

Genealogia Acadêmica da Relação Orientador-Orientado na Área de Ciência da Computação

Tales H. J. Moreira¹, Thiago M. R. Dias¹, Gray F. Moita¹

¹ Programa de Pós-graduação em Modelagem Matemática e Computacional
Centro Federal de Educação Tecnológica de Minas Gerais
Av. Amazonas, 7576, Nova Gameleira, 30510-000, Belo Horizonte, MG
{tales.info, thiagomagela}@gmail.com, gray@dppg.cefetmg.br

***Abstract.** Academic genealogy allows us to understand the guiding-oriented relationship, and to analyze how the knowledge diffuses through the counselor, especially in postgraduate programs, where most of the students intend to remain in the academic environment. Given this, this is an important means to analyze this relationship in several areas of knowledge. In this work, the data of orientations and supervisions coming from Lattes curricula related to the area of computer science are used, obtaining an overview of the genealogy of this group, and also, to produce genealogical trees, allowing to analyze how the knowledge has been propagated.*

***Resumo.** A genealogia acadêmica nos permite muito além de compreender a relação orientador-orientado, analisar como ocorre a difusão do conhecimento através do orientador, principalmente em programas de pós-graduação, onde grande parte dos alunos possuem a intenção de permanecer no meio acadêmico. Diante disso, este é um importante meio para análise desta relação em diversas áreas do conhecimento. Neste trabalho, são utilizados os dados de orientações e supervisões provenientes dos currículos Lattes, vinculadas à área de ciências da computação, obtendo uma visão geral da genealogia deste conjunto, e ainda, produzir árvores genealógicas, possibilitando analisar como o conhecimento tem se propagado.*

1. Introdução

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 pesquisadores.

Para Ferreira, Furtado, & 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. Além disso, os autores ressaltam que o aluno de PG é um pesquisador em potencial, em estágio avançado de desenvolvimento, ou seja, a caminho da autonomia científica, mas ainda dependente de um professor, o que justifica as atividades de orientação como efetivamente necessárias.

Adicionalmente aos dados básicos que caracterizam os vínculos sobre o processo de orientação, informações como publicações, área de atuação e projetos de pesquisa podem ser extraídos de todos os elementos que compõem a rede a ser analisada com o intuito de se obter dados quantitativos que representam a difusão do conhecimento a partir de um orientador em particular. Tendo em vista as possibilidades de visualização e entendimento do histórico de orientação e conseqüentemente a difusão do conhecimento de um determinado orientador, realizar a modelagem e caracterização de árvores genealógicas acadêmicas surge como uma alternativa interessante para a análise de como a ciência se propaga nas várias áreas de estudo.

Para isso, dados sobre orientações bem como de características individuais de cada um dos indivíduos que compõem a rede são analisados. Neste trabalho, os dados utilizados são os currículos que compõem a Plataforma Lattes sob a coordenação do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Neste caso, todos os currículos da Plataforma Lattes são processados, afim de encontrar todas as orientações da área de Ciências da Computação, com o objetivo de se obter uma visão geral do processo de orientação deste conjunto, e ainda, produzir árvores genealógicas acadêmicas, onde é possível verificar como o conhecimento tem se propagado. Além disso, é produzida uma floresta, onde todos os registros encontrados são modelados em forma de árvore. Dessa forma, é possível obter toda a genealogia acadêmica.

Posteriormente, as árvores geradas podem ser manipuladas com o intuito de melhorar a visualização dos elementos que a compõem e servirem de subsídios para análises diversas que visam compreender como a relação orientador-orientado tem acontecido neste conjunto de orientações.

2. Trabalhos Relacionados

Para [Sugimoto, 2014], genealogia acadêmica é o estudo quantitativo da herança intelectual adquirida através do relacionamento orientador-orientado. Logo, compreender tal relação pode proporcionar mecanismos para analisar como as orientações estão sendo conduzidas. Diante disso, este trabalho propõe uma abordagem baseada em árvores para a modelagem da genealogia acadêmica de pesquisadores orientadores, possibilitando visualizar todo o histórico de orientação de forma hierárquica, e conseqüentemente como o seu conhecimento tem se propagado.

Tuesta et al. (2012) apresentam 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.

Em Leite Filho & Martins (2006), são verificadas as influências da relação orientador-orientado no processo de produção de teses e dissertações dos PPGs em Contabilidade da cidade de São Paulo. Os autores citam como justificativas para o estudo a importância de se analisar aspectos que teriam ligação com a construção do conhecimento, especificamente em se tratando da área de Contabilidade e, a tentativa de sinalizar a importância da temática orientação.

Diversos outros trabalhos têm explorado a Plataforma Lattes como principal fonte de informações para caracterização e análise sobre orientações acadêmicas [Delgado et al., 2015; Dias & Moita, 2013; Miyahara, Mena-Chalco, & Cesar-Jr, 2011; Rossi & Mena-Chalco, 2014]. No entanto, a grande maioria dos trabalhos citados anteriormente atuam em pequenos conjuntos de dados previamente definidos, obtidos manualmente ou de forma limitada, diferentemente deste, que através de toda a base, seleciona os registros a serem utilizados, obtendo um conjunto completo de todas as orientações devidamente cadastradas na área de Ciências da Computação. Mesmo que o indivíduo não atue preferencialmente nesta área, qualquer registro de orientação que possuir esta área será analisado.

3. Metodologia

Este trabalho tem como principal fonte de dados os currículos da Plataforma Lattes. A Plataforma Lattes representa a experiência do CNPq na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações, se tornando um padrão nacional no registro da vida pregressa e atual dos estudantes e pesquisadores do país [CNPq, 2016].

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 [CNPq, 2016].

Diversos outros trabalhos para análise de dados científicos têm explorado a Plataforma Lattes como principal fonte de informações [Alves, Yanasse, & Soma, 2011a, 2011b, 2011c; Dias & Moita, 2013; Dias, Moita, Dias, Moreira, & Santos, 2013; Digiampietri et al., 2012; Mena-Chalco, Digiampietri, & Cesar-Jr, 2012], tendo a maioria, foco em extração e análise de colaboração científica. Existem trabalhos relacionados à orientação acadêmica que utilizam dados da Plataforma Lattes, porém, conforme dito na sessão anterior, a maioria atua em pequenos conjuntos, obtidos manualmente ou de forma limitada devido a dificuldades na extração dos dados.

Apesar de os dados dos currículos da Plataforma Lattes serem disponibilizados livremente, esses são visualizados através de interface de consulta que apresenta os currículos individualmente. Diante disso, técnicas e ferramentas para a extração e integração dos dados se fazem necessárias [Dias et al., 2013]. Como o enfoque deste

trabalho é caracterizar árvores genealógicas científicas a partir de dados da Plataforma Lattes e não dedicar esforços na extração dos currículos, logo, optou-se por utilizar o framework desenvolvido por [Dias et al., 2013] para extração dos dados a serem analisados, que possibilita a extração de todos os currículos da plataforma armazenando-os em formato XML (Extensible Markup Language) conforme Figura 1.

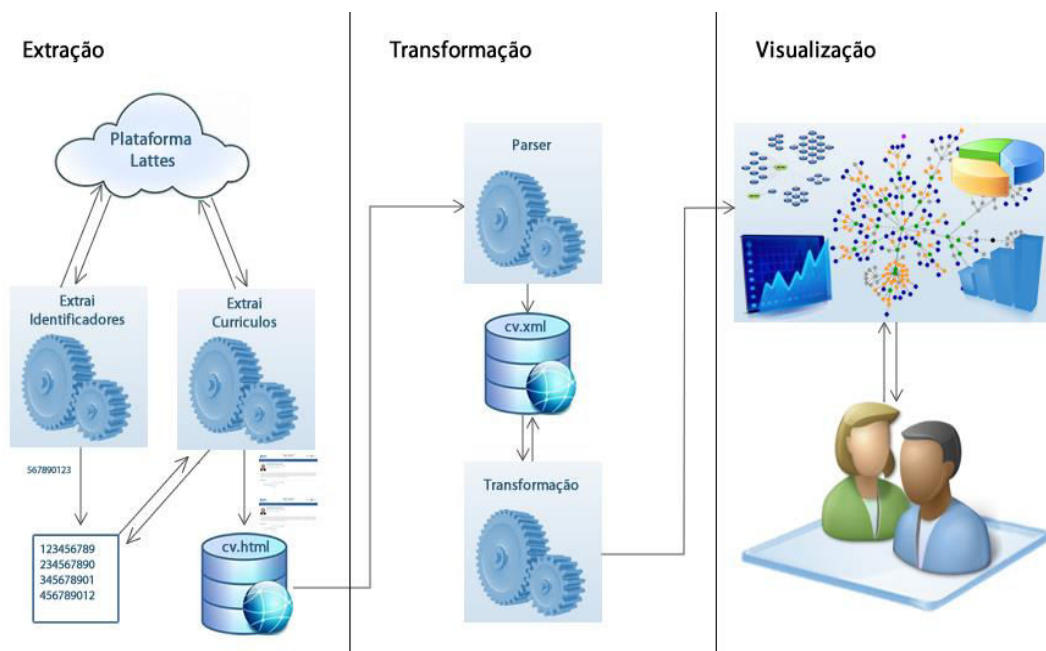


Figura 1. Arquitetura do framework de Extração e Integração de Dados Científicos.

De posse dos dados, o módulo de seleção do framework é aplicado a fim de se obter uma lista com todas as orientações da área de Ciências da Computação. Tais registros são transformados em árvores genealógicas e ordenadas de acordo com a quantidade de descendentes de forma decrescente, retornando assim os pesquisadores deste conjunto que mais orientaram em forma de raiz.

Para a construção das árvores, uma estratégia recursiva é utilizada, na qual uma raiz é previamente selecionada, fazendo a identificação de todas as orientações e supervisões concluídas, obtendo uma lista de orientados que estão diretamente vinculados ao nó raiz. Este processo é repetido para cada item da nova lista, identificando assim os vários níveis de descendentes do nó raiz. Conseqüentemente, currículos que não possuem orientações concluídas vão se tornando nós folhas da árvore. Esse processo acontece até que todos os descendentes sejam processados, encontrando por fim apenas nós folhas.

Portanto, cada currículo encontrado com orientações em Ciências da Computação passa a ser o nó raiz a cada iteração, gerando uma nova árvore caso o mesmo não tenha sido inserido na árvore de outro pesquisador, ou seja, não é um descendente de outro pesquisador.

4. Resultados

A Plataforma Lattes conta hoje com mais de 4.500.000 currículos cadastrados, contendo dados de diversas naturezas e áreas do conhecimento. A seleção para este trabalho foi realizada obtendo apenas registros de orientações que possuem como área do conhecimento Ciências da Computação, pertencente à grande área de Ciências Exatas e da Terra, além de ser proveniente de cursos de pós-graduação, neste caso, sendo selecionadas através de sua natureza. Diante disto, foram selecionados 13.103 registros de orientações ou supervisões concluídas em cursos de pós-graduação e na área a ser analisada. De posse dos dados citados acima, possibilitou-se ranquear as instituições com maior quantidade de orientações na área de Ciências da Computação, objeto de estudo deste, conforme é apresentado na Tabela 1.

Tabela 1. Ranking de instituições com orientadores mais relevantes na área

#	Quantidade de Descendentes	Número de Gerações	Instituição
1	234	4	UERJ
2 *	230	5	COPPE-UFRJ
3	189	3	PUC-RJ
4	158	5	UNICAMP
5	147	4	UFF
6	138	5	-
7	138	4	UFRJ
8	135	4	UFPE
9	124	3	PUC-RJ
10 **	115	3	UFRJ

* Bolsista de Produtividade em Pesquisa do CNPq - Nível 1B - CA CC - Ciências da Computação

** Bolsista de Produtividade em Pesquisa do CNPq - Nível 1D - CA CC - Ciências da Computação

Ao analisar os dados obtidos, observou-se que dois dos pesquisadores mais relevantes (2º e 10º) são bolsistas de Produtividade em Pesquisa do CNPq (Nível 1B e 1D respectivamente - CA CC - Ciências da Computação). Diante disso, podemos perceber a importância deste não somente para a área de Ciências da Computação, mas também como orientador, repassando seu conhecimento e obtendo uma das maiores árvores encontradas.

Outro fato importante encontrado na Tabela 1, é o fato de um dos indivíduos (6º) constar na lista por um erro no cadastro do currículo. Ao analisa-lo, percebeu-se que este cadastrou sua formação na área correta e repetiu o registro de forma incorreta na seção de orientações. Ou seja, registrou em seu currículo que orientou seu próprio orientador, fazendo que todos os descendentes de seu orientador fossem vinculados a ele. Diante disto, este deve ser ignorado na lista.

Além do ranqueamento por quantidade de descendentes, analisou-se os indivíduos por quantidade de gerações de descendentes. Ou seja, pela altura da árvore (Tabela 2).

Tabela 2. Ranking de instituições de orientadores com maior número de gerações (altura)

#	Quantidade de Descendentes	Número de Gerações	Instituição
1 *	230	5	COPPE-UFRJ
2	158	5	UNICAMP
3	138	5	-
4	234	4	UERJ
5	147	4	UFF
6	138	4	UFRJ
7	135	4	UFPE
8	103	4	USP
9	97	4	UNICAMP
10	56	4	UNICAMP

* Bolsista de Produtividade em Pesquisa do CNPq - Nível 1B - CA CC - Ciências da Computação

Neste caso, percebemos que o indivíduo bolsista de produtividade passou da 2ª posição para a 1ª, possuindo a árvore mais alta de toda a floresta encontrada. O segundo bolsista, apesar de possuir uma grande árvore, não teve seu conhecimento repassado a várias gerações, não permanecendo na segunda tabela.

Além disso, a Tabela 2 apresenta uma particularidade, onde indivíduos com redes relativamente menores se comparadas as da Tabela 1, porem com maior altura, ou seja, com uma maior quantidade de gerações. Constata-se que grande número de orientações não implica em árvores mais altas, pois este aspecto é influenciado por vários fatores, como por exemplo a natureza da orientação. Orientações em cursos de pós-graduação tendem a possuir as árvores mais altas, já que seus alunos tendem a permanecer no meio acadêmico, se tornando novos orientadores e aumentando a árvore verticalmente.

Para melhor visualização, também foram produzidas as árvores através de um meio visual. Diante disso, podemos observar na Figura 2 o orientador com maior quantidade de descendentes na área (indivíduo número 1 da Tabela 1 e ao mesmo tempo 4º da Tabela 2).

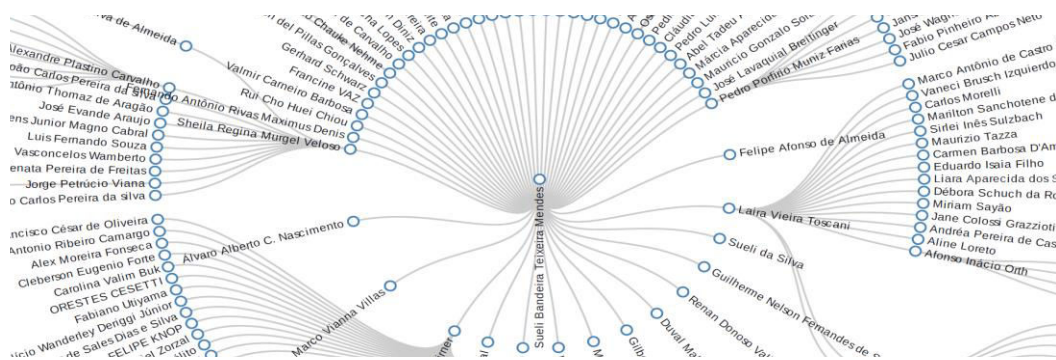


Figura 2. Indivíduo com maior quantidade de descendentes, focando o nó central.

Através da Figura 3, o mesmo pode ser visto, porém, sendo exibida toda a árvore deste orientador.

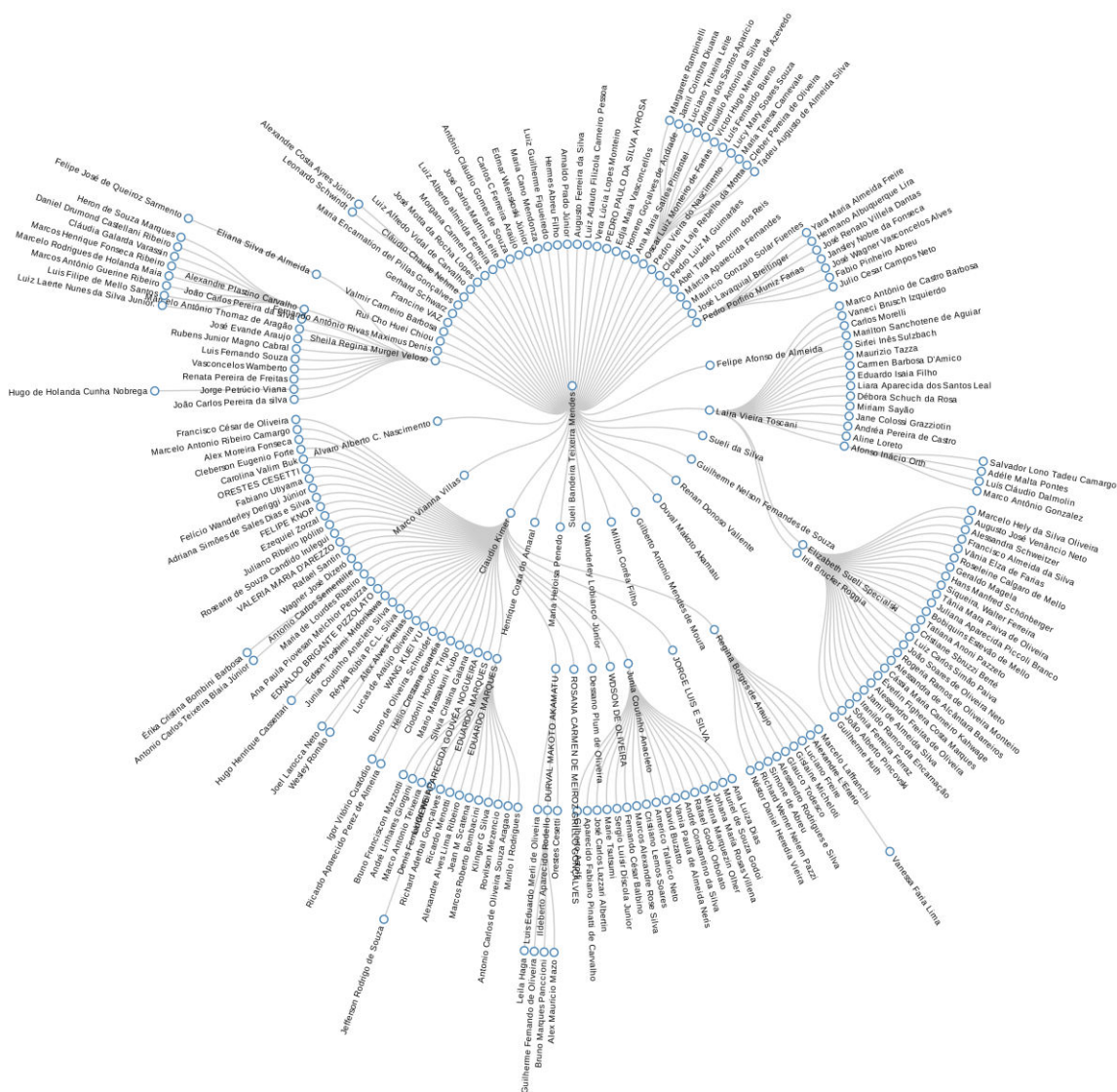


Figura 3. Indivíduo com maior quantidade de descendentes.

Também foram produzidas as redes das árvores mais altas, demonstrando-as de forma visual, onde é possível perceber de forma visual como o conhecimento tem se propagado na área de Ciências da Computação. Apesar disso, a Figura 4 apresenta uma árvore em que apesar de ser uma das mais altas, possui menor quantidade de descendentes diretos (orientações) se comparada a Figura 3.

Diante disto, observamos que um grande volume de orientações não resulta necessariamente em grandes redes, mas sim a qualidade e natureza das orientações. Também foi produzida a rede do 2º indivíduo da Tabela 2 que, juntamente com o indivíduo acima conta com cinco níveis de descendentes, sendo estas as árvores mais altas. O indivíduo de número 3 foi ignorado devido ao erro encontrado no registro de suas orientações.

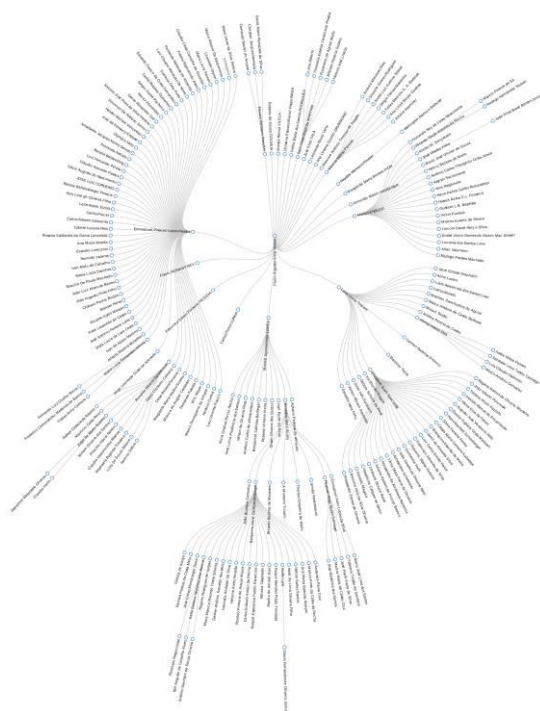


Figura 4. Indivíduo com maior quantidade de gerações de descendentes.

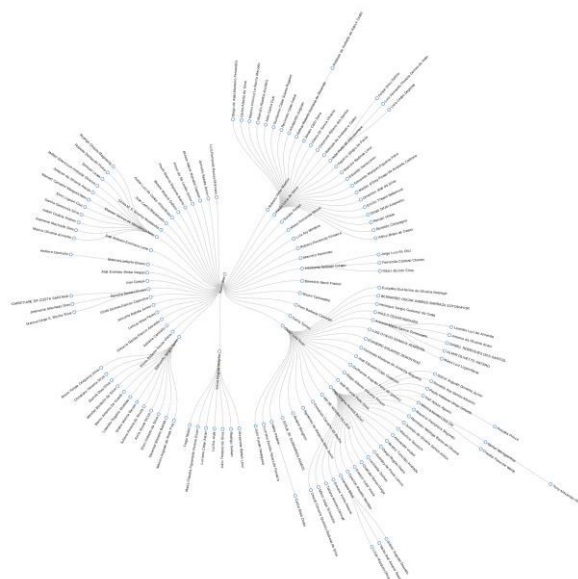


Figura 5. Indivíduo com maior quantidade de gerações de descendentes (2).

Todos os dados e imagens produzidos neste trabalho foram obtidos através de uma extensão criada a partir do framework iniciado por [Dias et al., 2013] voltada para genealogia acadêmica. Essa extensão possibilita a produção de dados para análise em ferramentas de terceiros, como o Pandas¹, ou visualização, como por exemplo o d3²,

¹ <http://pandas.pydata.org/>

² <http://d3js.org/>

desenvolvida em javascript responsável neste caso pela produção de imagens das árvores genealógicas.

5. Conclusões

Diferentemente de outros trabalhos citados aqui, este utiliza um conjunto maior de dados para análise de genealogia acadêmica, não existindo uma forma de compará-los. Neste trabalho foram processados todos os currículos da plataforma, afim de se obter uma visão geral da relação orientador-orientado na área de Ciências da Computação.

A grande maioria destes não possuem orientações identificadas pela plataforma, sendo estes transformados em nós folhas das árvores geradas. Ou seja, não se sabe exatamente qual indivíduo da plataforma foi orientado em grande parte dos casos. Isto se deve à falta de vínculos definidos pelos usuários dentro da plataforma Lattes.

Existe uma grande dificuldade de identificar tais indivíduos, já que não existe vínculo entre este e seu orientador. Quanto maior o número de orientações, maior é a rede de colaboração e sua produção passa a ser realizada de forma mais acentuada com os orientados.

Foi possível encontrar características de determinados orientadores de acordo com as classificações aplicadas, identificando inclusive erros no registro das orientações.

Foi ainda possível gerar árvores de importantes orientadores de acordo com a classificação realizada, permitindo constatar as características citadas através de um meio visual.

Como trabalhos futuros, florestas de árvores genealógicas individuais podem ser geradas para possível aplicação de estratégias baseadas em árvores e análise do conjunto.

Referências

- Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2011a). LattesMiner: uma linguagem de domínio específico para extração automática de informações da Plataforma Lattes. In *XII Workshop de Computação Aplicada - WORCAP* (p. 6). Retrieved from http://mtc-m18.sid.inpe.br/col/sid.inpe.br/mtc-m18/2013/01.15.16.10/doc/worcap2012_submission_61 - Alexandre D. Alves.pdf
- Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2011b). Perfil dos Bolsistas PQ das Áreas de Engenharia de Produção e de Transportes do CNPq: Enfoque na Subárea de Pesquisa Operacional. In *XLIII Simpósio Brasileiro de Pesquisa Operacional* (p. 12). Retrieved from <http://www.din.uem.br/~ademir/sbpo/sbpo2011/pdf/88144.pdf>
- Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2011c). SUCUPIRA: um Sistema de Extração de Informações da Plataforma Lattes para Identificação de Redes Sociais Acadêmicas. In *Actas da 6ª Conferência Ibérica de Sistemas e Tecnologias de Informação - CISTI* (pp. 371–376). Retrieved from <http://mtc-m19.sid.inpe.br/col/sid.inpe.br/mtc-m19/2011/12.14.17.28/doc/05974195.pdf>

- CNPq. (2016). Sobre a plataforma Lattes. Retrieved June 13, 2016, from <http://lattes.cnpq.br/>
- Delgado, K. V, Tuesta, E. F., Digiampietri, L. a, Mugnaini, R., Mena-Chalco, J. P., & Pérez-Alcázar, J. J. (2015). Análise Comparativa da Produtividade dos Pares Orientador-Orientado em Ciência da Computação. *Revista Eletrônica de Sistemas de Informação ISSN 1677-3071 Doi: 10.5329/RESI, 14(1)*.
- Dias, T. M. R., & Moita, G. F. (2013). Extração e Modelagem de Redes de Colaboração Científica. In *Conferência IADIS Ibero-Americana WWW/Internet* (Vol. 1).
- Dias, T. M. R., Moita, G. F., Dias, P. M., Moreira, T. H. J., & Santos, L. R. F. (2013). Modelagem e Caracterização de Redes Científicas: Um Estudo Sobre a Plataforma Lattes. In *BRASNAM - II Brazilian Workshop on Social Network Analysis and Mining* (pp. 10–20).
- Digiampietri, L. a, Mena-Chalco, J. P., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K. V, Mugnaini, R., & Silva, G. S. (2012). Minerando e Caracterizando Dados de Currículos Lattes. In *I Brazilian Workshop on Social Network Analysis and Mining (BrasNAM)* (Vol. d, p. 12).
- Ferreira, L. M., Furtado, F., & Silveira, T. S. (2009). Relação Orientador-Orientando. O Conhecimento Multiplicador. *Acta Cirúrgica Brasileira, 24(3)*, 170–172. <https://doi.org/10.1590/S0102-86502009000300001>
- Leite Filho, G. A., & Martins, G. D. A. (2006). Relação orientador-orientando e suas influências na elaboração de teses e dissertações. *Revista de Administração de Empresas, 46(spe)*, 99–109. <https://doi.org/10.1590/S0034-75902006000500008>
- Mena-Chalco, J. P., Digiampietri, L. a, & Cesar-Jr, R. M. (2012). Caracterizando as redes de coautoria de currículos Lattes. In *BraSNAM - Brazilian Workshop on Social Network Analysis and Mining* (p. 12).
- Miyahara, E. K., Mena-Chalco, J. P., & Cesar-Jr, R. M. (2011). Genealogia Acadêmica Lattes. *Instituto de Matemática E Estatística*. São Paulo: Universidade de São Paulo.
- Rossi, L., & Mena-Chalco, J. P. (2014). Aos ombros de gigantes: um estudo de genealogia acadêmica dos matemáticos no Brasil. In *Simpósio de Pesquisa do Grande ABC (SPGABC)* (pp. 1–2). <https://doi.org/10.13140/RG.2.1.2193.7762>
- Sugimoto, C. R. (2014). Academic genealogy. In *Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact* (pp. 365–382). MIT Press. Retrieved from <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6828316>
- Tuesta, E. F., Delgado, K. V, Digiampietri, L. a, Pérez-Alcázar, J. J., Mugnaini, R., & Mena-Chalco, J. P. (2012). Análise temporal da relação orientador-orientado: um estudo de caso sobre a produtividade dos pesquisadores doutores da área de Ciência da Computação. In *Proceedings of the Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)* (p. 11).