

ADOÇÃO DE TÉCNICAS DE MINERAÇÃO DE TEXTO PARA CARACTERIZAÇÃO E ANÁLISE DE REDES DE COLABORAÇÃO CIENTÍFICA

Rafael Aparecido Santos

Universidade do Estado de Minas Gerais, Belo Horizonte, MG, Brasil
santosrafael@live.com

Jhonatan Fernando de Oliveira

Universidade Federal de São João del-Rei, São João del-Rei, MG, Brasil
jhonatan.oliveira@uemg.br

Thiago Magela Rodrigues Dias

Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG, Brasil
thiagomagela@gmail.com

1 INTRODUÇÃO

No início da ciência moderna, em meados do século XVI, as descobertas eram compartilhadas entre os pesquisadores através de cartas. A partir da segunda metade do século XVII foram criados os primeiros periódicos, que formalizaram o processo de comunicação no meio científico (AQUINO, 2002). Desde então, com o aumento do número de pesquisadores e, conseqüentemente, da produção científica, foi sendo notada a importância da preservação e armazenamento de documentos notáveis para a ciência.

De acordo com Borko (1968), a ciência da informação é a disciplina que investiga as propriedades e o comportamento informacional e se preocupa com a origem, coleção, organização, armazenamento, recuperação, interpretação, transmissão, transformação e utilização da informação.

Um dos aspectos que podem ser analisados dentro da ciência da informação é a colaboração entre os pesquisadores. Essa análise é interessante, pois segundo Nudelman e Landers (1972), artigos publicados



com autoria conjunta têm maior credibilidade junto à comunidade científica do que artigos de autoria única.

Entretanto, as informações necessárias para tal análise não estão acessíveis de forma simples. Elas precisam ser encontradas em meio às publicações. Uma forma de extrair informações relevantes em grandes bases de dados textuais é através da mineração de texto.

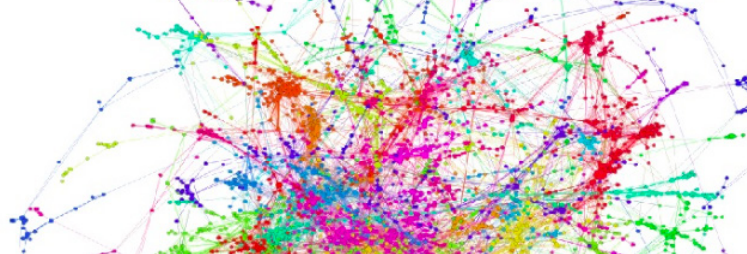
Na obra de Dias et al. (2016), os autores utilizam técnicas de mineração de texto para analisar as redes de colaboração científica entre os currículos cadastrados na Plataforma Lattes. Já na obra de Mattos e Job (2008), as autoras analisaram manualmente a participação dos autores brasileiros nas publicações da revista *Scientometrics*. Neste trabalho, utilizou-se técnicas de mineração de texto para extrair e processar dados de toda a base de artigos da revista *Scientometrics* e em seguida, gerou-se grafos que facilitam a visualização e análise das redes de colaboração científica entre as instituições que nela publicaram.

2 DESENVOLVIMENTO

Inicialmente foi realizada a extração dos dados no site do periódico *Scientometrics*. Foram extraídos de cada artigo o código identificador da publicação (*Digital Object Identifier*), os nomes dos autores, os nomes das instituições vinculadas, a cidade e o país de cada instituição.

Um dificultador identificado está relacionado aos nomes das instituições em cada artigo não vêm de uma fonte unificada, acarretando em uma mesma instituição sendo referenciada de formas diferentes, como por exemplo, a instituição “UFSCar – Universidade Federal de São Carlos” é a mesma que “Universidade Federal de São Carlos”.

Como cada nome de instituição diferente gera um nó na rede de colaboração, se fossem feitas comparações simples entre as cadeias de caracteres, seriam gerados mais de um nó representando a mesma instituição. Para evitar este problema, foi feito um processo de desambiguação, utilizando a função *SequenceMatcher*, do módulo *difflib* disponível de forma nativa na linguagem *Python*. Segundo Hellmann (2011), a função *SequenceMatcher* utiliza o algoritmo proposto por Ratcliff e Metzener



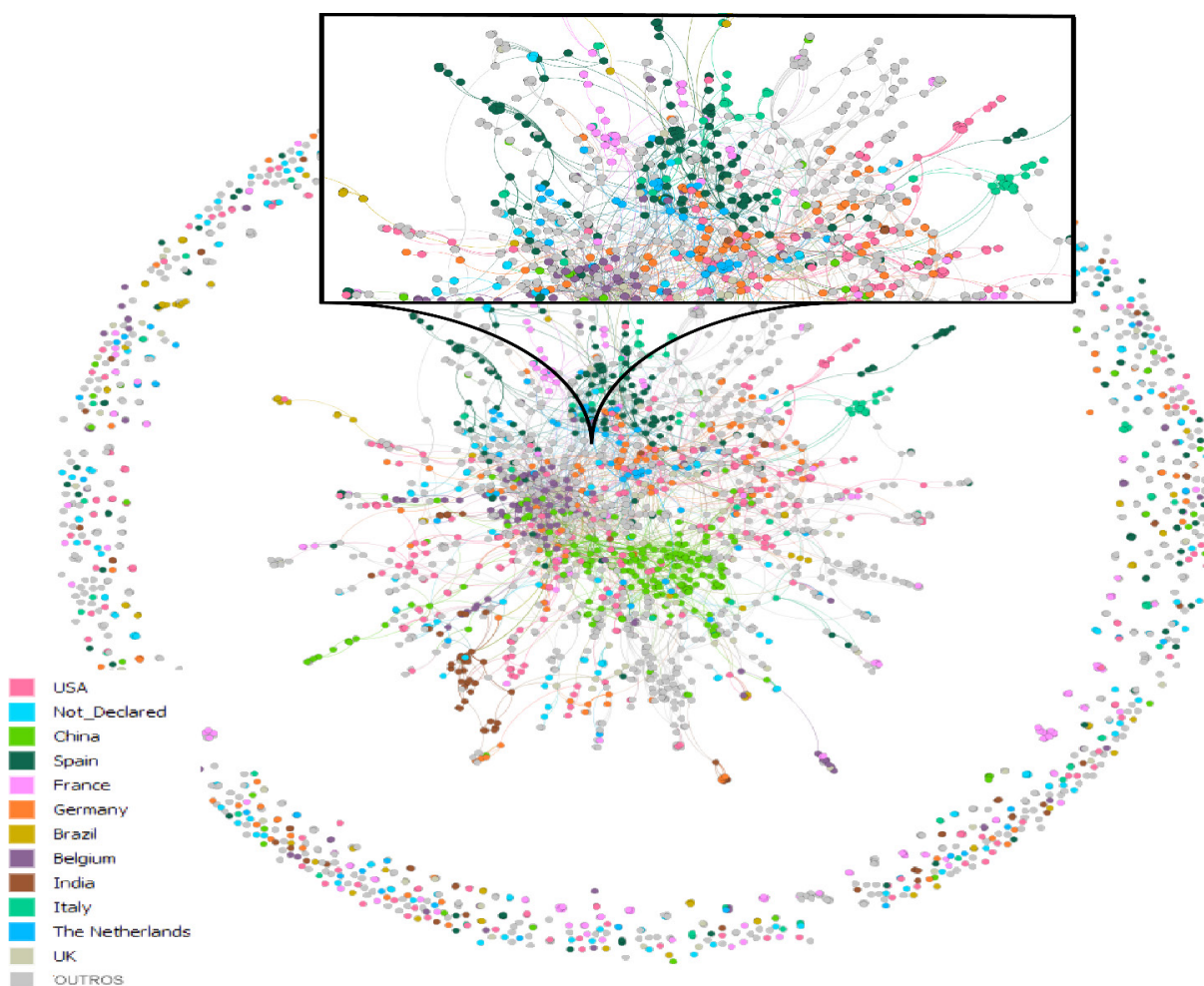
(1988) para retornar um valor entre 0 e 1 que representa o quão semelhante uma sequência é de outra, sendo 1, o resultado da comparação de sequências exatamente iguais.

Para o trabalho proposto, foram consideradas iguais as instituições cujos nomes tinham 90% ou mais de semelhança, ou seja, comparações em que a função *SequenceMatcher* retornou 0,9 ou mais. Os casos em que as semelhanças entre os nomes das instituições eram acima de 85% foram re-comparados. Nestes casos, as instituições em que a sequência de caracteres composta pelo nome, a cidade e o país da instituição juntos apresentavam semelhança acima de 90% também foram consideradas as mesmas instituições. Esses números foram definidos após a realização de testes por amostragem.

Já para a geração de arestas, foi considerada a vinculação de diferentes instituições em um mesmo artigo, ou seja, para cada publicação onde havia colaboração entre instituições, foi armazenado no arquivo de arestas.

3 RESULTADOS

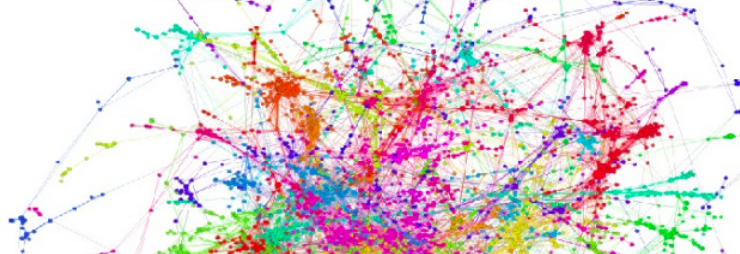
Foram extraídos do site da revista, dados referentes a todos os 5.036 artigos disponíveis no site até 18 de outubro de 2017, data em que os dados foram extraídos. O processamento gerou o número de 3.339 nós (instituições) e 5.071 arestas (colaborações). Como nem todos os artigos publicados possuem colaboração, foram encontrados 717 nós que não tinham arestas, esse número equivale a 21,47% do total de nós.

**FIGURA 1 - GRAFO APRESENTANDO TODOS OS NÓS E ARESTAS ENCONTRADOS**

Fonte: Elaborado pelos autores, 2018.

A Figura 1 apresenta todos os nós e arestas encontrados. Os Estados Unidos da América foi a nação com o maior número de instituições vinculadas, foram 365 no país, representando 10,93% do total. O segundo país com maior número de instituições foi a China, com 223 elementos, representando 6,68%.

Aplicando alguns filtros, foi possível visualizar as redes de colaborações de forma mais clara. A rede apresentada na Figura 2 foi gerada utilizando o filtro de intervalo de grau, onde são exibidas todas as instituições com mais de 20 arestas, ou seja, 20 colaborações.



4 CONCLUSÃO

Através do uso de técnicas de mineração de texto foi possível obter e padronizar informações contidas em todo o repositório de artigos da revista *Scientometrics*. Com o sucesso desse processo, foi possível visualizar, através da representação por grafos, as redes de colaboração entre as instituições com publicações na revista, conforme proposto pelo trabalho. Além disso, pode-se concluir que as colaborações tendem a ocorrer com maior frequência entre instituições que são geograficamente próximas.

REFERÊNCIAS

AQUINO, M. D. A. O campo da Ciência da Informação: gênese, conexões e especificidades. **Universitária/UFPB**, João Pessoa, p. 61-86, 2002.



BORKO, H. Information Science: what is it? **American Documentation**, Santa Monica, v. 19, n. 1, p. 3-5, Jan. 1968.

DIAS, M. R. et al. Um processo para caracterização e análise de redes de colaboração científica institucional. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 48, p. 59-72, 2016.

HELLMANN, D. **The Python standard library by example**. [S.l.]: Addison-Wesley Professional, 2011.

MATTOS, A. M.; JOB, I. A produção científica brasileira no periódico Scientometrics de 1978 até 2006. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v. 13, n. 26, p. 47-61, 2008.

NUDELMAN, A. E.; LANDERS, C. E. The failure of 100 divided by 3 to equal 33 1/3. **The American Sociologist**, v. 7, n. 9, p. 9, Nov. 1972.

RATCLIFF, J. W.; METZENER, D. E. Pattern-matching-the gestalt approach. **Dr Dobbs Journal**, 411 BOREL AVE, SAN MATEO, CA 94402-3522, v. 13, n. 7, p. 46, 1988.