Uma Ferramenta para Extração Semiautomática e Análise de Relevância de Artigos Científicos

Moacir Lopes de Mendonça Junior[1], Thiago de Abreu Lima[1], Alisson V. Brito[1], Alexandre N. Duarte[1]

[1]Programa de Pós-Graduação em Informática - Centro de Informática – Universidade Federal da Paraíba (UFPB) – João Pessoa – PB – Brasil

{moacir.lopes.jr, thiago.abreulima}@gmail.com, {alisson, alexandre}@ci.ufpb.br

Abstract. *This paper presents the results of an analysis of the influence and reputation of publications on Social Network Analysis. Its purpose is, through Social Network Analysis metrics, analyze if is possible to identify the most relevant publications on a given field by conducting automatic searches on scientific digital libraries. We conduct a literature survey on the subject in focus and an analysis of the interactions between publications and their citations.. Additionally, we conducted a quantitative analysis of the journals with the highest number of publications and the growth of publications over the years.*

Resumo. *O presente artigo apresenta os resultados da análise de influência e reputação de publicações sobre Análise de Redes Sociais. Seu propósito consiste em, por meio das métricas de Análise de Redes Sociais, analisar se através da busca de publicações em meios acadêmicos de indexação é possível identificar as publicações mais relevantes dentro daquele contexto. Para tanto, foi realizado um levantamento bibliográfico da literatura sobre o tema em foco e uma análise das interações entre as publicações e o suas citações. Adicionalmente, também foram realizadas análises quantitativas dos locais de publicação com maior número de publicações e a curva de crescimento de publicações ao longo dos anos.*

1. Introdução

Desde a sua origem, o homem tem buscado conhecimento em todas as suas manifestações. A construção do conhecimento se dá através de um processo social realizado a partir do trabalho e do esforço coletivo (Bourdieu, 2004; Burke, 2003; Meadows, 1999; Ziman, 1979) e, sendo assim, baseado em costumes, crenças, conceitos filosóficos e religiosos.

A partir da proposição de técnicas para análise, as redes sociais têm sido utilizadas para os mais variados focos (por exemplo, redes acadêmicas, de amizade, redes de contatos profissionais e de compartilhamento e discussões de fotos e vídeos) trazendo assim, novas oportunidades e desafios. Tais redes têm a capacidade de representar a diversidade social e a complexa formação dos relacionamentos entre indivíduos e os demais componentes de um grupo. São formadas pela composição de nós que estão interligados por um ou mais tipos específicos de interdependência (Berkowitz, 1982). E, por meio dessas ligações, vão construindo e reconstruindo uma estrutura social.

Segundo Chakrabarti (2003), análise de redes sociais (ARS) é "o mapeamento e medição de relações e fluxos entre pessoas, grupos, organizações, computadores, URLs, e outras entidades conectadas de informação/conhecimento". Esse campo de estudo, que se encontra na interseção entre sociologia e matemática, foca na relação entre os atores ao invés dos atributos de cada ator.

Neste contexto, o presente estudo tem como principal finalidade desenvolver uma ferramenta para a análise de influência de publicações de determinadas áreas de conhecimento de forma semi-automatizada, obtidos a partir de buscas em ferramentas acadêmicas. Com isso buscamos auxiliar a identificação de pontos importantes de uma área do conhecimento através da análise de autorias e citações, podendo assim ser

empregados em diversos estudos, como por exemplo, as revisões sistemáticas de literatura.

O artigo está organizado da seguinte forma: Na Seção 2 citamos os trabalhos relacionados ao assunto desta pesquisa. A Seção 3 apresenta a fundamentação teórica sobre análise das redes sociais. A Seção 4 introduz a ferramenta *Paper Crawler*, desenvolvida no contexto deste trabalho, seguida da Seção 5 onde apresentamos os resultados obtidos até o momento. O artigo é finalizado na Seção 6, como nossas conclusões e propostas para trabalhos futuros.

2. Trabalhos Relacionados

Recentemente pode-se verificar a existência de vários trabalhos que, assim como este, correlacionam análise de redes sociais com a análise de autorias e citações. Börner et al. (2005) analisam o impacto de grupos de pesquisa com base no número de publicações de seus integrantes e de suas respectivas citações em escala local e global.

Existem vários trabalhos de autores brasileiros relacionados a mineração de dados em bases acadêmicas e geração de redes sociais a partir dos dados coletados. Alves et al (2011) apresentam o Sucupira, sistema de extração de informações da Plataforma Lattes para identificação de redes sociais acadêmicas, que pode ajudar o Governo Federal a mapear informações sobre áreas do conhecimento especificas. Ströele et al (2011), apresentou uma análise interessante de colaborações científicas dentro da comunidade de pesquisa brasileira, utilizando um método de detecção de grupo para identificar as comunidades de pesquisa. Guedes e Duarte (2013) analisam os impactos dos relacionamentos sociais no mérito científico, tomando por base a comunidade formada pelos bolsistas de produtividade e pesquisa do CNPq.

Neste trabalho, o objetivo é através de uma ferramenta semi-automatizada, minerar publicações em ferramentas acadêmicas e analisar através de métricas de ARS a influência destas publicações em determinadas áreas de conhecimento.

3. Análise de Redes Sociais

De acordo com Berkowitz (1982), uma rede social pode ser definida como qualquer conjunto limitado de unidades sociais interligadas. Esta definição destaca três características: (1) as redes têm limites, ou seja, existe algum critério para determinar a associação na rede. (2) a definição de "conexão" em redes sociais. Para fazer parte da rede social, cada membro deve ter ligações para pelo menos um outro membro da rede. (3) a unidade social, ou seja, cada componente é único dentro deste conjunto.

A Análise de Redes Sociais (ARS) é baseada na descrição formal das redes através de estruturas denominadas grafos, estas são estruturas formadas por nós, arestas e os atributos (caso existam), que compõem cada uma destas subestruturas.

Através da ARS, é possível compreender e acompanhar de forma mais eficaz a disseminação de informações e a interação entre as pessoas que compõem a rede. Tal acompanhamento pode auxiliar na identificação de conectores de redes isoladas, melhorar a atuação de atores críticos, identificar a fragilidade da rede em relação à comunicação dos membros dentre outras (Simões, 2011).

3.1. PageRank

O *PageRank* é um algoritmo inicialmente proposto por Brin & Page (1998) para ordenar resultados de busca do Google que gera um peso numérico para cada nó, assim podendo estimar sua importância em relação ao grafo. O entendimento por trás do *PageRank* é

que uma página *Web* é importante se existem muitas páginas apontando para ela ou se existem páginas importantes apontando para ela.

4. Paper Crawler

Para obtenção e análise dos dados deste estudo foi desenvolvida uma ferramenta na linguagem Java denominada *Paper Crawler*[1] que a partir de uma cadeia de busca realiza uma consulta a bibliotecas digitais acadêmicas como *IEEEXplorer*, *ACM*, *Springer Link*, entre outros. Os resultados destas consultas são páginas HTML com as listagens dos resultados, que serão posteriormente analisadas no intuito de extrair os metadados dos artigos retornados, como, por exemplo, o título, autores, local de publicação, ano da publicação, entre outros.

Para cada artigo obtido, o *Paper Crawler* acessa a página HTML de detalhes deste, obtendo as referências utilizadas. Como as bibliotecas digitais acadêmicas, na maioria das vezes, não indexam estas referências, utilizamos dois analisadores de citações, o *Freecite*[2] e o *Paracite*[3], para transformar o texto descrito em informação que o *Paper Crawler* possa identificar. Esses dois sistemas recebem como entrada o texto das referências e retornam um arquivo XML com os metadados daquela citação.

Com todos os artigos e referências identificados, a ferramenta identifica os locais de publicação mais relevantes e, para finalizar, gera a rede de artigos e citações. Na Figura pode-se observar todo o processo utilizado nesta pesquisa para obtenção e análise dos dados.

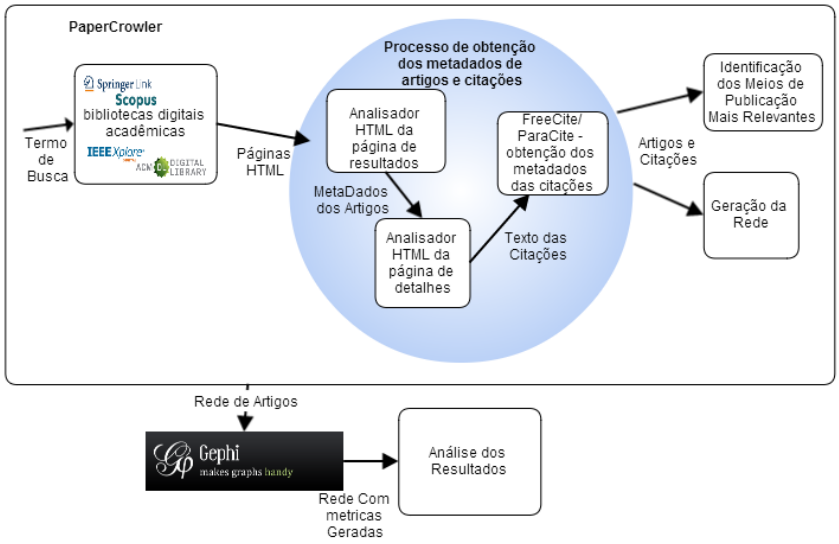

Figura 1: Processo de análise dos resultados

Como pode-se observar após a geração da rede de artigos e citações utilizou-se a ferramenta Gephi[4] para gerar uma visualização da rede e para calcular os valores das métricas de análise de redes sociais.

5. Resultados e Discussões

[1] Disponível em: https://appsnaauthorrank.googlecode.com/svn/trunk/
[2] http://freecite.library.brown.edu/welcome
[3] http://paracite.eprints.org/
[4] http://gephi.org

Esta seção é dedicada à apresentação dos resultados obtidos pela ferramenta *Paper Crawler* para a cadeia de busca *"Social Network Analysis" and "online"*. Neste primeiro experimento utilizamos apenas a base de publicações *IEEEXplorer*.

A partir da cadeia de busca a ferramenta consultou o *IEEEXplorer* e obteve 735 artigos. Em seguida, o sistema extraiu as referências de cada artigo recuperado, gerando assim uma rede de artigos com 10764 nós, incluindo os artigos 735 artigos iniciais e suas referências, e 11624 arestas. A partir desta rede utilizou-se a ferramenta Gephi para análise e cálculo das métricas com o intuito de identificar as publicações mais relevantes. Para isso utilizou-se duas métricas, o *PageRank,* a partir do qual podemos identificar nós que são mais citados por artigos também muito citados, e o Grau de Entrada, que neste caso representa o número de citações recebidas por um artigo. Desta rede a publicação mais relevante encontrada, considerando tanto o *PageRank* como o Número de Citações, foi *Social Network Analysis: Methods and Applications* do ano de 1994, está obteve um número de citações igual a 69 e um PageRank igual a 4,03E+11.

A partir da rede de artigos e citações identificou-se que estes foram publicados em cerca de 2559 locais de publicação e a nossa ferramenta foi utilizada para calcular o número de publicações para cada veículo. Vale salientar que não foi possível identificar o veículo de publicação em cerca de 38,63% nas referências, pois nenhum dos analisadores de citações utilizados conseguiu identificar isto no texto da citação. O meio de publicação mais relevante identificado foi o *Annual Service Research and Innovation Institute Global Conference,* com o número de publicações igual a 241.

A partir da rede de publicações gerada também foi possível analisar a evolução do número de publicações na área de Análise de Redes Sociais e áreas de publicações citadas ao longo dos anos, como pode ser ilustrado na Figura 2. Pode-se observar que depois do ano de 2009 houve uma queda no número de publicações.

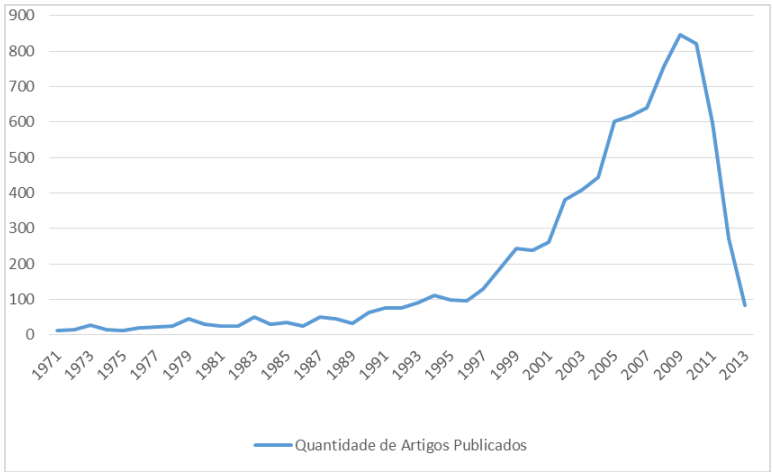

Figura 2: Número de publicações ao longo dos anos

Além disso, analisamos os artigos indexados pelo *IEEEXplorer* retornados na busca inicial, sem considerar suas referências. Neste caso obteve-se uma rede formada por 735 nós e 31 arestas. Esta rede se mostrou pouco conectada, significando que dentre os resultados há poucas citações. Para tentar identificar quais deles são mais relevantes na rede, primeiramente considerou-se a métrica Grau de Entrada, já que com ela poderíamos obter a quantidade de citações de cada publicação entre si. O maior número de citações encontrado foi 2. E observamos que houve quatro publicações empatadas com o mesmo nível de relevância sendo elas: "*Analyzing Online Asynchronous Discussion Using Content and Social Network Analysis", "Social computing and*

weighting to identify member roles in online communities", "Using egocentric networks to understand communication", "A Generic Architecture for a Social Network Monitoring and Analysis System".

Agora considerando a métrica *PageRank*, pode-se identificar os artigos que receberam mais citações de artigos também muito citados na rede. A publicação mais relevante considerando esta métrica foi a *Local Community Identification in Social Networks* que obteve um *PageRank* igual a 3,86E-03.

Também foi possível analisar o crescimento de publicações na área de Análise de Redes Sociais ao longo dos anos para os artigos indexados pelo *IEEEXplorer* retornados na busca inicial, obtendo-se a curva de crescimento ilustrada na Figura .

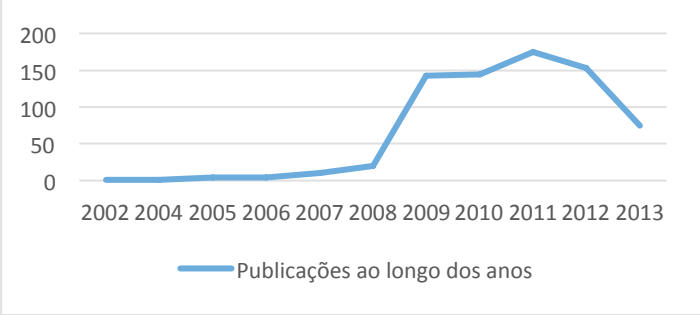

Figura 3: Publicações ao longo dos anos

Assim, pode-se observar que, considerando somente os resultados iniciais retornados pelo *IEEEXplorer*, o crescimento é muito mais sinuoso nos anos 2000, que coincide com o crescimento das Redes Sociais Online.

A partir da rede de artigos dos resultados iniciais retornados pelo *IEEEXplorer* identificou-se que estes foram publicados em cerca de 267 locais de publicação. Na Figura estão ilustrados os dezoito meios de publicação com mais publicações para a área de Análise de Redes Sociais.

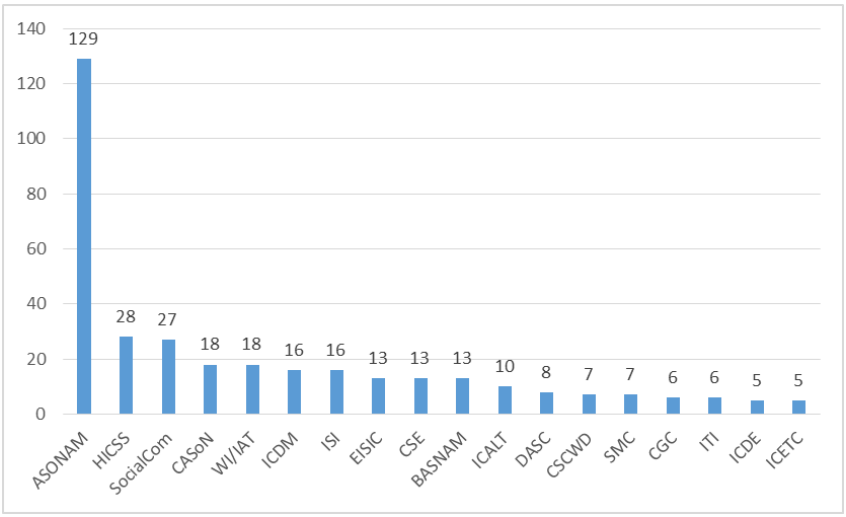

Figura 4: Meios de publicação mais relevantes

Pode-se observar que o evento que foi identificado com o maior número de publicações foi o ASONAM (*International Conference on Advances in Social Networks Analysis and Mining*), com 129 publicações encontradas.

6. Conclusões e Trabalhos Futuros

Pode-se afirmar que este estudo conseguiu apontar que é possível identificar o nível de relevância de publicações de determinadas linhas de pesquisa através da utilização de métricas de análise de redes sociais.

Os analisadores de citação utilizados, o *Freecite* e o *ParaCite*, apesar de serem de grande ajuda, ainda apresentam grandes problemas a serem resolvidos na resolução das citações. Ainda há muitas inconsistências nos dados retornados e, além disso, o tempo de resposta destes torna a ferramenta criada para este estudo muito lenta. Para a geração da rede citada nos resultados foi necessário um tempo médio de 2 horas e 30 minutos só para obter os dados em formato XML. Além disso, muitas publicações possuem referências fora do padrão, o que dificulta ainda mais o trabalho dos analisadores.

Ao usar os analisadores a taxa de erro foi bastante diminuída pois foi realizada uma combinação dos recursos do *Freecite* e do *Paracite,* já que identificou-se que no *Freecite* havia o problema de algumas informações virem erradas.

Como trabalhos futuros pretendemos realizar uma análise sobre quais autores são mais influentes dentro de um conjunto de artigos a partir da criação de uma métrica para análise de relevância dos autores e co-autores baseada no h-index mas que irá se resumir a coleção retornada pelas bases de publicação científica. Planejamentos também aprimorar a ferramenta *Paper Crawler* acrescentando um analisador para os arquivos HTML dos outros ambientes acadêmicos como o *ACM*, *Springer Link* e *Scopus*, permitindo assim uma maior abrangência na pesquisa. Finalmente, desejamos realizar uma análise dos dados com outras métricas para tornar este estudo mais preciso e finalmente concretizar a classificação de um conjunto de artigos em áreas.

Referências

Alves, A.D, Yanasse, H.H., Soma, N.Y. (2013). SUCUPIRA: a System for Information Extraction of the Lattes Platform to Identify Academic Social Networks, Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on.

Berkowitz, S. D., 1982. An Introduction to Structural Analysis: The Network Approach to Social Research. Toronto: Butterworth.

Bourdieu, P. Os usos sociais da ciência: por uma sociologia clínica do campo científico. São Paulo: UNESP, 2004. 86 p.

Brin S. and Page L. The anatomy of a large-scale hypertextual web search engine.Computer Networks and ISDN Systems, 30(1-7):107–117, 1998.

Burke, P. Uma história social do conhecimento: de Gutenberg a Diderot. Rio de Janeiro: Jorge Zahar, 2003.

Chakrabarti, S. (2003). Social Network Analysis, Mining the Web, Morgan Kaufmann, pp. 203-254.

Guedes, A. C. S. M.; Duarte, A. N. . Um Estudo sobre os Impactos dos Relacionamentos Sociais na Avaliação do Mérito Científico. In: Workshop de Teses e Dissertações (WTD), 2013, Salvador. Anais of the 19th Brazilian Symposium on Multimedia and the Web, 2013. v. 1

Meadows, A. J. A comunicação científica. Brasília: Briquet de Lemos, 1999. 268 p.

Ströele, V.; Silva, R.; Souza, M. F. DE; et al. Identifying Workgroups in Brazilian Scientific Social Networks. Journal of Universal Computer Science, v. 17, 2011.

Ziman, J. M. Conhecimento público. Belo Horizonte: Itatiaia, 1979.