparações realizadas, já na Seção IV, é apresentado os dados pertinentes a ordem de complexidade do estudo realizado, em seguida nas Seções V e VII apresentamos os experimentos realizados, resultados e as conclusões obtidas por meio deste estudo respectivamente.

II. Trabalhos Relacionados

Na literatura existem inúmeros estudos de grande importância para computação que envolvem grafos como a necessidade de saber se existe alguma forma de conexão entre dois pontos, qual o caminho mínimo entre esses pontos ou quais pontos podem ser atingidos a partir de um ponto inicial.

Quando tratamos de Internet é possível encontrar diversas outras aplicações para modelagem de grafos envolvendo coletas como por exemplo, realizar estimativas referentes ao tamanho da Web, identificar opiniões e tendências sociais ou descobrir quando um determinado individuo, conteúdo ou produto é capaz de influenciar na opinião da sociedade. Portanto uma modelagem envolvendo grafos feita de forma correta pode apresentar resultados de grande importância

levando problemas do cotidiano as soluções de qualidade significativa.

Como forma de representar a relação da modelagem de grafos com problemas reais, podemos citar a necessidade de uma transportadora realizar suas entregas, onde quanto menor a distância percorrida por seus veículos menor são as despesas que envolvem a entrega.

Tratando de informações na Web, a modelagem de grafos pode proporcionar a uma empresa conhecer determinados pontos onde seus produtos e serviços são pouco consumidos e assim identificar os possíveis motivos causadores deste fato além de identificar quem é seu público alvo determinando desta forma para onde pode ou não expandir seus recursos.

Por meio da coleta de informações na Web, também é possível identificar formadores de opiniões que no caso de disputas políticas podem realizar mudanças expressivas nos resultados finais.

Diante de todas essas possibilidades de uso da modelagem de grafos é preciso saber que o critério adotado para o inicio da coleta é de extrema importância para os resultados finais deste processo.

III. MÉTODOS

As características descritas na Tabela I são pertinentes a quantidade de vértices, arestas e datas das bases que aqui utilizamos para realizar as comparações propostas para este estudo. As bases utilizadas foram disponibilizadas pela *Stanford University* de forma gratuita para fins comerciais e acadêmicos por meio da biblioteca SNAP, acessível para consulta e utilização da sociedade [2].

Diferentes segmentos demandam de estudos e pesquisas dos mais diversos conteúdos. Por meio da Internet é possível obter todo conteúdo desejado. Entretanto não é sempre que esses conteúdos se encontram organizados de forma trivial para acesso, fazendo com que seja preciso realizar coletas para conseguir o que buscamos. Para se obter tais informações, existem diversas formas de abordagens que são capazes de coletar tais dados. Dentre tais abordagens podemos citar a busca em largura, a busca em profundidade e a busca a partir do grau.

Tabela I
DESCRIÇÃO DAS BASES UTILIZADAS

	Nós	Arestas	Data da Base
p2p-Gnutella05	8.846	31.839	5 de agosto de 2002
p2p-Gnutella06	8.717	31.525	6 de agosto de 2002
p2p-Gnutella08	6.301	20.777	8 de agosto de 2002
p2p-Gnutella09	8.114	26.013	9 de agosto de 2002
Total	31.978	110.154	

Para a abordagem de busca em largura existem os métodos de ***BFS (breadth-frist search)*** e ***Snowball***, para busca em profundidade existem os métodos de ***DFS (Depth-First Search)*** e ***Forest Fire*** [3], já para busca por grau existem os métodos de ***Greedy*** e ***Lottery*** [4].

O ***BFS (breadth-frist search)*** é um algoritmo que percorre por completo um grafo em largura, a partir de um determinado nó dado como raiz, expandindo e examinando todos os nó vizinhos até que todos os nós possíveis sejam coletados.

O ***Snowball*** possui um processo de coleta similar ao BFS realizando uma busca em largura. A diferença entre ambos esta na condição de parada que para o Snowball ocorre quando todos os nós descobertos na raiz são coletados, ou quando determinamos um ponto de corte, ou seja se quais níveis desejamos coletar utilizando esta abordagem.

O ***Forest Fire*** realiza a coleta em largura assim como o BFS e o Snowball, porém de forma aleatória onde cabe ao algoritmo decidir se deve ou não explorar um determinado nó vizinho, como consequência é possível que o processo de coleta seja finalizado antes que todos os nós candidatos sejam visitados. Assim para que essa técnica se torne comparável com as demais é preciso que o processo seja realizado a partir de um nó aleatório.

O ***DFS (Depth-First Search)*** é um método que percorre por completo um grafo em profundidade, a partir de um determinado nó dado como raiz.

O ***Greedy*** é método que seleciona o vértice de maior grau observando se o vértice ainda não foi coletado.

O ***Lottery*** é um método que o vértice de maior probabilidade quanto ao grau, dando preferência para os vértices de maior grau.

Com base nessas técnicas de coletada que o estudo apresenta, pretendemos implementar os métodos de BFS e Snowball, que são ambos técnicas de busca em largura, onde dado um grafo $G(V, A)$ e um vértice de origem, são buscados os demais vértices alcançáveis a partir do vértice de origem, expandindo a fronteira entre os vértices descobertos e não descobertos de maneira uniforme [5], conforme o exemplo mostrado na Figura 1, lembrando que que o BFS realiza a busca em todos os níveis possíveis e que o Snowball apenas nos níveis determinados.

Mediante os métodos de busca em largura aqui implementados, pretendemos comparar a quantidade de vértices e arestas encontradas, apresentando também o caminho percorrido pela busca em largura em cada base e métodos utilizados. Este caminho será apresentado em um arquivo a parte onde juntamente com os valores pertinentes ao número de vértices e arestas encontradas em cada busca realizada.

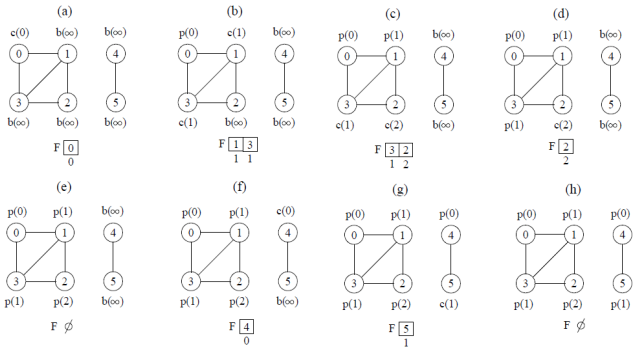

Figura 1. Exemplo de Busca em Largura [5]

IV. Análise de Complexidade

O método *BFS (breadth-frist search)* que utilizamos foi implementado a partir do processo necessário para uma busca em largura, tendo como complexidade de tempo para $O(|n|+|V^2|)$ visto que $|n|$ é custo para ler o arquivo de base e gerar o grafo e $|V^2|$ para realizar a busca em largura, onde que n o número de linhas a serem lidas no arquivo e V o número de vértices existentes no grafo. A complexidade de espaço deste algoritmo é $(m+3V)$, sendo m o tamanho da matriz responsável por armazenar o grafo e $3V$ os vetores a serem utilizados na busca.

for $i = 0$ **to** $i = V$ **do**
  for $j = 0$ **to** $j = V$ **do**
    if $grafo_{ij} == 1$ **then**
      $grafo_{ij} \leftarrow 2$
      $QtdAresta++;$
      if $visit_j \; != j$ **then**
        $visit_j \leftarrow j;$
        $fila_x \leftarrow j;$
        $x++;$
        $no++;$
        $n++;$
      end if
    end if
  end for
  $nivel_{i+1} \leftarrow no$
  $no \leftarrow 0$
end for

O método referente ao processo de Snowball que utilizamos também foi implementado seguindo a abordagem necessário para se realizar um busca em largura porém levando em consideração a necessidade de determinar a condição de parada do mesmo, tendo assim uma complexidade de tempo um pouco menor quando comparado com o método implementado para o BFS, o Snowball implementado tem complexidade de tempo $O(|n|+|Corte \times V|)$, já complexidade de espaço deste é a mesma do BFS $(m+3V)$, sendo m o tamanho da matriz responsável por armazenar o grafo e $3V$ os vetores a serem utilizados na busca.

for $i = 0$ **to** $i = Corte$ **do**
  for $j = 0$ **to** $j = V$ **do**
    if $grafo_{ij} == 1$ **then**
      $grafo_{ij} \leftarrow 2$
      $QtdAresta++;$
      if $visit_j \; != j$ **then**
        $visit_j \leftarrow j;$
        $fila_x \leftarrow j;$
        $x++;$
        $no++;$
        $n++;$
      end if
    end if
  end for
  $nivel_{i+1} \leftarrow no$
  $no \leftarrow 0$
end for

V. Experimentos

Lançando mão das bases aqui utilizadas, das técnicas implementadas e das comparações realizadas, com esse trabalho é possível perceber o quanto determinadas estratégias de amostragem utilizadas para coletar os dados podem ser tendenciosas [6], [7], [8].

Observando as Figuras 2, 3, 4, 5 e comparando-os com a Tabela I podemos perceber que apenas o método BFS foi capaz de encontrar a quantidade total de arestas existentes nos grafos. Porém nenhuma das estratégias (BFS e Snowball em diferentes níveis) abordadas neste estudo foram capazes de encontrado todos os vértices existentes nos grafos das bases *Peer to Peer* que são bases de maior escala, visto que os vértices existentes não foram encontrados em sua totalidade pelos métodos implementados, não é possível afirmar que o caminho percorrido pelo mesmo esta correto.

Porém ressaltamos que antes de dar inicio aos testes utilizando as bases disponibilizadas por meio da biblioteca SNAP, os algoritmos implementados em questão. Foram testados em 3 bases independentes de menor escala geradas aleatoriamente seguindo o mesmo formato das bases Peer to Peer, possuindo estas 10 vértices e 30 arestas cada.

Logo utilizando as bases de menor escala, foi possível alcançar de forma total todos os vértices e arestas possíveis. Ainda tendo apresentado o correto caminho percorrido para atingir os resultados obtidos, é preciso ressaltar que o ponto inicial da coleta também podem influenciar nos resultado. Portanto todos os testes realizados para este estudo foram feitos partindo do mesmo ponto inicial.

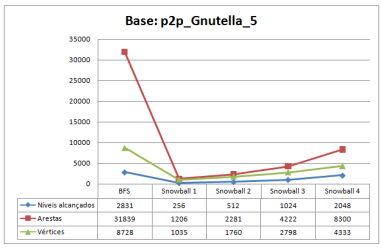

Figura 2. Resultado das buscas em Largura na base p2p-Gnutella05

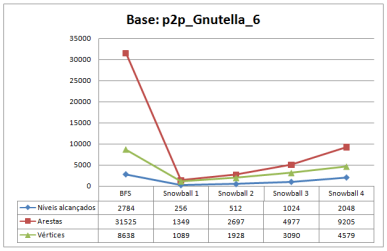

Figura 3. Resultado das buscas em Largura na base p2p-Gnutella06

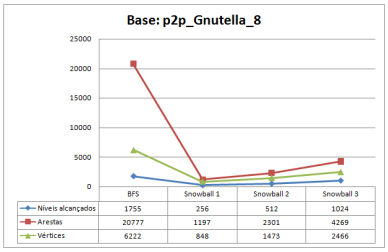

Figura 4. Resultado das buscas em Largura na base p2p-Gnutella08

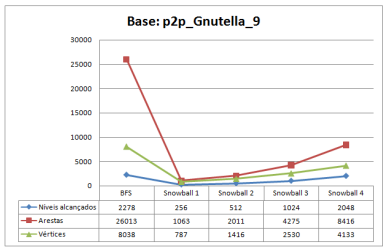

Figura 5. Resultado das buscas em Largura na base p2p-Gnutella09

VI. Trabalhos Futuros

Para trabalhos futuros pretendemos melhorar o processo de busca aqui realizado, procurando assim obter resultados mais satisfatórios como: A coleta total dos nós e arestas possíveis; apresentar os sub-grafos gerados apartir de cada nó; melhorar custo dos algoritmos apresentados; implementar novos métodos de busca, visando assim comparar os resultados obtidos neste estudo com os possíveis resultados gerados a partir de novas implementações.

VII. Conclusão

Neste trabalho foram implementadas duas técnicas de amostragem para coleta em grafos. Os algoritmos propostos são compostos de duas fases.

A primeira sendo responsável por efetuar a leitura dos dados em um determinado arquivo base gerando o grafo a ser utilizado. A segunda fase efetua a busca em largura no grafo, permitindo descobrir o número de arestas, vértices e o caminho percorrido pelo algoritmo. Entretanto por meio deste estudo percebemos que os métodos aqui apresentados não foram capazes de atingir em um todo a coleta dos dados pertinentes aos valores de vértices e arestas para bases de maior escala.

Diante da complexidade do algoritmo apresentado é necessário que esta seja melhorada, uma vez que a literatura apresenta algoritmos de melhor complexidade [5] para que assim o uso dos algoritmos apresentados seja viável. Embora o tempo de resposta para os resultados revelados para este estudo seja em media de 7 segundos.

Referências

[1] F. Benevenuto, “Redes sociais online: Técnicas de coleta, abordagens de medição e desafios futuros.” Belo Horizonte, Brasil: Short course on the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC), October 2010, p. 30.

[2] J. Leskovec. (2002) Stanford large network dataset collection. [Online]. Available: http://snap.stanford.edu/data/

[3] M. Kurant, A. Markopoulou, and P. Thiran, “On the bias of breadth first search (bfs) and of other graph sampling techniques,” 2010, pp. 1–8. [Online]. Available: fileadmin/ITCBibDatabase/2010/kurant10.pdf

[4] R. Minhano. (2010) Coletando relações sociais na rede orkut. [Online]. Available: http://www.slideshare.net/rogeriominhano/crawling-orkut

[5] N. Ziviani, *Projeto de Algoritmos: com implementações em Pascal e C*, 1st ed. São Paulo: Cengage Learning (Thomson / Pioneira), 2004.

[6] S. Lee, P. Kim, and H. Jeong, “Statistical properties of sampled networks,” vol. 73, no. 1, pp. 102–109.

[7] Y.-Y. Ahn, S. Han, H. Kwak, Y.-H. Eom, S. Moon, and H. Jeong, “Analysis of topological characteristics of huge online social networking services,” in *Proceedings of the 16th international conference on World Wide Web (WWW07)*. ACM, 2007, pp. 835–844.

[8] F. Benevenuto, J. M. Almeida, and A. S. Silva, “Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações.” Campo Grande, Brasil: Short course on the Brazilian Symposium on Computer Networks and Distributed Systems (SBRC), May 2011, p. 40.