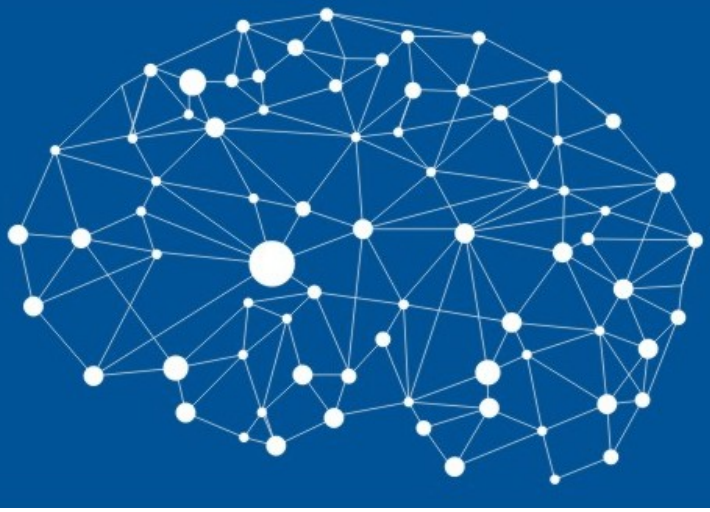


Um processo para identificação de colaborações em repositórios de publicações científicas

Thiago Magela Rodrigues Dias[1]; Patricia Disa[2]; Gray Moita[3]

DIAS, T. M. R.; DISA, P.; MOITA, G.. Um processo para identificação de colaborações em repositórios de publicações científicas In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A98

1, 2, 3
;

UM PROCESSO PARA IDENTIFICAÇÃO DE COLABORAÇÕES EM REPOSITÓRIOS DE PUBLICAÇÕES CIENTÍFICAS

Eixo temático: Colaboração

Modalidade: Apresentação oral

1 INTRODUÇÃO

Nos últimos anos, além da produção científica, tem havido um constante crescimento no estudo das redes em relação às diversas disciplinas que vão desde a ciência da computação a áreas como a economia e sociologia. Uma rede pode ser caracterizada como um grafo, que consiste de um conjunto de nós (vértices) e ligações (arestas) entre os nós. Estas ligações podem ser, direcionadas ou não direcionadas, e podem, opcionalmente, ter um peso associado. A Internet, por exemplo, pode ser considerado um exemplo de uma rede importante e amplamente estudada em diversas áreas atualmente. Entre os vários tipos de redes, existem as redes sociais. Uma rede social é um conjunto de pessoas ou grupos que têm algum tipo de relação entre eles (Newman, 2001a).

No domínio científico, um exemplo de uma rede social é a rede de colaboração científica que pode ser observada como um grafo no qual os vértices correspondem aos autores de publicações científicas e as arestas correspondem a relação de co-autoria. Neste tipo de rede, as arestas podem ou não ser ponderadas. Vários trabalhos tem objetivado analisar redes de colaboração científica para compreensão de como grupos de pesquisa tem realizados seus trabalhado ou como determinada área realiza seus estudos (Barabási, 2004; Newman, 2004; Newman, 2001b; Newman, 2001c).

Aliado a isso, a presença de dados de publicações disponíveis em diferentes formatos e em diferentes repositórios dificulta a realização de consultas por parte de usuários que necessitam de uma visão unificada desses dados ou da identificação de grupos de indivíduos que estejam trabalhando com determinado tema em diferentes instituições e regiões.

Diante disso, este trabalho tem como objetivo apresentar um método para a identificação de colaborações científicas em grandes repositórios de dados com baixo custo computacional. Para o estudo de caso serão analisados os currículos Lattes como principal

fonte de dados. Para caracterizar a relação de colaboração entre pares de pesquisadores, são analisados trabalhos que 2 (dois) ou mais pesquisadores realizaram em conjunto.

2 TRABALHOS RELACIONADOS

Para Revoredo et al. (2012), redes como as de comunidades científicas de formação recente possuem poucos parâmetros para classificação de assuntos de interesse e pouco entendimento tanto da existência como o potencial de relações de colaboração. Nestas comunidades, a compreensão de sua composição e tendências de interesse se beneficia de técnicas de descoberta de conhecimento a partir de seus artefatos principais de produção – publicações.

No trabalho de (Newman 2001a) é apresentada uma avaliação extensa sobre características sociais das redes de co-autoria em redes científicas de computação, biologia, física e medicina no período de 1995 a 1999. Já em (Newman 2004), o autor faz análise de co-autoria para identificar propriedades estatísticas, procurando por padrões comuns.

Em Sampaio et al. (2012) são apresentados diversos trabalhos de análise de redes científicas brasileiras que utilizam técnicas de mineração de dados e técnicas de data warehouse para análise dos dados.

Apesar das dificuldades citadas, as redes de colaboração científica têm sido alvo de vários trabalhos devido a sua riqueza de dados disponíveis nos mais diversos formatos e repositórios. (Barbosa et al. 2011; Menezes et al. 2012).

3 DESENVOLVIMENTO

Para a caracterização da rede de colaboração apresentada neste trabalho, foi utilizado a Plataforma Lattes como principal fonte de informação. A Plataforma Lattes é de grande relevância para a produção técnico-científica brasileira, e ganhou nos últimos anos projeção internacional, através de parcerias estabelecidas com países da América e Europa, formando a Rede ScienTI (Silva e Nascimento 2006).

Os currículos Lattes se tornaram um padrão nacional utilizado na avaliação individual das atividades científicas, acadêmicas e profissionais, agrega dados de pesquisadores de todas as áreas do conhecimento, tornando a Plataforma uma fonte extremamente rica para investigar e compreender o comportamento de diversos grupos de pesquisa (Digiampietri et al. 2012).

Todo o processo de extração dos dados é realizado pela plataforma proposta por Dias e Moita (2014). Todo o processo de extração e integração dos dados é dividido em três partes principais denominadas de Extração, Processamento e Visualização. Porém, neste trabalho só foram utilizados os resultados da etapa de extração dos currículos Lattes, tendo em vista, que a identificação das colaborações utiliza-se dos dados das publicações cadastradas em cada um dos artigos, como título e autores.

Para a identificação das colaborações tendo em vista a grande quantidade de autores e publicações a serem analisados, um algoritmo que tenha baixo custo computacional e com alta taxa de precisão foi proposto. O intuito é que seja possível realizar a identificações de colaborações científicas de grande quantidade de dados em tempo hábil. Algoritmo 1.

Algoritmo 1 – Algoritmo para identificação de colaboração

Identification-Collaboration

1. *n* ← number of articles author
2. **for** *i* ← 1 **to** *n*
3. *x* ← *string[i]* // x is article title [i]
4. *x* ← *stopword[x]* // removes token without semantic value
5. *x* ← *normalization[x]* // remove whitespace and accentuation
6. *x* ← *lowercase[x]*
7. **if** *hash[x]* **in** *dictionary* // checks whether x is in the dictionary
8. *dictionary[x]* ← *id_author*
9. **else** *dictionary* ← *x, id_author*

Fonte – Próprios autores

Para realizar as identificações das colaborações entre os autores, cada título de um trabalho cadastrado em um determinado currículo passa por uma transformação que tem como objetivo obter o título sem as palavras que não tenham valor semântico, sem acentuação e sem os espaços entre elas. Esta estratégica tem como intuito minimizar a ocorrência de erros gramaticais que podem estar inseridos nos títulos dos artigos. Consequentemente, todo o texto é padronizado em letras minúsculas e a string resultante é concatenada com o ano da publicação para posteriormente ser transformada em uma chave que representa o trabalho em análise, passos 2 a 6 do algoritmo.

Após a transformação verifica-se no dicionário utilizado para a caracterização da rede de colaboração se a chave já está presente. Caso a chave já exista no dicionário, o

identificador do autor do currículo em análise é vinculado à chave, caso contrário, são inseridos a chave e o identificador no referido dicionário. Tabela 1.

Tabela 1 – Exemplo de dicionário construído pelo algoritmo de identificação

Chave do Dicionário	Autores
modelagemcaracterizacaoredescientificasestudosobreplataformalattes2013	Id01, Id25
studyaboutinfluenceacademicperformancestudentsuserssocialnetworks2013	Id25, Id175, Id98
analysiscollaborationnetworksscientificpublications2013	Id01, Id98, Id67
...	
processoparaidentificacaocolaboradoresredescientificas2013	Id01, Id28, Id174

Fonte: Próprios autores

Importante destacar que cada um dos currículos cadastrados na Plataforma Lattes possui identificador único. Este identificador é utilizado tanto para caracterizar o usuário da plataforma como também para permitir o acesso ao currículo individual de cada usuário.

Uma dificuldade encontrada pelos autores no cadastro de uma publicação está no momento da citação de seus colaboradores. Devido à falta de padronização e possibilidade de ambiguação no nome de citação dos colaboradores, não é possível garantir que o nome informado na lista de colaboradores pertença a somente um usuário da plataforma, o que dificulta a caracterização da rede de colaboração pelos nomes de citação informados em cada uma das publicações cadastradas. Destacando a importância do método proposto.

4 RESULTADOS

Para efeito de comparação, foram utilizados os currículos dos professores integrantes de um Programa de Pós Graduação contendo 25 docentes. Logo, este grupo foi utilizado para avaliar o algoritmo proposto em comparação com uma solução que utiliza a estratégia de comparação entre títulos de artigos, verificando a equivalência entre títulos pelo cálculo da distância de Levenshtein. O grau real corresponde a quantidade de colaboradores de um determinado autor, calculado manualmente, para que dessa forma pudesse avaliar os métodos.

Como resultado, o algoritmo obteve melhores resultados já que conseguiu ter uma precisão de 100% nas identificações, ou seja, todas as colaborações indicadas são colaborações reais e um percentual de 97,61% de revocação, superando os 95,23% do método

utilizado para fins de comparação. A revocação indica o percentual de colaborações reais que foi possível identificar. Tabela 2.

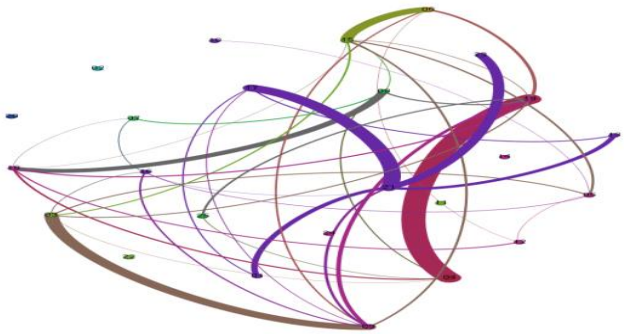
Tabela 2 – Análise comparativa do método proposto

Métrica	Método Comparado	Método Proposto
Precisão	100%	100%
Revocação	95,23%	97,61%

Fonte – Próprios Autores

Com a aplicação de métodos para a identificação das colaborações, utilizando elementos como os citados anteriormente, é possível a modelagem das redes de colaboração e diante disso, várias métricas para análise de redes podem ser aplicadas objetivando extrair conhecimento sobre como estes grupos estão estruturados, como colaboram, possibilitando alavancar a produção científica nacional com este conhecimento adquirido. Exemplo da rede de colaboração caracterizada pode ser observada na Figura 1.

Figura 1 – Rede caracterizada após o processo de identificação



Fonte – Próprios autores

Na Figura 1 cada vértice representa um pesquisador e o tamanho do vértice representa a quantidade de publicações que ele possui. Já as arestas entre os vértices são caracterizadas por publicações que pesquisadores realizaram em colaboração e a espessura das arestas indicam a quantidade de trabalhos que dois pesquisadores realizaram em conjunto. As cores dos vértices indicam as áreas de atuação de cada pesquisador. Diante disto, diversas métricas para classificação, agrupamento, ranqueamento, dentre outras podem ser aplicadas.

5 CONCLUSÕES

Este trabalho apresenta um processo para a identificação de colaborações científicas de currículos cadastrados na Plataforma Lattes. O método é eficiente no processo de identificação de colaboração entre os autores, além de possibilitar a caracterização de redes de colaboração com alto grau de precisão.

O algoritmo para identificação de colaborações apresenta excelentes resultados com relação a sua precisão por utilizar o identificador que representa de forma única um determinado autor. A grande vantagem da adoção deste método é com relação ao seu custo computacional. Como é realizada apenas uma comparação para cada título de artigo, é possível caracterizar a rede de colaboração a um custo linear $\theta(n)$, diferentemente dos métodos que trabalham com outras técnicas como, por exemplo, validação cruzada a um custo polinomial $\theta(n^2)$. Resultado disto é a viabilidade para se construir redes com um número muito grande de títulos, tornando o algoritmo proposto uma excelente alternativa para a identificação de colaboração em repositórios com grande quantidade de dados.

REFERÊNCIAS

BARABASI, A. L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, v. 5, n. 2, p. 101-113, 2004. ISSN 1471-0056.

BARBOSA, E. M.; MORO, M. M.; LOPES, G. R.; OLIVEIRA, J. P. M.(2011) VRRC: Uma Ferramenta Web para Visualização e Recomendação em Redes de Co-autoria. In: VIII Sessão de Demos, **Simpósio Brasileiro de Banco de Dados**, Florianópolis.

DIGIAMPIETRI, L. A. ; MENA-CHALCO, J. ; ALCAZAR, J. J. P. ; TUESTA, E. F. ; DELGADO, K. V.; MUGNAINI, R. ; SILVA, G. S. (2012). Minerando e Caracterizando Dados de Currículos Lattes. In: **Brazilian Workshop on Social Network Analysis and Mining** (BraSNAM), Curitiba, PR, Brasil.

DIAS, Thiago. M. R. ; MOITA, Gray F ; DIAS, Patrícia M.; MOREIRA, Tales H. J. . Identificação e Caracterização de Redes Científicas de Dados Curriculares. **iSys: Revista Brasileira de Sistemas de Informação**, Rio de Janeiro , v. 07, p. 05-18, 2014.

MENEZES, V.S.A.; SILVA, G. Z.; SOUZA, J. M. (2012) Análise de Redes Sociais Científicas: Modelagem Multi-relacional. In: **Brazilian Workshop on Social Network Analysis and Mining** (BraSNAM), Curitiba- PR.

NEWMAN, M. E. J. (2001). "Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality." **Physical Review** E 64(1): 016132.

NEWMAN, M. E. J. (2001). "Scientific collaboration networks.I. Network construction and fundamental results." **Physical Review** E 64(1): 016131.

NEWMAN, M. E. J. (2001). "The structure of scientific collaboration networks." **proceedings of the national academy of sciences** 98(2): 404-409.

NEWMAN, M. E. J. (2004). "Coauthorship networks and patterns of scientific collaboration." **proceedings of the national academy of sciences** 101(suppl 1): 5200-5205.

REVOREDO, k.,ARAÚJO R., SILVEIRA B.; MURAMATSU T. (2012). Minerando publicações científicas para análise da colaboração em comunidades de pesquisa.In: **Brazilian Workshop on Social Network Analysis and Mining** (BraSNAM), Curitiba- PR.

SAMPAIO, J. O.; FARIA, F. F.; PERORAZIO, R. A.; AQUINO, E. C. (2012) Análise da Produtividade da Rede Social de Computação do Brasil. In: **Brazilian Workshop on Social Network Analysisand Mining** (BraSNAM), Curitiba- PR.

SILVA, L. C. R.; NASCIMENTO, H. A. D. (2006) Visualizando Bases Curriculares. In: **Encontro de Tecnologia e Informática** - ETI, 2006, Goiânia. Encontro de Tecnologia e Informática - ETI/2006.