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 que 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 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 em cada base, por meio dos métodos de busca em largura implementados para obter os resultados descritos neste estudo.**

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

I. Introdução

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 direcionadas que juntas possuem 31.978 usuários (nós) e 110.154 ligações (arestas) conforme apresentados na tabela I.

A Seção II apresenta a relação que este estudo tem outros estudos, a Seção III apresenta a motivação para realização deste trabalho. Os objetivos são abordados na Seção IV, já na Seções V e VI é descrito o método utilizado para o desenvolvimento das comparações, testes realizadas e resultados que esperados para esse trabalho .

II. Trabalhos Relacionados

Na literatura existem inúmeros estudos de grande importância para computação que envolvem grafos como a necessidades 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 aplicações para modelagem de grafos envolvendo coletas como por exemplo, realizar estimativas referentes ao tamanho da Web, identificar opiniões e tendências sociais ou descobrir o quando um determinado individuo, conteúdo ou produto é capaz de influenciar na opinião da sociedade, por tanto uma modelagem envolvendo grafos feita de forma correta pode apresentar resultados de grande importância levando problemas do cotidiano as soluções de qualidade significativa.

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. Justificativas

É visível o grande crescimento do uso dos serviços e atrativos oferecidos pela Internet atualmente, 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 que tem como objetivo a coletada de dados utilizando estratégias de amostragem. Diversos desses estudos geram resultados que podem ser tendenciosos devido à estratégia de amostragem utilizada durante a coleta dos dados.

IV. Objetivos

Assim como a Internet, os serviços de redes sociais tais como Orkut, Facebook, Twitter e inúmeros outros, também têm aumentado a cada dia o número de usuários adeptos a este meio de socialização gerando novas informações e dados. A interação que ocorrer entre os usuários das redes sociais são trabalhadas como grafos, onde os usuários são considerados vértices e as ligações entre estes usuário as arestas [1]. Juntamente com este crescimento existe a necessidade de se coletar informações. Por tanto para este estudo é pretendido implementar diferentes algoritmos de amostragem e quantificar o quanto cada algoritmo pode levar a obtenção de resultados tendenciosos. Como forma

de comparar resultados obtidos vamos utilizar 4 bases peer to peer network, que juntas possuem 31.978 usuários (nós) e 110.154 ligações (arestas) entre os usuários conforme apresentados na tabela I.

V. METODOLOGIA

As bases de dados peer to peer network aqui utilizada para realizar as comparações descritas neste trabalho, foram disponibilizadas pela Stanford University de forma gratuita para fins comerciais e acadêmicos por meio da biblioteca SNAP, estando acessível para consulta e utilização da sociedade [2].

Tabela I
DESCRIÇÃO DAS BASES GNUTELLA PEER TO PEER 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	

Dentre os diversos para coleta de dados, existem as abordagens de buscas em largura, busca em profundidade e busca por grau dentre as diversas abordagens existentes para realizar coleta de dados na Web, temos o **BFS (breadth-frist search)**, o **Snowball**, o **DFS (Depth-First Search)** e o **Forest Fire** [3], o **Greedy** e o **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.

O **Forest Fire** realiza a coleta em largura assim como o BFS e o Snowball, porem 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

O **DFS (Depth-First Search)** é um algoritmo 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, lembrando que que o BFS realiza a busca em todos os níveis possíveis e que o Snowball apenas nos níveis determinados. Assim sendo por meio dos métodos implementados pretendemos comparar a quantidade de vértices e arestas encontradas com o BFS e com o Snowball em diferentes níveis, apresentando também o caminho percorrido pela busca em largura em cada base e métodos utilizados, sendo que 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.

VI. RESULTADOS ESPERADOS

Lançando mão das bases de dados peer to peer aqui utilizada, das técnicas implementadas e das comparações realizadas, pretendemos com esse trabalho verificar se realmente a escolha de uma determinada estratégia de amostragem utilizada para coletar os dados pode ser tendenciosas na obtenção de resultados [5], [6], [7], apresentando os resultados obtidos e a complexidade dos algoritmos que foram utilizados durante todo o processo e o caminho percorrido para gerar tais resultados.

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] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," in *Physical Review E,73*, pp. 102–109.

[6] 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.

[7] 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.