

UMA ESTRATÉGIA BASEADA EM ANÁLISES DE REDES PARA ESTUDO DO PROCESSO DE ORIENTAÇÃO NO BRASIL

Tales Henrique José Moreira

Mestre em Modelagem de Matemática e Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: tales.info@gmail.com

Thiago Magela Rodrigues Dias

Doutor Modelagem de Matemática e Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: thiagomagela@cefetmg.br

Patrícia Mascarenhas Dias

Doutoranda em Modelagem de Matemática e Computacional
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET)
E-mail: patriciamdias@gmail.com

Adilson Luiz Pinto

Doutor em Ciência da Informação
Universidade Federal de Santa Catarina (UFSC)
E-mail: adilson.pinto@ufsc.br

INTRODUÇÃO

Tendo em vista as possibilidades de visualização e entendimento do histórico de orientação e, conseqüentemente, a difusão do conhecimento, realizar a modelagem e a caracterização de redes genealógicas acadêmicas surge como uma alternativa interessante para a análise de como a ciência brasileira tem se propagado utilizando-se para tanto, dados de orientações. Para isso, as redes genealógicas acadêmicas podem também ser caracterizadas, facilitando o entendimento do processo de formação com a análise destas redes. De modo geral, a caracterização de uma rede se dá por um grafo composto por um conjunto de nós (vértices) e conexões (arestas) entre os nós.

Árvores genealógicas podem ser definidas como uma estrutura que representa todo ou parte do histórico dos antepassados de um indivíduo. Trata-se de uma representação gráfica que apresenta de forma hierárquica os antepassados, podendo ou não ter infor-

mações complementares que visam permitir um melhor entendimento do histórico de um indivíduo.

Diante disto, as árvores genealógicas acadêmicas são caracterizadas como árvores que representam hierarquicamente o histórico de um orientador e todos os seus orientados. Logo, caracterizando uma árvore genealógica acadêmica, é possível observar como o seu conhecimento foi repassado ao longo do tempo.

Neste trabalho, o objetivo principal é utilizar os currículos que compõem a Plataforma Lattes sob a coordenação do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) para a caracterização e análise de árvores e florestas genealógicas acadêmicas. Para tanto, os currículos cadastrados na Plataforma Lattes dos indivíduos que têm atuado como orientadores são analisados com o objetivo de se extrair informações de interesse e, dessa forma, obter uma visão geral do processo de orientação deste conjunto e ainda produzir árvores genealógicas e florestas acadêmicas, através das quais é possível verificar como o conhecimento tem se propagado nas diversas áreas do conhecimento.

TRABALHOS RELACIONADOS

Um dos primeiros trabalhos encontrados que tratam a genealogia acadêmica com base nos dados da Plataforma é proposto por Miyahara (2011). Nele, são construídas árvores genealógicas de pesquisadores, considerando para isso as relações de orientações. Para o autor, as árvores genealógicas podem indicar todo o histórico de um determinado pesquisador. A ferramenta proposta utiliza os currículos para a caracterização das árvores.

No trabalho de Mena-Chalco, Digiampietri e Cesar-Junior (2012), é proposta uma ferramenta para a caracterização de árvores genealógicas utilizando para isso dados curriculares da Plataforma Lattes. São consideradas informações sobre orientações e de forma automática é gerada a árvore de um pesquisador em particular, podendo ser definida a quantidade de níveis a ser considerada.

Já Tuesta e demais autores (2012) seguem em uma linha diferente, apresentando uma análise temporal da relação orientador-orientado, com um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação, extraindo os dados de análise dos currículos Lattes. No trabalho são analisadas as principais características do grupo e as relações de coautoria.

Trabalhos com base em genealogia acadêmica podem ter como objetivo propor técnicas para mensurar quantitativamente e/ou qualitativamente as orientações realizadas e conseqüentemente, as árvores genealógicas. Rossi e Mena-Chalco (2015) propõe um índice-h genealógico expandido com base na medida índice-h conhecida

na Bibliometria/Cientometria. Esta derivação tem como objetivo caracterizar grafos de orientações permitindo o estudo em função do seu desempenho no papel de formador de recursos humanos.

Diferentemente dos trabalhos anteriores, a proposta deste estudo é propor uma técnica que possa analisar e tentar identificar de forma automática, em todo o repositório de currículos da Plataforma Lattes, todas as orientações realizadas, incluindo aquelas que não estão implicitamente vinculadas nos currículos dos orientadores.

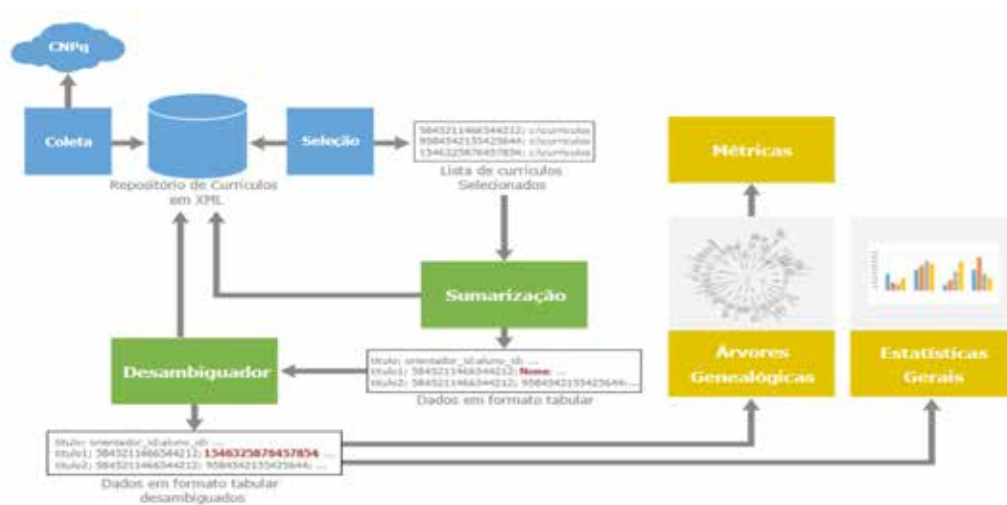
DESENVOLVIMENTO

Este estudo tem como principal fonte de dados os currículos da Plataforma Lattes. O Currículo Lattes se tornou um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país, e é hoje adotado pela maioria das instituições de fomento, universidades e institutos de pesquisa do país. Por sua riqueza de informações e sua crescente confiabilidade e abrangência, se tornou elemento indispensável e compulsório à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia.

A identificação de relacionamentos em orientações não é uma tarefa trivial. Atualmente, os registros de orientações dos currículos possuem uma opção de se realizar a vinculação manual do nome do orientado ou dos coautores a seus identificadores únicos na Plataforma Lattes. No entanto, tal vínculo não é automático e, em geral, relacionamentos antigos permaneceram sem seus vínculos com os identificadores, exibindo apenas o nome no registro de orientação ou coautoria. Diante disso, uma estratégia de identificação de relacionamento se faz necessária para que se possa caracterizar redes com a maior quantidade possível de indivíduos.

Um dos propósitos deste estudo é a realização da coleta de todos os currículos da Plataforma Lattes utilizando técnicas de extração de dados da web. De posse dos currículos, é possível aplicar a eles diversas transformações e processá-los a fim de se obter resultados importantes. Para a coleta dos currículos que compõem a Plataforma Lattes, um arcabouço denominado LattesDataXplorer (DIAS, 2016), foi utilizado. Diante do arcabouço utilizado, foi realizada uma expansão do mesmo para atender as necessidades deste estudo (Figura 1).

Figura 1 - Visão geral do LattesDataXplorer expandido



Fonte: elaborado pelos autores.

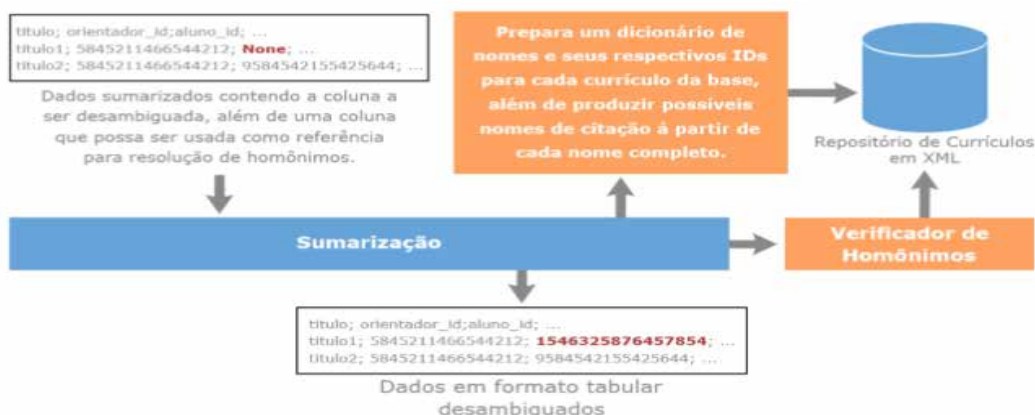
De posse dos currículos, os dados podem ser selecionados de acordo com as necessidades e sumarizados, possibilitando a produção de conjuntos de dados distintos dependendo da necessidade das análises a serem realizadas. Todos estes dados são armazenados em formato tabular, facilitando a sua posterior leitura. Novos conjuntos podem ser definidos de acordo com a necessidade.

Os arquivos resultantes do processo são armazenados em formato CSV (Comma Separated Values), simplificando a leitura dos dados. Desta forma, o processo de abrir cada um dos currículos e aplicar uma consulta em XPath (XML Path Language) para obtenção dos dados será realizado apenas uma vez.

A identificação de relacionamentos em orientações não é uma tarefa trivial. Atualmente, os registros de orientações dos currículos possuem uma opção de se realizar a vinculação manual do nome do orientado ou dos coautores a seus identificadores únicos na Plataforma Lattes. No entanto, tal vínculo não é automático e, em geral, relacionamentos antigos permaneceram sem seus vínculos com os identificadores, exibindo apenas o nome no registro de orientação ou coautoria. Diante disso, uma estratégia de identificação de relacionamento se faz necessária para que se possa caracterizar redes com a maior quantidade possível de indivíduos.

A primeira etapa da estratégia de identificação proposta neste trabalho é a obtenção dos dados pessoais de cada indivíduo com currículo cadastrado na Plataforma Lattes. Através destes dados, é possível obter informações como nome completo e nome em citações bibliográficas, utilizadas para produção do dicionário, conforme a Figura 2. Além de nomes de citações presentes nos currículos, também são produzidos, automaticamente, possíveis nomes de citação a partir do nome completo.

Figura 2 - Processo de desambiguação de nomes da Plataforma Lattes



Fonte: elaborado pelos autores.

O desambiguador é inicializado, gerando um dicionário com informações de todos os indivíduos (cerca de 20.000.000 de nomes de citações distintos) e, posteriormente, aplicado em todo o conjunto de orientações a serem desambiguadas, resolvendo o nome de orientados não vinculados em cada orientação concluída informada nos currículos dos orientadores. Em caso de um único identificador para este orientado no dicionário, o identificador é atribuído a ele. Em situações de homônimos, outras informações como nome e identificador do orientador são utilizadas para tentar localizar o correto identificador do orientado. Isso possibilita a geração de redes com um número maior de elementos, obtendo assim um melhor resultado.

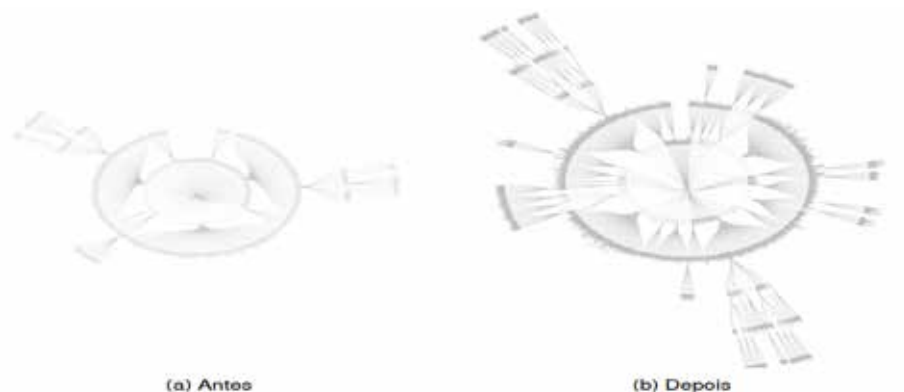
Consequentemente, diante do exposto, nem todos os relacionamentos podem ser identificados, seja por erros de digitação na definição do nome dos alunos orientados ou mesmo pela inexistência do currículo na Plataforma Lattes. Neste caso, estes serão transformados em nós folhas e seus descendentes mesmo que existam não serão incorporados, já que não é possível analisar seus currículos pela falta do identificador. Cabe ressaltar que o mesmo desambiguador, além de orientações, pode ser aplicado em qualquer sessão do currículo, como formações acadêmicas, produções bibliográficas, entre outras.

Os dados analisados foram coletados no final de 2017 com aproximadamente 5.200.000 currículos, totalizando cerca de 8.449.497 orientações independentemente de sua natureza. Antes da execução do método de identificação proposto neste trabalho, apenas 621.384 possuíam relacionamentos vinculados implicitamente. Ou seja, apenas estas orientações foram devidamente vinculadas com os orientados pelo professor orientador em seus currículos.

Após a execução do método proposto, 3.804.063 relacionamentos novos foram identificados, um total de 45,03%, valor muito superior aos 7,35% encontrados antes do

processo de desambiguação. A Figura 3 apresenta a árvore genealógica acadêmica do orientador com a maior quantidade de gerações identificada, antes e após a aplicação do desambiguador.

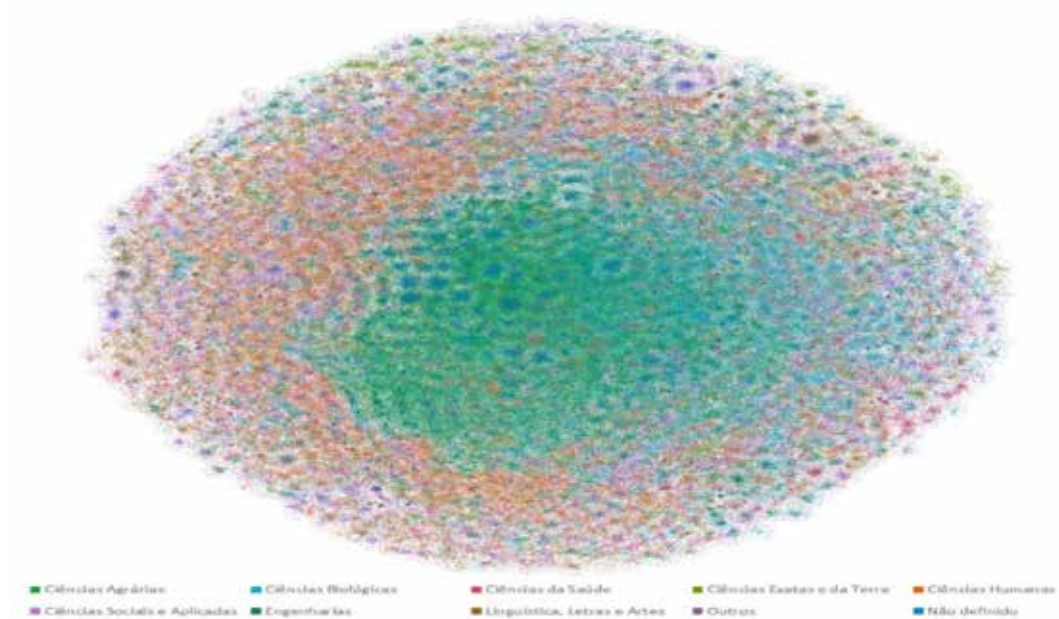
Figura 3 - Árvores antes (a) e depois (b) do processo de desambiguação



Fonte: elaborado pelos autores.

Além disso, também foi possível caracterizar a floresta com floresta com as 100 maiores árvores identificadas conforme processo descrito anteriormente (Figura 4). Logo, foram considerados as maiores árvores tendo em vista a quantidade de elementos.

Figura 4 - Floresta com as 100 maiores árvores identificadas



Fonte: elaborado pelos autores.

As cores dos nós são referentes as suas grandes áreas de atuação. Dentre as 100 maiores árvores identificadas, 14 raízes dessas árvores são de orientadores estrangeiros. Ressalta-se ainda que deste conjunto um grande percentual dos orientadores que são raízes das principais árvores, tem atuação nas grandes áreas de Ciências Biológicas e Agrárias.

Ao analisar as 100 maiores árvores caracterizadas, essas são responsáveis por englobar aproximadamente 3,8% de todos os indivíduos com currículos cadastrados na Plataforma Lattes. Este conjunto possui um total de 206.573 conexões, que neste estudo significam orientações, apresentando, como era de se esperar, uma pequena densidade da floresta (0,00000548) e grau médio próximo de 1 (1,064). O diâmetro da rede é igual a 15, mas com um caminho mínimo médio igual a 3,752, o que significa que dentre o conjunto analisado, um determinado indivíduo está em média muito próximo aos outros.

CONSIDERAÇÕES

Existe atualmente uma grande dificuldade de identificar tais indivíduos, já que não existe vínculo entre estes e seus orientadores, sendo necessária a aplicação de técnicas de desambiguação de nomes. Com a aplicação do método proposto, foi possível identificar orientados não vinculados a seus orientadores, permitindo, desta forma, a caracterização de redes com maior precisão.

REFERÊNCIAS

- DIAS, T. M. R. *Um Estudo Sobre A Produção Científica Brasileira A Partir De Dados Da Plataforma Lattes*. 2016. Tese (Doutorado em Modelagem Matemática e Computacional) - Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.
- MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JUNIOR, R. M. Caracterizando as redes de coautoria de currículos Lattes. In: BRASNAM - BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 1., 2012, Curitiba. *Anais [...]*. [S. l.]: SBC, 2012.
- MIYAHARA, E. K. *Genealogia Acadêmica Lattes*. 2011. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2011.
- ROSSI, L.; MENA-CHALCO, J. P. Índice-h genealógico expandido: Uma medida de impacto em grafos de orientação acadêmica. In: BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 4., 2015, Recife. *Anais [...]*. [S. l.]: SBC, 2015.
- TUESTA, E. F.; DELGADO, K.; DIGIAMPIETRI, L. A. et al. Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação. In: BRASNAM - BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 1., 2012, Curitiba. *Anais [...]*. [S. l.]: SBC, 2012.

