Caracterização de redes de coautoria: um repositório de informações topológicas

Thiago C. Cunha[1], **Natalie C. Araujo**[2], **Orlando A. Gomes**[1]

[1]Faculdade de Ciências Empresariais (FACE) – Universidade FUMEC
Belo Horizonte – MG – Brasil

[2]Diretoria de Gestão Empresarial – CEMIG
Belo Horizonte – MG – Brasil

thiagocunha@fumec.edu.br, natalie.araujo@cemig.com.br, orlando@fumec.br

Abstract. *Choosing a dataset for social network research like recommendation systems or link prediction could require the execution of several preliminary experiments. This paper describes the research that creates a co-authorship network characteristics database. The first public version of this repository will have more than ten thousand subsets, including some of the major public academic networks available. A cloud computing infrastructure has been used to make this work possible. The created metrics repository may help data mining and social network researchers, especially those interested in academic or co-authorship networks.*

1. Introdução

Uma rede pode ser caracterizada como um conjunto de itens, que chamamos de vértices ou nós e suas conexões, chamadas de arestas [Newman 2003]. As redes de coautoria são definidas pelos autores e suas ligações que representam as publicações feitas em parceria com outros pesquisadores. Um benefício em se utilizar esse tipo de rede é a alta disponibilidade de redes confiáveis [Newman 2003, Wang et al. 2015].

A caracterização de redes de coautoria foi abordada em trabalhos como [Liu et al. 2005, Mena-Chalco et al. 2012]. Apesar da relevância dos trabalhos, cada um busca dar a visão geral das características de uma determinada rede de coautorias. Já este trabalho busca mapear, através de um método único, um conjunto de bases de coautoria, bem como milhares de seus subconjuntos, criando um repositório de autorias de publicações científicas com suas métricas focadas na topologia das redes. O repositório criado será disponibilizado publicamente, favorecendo o uso de dados padronizados em diferentes pesquisas relacionadas à análise de redes sociais, especialmente as focadas em redes de coautoria.

2. Solução Proposta

Foram escolhidos quatro repositórios com informações de coautorias conforme resumo na Tabela 1. A divisão dos subconjuntos foi baseada em diferentes critérios. Nos repositórios *mag* e *dblp*, as publicações foram divididas por periódico. Já nos repositórios *lattes* e *arxiv*, a divisão foi feita por área de atuação do pesquisador.

Tabela 1. Quantidades de subconjuntos utilizados por repositório.

	mag	*latttes*	*arxiv*	*dblp*
Número de subconjuntos	23404	1398	161	1470
Formato dos dados	*CSV*	*XML*	*XML*	*XML*

Após a obtenção dos dados de cada repositório original, os dados foram estruturados para otimizar os experimentos. Nessa etapa os dados foram importados em bancos de dados relacionais, foram criados índices de acesso e dados desnecessários foram excluídos. A otimização é necessária tanto por questões de desempenho, quanto por custos, já que cada repositório original possui até 150 GB de dados brutos e os experimentos foram executados em uma plataforma de computação em nuvem, tarifada por recursos consumidos. A Figura 1 representa o processo completo da caracterização.

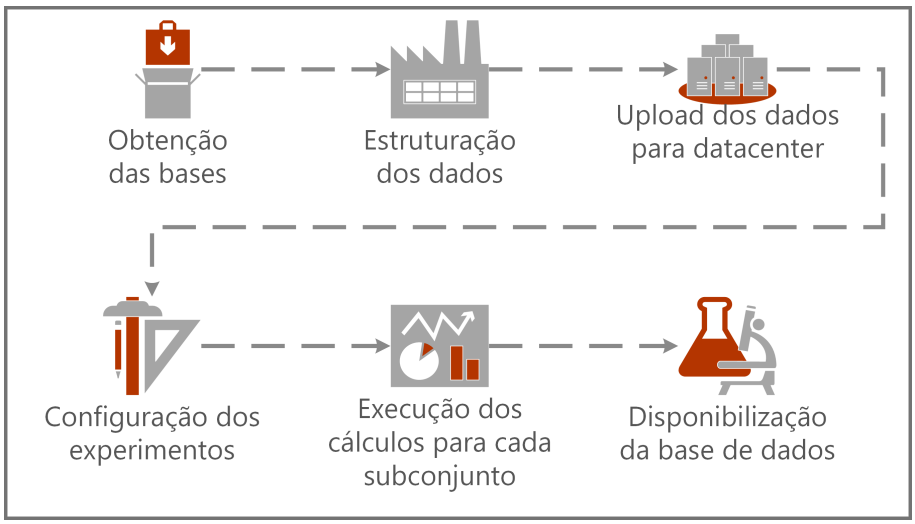

Figura 1. Visão geral do processo de caracterização das bases.

A escolha das métricas coletadas foi feita a partir de um levantamento de artigos relacionados à análise de redes sociais, predição de links e redes de coautoria. São exemplos de métricas que foram coletadas: quantidade de nós, grau médio, densidade, área de atuação dos pesquisadores, diâmetro, quantidade de pares, entre outras.

Para realização dos experimentos descritos foram desenvolvidas ferramentas em linguagem *Python* e *C#*. Para cálculo das métricas de rede foi utilizada a biblioteca *NetworkX*. A plataforma de computação em nuvem utilizada foi a *Microsoft Azure*. Os dados foram armazenados em bancos de dados *MySQL* e em repositórios do tipo blob, na *Azure*. Foram utilizadas sete instâncias de máquinas otimizadas para uso de memória, com diferentes configurações de hardware, disponibilizando de 7 GB a 224 GB de RAM para cada experimento. A automação do processo favorecerá a atualização da base com novos conjuntos e métricas, que será feita de acordo com a necessidade das pesquisas relacionadas e sugestões recebidas.

3. Considerações Finais

Apesar da grande disponibilidade de bases de coautoria, há uma carência em levantamentos de caracterização dessas redes. Este trabalho irá contribuir com pesquisas relacionadas a áreas como análise de redes sociais e predição de links que precisam escolher conjuntos adequados para seus experimentos. O uso da computação em nuvem favoreceu a análise de milhares de conjuntos, sendo alguns com poucas unidades e outros com mais de duzentos mil autores. Ao término desta pesquisa, será disponibilizado um repositório público de redes de coautoria com suas principais métricas de rede.

Referências

Liu, X., Bollen, J., Nelson, M. L., and Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing & management*, 41(6):1462–1480.

Mena-Chalco, J. P., Digiampietri, L. A., and Cesar-Jr, R. M. (2012). Caracterizando as redes de coautoria de currículos lattes. In *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 1–12.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Wang, P., Xu, B., Wu, Y., and Zhou, X. (2015). Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38.