Redes de propagação do conhecimento: adoção de análises de redes sociais em dados de orientações acadêmicas

Tales Henrique José Moreira[1]; Gray Farias Moita[2];
Thiago Magela Rodrigues Dias[3]; Patrícia Mascarenhas Dias[4]

Fundamentações

A evolução das pesquisas científicas tem forte influência do processo de formação, onde pesquisadores orientadores inserem novos pesquisadores que contribuem para que novos estudos sejam realizados em diversas áreas do conhecimento. Grande parte dos trabalhos realizados com orientação no Brasil são decorrentes de Programas de Pós–Graduação (PPGs), impulsionados pela necessidade de capacitação e titulação de docentes e de pesquisadores. Nesse

1 Mestre, Centro Federal de Educação Tecnológica de Minas Gerais – CEFET–MG, Brasil, tales.info@gmail.com.

2 Doutor, Centro Federal de Educação Tecnológica de Minas Gerais – CEFET–MG, Brasil, gray@dppg.cefetmg.br.

3 Doutor, Centro Federal de Educação Tecnológica de Minas Gerais – CEFET–MG, Brasil, thiagomagela@cefetmg.br.

4 Mestre, Centro Federal de Educação Tecnológica de Minas Gerais – CEFET–MG, Brasil, patriciamdias@gmail.com.

contexto, a bibliometria e cientometria são importantes para a compreensão e análise das atividades de orientação e de produção acadêmica (ARAÚJO, 2006).

Dados sobre a orientação acadêmica se caracterizam como um novo e importante objeto de estudo para compreender o processo de formação através da genealogia acadêmica, já que possibilita compreender e analisar a propagação do conhecimento. Para Sugimoto (2014), genealogia acadêmica é um estudo quantitativo da herança intelectual através da relação orientador–orientado. Já para Ferreira, Furtado e Silveira (2009), o binômio (ou díade) orientador–orientado é indubitavelmente a base dos PPGs, o que determina o crescimento e a expansão dos cursos de Pós–Graduação (PG) e a demanda de orientação.

Neste contexto, as redes genealógicas acadêmicas podem 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 informações complementares que visam permitir um melhor entendimento do histórico de um indivíduo.

Diante disto, as árvores genealógicas acadêmicas são caracterizadas como grafos que representam hierarquicamente o histórico de um orientador e todos os seus

orientados. Logo, caracterizando uma árvore genealógica acadêmica, é possível a adoção de métricas de análises de redes sociais para compreender como o conhecimento no processo de orientação acadêmica foi repassado ao longo do tempo.

Objetivos

Neste trabalho, o objetivo principal é utilizar os currículos que compõem a Plataforma Lattes sob a coordenação do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) para a caracterização e análise de árvores genealógicas acadêmicas. Para tanto, os currículos cadastrados na Plataforma Lattes dos indivíduos que têm atuado como orientadores em PPGs 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 acadêmicas, onde é possível aplicar métricas de análises de redes sociais para verificar como o conhecimento tem se propagado nas diversas áreas do conhecimento.

Método

A escolha da Plataforma Lattes para a extração e integração juntamente com outras fontes de dados está relacionada ao fato de que a Plataforma Lattes é a única quando se trata da integração de dados científicos e acadêmicos de indivíduos vinculados a instituições da área de C&T, registrando os dados acadêmicos e as produções científicas dos

pesquisadores e instituições, permitindo que a atualização dos dados seja realizada pelos próprios pesquisadores.

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).

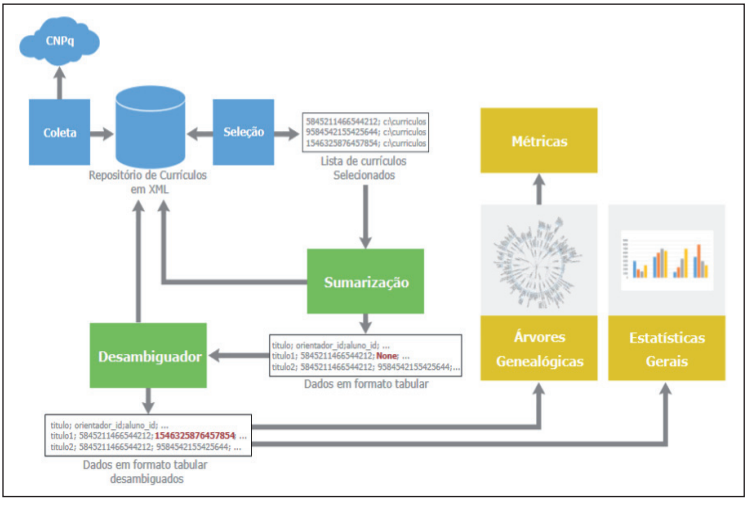

Figura 1 – Visão geral do *LattesDataXplorer* expandido.

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.

Resultados

Os dados apresentados na Tabela 1 correspondem a todo o conjunto de currículos da Plataforma Lattes no momento da extração utilizada, se revelando uma grande fonte de dados que pode ser utilizada na área da descoberta de conhecimento. Um destes é a genealogia acadêmica. Esta pode ser identificada através dos dados de orientações, possibilitando a análise da vida pregressa do indivíduo. Além disso, tais dados podem ser associados aos dados de produção bibliográfica.

Descrição	Quantidade Total
Dados de informações pessoais	4.591.941
Dados de formação acadêmica	10.488.174
Dados de produção bibliográfica (periódicos e anais de congresso)	19.335.882
Dados de orientações concluídas	8.138.267

Tabela 1 – Resultado da sumarização de todos os currículos analisados.

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.

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 2 apresenta a árvore genealógica acadêmica do orientador com a maior quantidade de gerações identificada, antes e após a aplicação do desambiguador.

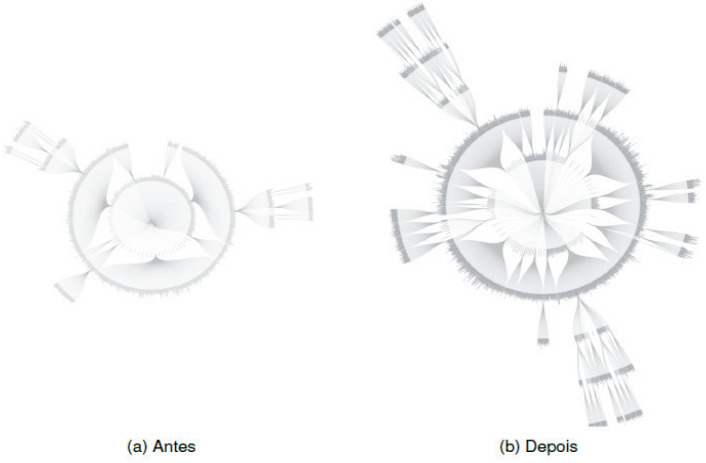

Figura 2 – Árvores antes (a) e depois (b) do processo de desambiguação.

Exemplos de estatísticas e métricas que podem ser implementadas são: média de orientações por indivíduo, frequência de orientações de uma determinada área e orientadores mais representativos considerando seus descendentes, entre outras. Todas essas métricas possibilitam explorar os dados extraídos da Plataforma Lattes e proporcionam uma ampla visão da genealogia acadêmica brasileira, tendo como fonte de dados principal os registros de orientações cadastrados nos currículos.

Se selecionados apenas os registros de orientação a nível de pós–graduação, é possível ranquear os indivíduos do conjunto de acordo com a quantidade de orientações concluídas (Tabela 2). Assim, são apresentadas na tabela um ranking dos indivíduos com as maiores quantidades de orientações diretas encontrados na Plataforma Lattes, além de outras informações relacionadas às árvores genealógicas, como o tamanho da rede e a quantidade de gerações.

N	Instituição	Quantidade de Orientações Diretas	Gerações	Tamanho de Rede
1	UFSC	399	5	1.396
2	PUC-SP	351	5	1.219
3	UFV	323	3	594
4	UFV	263	7	2.592
5	COPPE- UFRJ	259	5	1.542
6	UFRJ	246	5	1.129
7	PUC-Campinas	236	8	2.132
8	PUC-SP	229	4	1.541
9	UFV	224	5	1.275
10	UFV	223	3	494

Tabela 2 – Ranking dos indivíduos com maior quantidade de orientações de pós–graduação.

As maiores quantidades de orientações diretas nem sempre estão diretamente relacionadas à propagação de seu conhecimento para outras gerações. Em casos como o do indivíduo no 1 da Tabela 2, observa–se que há um grande volume de orientações. Inicialmente, pode–se presumir que seja uma das maiores redes encontradas. Porém, sua árvore genealógica possui 5 gerações de descendentes, bem inferior à maior encontrada, com 11 gerações (em nível de pós–graduação).

A Figura 3 apresenta a árvore genealógica do indivíduo com a maior quantidade de orientações diretas. Ele é exibido como o nó central e todos os seus descendentes estão distribuídos nos diferentes níveis da árvore. Pode–se observar nesta figura que apesar de seu nó raiz ter orientado a maior quantidade encontrada, seus orientados pouco orientaram, fazendo com que sua árvore não crescesse verticalmente ou seja, não possui grande quantidade de gerações.

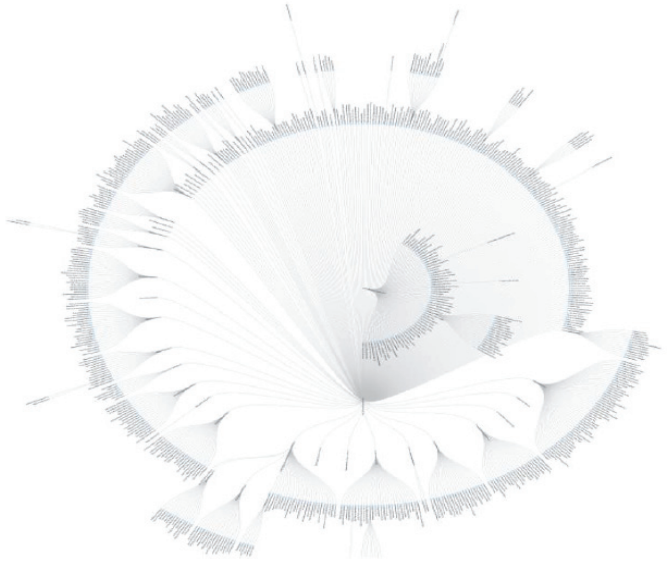

Figura 3 – Árvore genealógica do orientador com a maior quantidade de orientações.

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

Analisando 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.

Diante do exposto, percebe–se que, de certa forma, ao verificar as maiores árvores identificadas, elas estão interconectadas, resultando em uma floresta que contém uma única componente conexa. Tal fato, está relacionado a que os indivíduos nos últimos níveis das árvores possuíram algum tipo de orientação com indivíduos de outras árvores.

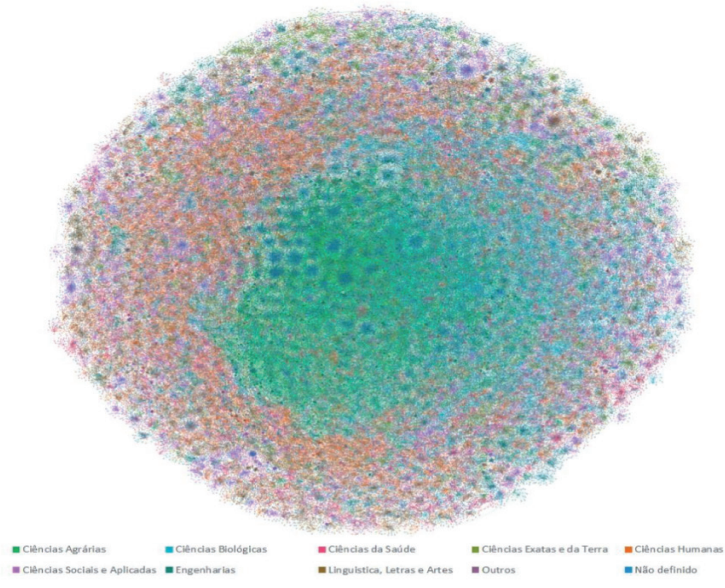

Figura 4 – Floresta com as 100 maiores árvores identificadas.

Conclusões

Neste trabalho foram analisados todos os currículos cadastrados na Plataforma Lattes a fim de se obter uma caracterização dos registros de orientações acadêmicas e do

comportamento da relação orientador–orientado quanto à produção bibliográfica e à geração de árvores genealógicas.

A grande maioria dos registros de orientações não possuem vínculos identificados nos currículos dos orientadores, logo, não é possível conhecer com exatidão quais os orientadores e orientados analisando os currículos e considerando apenas os nomes de citações. Isto se deve à falta de vínculos explícitos entre os nomes dos indivíduos a seus identificadores únicos na Plataforma Lattes.

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

Araújo, C. A. A. (2006) Bibliometria: evolução histórica e questões atuais. *Em questão*, 12 (1), 11–32

Ferreira, L. M.; Furtado, F.; Silveira, T. S. (2009) Relação Orientador–Orientando. O Conhecimento Multiplicador. *Acta Cirúrgica Brasileira*, 24 (3), 170–172.

Dias, T. M. R. (2016) *Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes*. 181 f. Tese (Doutorado) – Curso de Programa de Pós–graduação, Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte.

Sugimoto, C. R. (2014) Academic genealogy. In: Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact. MIT Press.