Finding the Needle in a Haystack: Who are the Most Central Authors Within a Domain?

Ionut Cristian Paraschiv[1], Mihai Dascalu[1,2(✉)], Danielle S. McNamara[2], and Stefan Trausan-Matu[1]

[1] Computer Science Department, University Politehnica of Bucharest, Bucharest, Romania
ionut.paraschiv@cti.pub.ro,
{mihai.dascalu,stefan.trausan}@cs.pub.ro
[2] Institute for the Science of Teaching and Learning, Arizona State University, Tempe, USA
dsmcnama@asu.edu

Abstract. The speed at which new scientific papers are published has increased dramatically, while the process of tracking the most recent publications having a high impact has become more and more cumbersome. In order to support learners and researchers in retrieving relevant articles and identifying the most central researchers within a domain, we propose a novel 2-mode multilayered graph derived from Cohesion Network Analysis (CNA). The resulting extended CNA graph integrates both authors and papers, as well as three principal link types: co-authorship, co-citation, and semantic similarity among the contents of the papers. Our rankings do not rely on the number of published documents, but on their global impact based on links between authors, citations, and semantic relatedness to similar articles. As a preliminary validation, we have built a network based on the 2013 LAK dataset in order to reveal the most central authors within the emerging Learning Analytics domain.

Keywords: Learning analytics · 2-mode multilayered graph · Co-authorship · Co-citation · Semantic similarity

1 Introduction

With the growing flow of information and emerging new inter-disciplinary research topics, it is becoming increasingly difficult to find and follow relevant publications and authors. Each research sub-domain (e.g., Learning Analytics or Educational Data Mining) usually starts from a few authors who introduce broad research questions or trending topics around which a community gradually evolves. Usually, the initial authors become central members in the research network, being cited in new publications. The research question that arises regards how can we identify the most important authors and publications within a sub-domain, and what are the metrics that can be effectively applied in order to obtain a relevant global view of the underlying research? In our previous research studies [1, 2], we have built a learning analytics engine capable of annotating a dataset of articles using their semantic context, and displaying them within a network of papers that highlights their semantic relations.

K. Verbert et al. (Eds.): EC-TEL 2016, LNCS 9891, pp. 632–635, 2016.
DOI: 10.1007/978-3-319-45153-4_79

In addition, our work has made extensive use of Cohesion Network Analysis (CNA) [3], a cohesion centered representation of discourse in which semantic similarity links between different text segments are combined into a multi-layered cohesion graph [4]. This graph provides valuable insights of local cohesion expressed in the semantic relatedness between adjacent or transition sentences, meanwhile transcending towards global cohesion when evaluating inter-paragraph cohesion flow. Having this background, we propose a new approach, an extended CNA *2-mode multilayered graph*, capable of facilitating the identification of the most important authors and publications from a research domain by applying various Social Network Analysis (SNA) metrics [5]. As an initial validation, we have used the model to identify the top central authors and articles from the LAK (Learning Analytics and Knowledge) Dataset [6], which includes publications from the Learning Analytics domain (652 LAK and EDM conference papers, 45 journal papers, and 1214 distinct authors) in RDF format (https://www.w3.org/TR/REC-rdf-syntax/), containing unique URIs for all authors, articles and citations.

2 The Extended CNA 2-Mode Multilayered Graph

Our model combines three different approaches to evaluate the importance of both authors and articles within a domain: *Co-citation Analysis*, *Co-authorship Networks* and *Semantic Similarity*. These three types of links are used to build a *2-mode multilayered graph* on which graph theory measures [7] are applied to identify the most central nodes, (i.e. authors, papers) from the input dataset. The generated graph represents an integrated view of articles and authors, where each layer contains links with scores computed using different approaches. By jointly indexing the two different sets of nodes contained in our 2-mode graph, co-occurrence patterns emerge [8], suitable for generating an overview of the domain.

Co-authorship links [9] represent the first layer of our extended CNA graph in which two papers are related if they have at least one common author. Usually, the same author is interested in similar topics, so we can assume that papers with at least one common author are related. At the second layer, *co-citations* are enforced, having as roots one of the first techniques developed to annotate a dataset of articles [2]. The idea is that two papers are related if they contain at least one common citation, meaning that they should have semantic resemblance. The increase in the number of common citations between two articles usually denotes a higher degree of similarity and a tighter coupling among them. Third, the *semantic similarity* layer shifts the focus towards the actual content of the papers by evaluating the degree of their relatedness. Our integrated framework, *ReaderBench* [3, 4], integrates the automated building process of the CNA cohesion graph in which multiple semantic models are combined: (a) cosine similarity in Latent Semantic Analysis (LSA) vector spaces, (b) Jensen-Shannon dissimilarity between Latent Dirichlet Allocation (LDA) topic distributions, and (c) semantic distances (e.g., path length, Wu-Palmer, Leacock-Chodorow) in lexicalized ontologies – WordNet [4].In addition, we take the analysis further by applying SNA metrics [5] to identify patterns and meaningful relations between nodes, in conjunction with the evaluation of

each node's centrality. First, *degree centrality* quantifies the importance of each node as the sum of the scores of all links connected to it. Second, *closeness* reflects the centrality of each node as the average sum of all shortest paths between the current node and all other nodes in the graph; closeness can therefore be considered a measure of speed in terms of spreading the information within the network [10]. Third, *betweenness* evaluates the number of times a given node acts as a bridge along all shortest paths between pairs of any two other nodes. In contrast to closeness, betweenness can be perceived as a measure of control for the linkage among other nodes [10].

3 Exploring the LAK Dataset

Our CNA *2-mode multilayered graph* was applied to the 2013 LAK dataset [6] that contains machine readable information in which each resource (author, article or citation) is uniquely identified. Table 1 depicts the top 10 authors in terms of betweenness centrality. The top 5 authors are *"Ryan Baker", "Neil Heffernan", "Joseph Beck", "Kenneth Koedinger"* and *"Jack Mostow"*, authors with a high impact in the broader Computer Education domain, as well as the Learning Analytics domain, having a total of 102 unique published papers and more than 33,000 collective citations according to Google Scholar. The top ten authors collectively reach more than 80,000 citations and 141 unique papers in the dataset. Of particular interest is "Jose Gonzales-Brenes" who does not have many citations ($n = 125$), but is a co-author in 5 out of 8 papers with "*Jack Mostow*" (ranked 5) and in one with "*Peter Brusilovsky*" (ranked only 25 in this data set, but with more than 20,000 citations worldwide). Therefore, Gonzales-Brenes is tightly connected to two highly influential researchers and creates a bridge between the two research communities.

Table 1. Top 10 authors from Learning Analytics ordered by their betweenness centrality.

Author	M1	M2	M3	P	CC	NP
Ryan Baker	43,191	0.9	2,817	36	5,968	2
Neil Heffernan	23,823	0.8	2,317	25	3,645	2
Joseph Beck	18,906	0.8	2,110	18	2,958	1
Kenneth Koedinger	17,938	0.8	2,274	23	17,317	1
Jack Mostow	15,689	0.8	1,943	16	3,773	0
Arthur Graesser	14,573	0.7	1,788	16	34,539	1
Zachary Pardos	12,920	0.8	2,149	13	857	0
Jose Gonzalez-Brenes	12,448	0.8	1,848	8	125	0
Sebastian Ventura	11,200	0.8	1,832	14	6,035	0
Cristobal Romero	10,312	0.8	1,810	15	5,077	0

* SNA Metrics: *M1* = Betweenness centrality; *M2* = Closeness centrality; *M3* = Degree; *P* = Number of published articles; *CC* = Citation count; *NP* = Number of papers from Top 10.

4 Conclusions

In this paper, we have introduced a *2-mode multilayered graph*, an extension of our Cohesion Network Analysis, which a represents a combination of multiple complementary perspectives in order to build a mixed *article-author* graph. In the context of hundreds, even thousands of publications within each research field every year, our approach provides valuable support in retrieving relevant resources, helping learners to *find the needle in the haystack*.

Our method can be further extended with additional SNA metrics, enhanced visualization tools, and the ability to check the evolution of a domain. Currently, the views are highly cluttered because of the large number of nodes - a potential solution would presume the creation of hierarchical clusters that group similar nodes. With such future modifications, we expect the CNA 2-mode multilayered approach to have a significant impact on information retrieval.

Acknowledgement. This work is partially funded by the 644187 H2020 RAGE (Realising an Applied Gaming Eco-System) http://www.rageproject.eu/project.

References

1. Paraschiv, I.C., Dascalu, M., Dessus, P., Trausan-Matu, S., McNamara, D.S.: A paper recommendation system with readerbench: the graphical visualization of semantically related papers and concepts. In: Li, Y., et al. (eds.) State-of-the-Art and Future Directions of Smart Learning. LNET, pp. 443–449. Springer, Germany (2015)
2. 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: DS-CSCL-2015/CSCS20, pp. 759–764. IEEE, Bucharest (2015)
3. Dascalu, M., Trausan-Matu, S., McNamara, D.S., Dessus, P.: ReaderBench – automated evaluation of collaboration based on cohesion and dialogism. Int. J. Comput. Supported Collaborative Learn. **10**(4), 395–423 (2015)
4. Dascalu, M.: Analyzing discourse and text complexity for learning and collaborating. Studies in Computational Intelligence, vol. 534. Springer, Cham (2014)
5. Scott, J.: Social Network Analysis. SAGE Publications Ltd., Thousand Oaks (2012)
6. Arora, R., Ravindran, B.: Latent Dirichlet Allocation based multi-document summarization. In: 2nd Workshop on Analytics for Noisy Unstructured Text Data, pp. 91–97. ACM, Singapore (2008)
7. Biggs, N., Lloyd, E., Wilson, R.: Graph Theory, 1736-1936. Oxford University Press, Oxford (1986)
8. Borgatti, S.: 2-mode concepts in social network analysis. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and System Science, pp. 8279–8291. Springer, New York (2009)
9. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. In: Mapping Knowledge Domains. Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering, Irvine (2003)
10. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Soc. Netw. **27**, 39–54 (2005)