A Paper Recommendation System with ReaderBench: The Graphical Visualization of Semantically Related Papers and Concepts

Ionut Cristian Paraschiv, Mihai Dascalu, Philippe Dessus, Stefan Trausan-Matu, Danielle S. Mcnamara

A Paper Recommendation System with *ReaderBench*: The Graphical Visualization of Semantically Related Papers and Concepts

Ionut Cristian Paraschiv[1], Mihai Dascalu[1], Philippe Dessus[2], Stefan Trausan-Matu[1], Danielle S. McNamara[3]

[1] University Politehnica of Bucharest, Computer Science Department, Romania
ionut.paraschiv@cti.pub.ro, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro
[2] LSE, Univ. Grenoble Alpes, France
philippe.dessus@upmf-grenoble.fr
[3] LSI, Arizona State University, USA
dsmcnama@asu.edu

Abstract. The task of tagging papers with semantic metadata in order to analyze their relatedness represents a good foundation for a paper recommender system. The analysis from this paper extends from previous research in order to create a graph of papers from a specific domain with the purpose of determining each article's importance within the considered corpus of papers. Moreover, as non-latent representations are powerful when used in conjunction with latent ones, our system retrieves semantically close words, not present in the paper, in order to improve the retrieval of papers. Our previous analyses used the semantic representation of papers in different semantic models with the purpose of creating visual graphs based on the semantic relatedness links between the abstracts. The current analysis takes a step forward by proposing a model that can suggest which papers are of the highest relevance, share similar concepts, and are semantically related with the initial query. Our study is performed using paper abstracts in the field of information technology extracted from the Web of Science citation index. The research includes a use case and its corresponding results by using interactive and exploratory network graph representations.

Keywords: paper recommendation system, scientometrics, semantic similarity, discourse analysis.

1 Introduction

As more and more papers are being published, the need grows for creating a semantic repository with automatically tagged resources facilitating information access. Researchers and learners alike need to stay up-to-date and constantly search for new papers on certain topics as part of their daily activities. Since the daily retrieval of documents from the Internet leads to a data overflow, it is worthwhile to consider new approaches for a more comprehensive analysis of a database of articles.

adfa, p. 1, 2011.

We propose a model that begins with a corpus of paper abstracts that are automatically tagged and whose results are used for a semantic database for user defined queries. Once a user inputs a query in natural language text, a graphic visual representation of the query and all the related papers is displayed along with a list of related papers ordered by their level of similarity to the input text. In addition, a list of similar topics demonstrating high semantic overlap with the query is provided in order to stimulate the user in his/her research task.

In this paper, we begin by describing related studies that similarly discuss building network graphs for scientific papers. We then describe the methods used to implement the current system as well as a use case, which demonstrates the potential for our system. We conclude by describing possible future improvements.

2 Related Work

Mainstream database software based on keyword matching such as *Mendeley* or *DevonThink* can be considered as research paper recommendation systems, whereas more sophisticated approaches already exist [1]. Leaving aside the systems that rely on traditional information retrieval techniques [2], we expose two opposing approaches for analyzing the content of scientific papers: co-citation analysis and semantic analysis. *Co-citation analysis* [3] is a technique that uses citations between different papers to generate a network graph of all the articles from a domain. Two papers are connected within the graph if they share a common citation, while their corresponding links are weighted by the number of related citations. Different algorithms can be applied to the resulting graph in order to determine citation patterns, as well as central and important articles. There are two main advantages to this method: 1) it is very fast to process because the citations are created by the authors, and 2) it can infer the most important articles from a dataset. However, the method does not consider the semantic content of a paper; hence, the results can be misleading when considering that many citations for the same paper can refer to different parts of it, thus reducing the semantic relevance of each citation link [4]. Nevertheless, this method remains a benchmark for the analysis of articles within particular domains as it is widely used in scientometric analyses [5].

Latent Semantic Indexing [6] creates a semantic representation of words and concepts by establishing associations between terms that co-occur in similar contexts. Based on an initial training corpora consisting of texts collections, patterns are captured as relationships between terms and concepts contained in similar documents. Therefore, starting from a term-document matrix, a Singular Value Decomposition (SVD) is applied in order to reduce the dimensionality of the representation. Within the resulting vector space, semantic relatedness is measured through cosine similarity between the vector representations of both words and documents. We take this approach even further when semantic similarity is computed within our system as an aggregated cohesion score [7] based on Latent Semantic Analysis (LSA) [8] cosine similarity, Latent Dirichlet Allocation (LDA) [9] Jensen-Shannon dissimilarity of topic distributions, and WordNet semantic distances [10].

3 The Implemented Paper Recommendation System

Our aim here was to extend on the semantic views described in [11] and to create a paper recommender system that enables users to define queries in natural language and retrieve the most relevant papers. This extended model detects the most similar papers within the dataset, based on semantic cohesion, that resemble the input query [7]. In addition, we introduce the idea of generating highly cohesive concepts to the initial query by considering the most relevant keywords from other documents that have the highest semantic relatedness to the input text, a query expansion technique.

In terms of technical implementation, the paper recommendation system relies on *ReaderBench* [7, 12], an advanced text processing tool that has many components including a text processing module that creates a layered cohesion graph used as an underlying discourse structure. *ReaderBench* represents a good starting point as it already has a fully functioning natural language processing pipeline [7, 13] and multiple integrated semantic models covering LSA vector spaces [8], LDA topic distributions [9], and WordNet semantic distances [10], as well as specific Social Network Analysis tools and metrics used for visualization [14].

When a user enters a query in natural language, the input text undergoes the same pipeline as general documents: text preprocessing and cleaning, lemmatization, part of speech tagging, syntactic dependency analysis, and topics extraction [7, 13]. After the input query is represented as semantic vectors in LSA and LDA models, its semantic similarity with each document from the dataset is computed. The resulting cohesion scores are used as measures for creating the links within the interactive and explorative displayed graph (see Figure 1).

Furthermore, in order to stimulate the learner's creativity in terms of query generation, the system also includes a module that extracts the most important topics of semantically related documents from the dataset (not present in the initial query) and computes the semantic distances between them and the input text.

4 Use Case

In order to demonstrate the adequacy of the proposed methods, a query example and its corresponding results are described in this section. The database of documents used in the experiments consists of article abstracts published between the years 2000 and 2004 from the Web of Science citation index for the Education and Educational Research [15] domain. From all these abstracts, a subset of 500 papers containing one of the following keywords: "*IT*", "*technology*" or "*computer science*" was extracted. In the present example, a user inputs the text "*electronic learning and information technology*", with the intention of finding papers that are about informational systems within the educational sciences domain. Figure 1 depicts a sub-graph with the most semantically related articles for an imposed semantic similarity threshold of 60%.

Figure 1. Network graph of the semantically related documents to the input query.

The three most similar documents from the dataset are "*Trajectories and tensions in the theory of information and communication technology in education*", "*Information technology and education in the information age*" and "*An interactive dynamic model for integrating knowledge management methods and knowledge sharing technology in a traditional classroom*". In contrast to traditional information retrieval systems that focus on identifying the occurrences of each query concept within the retrieved documents and weighting their underlying frequencies, we rely on the similarity of the semantic contexts. This provides a more exploratory approach based on the overall similarity between the query and document representations within each semantic model.

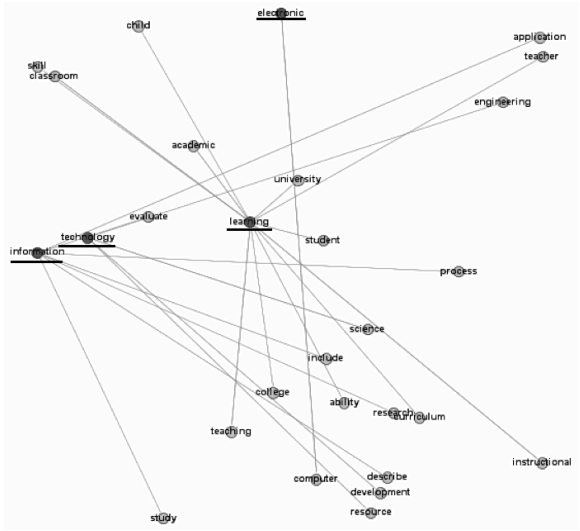

Figure 2. Concept map of query concepts and semantically related words.

The results of our use case scenario are in the same semantic context as the query and are good indicators that the methods used inside the recommender system can provide reliable and semantically relevant outputs. Moreover, the model recommends additional search terms for the current example such as “learner”, “teacher”, “science”, “curriculum”, “research”, “process”, and “development” (see Figure 2), which are clearly relevant inferred concepts for the query at hand.

In addition, we emphasize a major benefit of our method, derived from the use of semantic models for representing each document and the input query. The central concepts from the query become self-emergent as more text is presented and as the semantic context is more clearly specified. In contrast to traditional information retrieval systems in which the user needs to be specific while defining the keywords of the query, this system enables a refined search of semantically related documents and of similar concepts. Therefore, natural language queries describing the context in detail are encouraged in contrast to simple, keyword centered inputs.

5 Conclusions

While publications appear online at an increasing rate, our paper recommendation model can have a beneficial impact for anyone interested in the study of a specific subject or domain and can support the research communities in their endeavors. Moreover, users can further refine their searches by checking various related articles containing keywords that they may not have initially thought of. Therefore, through successive iterations using our recommendation system, the user can become more productive by exploring semantically related articles with diverse underlying concepts. In contrast to information retrieval systems centered on keywords identification, the integrated semantic representations provide a broader view of similar contexts, which in turn have the potential to stimulate creativity.

For future developments, we consider it appropriate to create a topics time-modeling system that generates a temporal view for the evolution of the most relevant paper concepts for a given timeframe and the articles' theme. As a drawback, the current model provides timely responses for hundreds or thousands of papers, but does not scale well with a large database of papers because an iterative search is performed throughout all potential documents. Therefore, further system performance enhancements are envisioned by considering clusters of similar papers as well as the implementation a hierarchical search.

Acknowledgements. The work presented in this paper was partially funded by the FP7 2008-212578 LTfLL project, by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398, as well as by the NSF grants 1417997 and 1418378 to Arizona State University. We also thank Pablo Jensen and Sebastian Grauwin for providing the initial corpus of paper abstracts, and we are grateful to Cecile Perret for her help in preparing this paper.

References

1. Joeran, B., Langer, S., Genzmehr, M., Gipp, B., Breitinger, C., Nürnberger, A.: Research Paper Recommender System Evaluation: A Quantitative Literature Survey. In: Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys'13). ACM, Hong Kong, China (2013)
2. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Vol. 1. Cambridge University Press, Cambridge, UK (2008)
3. Boyack, K.W., Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology, 61(12), 2389–2404 (2010)
4. Sword, H.: Stylish Academic Writing. Harvard University Press, Cambridge, Massachusetts & London, England (2012)
5. Gipp, B., Beel, J.r., Hentschel, C.: Scienstein: A Research Paper Recommender System. In: Int. Conf. on Emerging Trends in Computing (ICETiC'09), pp. 309–315, Virudhunagar, India (2009)
6. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., Beck, L.: Improving Information Retrieval with Latent Semantic Indexing. In: 51st Annual Meeting of the American Society for Information Science 25, pp. 36–40 (1988)
7. Dascalu, M.: Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence, Vol. 534. Springer, Switzerland (2014)
8. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104(2), 211–240 (1997)
9. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5), 993–1022 (2003)
10. Budanitsky, A., Hirst, G.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics, 32(1), 13–47 (2006)
11. Paraschiv, I.C., Dascalu, M., Trausan-Matu, S., Dessus, P.: Analyzing the Semantic Relatedness of Paper Abstracts - An Application to the Educational Research Field. In: 2nd Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2015), in conjunction with the 20th Int. Conf. on Control Systems and Computer Science (CSCS20), pp. 759–764. IEEE, Bucharest, Romania (2015)
12. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learners productions and strategies with ReaderBench. In: Peña-Ayala, A. (ed.) Educational Data Mining: Applications and Trends, pp. 335–377. Springer, Switzerland (2014)
13. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Pearson Prentice Hall, London (2008)
14. Bastian, M., Heymann, S., Jacomy, M.: Gephi: An open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media, pp. 361–362. AAAI Press, San Jose, CA (2009)
15. Grauwin, S., Jensen, P.: Mapping scientific institutions. Scientometrics, (2011)