Modelling the User Modelling Community (and Other Communities as Well)

Dario De Nart(✉), Dante Degl'Innocenti, Andrea Pavan, Marco Basaldella, and Carlo Tasso

Artificial Intelligence Lab Department of Mathematics and Computer Science, University of Udine, Udine, Italy
{dario.denart,carlo.tasso}@uniud.it,
{dante.deglinnocenti,pavan.andrea.1,basaldella.marco.1}@spes.uniud.it

Abstract. Discovering and modelling research communities' activities is a task that can lead to a more effective scientific process and support the development of new technologies. Journals and conferences already offer an implicit clusterization of researchers and research topics, and social analysis techniques based on co-authorship relations can highlight hidden relationships among researchers, however, little work has been done on the actual content of publications. We claim that a content-based analysis on the full text of accepted papers may lead to a better modelling and understanding of communities' activities and their emerging trends. In this work we present an extensive case study of research community modelling based upon the analysis of over 450 events and 7000 papers.

1 Introduction

Tracking the activity of research communities and discovering trends in research activity is a complex task which can produce great benefits for future research and is essential for research evaluation. Most commonly used evaluation techniques rely on Social Network Analysis (herein SNA), more specifically on the analysis of the co-authorship relation between scholars. Such methods can lead us to interesting insights about existing research communities and their evolution over time; however we believe that a more semantic approach can lead us to even more interesting insights about the actual topics dealt by a community, the emerging research themes and buzzwords, and also the existence of complementary research communities. Since SNA does not take into account the actual content of papers (which is the reason behind a collaboration) it does not allow discovering communities that deal with the same topics but do not know each other yet. On the other hand, knowing the content of research papers can lead to such information. Manually reading and understanding all the literature produced by a community such as the Computer Science one is simply not feasible due to huge amount of time and resources required. On the other hand, there exist automatic knowledge extraction tools able to extract meaningful concepts from unstructured text with enough precision for this purpose [3]. We claim

F. Ricci et al. (Eds.): UMAP 2015, LNCS 9146, pp. 357–363, 2015.
DOI: 10.1007/978-3-319-20267-9_31

that their combined usage with traditional SNA techniques can better model the research community, achieving a better description of its sub-communities and their relationships. In this paper we present in Sec. 3 an innovative approach towards modelling and discovery of research communities based on the combination of SNA and content-based analysis on the accepted contributions. Our approach is then used in Sec. 4 to analyse over 450 events and 7000 papers spanned over three years of activity of the Computer Science and ICT community. Due to copyright and data access restrictions, the analysis is limited to the proceedings published by *ceur-ws*[1]. Ceur-ws provides open access to a large number of Workshop, Poster session, and conference proceedings of events held all over the world, but mostly in Europe.

2 Related Work

The study of the connections between people and groups has a long research tradition of at least 50 years [1]. SNA is a highly interdisciplinary field whose traditional approach consists in selecting a small sample of the community and to interview the members of such sample. This approach has proved to work well in self contained communities such as business communities, academic communities, ethnic and religious communities and so forth [8]. However the increasing digital availability of big data allows to use all the community data and the relations among them. A notable example is the network of movie actors [12], that contains nearly half a million professionals and their co-working relationship [9]. Academic communities are a particularly interesting case due to the presence of *co-authorship* relations between their members. Several authors in the literature have analysed the connections between scholars by means of co-authorship: in [8] [9] a collection of papers coming from Physics, Biomedical Research, and Computer Science communities are taken into account in order to investigate cooperation among authors; in [1] a data set consisting of papers published on relevant journals in Mathematics and Neuroscience in an eight-year period are considered to identify the dynamic and the structural mechanisms underlying the evolution of those communities. *VIVO* [7] is a project of Cornell University that exploits a Semantic Web-based network of institutional databases to enable cooperation between researchers and their activities. In [6] the problem of content-based social network discovery among people who appear in *Google News* is studied: probabilistic Latent Semantic Analysis [5] and clustering techniques have been exploited to obtain a topic-based representation. The authors claim that the relevant topic discussed by the community can be discovered as well as the roles and the authorities within the community. The authors of [10] perform deep text analysis over the Usenet corpus. Finally the authors of [11] introduce a complex system for content-based social analysis involving NLP techniques which bears strong similarities with our work. The deep linguistic analysis is performed in three steps: (i) concept extraction (ii) topic detection

[1] http://ceur-ws.org/

using semantic similarity between concepts, and (iii) SNA to detect the evolution of collaboration content over time. However the approach relies on a domain ontology and therefore cannot be applied to other cases without extensive knowledge engineering work, whereas our work relies a domain-independent approach.

3 Proposed Method

In order to support our claims a testbed system was developed to provide access to CEUR volumes, integrate the concept extraction system presented in [2], and aggregate and visualize data for inspection and analysis. The SNA part of our study is performed in the following way: all contributing authors are considered as features of an event and form a vector-space model in which it is possible to estimate the similarity between events by means of *cosine similarity*. Non-zero similarity values are then considered to create, for each event, a neighborhood of similar events; such connections are then represented as a graph (*Author Graph* - AG). Communities of similar events in the AG are then identified with the Girvan-Newman clustering algorithm [4] which allows to cluster events corresponding to well connected communities. The Content-based part of the study, instead, is performed in a novel way: the usage of an automatic knowledge extraction tool allows us to finely model the topics actually discussed in a conference and to group events according to semantic similarities. Such topics are represented by means of *keyphrases* (herein KPs), i.e. short phrases of 1 to 4 words that represent a particular concept. Our knowledge extraction tool associates to each relevant KP a score named *keyphraseness* which is intended as an estimation of the actual relevance of a concept inside a long text such as a scholarly paper. By extracting a sufficient number of relevant KPs we can obtain a detailed representation of the main topics of a paper as well as the relevant entities therein mentioned, where more relevant concepts are associated to higher keyphraseness. For each CEUR event, all its papers are processed creating a pool of *event keyphrases*, where each KP is associated to the *Cumulative Keyphraseness* (*CK*) i.e. sum of all the keyphraseness values in the corresponding papers. By doing so a topic mentioned in few papers, but with an high estimated relevance (keyphraseness), may achieve a higher CK than another one mentioned many times but with a low average estimated relevance. For each KP an *Inverse Document Frequency* (*IDF*) index is then computed on event basis, namely we compute the logarithm of the number of considered events divided by the events in which the considered KP appears, as broadly used in Information Retrieval. Then, for each KP in each event, a CK-IDF score is computed by multiplying the IDF with the corresponding CK. This measure behaves in a manner that closely resembles the well known TF-IDF measure; however there is a substantial difference: the CK part of the formula takes into account features more complex than just term frequency[2]. All extracted topics, with their related CK-IDF values, are considered as features of an event and form a vector-space model in which is possible to estimate the semantic similarity between events by means of *cosine similarity*. Due to the high number of non-zero values, only the highest 10% of these similarity values is considered. Within such

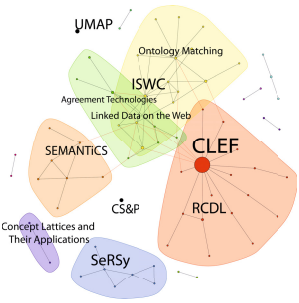

Fig. 1. 2012 AG

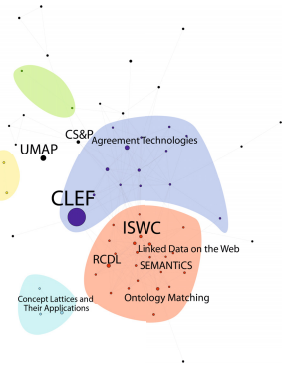

Fig. 2. 2012 TG

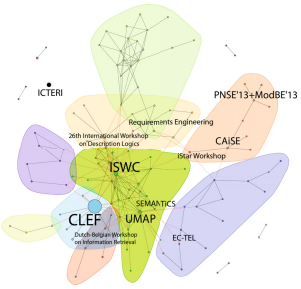

Fig. 3. 2013 AG

a value range, for each event a neighborhood of significantly similar events is identified; such connections are then represented as a graph named *Topic Graph* (herein TG). Communities of similar events in the graph are then identified, as in the previous scenario, with the Girvan-Newman clustering algorithm. Finally, complementary communities analysis is performed by comparing the AG and the TG: events that are connected in the TG and have no direct connections in the AG are potentially complementary communities. To detect such situations a simple metric called *Complementarity*, evaluated as the difference between topic similarity and author similarity, is proposed. Positive values suggest that the considered events bear a strong topic similarity and a low author similarity, meaning that, even though the topics discussed are similar, the contributing authors have little or no overlap.

4 Results

In this section we present and compare the results of our analysis on the last three years of CEUR volumes. Only proceedings available in CEUR were considered, therefore, for conferences such as UMAP, ISWC, and CAiSE we are considering only the part of their proceedings published on CEUR (usually poster and demo sessions, workshops, and doctoral consortia). Fig. 1 and Fig. 2 represent our model of the 2012 events whose proceedings were published on CEUR. The clusters obtained in the AG and in the TG are notably different: in the AG there are several clusters, while in the TG most of the events belong to two large clusters with one of them clearly including all Semantic-Web related events. Fig. 3 and Fig. 4 represent our model of the 2013 events. Again, the AG presents more clusters than the TG, however the four clusters found identify different groups of events: compared to Fig. 4 the upper one includes Data Science related events, the lower one Software Engineering related events, the right one E-Learning related events, and the left one theoretical Computer Science related

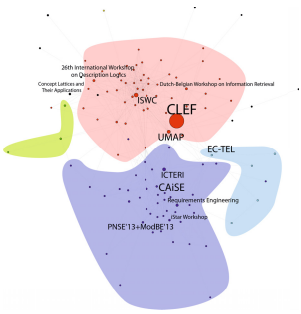

Fig. 4. 2013 TG

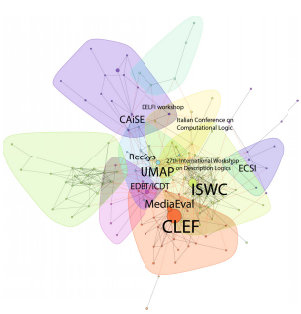

Fig. 5. 2014 AG

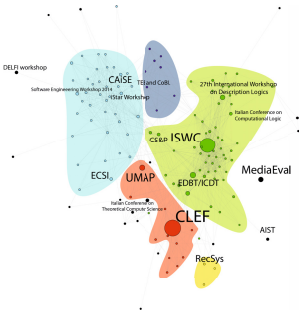

Fig. 6. 2014 TG

events. Finally, Fig. 5 and Fig. 6 represent our model of the 2014 events. Due to the large number of 2014 events published on CEUR it was possible to identify more clusters, however, there are still more clusters in the AG than in the TG. In both graphs is clearly recognizable a large Semantic-Web related cluster, which in the TG includes also theoretic Computer Science events. The comparative analysis of the 2012, 2013, and 2014 models of CEUR events can provide insights on the evolution of the involved research communities. For instance, we can observe how CLEF and ISWC always attract different authors communities, even though there has been some topic overlap in the history of such events; moreover, we can observe how UMAP, in the considered period, was thematically closer to CLEF than ISWC, but it attracted part of the ISWC community as well. Another interesting insight about what research communities actually debate can be obtained by looking at the extracted concepts with the lowest IDF, which means the most widely used in the considered data set (listed in Table 1). The term "Semantic Web" appears in all the three considered years of CEUR proceedings on a large part of the published papers (spanning from 20% to 35% of the considered proceedings) which is far larger than the part covered by the identified Semantic Web event cluster. Finally, the complementary communities analysis highlights how every considered event has at least a potentially complementary event. Since listing all the pairs of potentially complementary events would require too much space, we are only reporting, in Table 2 the potentially complementary events for the UMAP community held in 2014.

5 Conclusions and Future Work

In this paper we presented a new approach towards scientific communities modelling based on a twofold view with the aim of discovering shared interests, spotting research communities and, hopefully, help scientist to address the problem of finding the right venue for their work. The ability of identifying potentially complementary communities is, in our opinion, the most notable feature of our approach: traditional SNA can detect existing communities, but is unlikely to

Table 1. Most widespread buzzwords and their frequency in the corpus

2012		2013		2014	
buzzword	frequency	buzzword	frequency	buzzword	frequency
system	0.30	system	0.521	system	0.662
data	0.291	data	0.416	model	0.607
computer science	0.261	model	0.385	data	0.576
model	0.246	computer science	0.378	information	0.533
information	0.231	information	0.335	computer science	0.478
ontology	0.201	Semantic Web	0.248	Semantic Web	0.355
knowledge	0.201	ontology	0.242	research	0.294
Semantic Web	0.201	Natural Language Processing	0.192	language	0.294
Natural Language Processing	0.134	Software	0.186	Natural Language Processing	0.282

Table 2. Most complementary events to UMAP 2014

Event	**Complementarity score**
the Workshops held at Educational Data Mining 2014	0.267
the Workshops of the EDBT ICDT 2014 Joint Conference	0.238
the 16th LWA Workshops KDML IR and FGWM	0.235
Workshop on Semantic Matching in Information Retrieval	0.223

identify communities that should talk each other, meet or join. On the other hand, our approach exploits state of the art knowledge extraction techniques to investigate the topics actually dealt by a community and can easily identify communities that deal with the same topics, but have little or no social overlap at all. Our future work includes extending our models of the research community activity with more data coming from different sources than CEUR, as well as employing our modelling techniques to other domains and with different goals. Finally, we believe that the presented modelling technique can be exploited to provide personalized services, for instance scientific papers or conference recommender systems.

References

1. Barabsi, A., Jeong, H., Nda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications **311**(34), 590–614 (2002)
2. Degl'Innocenti, D., De Nart, D., Tasso, C.: A new multi-lingual knowledge-base approach to keyphrase extraction for the italian language. In: Proc. of the 6th Int.l Conf. on Knowledge Discovery and Information Retrieval. SciTePress (2014)
3. Gangemi, A.: A Comparison of Knowledge Extraction Tools for the Semantic Web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013)
4. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99**(12), 7821–7826 (2002)
5. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval, SIGIR 1999, pp. 50–57. ACM, New York (1999)

6. Joshi, D., Gatica-Perez, D.: Discovering groups of people in google news. In: Proceedings of the 1st ACM International Workshop on Human-Centered Multimedia, pp. 55–64. ACM (2006)
7. Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J., et al.: Vivo: Enabling national networking of scientists. In: Proceedings of the Web Science Conference 2010, pp. 1310–1313 (2010)
8. Newman, M.: Scientific collaboration networks. network construction and fundamental results. Phys. Rev. E **64**, 016131 (2001)
9. Newman, M.: The structure of scientific collaboration networks. Proc. of the National Academy of Sciences 98(2), 404–409 (2001)
10. Sack, W.: Conversation map: a content-based usenet newsgroup browser. In: From Usenet to CoWebs, pp. 92–109. Springer (2003)
11. Velardi, P., Navigli, R., Cucchiarelli, A., D'Antonio, F.: A new content-based model for social network analysis. In: ICSC, pp. 18–25. IEEE Computer Society (2008)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of "small-world" networks. Nature **393**(6684), 440–442 (1998)