A Tool for Analyzing Academic Genealogy

Gabriel Madeira, Eduardo N. Borges(✉), Giancarlo Lucca, Helida Santos, and Graçaliz Dimuro

Centro de Ciências Computacionais, Universidade Federal do Rio Grande-FURG, Rio Grande, Brazil
{gabrielmadeira,eduardoborges,giancarlo.lucca,helida, gracaliz}@furg.br

Abstract. Academic genealogy investigates the relationships between student researchers and advisors and has been used as a resource to analyze the spread of scientific knowledge. This work presents the development of a system that creates academic genealogy trees of researchers from the Brazilian library of theses and dissertations. The proposed system allows users to query and track information about researchers available on the database and retrieve their academic trees with the any desirable depth. This paper extends a previous work presenting new analyzes including the temporal distribution of documents and the number of advisors as a function of advising relationships.

Keywords: Academic genealogy · Genealogy trees · Data integration · Information visualization

1 Introduction

Currently, there are a large number of scientific publications and academic papers available in various Web repositories. Each research institution or university publishes the results achieved in its own institutional repository. In this way, scientific publications are cataloged and organized in a dispersed manner. These data altogether contain the major scientific contributions and collaborations among researchers over time. Analyzing the metadata from multiple publications allows one to map and understand how the relationships between researchers affect the advancement of knowledge in several areas of science.

Genealogy is an auxiliary field of history that studies the origin, evolution and spread of family groups [13]. This evolution is often represented using a structured diagram in the form of family trees [5]. Genealogy trees are well-known structures that organize, through kinship ties, the whole history of an individual's ancestors. Using this structure we can analyze the origin and development of a family lineage over time.

Genealogy trees can be used in academia to analyze relationships between professors, students and researchers. Figure 1 [12] shows an example of an academic genealogy tree. Any metadata sets that describe the elements or their relationships can be used.

The drawing of an academic genealogy tree allows one to see who advised a researcher and how this researcher influenced others over time. A forest (set of trees)

J. Filipe et al. (Eds.): ICEIS 2019, LNBIP 378, pp. 443–456, 2020.
https://doi.org/10.1007/978-3-030-40783-4_21

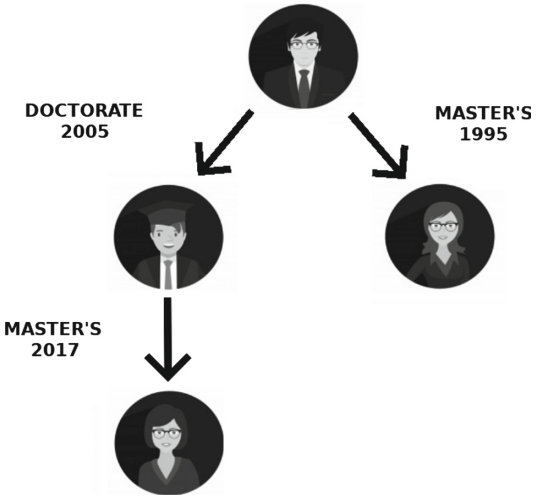

Fig. 1. An academic genealogy tree representing advising relationships in graduate programs. Each edge starts from the advisor and presents the year of master's or doctorate degree [12].

depicts the description a research area using metrics that let, through statistical analyzes and data mining, to extract relevant knowledge for the area under study [3]. Therefore, these structures allow us to analyze how knowledge is spreading across generations of scientists and how these links affect the development of science.

Aiming at visualizing academic genealogy trees created from a set of metadata extracted and integrated from multiple sources, the Information Management Research Group developed at Centro de Ciências Computacionais at Universidade Federal do Rio Grande (FURG) an information system called The Gold Tree [12]. This system allows a researcher to query and track information about his or her advisors and graduate students at any level. A case study was explored to validate the system using data from more than 570 thousand PhD theses and masters dissertations. In addition to including recent related work in the study, this paper extends previous work reporting new data analysis: an overview of the researchers' graph, the temporal distribution of relationships, and the number of advisors as a function of advising relationships.

The rest of this paper is organized as follows. In Sect. 2, we discuss related work. Section 3 presents the methodology to develop the proposed solution. Details on the obtained results are given in Sect. 4. Finally, in Sect. 5, we draw our conclusions and point out some directions for future work.

2 Related Work

In recent years, several studies have explored the visualization of academic collaboration data. While some platforms such as ResearchGate [17], Google Citations, and the Web of Science (WoS) classify registered researchers by citation indexing their articles and papers [1], other tools such as Pajek [2] and PubNet [8] are only concerned with

viewing the research networks. Furthermore, we point out that there are also solutions that use specific data sources to extract information and generate knowledge from co-authoring relationships [10,14]. The following subsections present in details the works used as baseline in the validation of the proposed system.

2.1 Academic Family Tree

Neurotree is a Web database created to document the lineage of academic mentorship in neuroscience [5]. The authors present a temporal analysis of the database growth in a period of seven years. The following metrics were performed: the number of researchers and relationships, the monthly growth rate, the fraction of researchers linked in the main graph, the average distance between researchers, and the average number of connections per researcher. In addition, they report the accuracy of related data in Neurotree with data reported on Web sites of five research groups. Finally, in order to study the relationship between mentorship groups and research areas within neuroscience, they provide a clustering analysis.

This tree exists as a part of the larger Academic Family Tree,[1] which seeks to build a genealogy across multiple academic fields, building a single, interdisciplinary academic genealogy. Figure 2 [12] presents the result of a query by the name of the researcher.

The contents of the database are entirely crowd-sourced. So it is totally dependent on human effort. This feature makes it very susceptible to field fill errors, always presenting incomplete data as well. Any Web user can add information concerning researchers and the connections between them, which can leave the database with a poor quality and/or with false information.

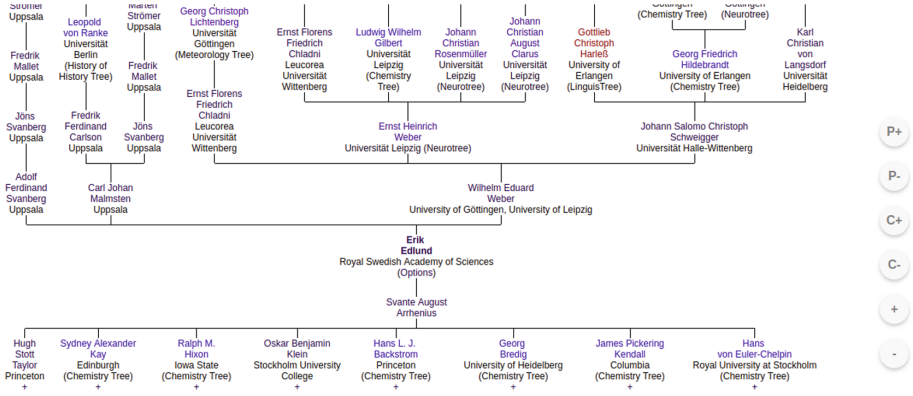

Fig. 2. Result of the query "Erik Edlund" using the Academic Family Tree [12].

[1] https://academictree.org.

2.2 Acácia Platform

Acácia Platform[2] [4] is a system created in 2017 for documenting the formal relations of advising in the context of Brazilian graduate programs. The system uses data registered in the Lattes Platform,[3] which is a database of Brazilian researchers' curricula maintained by the Ministry of Science and Technology and Innovation. Currently, Acacia Platform has over 1 million vertices and relationships. Each vertex represents a researcher and each edge an advising relation completed between two researchers (advisor and student). Figure 3 [12] presents the result of a query by the researcher's name. The system shows some bibliometric indexes as the number of direct and indirect descendants and information about the advising relationships.

Fig. 3. Result of the query "José Palazzo Moreira Oliveira" using Acácia Platform. For each relationship, the name, academic degree and year of conclusion are presented [12].

2.3 Science Tree

Created in 2015, the Science Tree[4] application collects metadata of academic genealogy from many countries [6]. The authors are crawling data from a variety of sources, including the Networked Digital Library of Theses and Dissertations (NDLTD), which has more than 4.5 million theses and dissertations from around the world. They developed a framework to extract academic genealogy trees from these data, providing a series of analyses that describe the main properties of the academic genealogy tree. Figure 4 [12] presents the result of a query using the same researcher of Fig. 3.

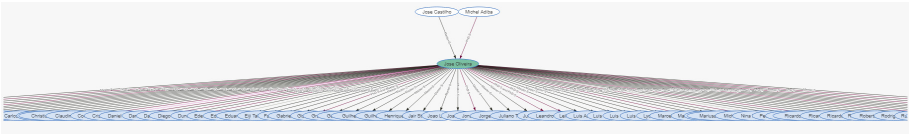

Fig. 4. Result of the query "José Palazzo Moreira Oliveira" using Science Tree [12].

[2] http://plataforma-acacia.org.
[3] http://lattes.cnpq.br.
[4] http://www.sciencetree.net.

2.4 The Brazilian Research Community

Aiming to reconstruct advisor-advisee relationships from records from many institutions around the world and from distinct disciplines, Dores et al. [7] also build academic genealogy trees from the Lattes Platform. In this paper, 222,674 curricula vitae from researchers holding a PhD degree were collected and then processed.

The proposed algorithm, in the first step, orders the curricula by the year of the PhD conclusion. Then, for each researcher with a curriculum on Lattes Platform it is created or updated a vertex which refers to itself on the graph. After that, for each advising relationship (masters and/or PhD) an edge is established between this researcher vertex and the corresponding students' vertices. Thus the relationships between students and advisors can be gradually built. Figure 5 shows the tree of Marcos André Gonçalves, where the node colors represent the tree levels.

Among the reported results, the authors presented 903,183 vertices, 1,144,051 edges, 70,610 trees, and 22,061 distinct components. The average tree size was 40.19 while the average tree width was 3.81. The 10 largest trees have more than 5,000 nodes, although 80% of them have less than 20 nodes. Regarding the distribution of the trees depth, 50% of them had only one level.

They also report that several researchers opt to study abroad. Portugal followed by United States were the countries most chosen by Brazilian researchers to conclude their masters and/or doctorate studies.

Fig. 5. Example of an academic genealogy tree built from Lattes [7].

3 Methodology

This section presents the methodology adopted in this paper. The proposed approach is divided into the following steps: data source definition, harvesting and pre-processing, data modeling and indexing, web information system construction, and data analysis.

Unlike the Academic Family Tree presented in Sect. 2.1, which is user dependent, we have chosen to collect official data available in digital libraries. These data sources must support some interoperability feature such as the OAI-PMH protocol [11] and the Dublin Core[5] format. This choice solves the problem of cold start and registration of false information.

In order to evaluate the proposed system we use Brazilian Digital Library of Theses and Dissertations[6] (BDTD). It is developed and managed by the Brazilian Institute of Information in Science and Technology (IBICT), and integrates the repositories of educational and research institutions in Brazil, and also stimulates the registration and publication of theses and dissertations electronically. At the time of our harvesting more than 570,000 documents were indexed.

The main metadata fields collected were:

- author;
- advisor;
- name of the educational or research institution;
- acronym of the educational or research institution;
- title;
- topics;
- URL of the document in the original repository;
- PhD thesis or Master's dissertation;
- URL of the author curriculum at Lattes Platform;
- citation;
- year of publication.

After the data selection and harvesting, a set of cleaning operations were applied. Several analyzes were performed to identify anomalies, so errors could be corrected or eliminated. Some transformation operations were applied in the author and advisor fields. The most common ones include: inverting last names based on the comma character, removing institution acronyms, and removing structural prefixes present in the content of a metadata. In addition, duplicate tuples from more than one digital repository were removed. At the end of the cleaning process, the number of theses and dissertations decreased to 465,847.

In Computer Science, trees are data structures widely used to represent elements hierarchically organized. However, the academic genealogy trees are represented by graphs, because a researcher may have more than one ancestor (one for each dissertation or thesis) and because there may be cycles. That is the reason why we transformed the cleaned data to store the complete graph in the relational model. Figure 6 shows the data model [12]. The first table contains all the researchers and their properties. The second

[5] http://dublincore.org.
[6] http://bdtd.ibict.br.

keeps M:N directed advising relationships between pairs of academics. The Database Management System (DBMS) used for storage was PostgreSQL. All the process of harvesting and pre-processing were developed in PHP programming language.

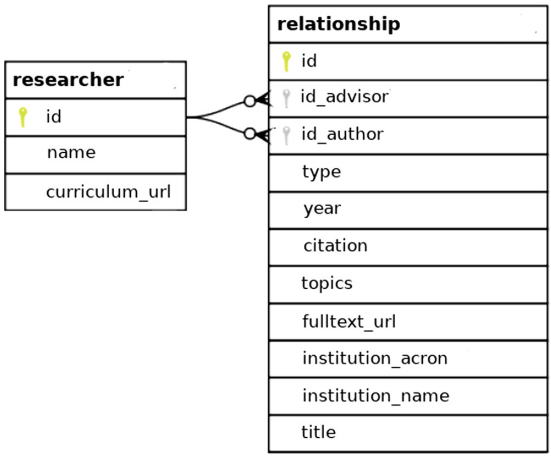

Fig. 6. Relational data model representing the genealogy academic trees [12].

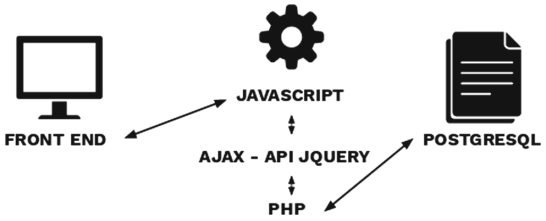

Fig. 7. Data flow between front and back ends [12].

The Web application interface was developed using Javascript, HTML, and CSS. The queries are sent to the back-end using Jquery library [15]. Methods implemented in PHP retrieve the subgraph from data stored in the DBMS. We have used the dagre-d3 library [16] to draw the academic trees with the selected data. Figure 7 shows the data flow between front and back-end [12].

From the developed system we analyze some properties of the genealogy graph, such as the giant component, the advising distribution by year and the advising density.

4 Results

4.1 The Developed Tool

Figure 8 shows the web interface of the proposed system, which is designed to be simple and intuitive. The button *Search Academic* opens a query field that allows the user to search by the name of the researcher.

The architecture presented in Fig. 7 [12] allowed to implement a dynamic search that suggests multiple researchers as the user types in the query field. For each character entered, the results are filtered and displayed on the screen. All substrings of author and advisor names with at least 3 characters have been indexed in the DBMS, so the user does not need to know the full name of a researcher, the order of names nor to complete each name. Figure 9 exemplifies this behavior while the user is querying by "avancini mar" [12].

Next to the button *Search Academic* the user can edit the depth level. The tree expands the number of levels both toward the leaves and toward the root. Thus, he or she can see the advising lineage and the graduated students. Figure 8 shows the result tree setting two levels and selecting "Rita Maria Pereira Avancini" from the five returned researchers.

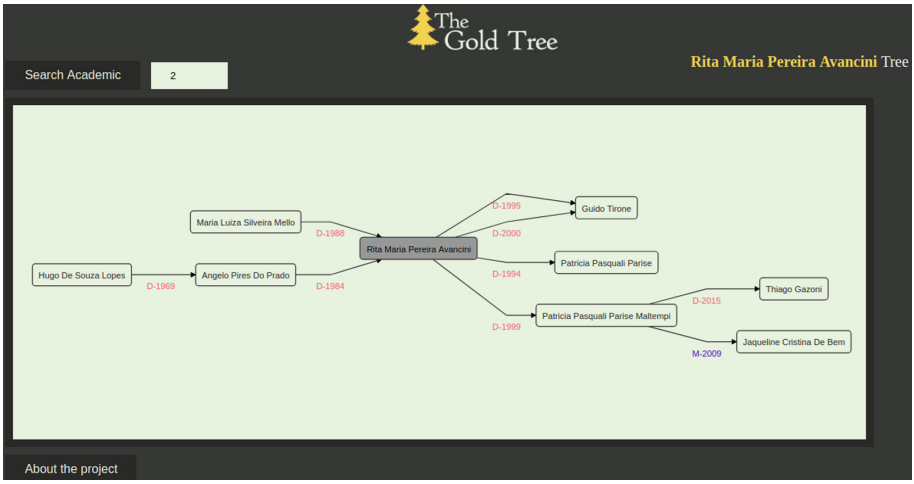

Fig. 8. Web interface of the proposed system showing the result selecting "Rita Maria Pereira Avancini" with 2 levels depth [12]. (Color figure online)

Fig. 9. Example of the dynamic search feature, that suggests multiple researchers as the user types in the query field [12].

Each researcher is represented by a vertex that contains his or her full name. Master's advising relationships are edges with M-YEAR labels in blue. Doctoral advising relationships have D-YEAR labels in red.

The user can freely move the tree in the rectangular area in which it appears and zoom in/out using the mouse controls. By clicking on a relationship, a window opens displaying the information available in the thesis or dissertation metadata (Fig. 10 [12]). Also, when you click on a vertex, a new tree is generated using the selected researcher as the target of the query.

The information system developed is available online.[7]

4.2 An Example Case of Brazilian Academic Genealogy

Figure 11 shows the genealogy of the researcher with the highest number of descendants (274 masters and 159 doctorates, with a total of 433 advising relationships), considering 3 levels of depth. This view contains 3726 vertices and 4192 edges, which represents almost 1% of the BDTD. We have applied the Yifan Hu algorithm [9], named ego network, to organize the vertices and then draw this researcher tree.

Fig. 10. Metadata describing the thesis or dissertation presented when clicking on an advising relationship [12].

[7] http://thegoldtree.c3.furg.br/.

Expanding the ego network to the maximum depth, we found the giant component of the BDTD graph with 305,733 (68.9%) vertices connected by 320,591 (68.8%) edges. These properties are summarized in Table 1.

Table 1. Properties of the Brazilian research network extracted from BDTD.

Property	Value
Vertices	443,654
Edges	465,847
Vertices in the giant component	305,733
Edges in the giant component	320,591

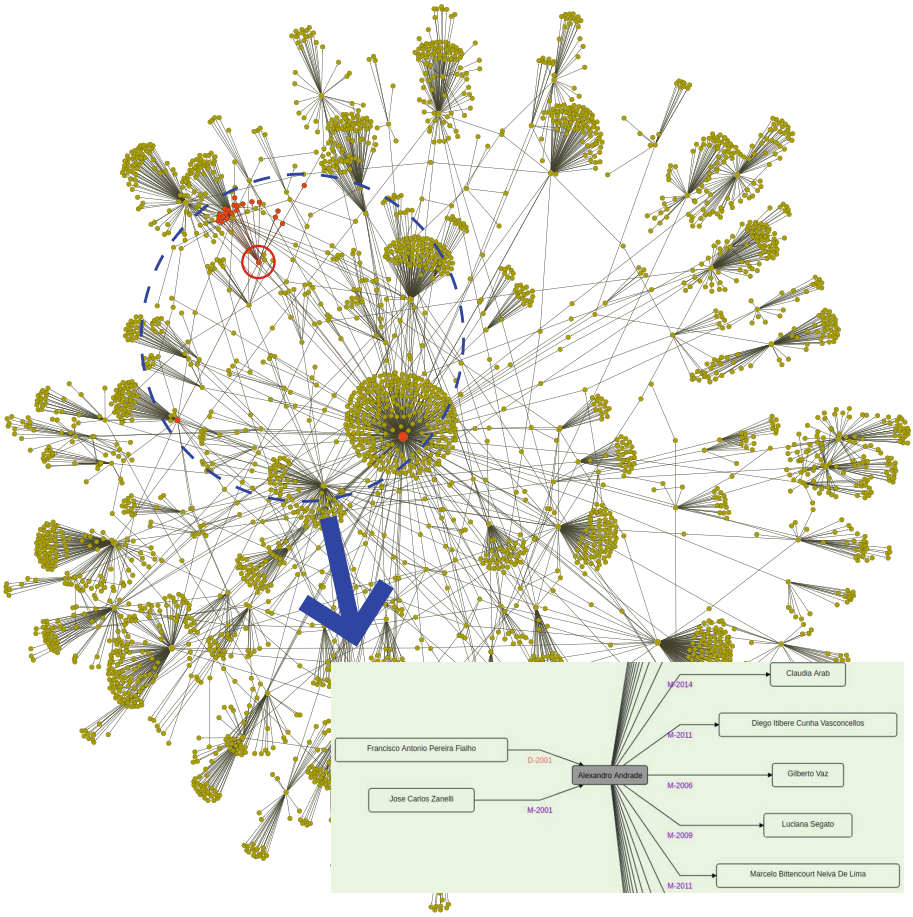

Fig. 11. An ego network of the advisor with the greatest number of descendants.

Analyzing the temporal distribution of the amount of published theses and dissertations (Fig. 12, we can see a constant and timid growth rate up to the year of 2000. The amount increases significantly between 2001 and 2013, not only because of the popularity of the digital institutional repositories, but also due to the expansion of Brazilian universities and graduate programs. However, the publication of documents has been decreasing in recent years.

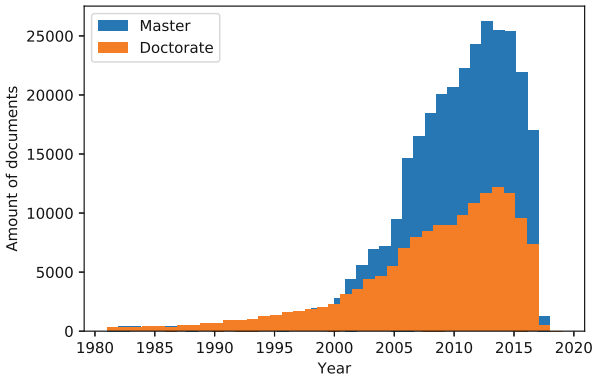

Fig. 12. Distribution of theses and dissertations published over time.

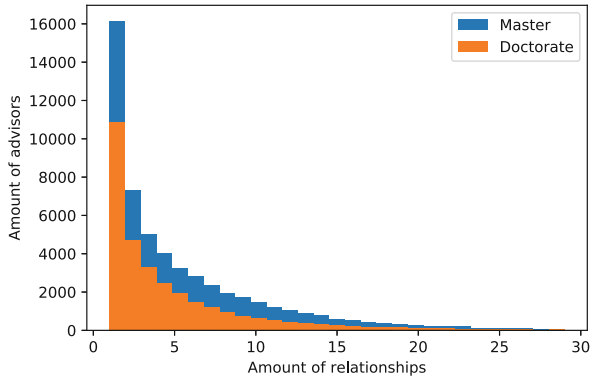

Fig. 13. Distribution of advising relationships.

Figure 13 shows the density of advising relationships. On a logarithmic scale, Fig. 14 shows the equivalent power law. The vast majority of researchers has mentored few master's or doctoral students. Only 396 (1.2%) researchers advised more than 30 students.

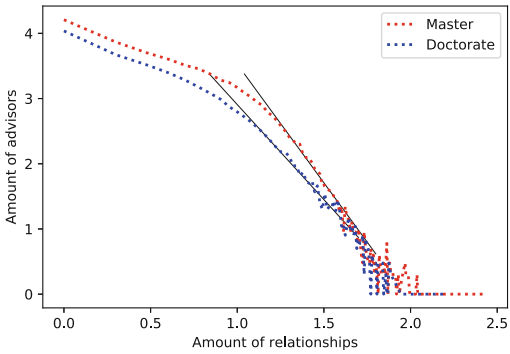

Fig. 14. Distribution of advising relationships in a logarithm scale.

5 Conclusion

This paper carries on studies conducted in a previous work [12], which describes the development of an academic tree visualization system called The Gold Tree. Such tool is based on the Brazilian Digital Library of Theses and Dissertations (BDTD). Despite the results presented in our work used a single database, the data source chosen could be easily replaced by one that publishes the metadata using the Dublin Core standard and the OAI-PHM protocol.

The Gold Tree would aid the study of Brazilian academic genealogy, since it would allow a user to query about a researcher on the database, defining the level of the tree depth in order to visualize the academic tree. The database used could be more understood as some studies of the features of the database were conducted, such as the number of researchers and connections, information on which was the largest component, the distribution of documents (theses and dissertations) per year, and the distribution of the amount of advising relationships.

Compared to related and similar systems, The Gold Tree is able to handle some limitations of others, as described as follows. Our proposed system uses an official source of data unlike Academic Family Tree, which is fed by users with unofficial confirmation. On Acácia Platform, it is not possible to visualize the academic tree. And on Science Tree, there is a limitation on the tree level of depth and also it is not possible to have a dynamic search. Both of these issues are solved in our proposal. Comparing the results obtained previously in [7], while the analysis is done only on the data of the Lattes Platform, in this paper we focus on data extracted from the Brazilian Digital Library of Theses and Dissertations, completing the available information about the Brazilian academic genealogy.

Future work concern on the creation of a recommendation system of advisors, which based on an abstract of a thesis or masters proposal. The system would be able to extract several features from the titles and abstracts of the theses and dissertations of all the descendants of a researcher to set up the advising profile, based on the vector space model. An machine learning approach would rank a list of the most apt academic advisors, using multiple classification algorithms with a high diversity.

Acknowledgment. This study was supported by the Fundação de Amparo à Pesquisa do Estado do RS (FAPERGS) [grant numbers TO 17/2551-0000872-3, TO 19/2551-0001279-9, and TO 19/2551-0001660], Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [grant number 305882/2016-3], and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) [grant number 88887.464880/2019-00].

References

1. Barabâsi, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A **311**(3–4), 590–614 (2002)
2. Batagelj, V., Mrvar, A.: Pajek—analysis and visualization of large networks. In: Mutzel, P., Jünger, M., Leipert, S. (eds.) GD 2001. LNCS, vol. 2265, pp. 477–478. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45848-4_54
3. Chang, S.: Academic Genealogy of Mathematicians. World Scientific, Singapore (2011)
4. Damaceno, R.J.P.; Rossi, L.M.C.J.P.: Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na plataforma lattes. In: Proceedings of the 32nd Brazilian Symposium on Databases, Uberlândia (2017)
5. David, S.V., Hayden, B.Y.: Neurotree: a collaborative, graphical database of the academic genealogy of neuroscience. PLoS ONE **7**(10), e46608 (2012). https://doi.org/10.1371/journal.pone.0046608
6. Dores, W., Benevenuto, F., Laender, A.H.: Extracting academic genealogy trees from the networked digital library of theses and dissertations. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp. 163–166. ACM, New York (2016). https://doi.org/10.1145/2910896.2910916
7. Dores, W., Soares, E., Benevenuto, F., Laender, A.H.F.: Building the Brazilian academic genealogy tree. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) Research and Advanced Technology for Digital Libraries, pp. 537–543. Springer International Publishing, Cham (2017)
8. Douglas, S.M., Montelione, G.T., Gerstein, M.: PubNet: a flexible system for visualizing literature derived networks. Genome Biol. **6**(9), R80 (2005). https://doi.org/10.1186/gb-2005-6-9-r80
9. Hu, Y.: Efficient, high-quality force-directed graph drawing. Math. J. **10**(1), 37–71 (2005)
10. Laender, A., Moro, M., Silva, A., et al.: Ciência brasil-the brazilian portal of science and technology. In: Seminário Integrado de Software e Hardware, pp. 1366–1379. Sociedade Brasileira de Computação (2011)
11. Lagoze, C., Van de Sompel, H.: The open archives initiative: building a low-barrier interoperability framework. In: Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital libraries, pp. 54–62. ACM, New York (2001). https://doi.org/10.1145/379437.379449
12. Madeira, G., et al.: The gold tree: an information system for analyzing academic genealogy. In: Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS, INSTICC, pp. 114–120. SciTePress (2019). https://doi.org/10.5220/0007758401140120
13. Malmgren, R.D., Ottino, J.M., Nunes, L.A.: The role of mentorship in protégé performance. Nature Int. J. Sci. **465**. 7298, 622 (2010). https://www.nature.com/articles/nature09040
14. Mena-Chalco, J.P., Cesar-Jr, R.M.: Prospecção de dados acadêmicos de currículos Lattes através de scriptLattes, chap. Bibliometria e Cientometria: reflexões teóricas e interfaces, pp. 109–128. Pedro & João Editores, São Carlos (2013)
15. Osmani, A.: Learning JavaScript Design Patterns: A JavaScript and jQuery Developer's Guide. O'Reilly Media, Inc. (2012)

16. Roeder, L.: Dagre-d3 (2018). https://github.com/dagrejs/dagre-d3
17. Yu, M.C., Wu, Y.C.J., Alhalabi, W., Kao, H.Y., Wu, W.H.: Researchgate: an effective altmetric indicator for active researchers? Comput. Hum. Behav. **55**, 1001–1006 (2016). https://doi.org/10.1016/j.chb.2015.11.007