EXTRAÇÃO DE REDES DE COLABORAÇÃO DA PLATAFORMA LATTES: O CASO DOS PRIMEIROS AUTORES DE ARTIGOS DA ESCOLA REGIONAL DE BANCO DE DADOS [1]

Fernanda Maria de Souza [2], Rebeca Schroeder Freitas[3]

[1] Vinculado ao projeto "Redes Sociais Profissionais: Extração e Análise de Dados".
[2]Acadêmico (a) do Curso de Ciências da Computação – CCT – Bolsista PIVIC/UDESC.
[3] Orientador, Departamento de Ciências da Computação – CCT – rebeca.schroeder@udesc.br

A análise e a extração do conhecimento dos dados é uma área que vem crescendo fortemente nos últimos anos [1]. Neste âmbito, pesquisadores tentam desenvolver modelos que explicam o surgimento de várias características em redes de colaboração [2]. A rede de colaboração representa uma relação compartilhada pelos autores em certas áreas, em que o número de publicações conjuntas feitas pelos dois autores representa a força de seu relacionamento, enquanto o número de publicações ligadas ao autor fornece uma medida de seu sucesso de colaboração [2]. Neste trabalho, foram coletados dados a partir da plataforma Lattes, referentes aos primeiros autores de artigos de toda história da Escola Regional de Banco de Dados (ERBD), em suas 15 edições. O objetivo do trabalho é identificar e entender as tendências de colaboração de acordo com a atividade de publicações dos autores dessa comunidade específica, bem como o comportamento dos mesmos dentro da rede, com a extração, geração de grafos e análise desses dados.

Primeiramente, 167 de 221 primeiros autores tiveram seu currículo localizado no Lattes com a busca manual, devido à dificuldade de extração desses dados computacionalmente pela presença do sistema Captcha na plataforma de currículos. Os currículos desses pesquisadores foram transferidos em formato XML (Extensible Markup Language), uma linguagem de marcação que visa facilitar o compartilhamento e futura formatação de informações. Com os currículos de autores já transferidos, a análise e subdivisão em DataFrames para tabelas .csv foi feita com a linguagem de programação Python para identificar diversas informações desses pesquisadores: (i) atuações profissionais, (ii) dados sobre suas formações acadêmicas, (iii) periódicos publicados, (iv) trabalhos em eventos e (v) capítulos de livro.

Após, com as informações sobre os autores estabelecidas, o próximo passo foi a geração do grafo de coautoria. Um problema encontrado durante a elaboração do grafo foi a deduplicação de nomes e/ou citações de autores, devido a diferentes formas de cadastros que foram realizadas envolvendo um mesmo autor. Para resolver o problema, foi utilizado a biblioteca pandas_dedupe do Python, que a partir de um treinamento prévio do usuário, como na Figura 1, identifica padrões para resolver as duplicações de nomes equivalentes.

Com a deduplicação de nomes, o grafo foi refinado e gerado com sucesso, para após ser visualizado com o software Gephi, conforme a Figura 2. Os autores foram representados como nós, ao todo 2931 envolvidos na rede de colaboração, enquanto que publicações em comum foram representadas como as ligações entre esses autores (arestas), com 12157 arestas no total. As publicações em comum consideradas foram periódicos (arestas em azul) – 72,02% e trabalhos em eventos (arestas em rosa) – 27,98%.

Através de métricas de Teoria dos Grafos, foi possível aplicar diferentes cálculos sobre a rede, como: (i) PageRank (identificar os autores mais importantes da comunidade), (ii) Clustering

Apoio:

Coefficient (cálculo da tendência de determinados autores colaborarem entre si), (iii) Betweenness Centrality (detectar os autores que mais tem controle sobre a rede) e (iv) Comprimento Médio de Caminho (número de publicações para chegar de um autor à outro).

Portanto, como resultado final do trabalho, destaca-se a extração e análise de dados dos primeiros autores de artigos da ERBD visando representar como a comunidade se comporta de acordo com os autores que a compõem e suas publicações ao longo da carreira. Como trabalhos futuros, deseja-se identificar a evolução dos autores dessa rede a partir de suas formações e afiliações, aumentando também a comunidade analisada.

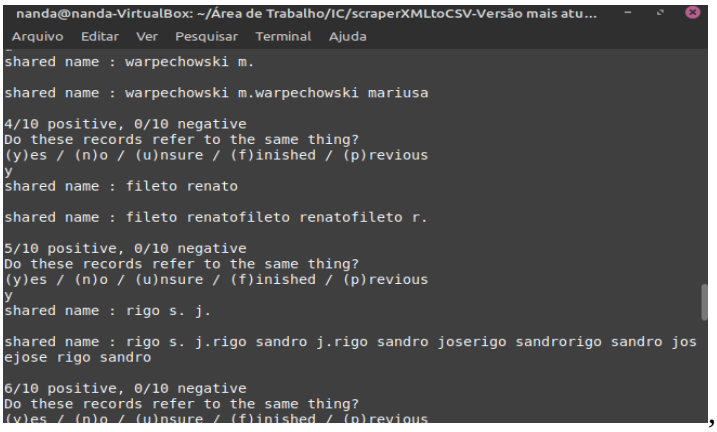

Figura 1. *Console para Deduplicação de nomes*

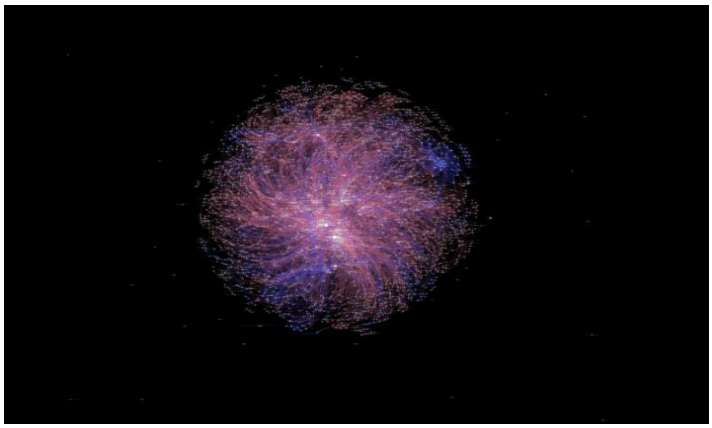

Figura 2. *Grafo de Coautoria dos primeiros autores de artigos da ERBD*

Palavras-chave: Escola Regional de Banco de Dados, Redes de Colaboração, Métricas de Análise de Dados.

[1] Stattner, E., Collard, M. 2017. Link Clustering for Extracting Collaborative Patterns in a Scientific Co-Authored Network. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (ASONAM '17).
[2] Pandey, A. et al. 2015. Analyzing Link Dynamics in Scientific Collaboration Networks: A Social Yield Based Perspective. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15).