A Content-Based Approach to Social Network Analysis: A Case Study on Research Communities

Dario De Nart[1], Dante Degl'Innocenti[1(✉)], Marco Basaldella[1], Maristella Agosti[2], and Carlo Tasso[1]

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

[2] University of Padua, Padua, Italy
agosti@dei.unipd.it

Abstract. Several works in literature investigated the activities of research communities using big data analysis, but the large majority of them focuses on papers and co-authorship relations, ignoring that most of the scientific literature available is already clustered into journals and conferences with a well defined domain of interest. We are interested in bringing out underlying implicit relationships among such containers and more specifically we are focusing on conferences and workshop proceedings available in open access and we exploit a semantic/conceptual analysis of the full free text content of each paper. We claim that such content-based analysis may lead us to a better understanding of the research communities' activities and their emerging trends. In this work we present a novel method for research communities activity analysis, based on the combination of the results of a Social Network Analysis phase and a Content-Based one. The major innovative contribution of this work is the usage of knowledge-based techniques to meaningfully extract from each of the considered papers the main topics discussed by its authors.

Keywords: Content-based · Social network analysis · Social semantic · Research communities · Text processing · Clustering · Scientific publishing

1 Introduction

Finding a suitable venue for presenting a research project is a critical task in the research activity, especially in a research community such as Computer Science, where there are several established conferences with very low acceptance rates. Conference venues typically aggregate researchers from a specific community (e.g.: Semantic Web, Digital Libraries, User Modelling, etc.) interested in

D. Calvanese et al. (Eds.): IRCDL 2015, CCIS 612, pp. 142–154, 2016.
DOI: 10.1007/978-3-319-41938-1_15

discussing their results, however it is hard for young researchers to identify the right venue to introduce their work, as well for experienced researcher to find new venues and communities that might be interested in their projects/results.

Social Network Analysis [20] (herein SNA) based on co-authorship can produce interesting insights on the activities of a research community, even if it does not take into account the actual content produced by the community. In the next section we illustrate how only few research works have explored the real contents of research papers in order to analyse trends emerging inside a scientific community, mostly because of the difficulties in gaining access to the full text of papers and to the complexity of Natural Language Processing (herein NLP) techniques required to extract meaningful concepts from unstructured text.

In this paper we propose a new approach to analyse the semantic and social relationship among scientific conferences, in order to discover shared topics, competences, trends, and other implicit relationships. More specifically we have experimented the proposed approach on two data sets: CEUR conference and workshop proceedings published from January 1st 2014 to December 1st of the same year and the proceedings of ten editions of the Italian Research Conference on Digital Libraries (herein IRCDL) from 2005 to 2014. CEUR[1] is a website that provides open access to a large number of Workshop and conference proceedings of events held all over the world, but mostly in Europe. Such resource is extremely valuable in order to gain a global view of the current interactions among different research communities. CEUR offers information about the conferences, the co-located events, and the contributing authors; such data can be used to perform analysis based upon author contribution and to group conferences according to their location and participating authors. On the other hand, ten editions of IRCDL proceedings represent a considerable amount of peer reviewed literature generated by a cohesive community over a relatively long span of time, allowing the identification of research trends over time.

The work presented in this paper presents two case studies of social and content-based analysis over a research community: the grouping of CEUR volumes according to contributing authors and topics covered and the analysis of topics dealt by the IRCDL community over ten years. We claim that both social and semantic analysis [3] can provide meaningful insights on the activity of scientific communities such as the ones publishing their proceedings on CEUR. On the social side, we are employing established techniques to group events according to the authors involved, while on the semantic side, we take advantage of advanced NLP techniques and tools that we have developed over the years ([5,16]) for analyzing the textual content of each article in each volume and to group events according to their shared topics.

The rest of the paper is organized as follows: in Sect. 2 we briefly introduce some related work, in Sect. 3 we present our original approach, in Sect. 4 the results of our analysis are discussed, and Sect. 5 concludes the paper and presents some planned future work.

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

2 Related Work

The study of the connections between people and groups has a long research tradition of at least 50 years [2,18,20,21]. Moreover, SNA is an highly interdisciplinary field involving sociology, psychology, mathematics, computer science, epidemiology, etc. [15] Traditional social networks studies have been performed in many fields. The traditional approach towards SNA 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 [12]. 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 [1,22], that contains nearly half a million professionals and their co-working relationship [14].

Academic communities are a particularly interesting case due to the presence of *co-authorship* relations between their members. Several authors in literature have analysed the connections between scholars by means of co-authorship: in [12–14] 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 [2] 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. Finally, the authors of [15] consider in their analysis the specific case of the SNA research community.

VIVO [9] is a project of Cornell University that exploits a Semantic Web-based network of institutional databases to enable cooperation between researchers and their activities. The system however is quite "ad-hoc", since it relies on a specific ontology and there is no automatic way to annotate the products of research with semantic information, requiring in such a way a huge preliminary effort to prepare the data. Another SNA tool that is used in the academic field is *Flink* [11]. The system performs the extraction, aggregation, and visualization of on-line social networks and it has been exploited to generate a Web-based representation of the Semantic Web community. In [8] the problem of content-based social network discovery among people who appear in *Google News* is studied: probabilistic Latent Semantic Analysis [7] and clustering techniques have been exploited to obtain a topic-based representation. Another system that exploits the full text of email messages between scholars is presented in [10]. 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 [17] perform deep text analysis over the Usenet corpus. However their tool is an exploratory system that serves for visualization purposes only. Finally the authors of [19] 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 using semantic similarity between concepts, and (iii) SNA to detect the evolution of cooperation 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 the work presented in this paper relies for content-based analysis on a knowledge-based domain-independent approach. Moreover our experiment has been performed on a much larger scale considering over 2100 research papers.

3 Proposed Methodology

In order to support our analysis a testbed system was developed to access documents, integrate the keyphrase extraction system presented in [5], and aggregate and visualize data with purposes of inspection and analysis. Our approach is twofold: we take into account social connections between events, considering the authors who contributed, and the semantic connections, analysing the topics discussed. These two different perspectives are then used to get a better overall picture of the considered research community.

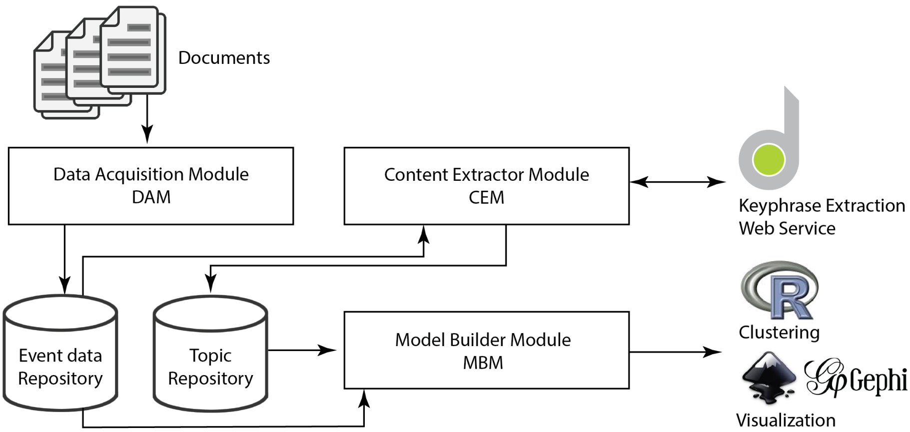

Fig. 1. System architecture overview.

The testbed system is constituted by three modules: the Data Acquisition (DA) module, the Content Extraction (CE) module, and the Graph Builder (GB) module, as shown in Fig. 1. The DA module reads the considered documents and populates the Event Data repository, that contains the list of considered events and their related data including contributing authors, venue, date, and links to full text papers. The CE module retrieves the full text of each considered paper and acts as interface for a Keyphrase Extraction system. Such system extracts a set of meaningful keyphrases (KPs) from each article's full text using the algorithm described in [5]. Keyphrases identify relevant concepts in the document and each of them is associated with an estimated relevance score called *keyphraseness*. Keyphraseness is evaluated using a knowledge-based approach that exploits different kinds of knowledge: Statistical Knowledge, Linguistic Knowledge, Meta/Structural Knowledge, and Semantic/Social Knowledge [4]. Keyphraseness therefore can be considered a fine estimation of the real relevance of a phrase inside a long text such as a scholarly paper. Associations between KPs and papers are then stored in the Keyphrase Repository.

The GB module, finally, handles the creation of the network models: the SNA-based one and the Content-based one. Clustering and Visualization are handled by external tools such as R and Gephi.

The SNA part of our study is performed by exploiting established and well known methods: an *Author Graph* (AG) is built where events are nodes and the fact that two events share some authors is represented by an undirected link between the corresponding nodes. Nodes are weighted according to the number of authors involved in the corresponding event, links are weighted proportionally to the number of authors shared. Communities of similar events in the graph are then identified applying the Girvan-Newman clustering algorithm [6] 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 automatic KP extraction allows us to model the topics actually discussed in a conference and to group events according to semantic similarities. For each considered event, all the accepted papers are processed creating a pool of *event keyphrases*, where each keyphrase is associated to the *Cumulative Keyphraseness* (CK) i.e. sum of the related keyphraseness values in the considered documents, as shown in Formula (1).

$$CK(k, event) = \sum_{paper \in event} Keyphraseness(k, paper) \tag{1}$$

By doing so a topic mentioned in few papers, but with an high estimated relevance, may achieve an higher CK than another one mentioned many times but with a low average estimated relevance. For each keyphrase an *Inverse Document Frequency* (IDF) index is then computed on event basis, namely we compute the logarithm of the number of events considered divided by the events in which the considered keyphrase appears, as shown in Formula (2).

$$IDF(k) = \log \frac{|AllEvents|}{|EventsContainingKPk|} \tag{2}$$

Intuitively, the larger the IDF, the least events are charcterized by the considered keyphrase. When a keyphrase is relevant in all the considered events, its IDF is zero. Such value is then combined with the CK, as shown in Formula (3) to create, for each KP in each event a $CK - IDF$ score.

$$CK - IDF(k) = CK(k) * IDF(k) \tag{3}$$

The $CK - IDF$ score promotes keyphrases that are relevant within an event and, at the same time, not widely used throughout the whole set of considered events. 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 mere term frequency. Subsequently, a *Topic Graph* (TG) is built, where events are represented by nodes and the fact that two events share some keyphrases is represented by an undirected link between such nodes. Nodes are weighted according to the number of different keyphrases extracted from their papers, and links according to the sum of $CK - IDF$ values of the keyphrases shared between two events. Communities of similar events in the graph are then identified, as in the previous scenario, with the Girvan-Newman clustering algorithm.

Both the social-based and the content-based graphs are then exported in different formats to allow visual inspection of the obtained graphs and clusters.

4 Results

In this section we present two case studies on research community analysis. In the first part of the section we present the analysis performed on the CEUR events published in 2014. The goal of such an analysis is to detect clusters of events that represent the meeting points of a specific research community (e.g. the Semantic Web community, the Recommender System one, the Digital Libraries one, and so on) and to identify groups of events dealing with similar or complementary topics. Once research communities are identified it is possible to further investigate their activities by analysing the evolution of the topics dealt with in the published papers. In the second part of this section we outline the methodology used to detect trending topics and provide examples built upon the second considered data set which includes the proceedings of ten IRCDL editions.

4.1 CEUR Proceedings Analysis

The first case study is based upon 2014 CEUR volumes, upon which both social and semantic analysis are performed, thus generating both an AG and a TG. The considered data set contains all CEUR volumes published before December the 1st 2014 that are proceedings of events held during 2014; it consists in 135 events with over 8400 contributing authors and over 2000 accepted papers.

To get an overview of both the AG and the TG, we are considering five features: the number of edges, the average degree, that is the average number of outgoing edges for each node, the network diameter, that is the longest path in the graph, the graph density, that is a measure of how well connected the graph is, spanning between 0 (all isolated nodes) and 1 (perfectly connected graph), and the average path length, that is the average length of a path connecting two distinct nodes. The number of nodes is omitted because we are assuming that each event is represented by a node and therefore their count is 135 in both cases.

At first glance the AG presents a sparse network structure, with a very low density as shown in Table 1, with a few isolated nodes, meaning that relatively

Table 1. Author Graph global statistics

# of edges	Average degree	Network diameter	Graph density	Average Path length
405	6	8	0.045	3.078

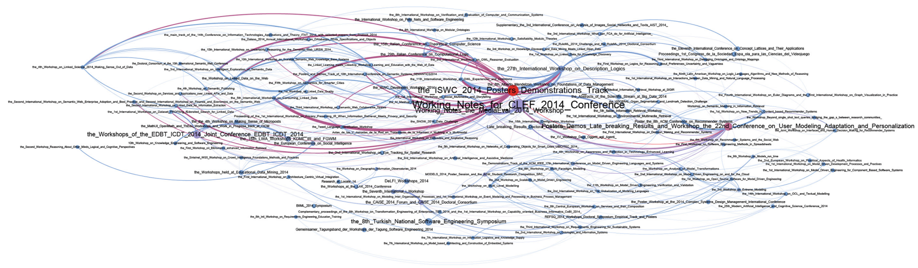

Fig. 2. Overview of the Author Graph.

few authors contribute to more than one conference and some events do not share authors with the others.

Figure 2 shows a visualization of the AG in which the size of the nodes is proportional to the number of authors who contributed to the event, and the colour depends on the *betweenness centrality* of the node (namely the number of shortest paths containing that node); edge size is proportional to the number of authors who contributed to both the events connected by the edge and edge color depends on the betweenness centrality. Nodes and edges with a high centrality are red, while low centrality ones are blue. The centrality value allows to identify the events that serve as hubs for different communities: events with a high centrality, in fact, might be interdisciplinary meetings where members of otherwise distinct communities get together. On the other hand, events with a low centrality might be more focused and therefore interested only for the members of a single community.

It can be noticed how the largest event in term of contributing authors (CLEF 2014) is not the most central one from a network perspective (which is the ISWC 2014 Poster and Demo Session), few events have an high centrality and some of them are relatively small in terms of number of contributing authors (such as the Workshops, Poster, and Demo Session of UMAP 2014), and, finally some large events in terms of contributing authors have an extremely low centrality (such as the Turkish Software Engineering Symposium or the International Workshop on Description Logics), meaning that they serve as the meeting point of a relatively closed community rather than a point of aggregation for diverse research areas.

In order to identify groups of events representing meeting points of wide research communities, a clustering step is performed, removing edges with an high betweenness centrality value. By doing so only groups of strongly interconnected events remain connected. The result of the clustering step is shown in Fig. 3, where all the isolated nodes are omitted.

Fig. 3. The three main clusters in the Author Graph.

Three clusters can be observed: the first and largest one groups, with little surprise, the ISWC 2014 Poster and Demo Session which is clearly a massively aggregating event, with all its co-located events and other Semantic Web related events as well; the other two clusters are much smaller and revolve around CLEF 2014 and the Workshops, Poster, and Demo Session of UMAP 2014. However, due to the sparsity of the graph, most of the events cannot be clearly clustered and therefore other kinds of correlations between events should be considered to get a better picture.

The TG, on the other hand is, as shown in Table 2 much more dense with a graph density of 0.94 and a diameter of 2. These data highlight how the papers presented at the considered events share a common lexicon, which is an expected result, since CEUR publishes only computer science proceedings.

Table 2. Topic Graph global statistics

# of edges	Average degree	Network diameter	Graph density	Average Path length
8543	126.56	2	0.94	1.041

The generated TG is therefore extremely well connected and, considered as-is, it does not provide useful insights.

After pruning low-weight edges, representing the sharing of low $CK - IDF$ terms between two events, and application of the Girvan-Newman clustering technique we obtain the clusters shown in Fig. 4 which are significantly different from the ones obtained by analyzing the AG. There is an higher number of clusters and, even though many events remain isolated, more events are grouped in a cluster. The largest cluster includes two of the most central events, namely CLEF and UMAP, meaning that, although merging different communities, they deal with similar or tightly related topics. ISWC, the most central event in the AG, however, in the TG is included in a relatively small cluster in which only few of its co-located events appear. The majority of the events that are included in the ISWC cluster in the AG are, indeed, in the TG included in the UMAP/CLEF cluster or form a cluster on their own, like the ISWC Developers' Workshop and

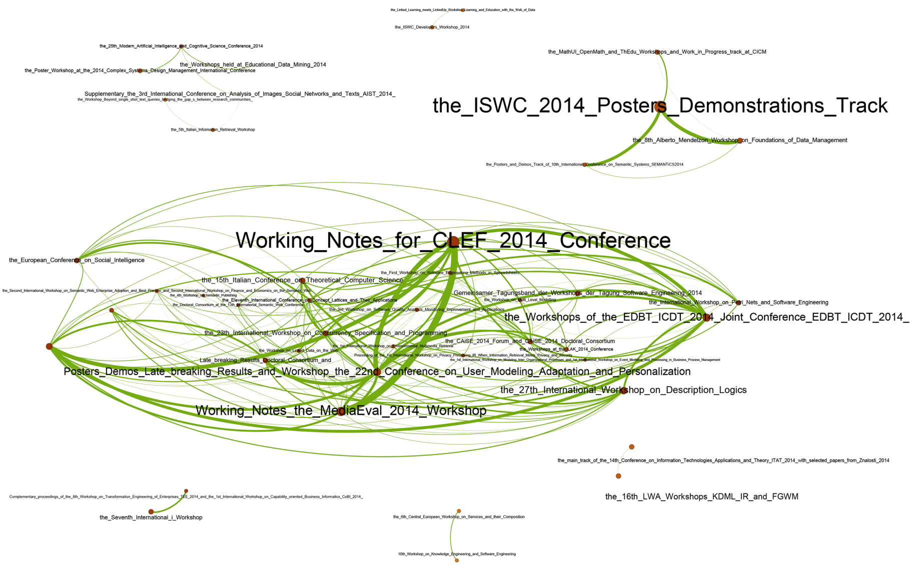

Fig. 4. Clusters obtained from the Topic Graph.

the LinkedUp Challenge. Several other small clusters are present, representing topics discussed only by a handful of events.

One final 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. They are listed in Table 3. Since we used the logarithm to the base 2, an IDF of 1 means that the considered concept is relevant in half of the considered conferences, and with an IDF of 0.5 in about 2/3. Even though all these concepts are relevant in most of the analyzed papers, their extremely broad adoption makes them nearly irrelevant when considered for differentiating and grouping events according to the discussed topics.

Most of these concepts are, as expected, very generic (such as "System" or "Model") in the field of Computer Science and Information Technology (to which all the considered events belong), however some of them are very specific and usually associated with a precise research community, such as Semantic Web, Machine Learning, and Natural Language Processing. Semantic Web, in particular, appears in almost half of the considered events, even if the Semantic Web research community identified by cluster analysis is far from including half of the considered events.

4.2 IRCDL Proceedings Analysis

The second case study is focused on the evolution over time of the academic debate within a single community. Since our interest is focused on topics, only the TG of these events is generated.

Table 3. Most commonly extracted keyphrases ranked by their IDF

Topic	IDF
system	0.427
model	0.474
data	0.601
information	0.671
computer science	0.700
semantic web	1.076
language	1.144
web	1.144
semantics	1.191
software engineering	1.241
natural language processing	1.267
machine learning	1.267

To achieve temporal modelling, papers are grouped by year, then using the approach described in Sect. 3 to model the TG, every group of papers is represented as a node in a network. The first relevant insight about how the scientific debate evolved over time is given by the mere distribution of extracted topics among the considered years: buzzwords come and go and their presence inside the full text of published papers reflects the trends in the research community. The fraction of papers including a specific term is a significant measure of how much widespread such term is at a specific time. In Fig. 5 we show the result of this kind of analysis over the 10-years-wise most relevant buzzwords found in the IRCDL proceeedings. It can be noticed how “Digital Libraries”, which is the focus of the conference, is by far the most widespread term and consistently appeared in accepted papers over the ten years. On the other hand, some growing and diminishing trends can be easily spotted: “Information Retrieval” was a widespread topic in the first editions, however in the more recent ones its presence diminished significantly; “Cultural Heritage”, instead has encountered a growing popularity in recent editions while is was somehow less relevant in first ones.

This analysis, however, does not provides actual insights on the topics that actually characterized a specific year or a given time frame in research. In other words it does not answers the question “what was that year about?”. To achieve this goal we must evaluate a time-wise IDF that allows us to set apart buzzwords consistently present in the domain and concepts that surfaced only in a certain time frame. Again, creating a Topic Graph where papers are grouped by year to form nodes allows this kind of analysis.

Figure 6 shows the Topic Graph built upon IRCDL accepted papers grouped by year annotated with the most significant topics for each node (i.e. the ones with the highest $CK - IDF$). In this graph nodes represent editions of the conference, the larger the node is drawn, the more distinct topics were extracted from its associated

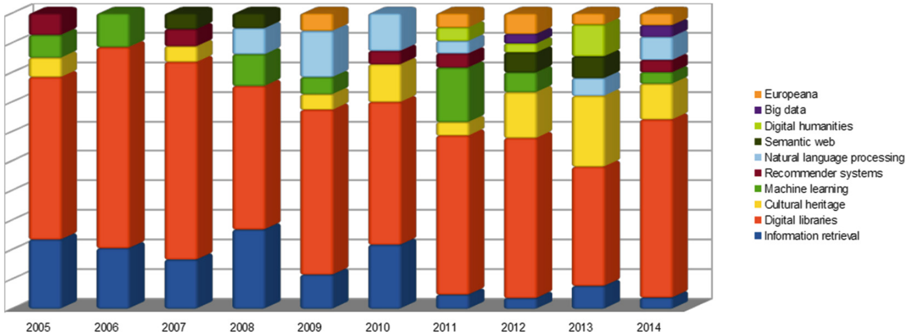

Fig. 5. Most frequent topics over ten editions of the IRCDL conference. (Color figure online)

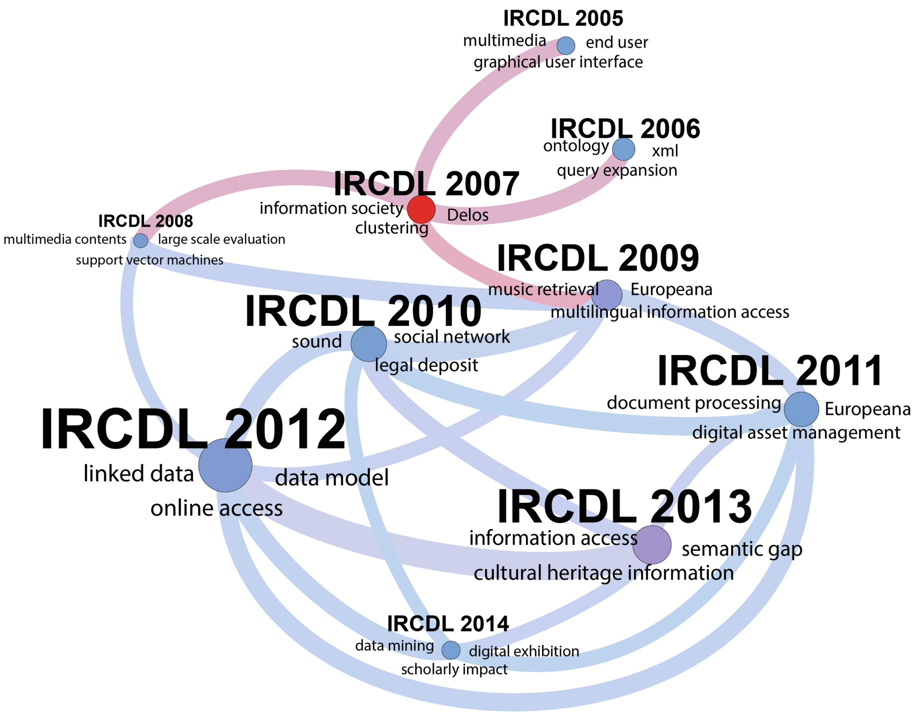

Fig. 6. Characterizing topics of each IRCDL edition.

papers; the presence of an arc between two nodes implies a significant overlap in the associated topics. Nodes in the figure are coloured according to their centrality with a "hot" color (tending to red) meaning a high centrality in the network. Highly central nodes, such as the 2007 edition are to be considered turning points in the history of the conference, since they represent a bridge between distinct groups of topic-wise similar editions. It is interesting to note how, in this example, the part of the graph representing the first editions of the conference is relatively sparse indicating little overlap aside from buzzwords, while more recent editions are much more connected, indicating a great deal of shared topics which implies that the conference has found a consistent core of topics.

5 Conclusions and Future Works

In this paper we presented a new approach to discover the semantic and social relations among scientific conferences with the aim of discovering shared interests, spotting research communities and, hopefully, help scientist addressing the problem of finding the right venue for their work. Moreover we have also shown how our approach can be applied to track the evolution of the academic debate over time as well.

The dual analysis on author participation and topics dealt by a conference is, in our opinion, the most notable feature of our approach: traditional social network based analysis can detect existing communities, but is unlikely to identify complementary communities that discuss the same topics and therefore 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, by comparing the topology of the SNA based model and the content-based model one can identify communities that deal with the same topics, but have little or no social connections at all. Identifying such communities, in our opinion, might help scholars to find relevant literature and, hopefully, to foster knowledge transfer from one community to another, improving the quality of research. Editors and organizers as well might obtain from temporal analysis meaningful insights over the trends within their community and exploit such information to provide a more attractive venue or tracks for authors.

However, we won't proceed further in these speculation and simply conclude this work remarking that, due to the modularity and domain independent nature of the system and methodology here proposed, the analysis presented could be easily applied to others languages, domains, and communities.

As future work, we are planning to apply our techniques over larger data sets, possibly with the collaboration of scientific journal editors, to model entire domains such as Computer Science, Mathematics, and Environmental Sciences, rather than relatively small subsets like in this introductory work.

References

1. Albert, R., Jeong, H., Barabási, A.L.: Internet: diameter of the world-wide web. Nature **401**(6749), 130–131 (1999)
2. Barabsi, A., Jeong, H., Nda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Phys. A Stat. Mech. Appl. **311**(34), 590–614 (2002)

3. Dattolo, A., Ferrara, F., Tasso, C.: On social semantic relations for recommending tags and resources using Folksonomies. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T. (eds.) Human – Computer Systems Interaction: Backgrounds and Applications 2. AISC, vol. 98, pp. 311–326. Springer, Heidelberg (2012)
4. De Nart, D., Tasso, C.: A domain independent double layered approach to keyphrase generation. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies, WEBIST 2014, pp. 305–312. SciTePress (2014)
5. Degl'Innocenti, D., De Nart, D., Tasso, C.: A new multi-lingual knowledge-base approach to keyphrase extraction for the Italian language. In: Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval, pp. 78–85. SciTePress (2014)
6. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999, pp. 50–57. ACM, New York (1999)
8. 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)
9. 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, vol. 2010, pp. 1310–1313 (2010)
10. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and role discovery in social networks. Computer Science Department Faculty Publication Series, p. 3 (2005)
11. Mika, P.: Flink: semantic web technology for the extraction and analysis of social networks. Web Semant. Sci. Serv. Agents World Wide Web **3**(2), 211–223 (2005)
12. Newman, M.: Scientific collaboration networks. I. network construction and fundamental results. Phys. Rev. E **64**, 016131 (2001)
13. Newman, M.: Scientific collaboration networks. II. shortest paths, weighted networks, and centrality. Phys. Rev. E **64**, 016132 (2001)
14. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. **98**(2), 404–409 (2001)
15. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. J. Inf. Sci. **28**(6), 441–453 (2002)
16. Pudota, N., Dattolo, A., Baruzzo, A., Tasso, C.: A new domain independent keyphrase extraction system. In: Agosti, M., Esposito, F., Thanos, C. (eds.) IRCDL 2010. CCIS, vol. 91, pp. 67–78. Springer, Heidelberg (2010)
17. Sack, W.: Conversation map: a content-based usenet newsgroup browser. In: Leug, C., Fisher, D. (eds.) From Usenet to CoWebs, pp. 92–109. Springer, New York (2003)
18. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications, London (2000)
19. 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)
20. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)
21. Watts, D.: Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton University Press, Princeton (1999)
22. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature **393**(6684), 440–442 (1998)