Extração e Processamento de Dados de Produção Científica: o caso da Escola Regional de Banco de Dados[1]

Otávio Almeida [2], Rebeca Schroeder Freitas [3]

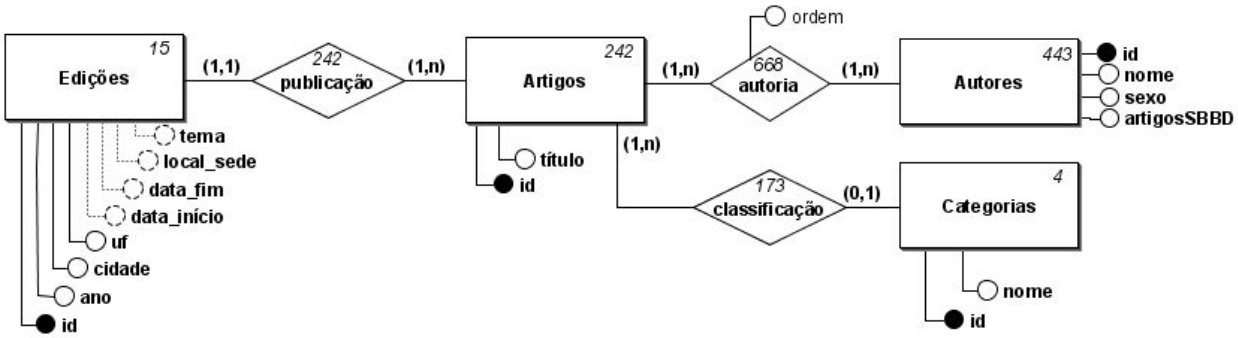
[1] Vinculado ao projeto "Redes Sociais Profissionais: Extração e Análise de Dados".
[2] Acadêmico do Curso de Ciências da Computação – CCT – Bolsista PROIP/UDESC.
[3] Orientador, Departamento de Ciências da Computação – CCT – rebeca.schroeder@udesc.br.

Palavras-chave: Rede de Coautoria, Escola Regional de Banco de Dados, Análise de Dados.

Em diversos trabalhos presentes na literatura, a análise de comunidades científicas vem sendo desempenhada através do mapeamento de relações de coautoria em artigos científicos [1][4]. Estas relações permitem representar uma rede de colaboração, onde diversas análises podem ser aplicadas para favorecer ao autoconhecimento destas comunidades, identificar padrões, e até mesmo prever o futuro das interações profissionais [3]. Neste trabalho, os dados referentes à rede de colaboração relacionada ao evento *Escola Regional de Banco de Dados* (ERBD) foram coletados e analisados. A ERBD é um evento anual promovido pela Sociedade Brasileira de Computação, que visa a divulgação científica na área de banco de dados.

Um dos principais desafios para este trabalho foi recuperar os registros de publicações dos 15 anos de edição do evento, visto que muitos dos anais não estavam disponíveis na Web. Assim, para algumas edições (anos) foi necessário recuperar os dados dos anais impressos do evento. Um banco de dados relacional foi construído para reunir os dados coletados. A Figura 1 apresenta o esquema conceitual deste banco, que recupera detalhes de cada edição, seus autores, artigos e categorias de artigos. Este esquema conceitual, definido através do modelo Entidade-Relacionamento [2], define que uma edição poderá ter vários artigos, cada qual com diversos autores. Os autores são relacionados a uma ordem, que representa a sequência do referido autor na lista de autores de um artigo. No total foram identificados 15 edições, 242 artigos, 443 autores, 668 relacionamentos de autores com artigos (autoria) e 4 tipos de categorias de artigos. Um problema tratado durante a extração de dados foi a deduplicação de nomes de autores. Este problema ocorre quando um autor produz vários artigos, mas que são listados com variações de seus nomes.

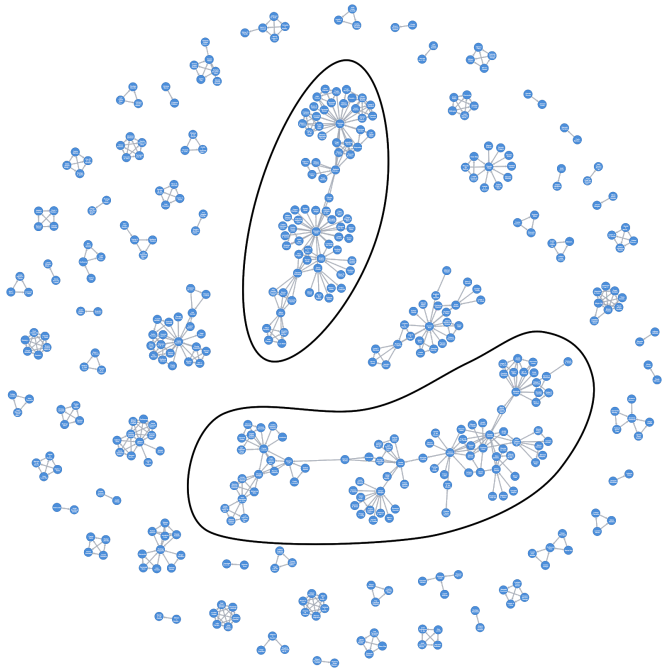
Figura 1. *Esquema Conceitual dos Dados Extraídos*



A partir dos dados coletados, foi possível extrair dados bibliométricos do evento, produzindo estatísticas e gráficos relacionados dos quais pode-se exemplificar: (i) Número de Artigos por Edição e respectivas Categorias de Artigos, (ii) Média de Autores por Artigo por Edição, (iii) Número de Autores por Sexo por Edição e (iv) Temas com maior repetição nos títulos dos artigos.

Adicionalmente, uma rede de coautoria foi construída utilizando o banco de dados Neo4J, o qual representa os autores como nós e as arestas que associam dois autores quando estes publicaram juntos em artigos. Através da análise desta rede foi possível identificar grupos de pesquisa, autores prolíficos e suas conexões. A Figura 3 apresenta a rede de coautoria da ERBD.

Figura3. *Rede de Coautoria da ERBD.*



Como resultado deste trabalho destaca-se a construção de uma base de dados sobre a ERBD, que pode ser utilizada para manter as informações do evento a ser atualizada com as próximas edições. Os dados contidos nesta base são capazes de representar a evolução da comunidade, podendo servir como auxílio para planejar novos rumos para o evento e favorecer ao auto-conhecimento desta comunidade. Como trabalhos futuros, deseja-se aumentar os dados coletados para considerar outros eventos que os autores da ERBD também publicaram, utilizando para isto publicações indexadas. Além disso, dados sobre o perfil dos autores estão sendo coletados da Plataforma Lattes, para analisar a formação e afiliação dos autores no futuro.

[1] Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., and San-toro, N. (2011). On the temporal analysis of scientific network evolution. In: CASoN, pages 169–174.
[2] Batini, C., Ceri, S., and Navathe, S. B. (1992).Conceptual DatabaseDesign: an Entity-relationship approach. Benjamin-Cummings Publishing Co.
[3] Brandão, M. A., de Melo, P. O. S. V., and Moro, M. M. (2017). Tie strength dynamics over temporal co-authorship social networks. In WI ’17, page 306–313.
[4] Júnior, P. S. P., Laender, A. H. F., and Moro, M. M. (2011). Análise da rede de coautoria do simpósio brasileiro de bancos de dados. In XXVI SBBD - Short Papers, pages 131–138. SBC.