

Identificação de grupos por similaridade nas redes de colaboração acadêmica das áreas do CNPq

Wellington Barbosa Rodrigues, Jesús P. Mena-Chalco

Universidade Federal do ABC (UFABC)
{wellington.rodrigues, jesus.mena}@ufabc.edu.br

Resumo: A divulgação científica pode acontecer através de produções em conjunto por vários pesquisadores. Logo, possibilitando identificar relações de colaboração acadêmica. Cada área da ciência tem suas próprias características de colaboração entre autores, podendo apresentar semelhanças entre elas. Neste trabalho apresentamos os resultados iniciais da identificação de similaridade entre diferentes áreas do conhecimento, usando como base a produção científica dos bolsistas de produtividade em pesquisa do CNPq. O agrupamento das áreas pode acontecer entre áreas pertencentes a mesma grande área ou de grandes áreas distintas.

1. Introdução

No processo de divulgação científica é comum que os autores realizem suas produções em conjunto com outros pesquisadores. Dessa maneira se tornar possível a identificação de relacionamentos de coautoria ou citações entre pesquisadores [1] e com isso permitindo o uso de técnicas de teoria dos grafos para análise dessas redes.

Como forma de organizar as produções científicas, órgãos brasileiros de fomento a pesquisa (e.g. CAPES e CNPq) adotam divisões de áreas do conhecimento. Com essa divisão é possível a categorização de pesquisadores em determinadas áreas, normalmente, as que eles tem maior contribuição. Fazendo uso dessas categorias é possível dividir as colaborações científicas em grupos de áreas do conhecimento.

Este é um trabalho inicial que busca identificar as semelhanças entre as distintas áreas do conhecimento do CNPq fazendo uso de métricas da topologia de redes de coautoria em produções acadêmicas. Como representantes de cada área do conhecimento foram escolhidos, para este trabalho, os pesquisadores bolsistas produtividade em pesquisa (PQ) do CNPq, tais pesquisadores são reconhecidos pelo CNPq como sendo pesquisadores influentes em suas áreas.

2. Método

Para geração dos dados foram baixados os currículos Lattes de todos os PQs ativos no CNPq e separados por área do conhecimento, totalizando 85 áreas. As áreas com menos de 50 pesquisadores foram descartadas do processo, ficando 71 áreas. Cada área foi processada e convertida para um grafo, os pesquisadores foram considerados nós e a relação de coautoria em publicações (artigos em periódicos, artigos em eventos, livros, e capítulos de livros) foram arestas. Com isso originando 71 grafos de coautoria, um para cada área.

Nesses grafos foram extraídas 4 informações de topologia do grafo: Grau médio, densidade, assortatividade e transitividade. Essas medidas foram escolhidas por serem medidas topológicas normalizadas [2], logo, não são fortemente influenciadas pelo tamanho do grafo.

Foi utilizada a técnica de agrupamento hierárquico aglomerativo [3], tendo como critério de agrupamento a média da distância Euclidiana das ligações entre os grupos.

3. Resultados

A Figura 1 mostra os agrupamentos das áreas do conhecimento realizados através da técnica escolhida. Na vertical são apresentadas todas as áreas consideradas neste trabalho e na horizontal a distância calculada entre cada área. Como critério de corte para definição dos grupos foi utilizada a distância 1,7, o que resultou em 4 grupos distintos, que podem ser identificados pelas cores. O menor grupo contém 5 áreas e o maior com 37 áreas.

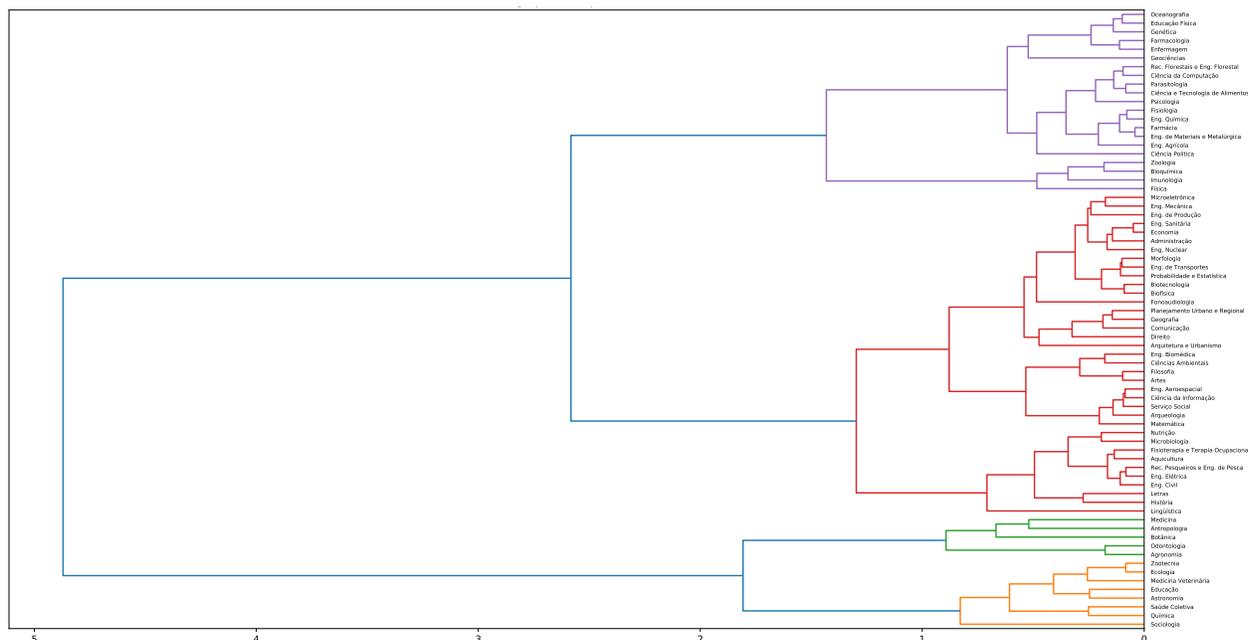


Figura 1. Agrupamento hierárquico das áreas de conhecimento do CNPq.

Os agrupamentos podem acontecer entre áreas da mesma grande área, como é o caso da Zoologia e a Bioquímica (distância de 0,1809) ou em grandes áreas diferentes como no caso de Engenharia Sanitária (Engenharias) e Economia (Ciências Sociais Aplicadas) que tem uma distância de 0,0490.

As duas áreas com menor distância entre si são Eng. de Materiais e Metalúrgica, pertencente à grande área Engenharias, e Farmácia, da grande área Ciências da saúde, com distância de 0,0414. Enquanto as duas áreas mais distantes são Eng. Biomédica (grande área de Engenharias) e Botânica (grande área de Ciências Biológicas) com distância de 8.5526. A distância média entre as 71 áreas é de 2,6245.

4. Conclusões

Com base nos dados obtidos através do uso de agrupamento hierárquico aglomerativo pode-se identificar que a grande área de uma pesquisa não impacta diretamente no comportamento das redes de colaboração entre os pesquisadores de suas áreas, pois cada área tem suas características, que podem ser similares a áreas que não lhe são correlatas.

Como trabalhos futuros, pretendemos utilizar técnicas mais robustas e avançadas de inteligência artificial para melhorar o agrupamento e definir de maneira mais precisa o perfil de cada área e grande área, além do cálculo de mais métricas topológicas viáveis para serem utilizadas como atributos das redes durante a análise.¹

5. Referências

- [1] CALAZANS, Marcos Moraes; FURTADO, Renata Lira; TOMAÉL, Maria Inês. Redes de Citação: estudo de rede de pesquisadores a partir da competência em informação. **Em Questão**, v. 21, n. 2, p. 181-202, 2015.
- [2] MENA-CHALCO, Jesús P.; DIGIAMPIETRI, Luciano A.; CESAR-JR, Roberto M. Caracterizando as redes de coautoria de currículos Lattes. **Anais do I Brazilian Workshop on Social Network Analysis and Mining**, SBC, 2012.
- [3] DAY, William HE; EDELSBRUNNER, Herbert. Efficient algorithms for agglomerative hierarchical clustering methods. **Journal of Classification**, v. 1, n. 1, p. 7-24, 1984.

¹O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.