








Using Academic Genealogy for Recommending Supervisors

Gabriel Madeira¹^a, Eduardo N. Borges¹^b, Giancarlo Lucca¹^c,
Washington Carvalho-Segundo²^d, Jonata C. Wiczynski¹^e, Helida Santos¹^f
and Graçaliz Dimuro^{1,3}^g

¹*Centro de Ciências Computacionais, Universidade Federal do Rio Grande, Rio Grande, RS, Brazil*

²*Instituto Brasileiro de Informação em Ciência e Tecnologia, Brasília, DF, Brazil*

³*Departamento de Estadística, Informática y Matemáticas, Universidad Pública de Navarra, Pamplona, Spain*

Keywords: Recommender Systems, Academic Genealogy, Academic Supervising, Nearest Centroid Classification.

Abstract: Selecting an academic supervisor is a complicated task. Masters and Ph.D. candidates usually select the most prestigious universities in a given region, investigate the graduate programs in a research area of interest, and analyze the professors' profiles. This choice is a manual task that requires extensive human effort, and usually, the result is not good enough. In this paper we propose a Recommender System that enables one to choose an academic supervisor based on his/her academic genealogy. We used metadata of different theses and dissertations and applied the nearest centroid model to perform the recommendation. The obtained results showed the high precision of the recommendations, which supports the hypothesis that the proposed system is a useful tool for graduate students.


1 INTRODUCTION


One of the first steps during the process of acquiring an academic degree, either Masters or Ph.D., is the choice of the theme to be investigated and then an academic supervisor, which will be in charge of aiding the student to achieve his/her goals. In the task of choosing an academic supervisor, the amount of experience regarding the theme's field of study is significant. However, this choice might not be so trivial. This job should include a thorough analysis of each professor's curriculum, including the list of scientific publications and all theses and dissertations advised, which can be available in multiple and distributed research repositories.


Ray and Marakas (Ray and Marakas, 2007) assert that students' usual criteria are professors' reputation, knowledge, and matching of interests, among others. However, this choice is often made in an unplanned


manner, which can become one of the reasons for regret, lack of motivation, and poor quality of research output. The authors proposed an analytical hierarchy process for selecting a thesis supervisor, which shows that the number of theses supervised is the least important criterion for both junior and senior graduate students. Besides matching interests, the professors' social network and relationship with other professors in the same institute and outside were pointed out as essential criteria.


In this paper, we developed a Recommender Systems (RS) that extracts knowledge from a set of descriptive metadata of theses and dissertations supervised throughout the advisors' career, considering social aspects extracted from their academic genealogy trees. Our methods can represent adequately the profile and research area of a young professor who mentored few or no students. When inputting the title and abstract of a thesis/dissertation proposal, the system returns a ranking of the most compatible advisors for the chosen theme. So, the major contributions of this paper are the following: a novel content-based recommendation approach for selecting academic supervisors; the use of academic genealogy trees (Sugimoto, 2014) (see an example in 4) to model the supervisors' profiles; and the experimental evaluation of the proposed RS using real data from a networked digital li-


^a  <https://orcid.org/0000-0001-8348-3498>


^b  <https://orcid.org/0000-0003-1595-7676>

^c  <https://orcid.org/0000-0002-3776-0260>

^d  <https://orcid.org/0000-0003-3635-9384>

^e  <https://orcid.org/0000-0002-8293-0126>

^f  <https://orcid.org/0000-0003-2994-2862>

^g  <https://orcid.org/0000-0001-6986-9888>

brary of theses and dissertations.

The experiments were conducted using a dataset containing more than 79,000 advisors from more than 600,000 theses and dissertations. Our system was able to recommend the correct advisors, on average, in the third position of the suggested ranking.

The rest of this paper is organized as follows. Section 2 presents the preliminary concepts necessary to understand our methods. In section 3 we discuss related work. Section 4 presents our approach to recommend academic supervisors. In Section 5, we discuss the obtained results. Finally, in Section 6, we draw our conclusions.

2 PRELIMINARY CONCEPTS

2.1 Recommender Systems

RSs provide suggestions for information related to several decision-making processes. The recommendations are offered as ranked lists of information items, which are personalized for each user. Besides filtering the most suitable information, RSs organize it with a high probability of relevance based on user preferences and constraints. (Ricci et al., 2011).

Among the features pointed by (Bobadilla et al., 2013) that define a RS, we highlight: type of data, e.g. ratings, content for items, social relationships and location-aware information; filtering algorithm, e.g. content-based, collaborative, context-aware or hybrid; techniques, e.g. probabilistic algorithms and fuzzy models; sparsity level of the database and the desired scalability; objective – predictions or top- n recommendations; quality evaluation, e.g. novelty, coverage and precision (Ge et al., 2010).

Content-based filtering (Salter and Antonopoulos, 2006) makes recommendations based on user past choices using the similarity between the content of these items and those to be recommended. Demographic filtering (Krulwich, 1997) performs the similarity among users, based on the principle that individuals with common personal attributes will also have common preferences. Examples of these attributes are gender, age, location and language. Collaborative filtering (Bobadilla et al., 2012) allows users to give ratings (explicit or implicitly) on information items, which can be used to recommend content for other users with similar profiles. Hybrid filtering (Chen et al., 2018) combines multiple filtering algorithms.

In this paper, we adapted a well-known content-based filtering neighborhood-based recommendation technique (Desrosiers and Karypis, 2011) for sug-

gesting academic supervisors, by making use of the content of theses and dissertations descriptive metadata and the advising relationships. We focus on system precision disregarding scalability to perform top- n recommendations.

2.2 Vector Space Model

Proposed by Salton in 1968 (Salton, 1968), Vector Space Model (VSM) is a classic information retrieval model implemented in many search engines. It uses bag of words representation and allows to retrieve documents ordered according to the query similarity. Let D be a collection of documents represented by vectors of weights associated with the terms contained in the collection vocabulary. The similarity between a query q and a document $d \in D$ is given by Eq.(1), which performs the cosine of the angle between the vectors.

$$\text{sim}(\vec{q}, \vec{d}) = \cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| |\vec{d}|} \quad (1)$$

The most common weighting scheme is called TF-IDF, represented by $tf \times idf$ (Manning et al., 2008) and defined by Eq.(2), where t is a term of the vocabulary V , d is a document, $n_{t,d}$ is the frequency of the term t in the document d , N is the size of the collection, i.e. the amount of documents, and df_t is the frequency of documents containing t .

$$tf_{t,d} \times idf_{t,d} = \frac{n_{t,d}}{\sum_{t' \in V} n_{t',d}} \times \log\left(\frac{N}{df_t}\right) \quad (2)$$

A VSM can be efficiently implemented using an inverted index, which maps each term of the vocabulary into a list of postings, which contain the identifier of the document containing the term and additional information such as frequency.

2.3 Nearest Centroid Classification

Nearest Centroid (Tibshirani et al., 2002; Manning et al., 2008) is a model that classifies test samples according to their distance to the centroid of data classes. For text classification, let n be the amount of documents in a set D , and let $s_i | 1 \leq i \leq n$ be a sample defined by (\vec{x}_i, y_i) , where \vec{x}_i is a document represented in the VSM using $tf \times idf$ weighting scheme, and let Y be the set of class labels, and $y_i \in Y$ is the class label of this sample.

In the training phase, the algorithm sets the centroids $\vec{\mu}(l)$ for each distinct class label $l \in Y$, computing the vector average or center of mass of its members. Equation (3) defines $\vec{\mu}(l)$, where D_l is the docu-

ment whose class label is l .

$$\vec{\mu}(l) = \frac{1}{|D_l|} \sum_{s_i \in D_l} \vec{x}_i \quad (3)$$

The prediction function reported in Eq.(4) returns the class label \hat{y} which minimizes the Euclidean distance between the associated centroid $\vec{\mu}_l$ and the test instance \vec{x} . Alternatively, Eq.(5) defines how the label can be predicted using the cosine similarity, previously presented in Eq.(1).

$$\hat{y} = \arg \min_{l \in Y} |\vec{\mu}_l - \vec{x}| \quad (4)$$

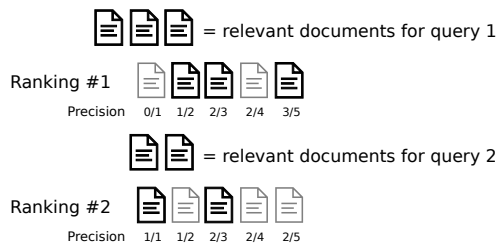
$$\hat{y} = \arg \max_{l \in Y} \text{sim}(\vec{\mu}_l, \vec{x}) \quad (5)$$

2.4 Mean Average Precision

The Mean Average Precision (MAP) (Manning et al., 2008) is a well-know metric for evaluating Information Retrieval systems. It performs the mean of the average precision scores calculated for several queries.

Figure 1 shows an example in which the MAP is performed. Relevant documents are represented in bold. For the first query, there are three relevant documents that are returned in the second, third and fifth positions of the ranking. The average precision for query 1 is $avgP_1 = (1/2 + 2/3 + 3/5)/3 = 0.59$. Query 2 retrieves two relevant documents in the first and third positions. The average precision for query 2 is $avgP_2 = (1/1 + 2/3)/2 = 0.83$. Therefore, $MAP = (avgP_1 + avgP_2)/2 = 0.71$.

In classification problems, the queries are the test instances submitted to the classification model.



average precision query 1 = $(1/2 + 2/3 + 3/5) / 3 = 0.59$

average precision query 2 = $(1 + 2/3) / 2 = 0.83$

mean average precision = $(0.59 + 0.83) / 2 = 0.71$

Figure 1: An example of MAP considering two queries.

3 RELATED WORK

(Husain et al., 2019) review the literature about expert finding systems between 2010 and 2019. These systems have been proposed in different domains and

environments, such as medicine, enterprise, question answering communities, and social networks. Academia was the largest domain, comprising 44 studies (65% of the sample). The majority of these systems were developed for specific academic tasks like paper reviewing, research collaborations, finding similar experts, and industry or university collaborations. Only one study addressed finding a suitable supervisor (Alarfaj et al., 2012). The authors proposed a simple database-driven approach that selects a supervisor from the university's academic staff, and a data-driven approach where candidates are extracted from pages returned by a web search engine.

(Hasan and Schwartz, 2018) developed RecAdvisor, a criteria-based Ph.D. supervisor recommender for Florida State University (FSU). The prototype collects information from four different sources: Microsoft Academic Graph, Computing Research and Education Association (CORE), professors' CVs and FSU's digital repository. The profiles are indexed using Elasticsearch (Gormley and Tong, 2015).

Selecting an academic supervisor is not a popular research theme. Most related work we could find were proposed for very specific scenarios, such as finding scientific articles and papers, recommending academic courses, and suggesting researchers for collaboration.

Docear's RS (Beel et al., 2013) is part of a literature management software. The system allows a researcher to search, read, make annotations and organize scientific articles, besides drafting manuscripts. Docear suggests citations from a digital library containing around 1.8 million research articles from various disciplines.

Champiri et al. (Champiri et al., 2015) published a survey analysing if incorporating contextual information in recommender systems is an effective approach to create more accurate and relevant recommendations in digital libraries. They highlight RSs with the purpose of exploring a research area and finding relevant research sources.

In order to recommend the most relevant courses to its users, the RARE system (Bendakir and Aïmeur, 2006) combines the benefits of both former students' experience learned in the data mining process and current students' ratings. It is a hybrid filtering approach based on association rules. Authors used the association algorithm Apriori implemented by Weka tool for training the model. With a similar purpose, O'Mahony and Smyth (O'Mahony and Smyth, 2007) developed a RS for an on-line enrolment application of Dublin's University College. Users can search by inserting keywords or specific core module IDs. The output is a list of elective modules which match

the search criteria and their profile. Authors used a item-based collaborative filtering algorithm (Karypis, 2001). Another strategy has recently been proposed to domain-aware grade prediction and top-n course recommendation (Elbadrawy and Karypis, 2016).

Rodrigues et al. (Rodrigues et al., 2018) use different strategies to suggest scientific collaboration for researchers based on their interest. The authors model the similarity between researchers using data from ResearchGate social network. They exploit co-authorship attributes and paper reading records with a hybrid approach, having both content-based and collaborative filtering. Experimental results showed that the content-based strategy outperforms neighborhood-based collaborative filtering strategies up to 21.16% regarding F-measure for the top-20 recommendation lists.

Mendonça et al. (Mendonça et al., 2020) present a systematic mapping of RSs based on scientific publications. They analysed that Machine Learning algorithms and Vector Space Model representation are the most used in content-based RSs for the academic field. On the other hand, for collaborative filtering approaches, common methods are based in neighborhood, such as k Nearest Neighbors. Databases frequently used were: CiteULike, DBLP, Microsoft Academic Search (MAS), CiteSeerx, PubMed and Web of Science.

The RS proposed in this paper differs from most related works in the following aspects. Instead of recommending collaborations based in coauthoring, we used the advising relationships to suggest academic supervisors. The supervisor profiles are learned using their academic genealogy trees build from Electronic Theses and Dissertations (ETDs) repositories. We quantitatively evaluate our system using data from more than 79,000 professors and 600,000 students.

4 A NOVEL ACADEMIC SUPERVISOR RS

In this section we present our novel approach for recommending academic supervisors. Figure 2 shows the architecture of the proposed RS.

The first process collects data and builds the academic genealogy. From a repository of ETDs, the RS selects a set of documents of interest. After that, the researchers, i.e. supervisors and authors, are extracted from the selected theses and dissertations. A deduplication method is applied to identify each unique research. From the relationships between unique researchers, the genealogy graph is built using the method proposed in (Madeira et al., 2020).

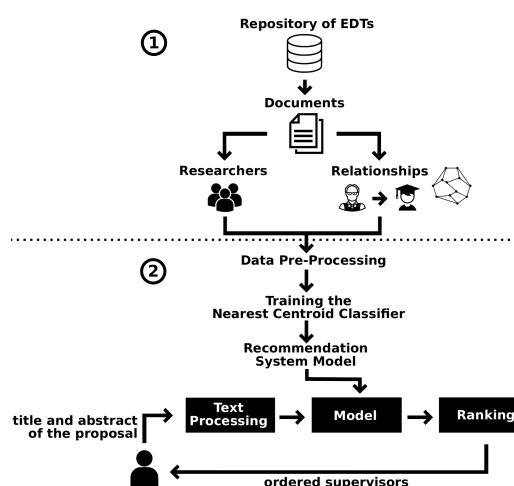


Figure 2: Architecture of the proposed academic supervisor RS.

Using machine learning, in the second process, we fit a model that will be used to perform recommendations. Textual data from the academic genealogy are pre-processed with standard operations. We transform data features into VSM using $tf \times idf$ term-weighting strategy, presented in Section 2.2. Next, using the transformed data, we train a classifier based on the Nearest Centroid algorithm (Section 2.3) using the cosine similarity and build the recommendation model.

Finally, users can query by inserting the title and abstract of the research proposal. This information is transformed using the same pre-processing scheme as before and then it is used as input of the classifier, which returns a final ranking of recommendations composed by the most suitable supervisors ordered by relevance. In the following subsections, each step is detailed.

4.1 Data Source

The proposed RS can handle different data sources. Repositories of ETDs must support some interoperability features, such as the OAI-PMH protocol (Devarakonda et al., 2011), or have an API available for harvesting metadata.

In this study, we used a Brazilian repository, known as (*Biblioteca Digital Brasileira de Teses e Dissertações* – BDTD)¹. This networked digital library contains metadata from more than 600 thousand documents. BDTD integrates and disseminates, in an unique website, the complete content of different theses and dissertations that are produced in Brazilian

¹Available in <http://bdttd.ibict.br>.

universities. Additionally, its access is open and free of any kind of charge.

This digital library also contributes to increase the content of Brazilian theses and dissertations on the internet, growing the visibility of the national technological and scientific production. Moreover, BDTD also provides major visibility and management of the investments done in graduate programs.

From the BDTD available metadata fields, we picked:

- network_acronym_str – acronym for the university;
- network_name_str – name of the origin repository;
- title – document’s title;
- description – document’s abstract;
- author – document’s author;
- advisor – document’s supervisor;
- author_lattes – URL of the author’s curriculum in Lattes Platform²;
- topic – related topics of the document;
- citation – how to cite the document;
- language – language of the document (mostly in Portuguese)
- publishDate – year the document was published;
- format – indicates if a document is a Ph.D. thesis or a Master dissertation;
- url – URL of the document in its original repository.

We collected 612,714 theses or dissertations from BDTD. For each document, we extracted the researchers (author and supervisor) and their relationship to build the academic genealogy. The giant component of the graph had more than 300 thousand vertices connected by more than 350 thousand edges.

4.2 Data Pre-processing

The amount of documents collected and the data volume stored can be large. Data can have different notations and language particularities. Thus, in order to standardize the information, a pre-processing is necessary. We applied the following operations in the textual data. First, the title, description, author and advisor metadata fields are tokenized. Tokens are normalized by turning characters to lowercase, removing

²An information system maintained by the Brazilian National Council for Scientific and Technological Development (CNPq) that integrates databases of curricula, research groups and institutions. Available in <http://lattes.cnpq.br>.

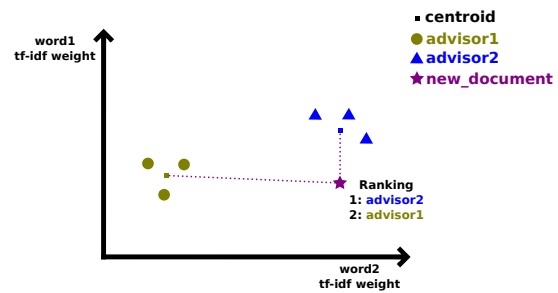


Figure 3: An example of how Nearest Centroid algorithm works with VSM.

accents and non-alphanumeric symbols. We also removed stopwords in Brazilian Portuguese using the NLTK toolkit (Loper and Bird, 2002). This operation reduces data volume and speeds up the system without affecting significantly the quality of the results. After that, the space of features of the title and description are transformed into the VSM, using $tf \times idf$ term-weighting strategy.

Besides, we applied a cleaning process, removing duplicates and documents with missing information. We also corrected misreported information, resulting in 579,486 pre-processed theses and dissertations.

Finally, an additional pre-processing operation sets the class label using the advisor metadata field, composing the training samples with the structure $s_i = (\vec{x}_i, y_i)$, where \vec{x}_i is the title and description $tf \times idf$ weights, and y_i is the class label of the sample i . The conducted experiments used only one level of the tree but the system can be parameterized to reach any depth, adding the vector components of the supervisors’ theses in the training samples.

4.3 Training the Classifier for Recommendation

Due to the good results presented in (Han and Karypis, 2000), we chose the Nearest Centroid classification algorithm to learn the supervising profiles.

Figure 3 shows an example of how the algorithm works regarding two candidate advisors and a new test instance (Ph.D. or Masters proposal). Each circle or triangle refers to a thesis or dissertation in the multidimensional term space. Dots are the centroids of the clusters formed by all works supervised by each advisor, i.e. the supervisors’ profiles. The user receives as recommendation a profile list, composed by the class labels (distinct advisors) of the $n = 2$ nearest centroids.

5 EXPERIMENTAL EVALUATION

In this section, we explain how we evaluate the quality of the proposed academic supervisor RS, we report implementation details, and we also present the obtained results.

5.1 Validation

To evaluate the proposed RS, we used k -fold cross-validation technique, which consists in splitting the available dataset in k folds and calculating the evaluation metrics k times, where in each interaction, one of the parts is used for testing and the others are used for training the model.

We used the evaluation metric Mean Average Precision (MAP), presented in Section 2.4 applied in the output of the predict function, which is a ranking with more than 79 thousand positions. Besides, a student may have two different advisors, one for the Masters dissertation and another for the Ph.D thesis. In this case, both were considered correct recommendations, because they could appear in distinct positions of the ranking.

5.2 Implementation

The implementation was coded in Python using the scikit-learn library (Pedregosa et al., 2011). The Nearest Centroid algorithm was adapted from the scikit-learn implementation to return an ordered list of centroids. This list is the ranking of the most suitable academic supervisors from BDTD for a Ph.D. or Masters proposal in the query.

We highlight that due to dataset volume, we needed to modify the source code of the python library. Precisely, the scikit-learn uses a float64 matrix to store the documents represented in the VSM. This approach slows down the system performance. So, to avoid this problem, we have changed it to *scipy.sparse.lil_matrix*, which implements row-based list of lists sparse matrix.

Experiments run on a dual-socket quad-core Intel® Xeon® L5420 2.5 GHz CPU with 32 GB of memory.

5.3 Results

In this section, we start by presenting a query example and the returned recommendations. After, we focus on the evaluation considering MAP and the frequency in which the correct advisor is well recommended.

Figure 4 shows the academic genealogy tree of the researcher named Marília Abrahão Amaral. Note that

she had two different advisors, one for Masters (M) and another for Ph.d (D). Relationships labels also include the thesis or dissertation’s publication year. Moreover, we can also observe that this person has already advised a Masters student in 2017.

Marília Abrahão Amaral Tree

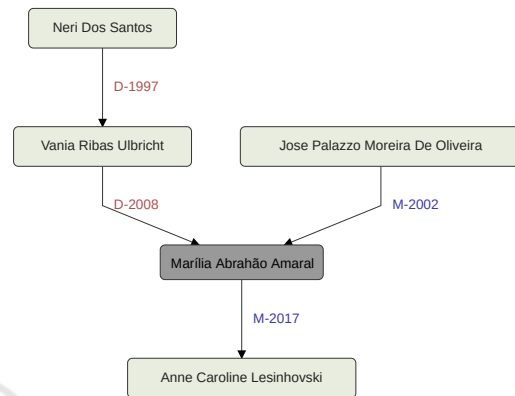


Figure 4: An example of an academic genealogy tree, where each arrow represents a Ph.D. (D) or Master (M) advising relationship.

In the proposed RS, using as input the title and abstract of her thesis, we obtain a ranking of supervisor recommendations. The first five positions are presented in Table 1.

Note that a correct answer (Marília’s Masters advisor, J. Palazzo) was returned in the fifth position. Considering that we have more than 79 thousand possible candidates, the obtained result is fairly good.

Moreover, if we take into account the remainder recommendations, all suggestions are indeed related to Marília’s research field and can also be considered exceptional recommendations. The first two researchers (J. Valdeni and R. Vicari), for instance, work in the same university of J. Oliveira. Those three professors had already been part of the same research group within the same graduate program. Therefore, any of them could have been Marília’s advisor as their profiles are strongly related to her topic of interest.

Table 1: Returned recommendations using the title and abstract of Marília Abrahão Amaral’s Ph.D. thesis.

Position	Name
1	Jose Valdeni de Lima
2	Rosa Maria Vicari
3	Alex Sandro Gomes
4	José Dutra de Oliveira Neto
5	Jose Palazzo Moreira de Oliveira
...	

Notwithstanding the above, the analysis of only one returned ranking can not be enough to evaluate the system quality for this query. Thus, considering all ten models fitted in cross-validation, the obtained mean position of this advisor was 5.2, reinforcing the quality of the recommendation.

To evaluate the general quality of the proposed system, we performed a study considering all 573,671 instances. Table 2 presents the results of the cross-validation process, achieving a MAP equals to 32.41%, meaning that our system was able to suggest the correct advisors, on average, in the third position of the recommended ranking, since $1/3 \approx 0.3241$.

Table 2: Cross-Validation evaluation considering the MAP metric.

Fold	MAP	Fold	MAP
1	0.3259	6	0.3245
2	0.3266	7	0.3256
3	0.3216	8	0.3223
4	0.3236	9	0.3242
5	0.3239	10	0.3228
		Avg.	0.3241

In order to clarify the effectiveness of the method, we present a graphical analysis of the general obtained results. Precisely, in Figure 5, we show a histogram containing the frequency of each position of a correct advisor returned in the ranking. In this figure, the x axis represents the first 100 positions of the ranking. In the y axis, there is the amount of correct recommendations for each position in the x axis.

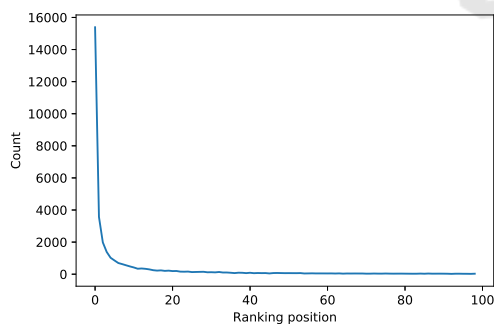


Figure 5: Number of queries that returned the correct advisor for each position of the ranking.

It can be noted that more than 15,000 queries have returned a correct advisor in the first position. Moreover, the majority of the queries have returned the advisor at least in the 10th position. After that, we observe that the system presents a stability. After the 100th position, the recommendations tend to be zero.

6 CONCLUSION

In this work we developed a recommender system that extracts knowledge from a set of descriptive metadata of theses and dissertations. We proposed a novel content-based recommendation approach for suggesting academic supervisors using academic genealogy to model their profiles.

Unlike most of the related work, which helps for finding scientific literature, academic courses, or researchers for collaboration, our system recommends supervisors for thesis and dissertation proposals. Taking into account that choosing an adequate advisor can be a hard task, such system seems to be an important assisting tool.

Experiments were conducted using realdata from a repository containing more than 600 thousand theses and dissertations. The evaluation shows that our system was able to recommend a correct advisor, on average, in the third position of the suggested ranking.

In future works, we intend to integrate an academic genealogy tree viewer with the recommender system in a Web platform. Lastly, additional filters will be included, such as the location and the university acronym.

ACKNOWLEDGMENTS

This study was supported by CAPES Financial Code 001, PNP/DCAPES (464880/2019-00), CNPq (301618/2019-4), and FAPERGS (19/2551-0001279-9, 19/2551-0001660).

REFERENCES

- Alarfaj, F., Kruschwitz, U., Hunter, D., and Fox, C. (2012). Finding the right supervisor: Expert-finding in a university domain. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, NAACL HLT '12, page 1–6, USA. Association for Computational Linguistics.
- Beel, J., Langer, S., Genzmehr, M., and Nürnberger, A. (2013). Introducing docear's research paper recommender system. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, page 459–460, New York, NY, USA. Association for Computing Machinery.
- Bendakir, N. and Aïmeur, E. (2006). Using association rules for course recommendation. In Beck, J. E., Aïmeur, E., and Barnes, T., editors, *Proceedings of the AAAI Workshop on Educational Data Mining*, pages 1–10, Palo Alto, California, USA. Association for the Advancement of Artificial Intelligence.

- Bobadilla, J., Hernando, A., Ortega, F., and Gutiérrez, A. (2012). Collaborative filtering based on significances. *Information Sciences*, 185(1):1–17.
- Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, 46:109–132.
- Champiri, Z. D., Shahamiri, S. R., and Salim, S. S. B. (2015). A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications*, 42(3):1743 – 1758.
- Chen, R., Hua, Q., Chang, Y.-S., Wang, B., Zhang, L., and Kong, X. (2018). A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access*, 6:64301–64320.
- Desrosiers, C. and Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In *Recommender systems handbook*, pages 107–144. Springer, Boston, MA.
- Devarakonda, R., Palanisamy, G., Green, J. M., and Wilson, B. E. (2011). Data sharing and retrieval using oai-pmh. *Earth Science Informatics*, 4(1):1–5.
- Elbadrawy, A. and Karypis, G. (2016). Domain-aware grade prediction and top-n course recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 183–190, New York, NY, USA. Association for Computing Machinery.
- Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, page 257–260, New York, NY, USA. Association for Computing Machinery.
- Gormley, C. and Tong, Z. (2015). *Elasticsearch: the definitive guide: a distributed real-time search and analytics engine*. O'Reilly Media, Inc., Sebastopol, CA, USA.
- Han, E.-H. S. and Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *European conference on principles of data mining and knowledge discovery*, pages 424–431, Department of Computer Science / Army HPC Research Center University of Minnesota, Minneapolis. Springer.
- Hasan, M. A. and Schwartz, D. G. (2018). Recadvisor: Criteria-based ph.d. supervisor recommendation. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1325–1328, New York, NY, USA. Association for Computing Machinery.
- Husain, O., Salim, N., Alias, R. A., Abdelsalam, S., and Hassan, A. (2019). Expert finding systems: A systematic review. *Applied Sciences*, 9(20):4250.
- Karypis, G. (2001). Evaluation of item-based top-n recommendation algorithms. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 247–254, New York, NY, USA. ACM.
- Krulwich, B. (1997). Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI Magazine*, 18(2):37.
- Loper, E. and Bird, S. (2002). NLTK: the natural language toolkit. *CoRR*, cs.CL/0205028.
- Madeira, G., Borges, E. N., Lucca, G., Santos, H., and Dimuro, G. (2020). A tool for analyzing academic genealogy. In Filipe, J., Śmiałek, M., Brodsky, A., and Hammoudi, S., editors, *Enterprise Information Systems*, pages 443–456, Cham. Springer International Publishing.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press, Cambridge, England.
- Mendonça, F. C., Gasparini, I., Schroeder, R., and Kemczinski, A. (2020). Recommender systems based on scientific publications: A systematic mapping. In *Proceedings of the 22nd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pages 735–742, Setúbal, Portugal. INSTICC, SciTePress.
- O'Mahony, M. P. and Smyth, B. (2007). A recommender system for on-line course enrolment: An initial study. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, page 133–136, New York, NY, USA. Association for Computing Machinery.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.
- Ray, S. and Marakas, G. (2007). Selecting a doctoral dissertation supervisor: Analytical hierarchy approach to the multiple criteria problem. *International journal of doctoral studies*, 2(1):23–32.
- Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, Boston, MA.
- Rodrigues, M. W., Brandão, W. C., and Zárata, L. E. (2018). Recommending scientific collaboration from researchgate. In *7th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 336–341, New York, NY, USA. IEEE.
- Salter, J. and Antonopoulos, N. (2006). Cinemascreen recommender agent: combining collaborative and content-based filtering. *IEEE Intelligent Systems*, 21(1):35–41.
- Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw Hill Text, New York, NY, USA.
- Sugimoto, C. R. (2014). Academic genealogy. In *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, pages 365–380. MIT Press, Cambridge, MA, USA.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.