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

Dario De Nart, Dante Degl'Innocenti, Marco Basaldella, and Carlo Tasso

Artificial Intelligence Lab
Department of Mathematics and Computer Science
University of Udine, Italy
{dario.denart,carlo.tasso}@uniud.it,
{dante.deglinnocenti,basaldella.marco.1}@spes.uniud.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 in particular 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 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 relationship. More specifically we have experimented the proposed approach by analysing the CEUR conference and workshop proceedings. 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. Moreover, open access to full papers allows to analyse textual contents by means of NLP techniques for conceptual or topic analysis, such as *Keyphrase Extraction*, making it possible to group events according to the actual contents of accepted contributions.

The work presented in this paper is aimed at grouping CEUR volumes according to contributing authors and topics covered. 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 [16] and [5] 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 Section 2 we briefly introduce some related work, in Section 3 we present our original approach, in Section 4 the results of our analysis are discussed, and Section 5 concludes the paper and presents some planned future work.

2 Related Work

The study of the connections between people and groups has a long research tradition of at least 50 years [2] [20] [18] [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.

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

This approach is 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 [22] [1], 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] [13] [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 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 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 more than 2000 research papers.

3 Proposed Methodology

In order to support our analysis a testbed system was developed to provide access to CEUR volumes, 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. CEUR provides open access to papers in PDF format as well as some interesting metadata, like the authors and the venue where the work was presented. In this work we decided to limit our analysis to 2014 events whose proceedings were published on CEUR before December the 1st 2014.

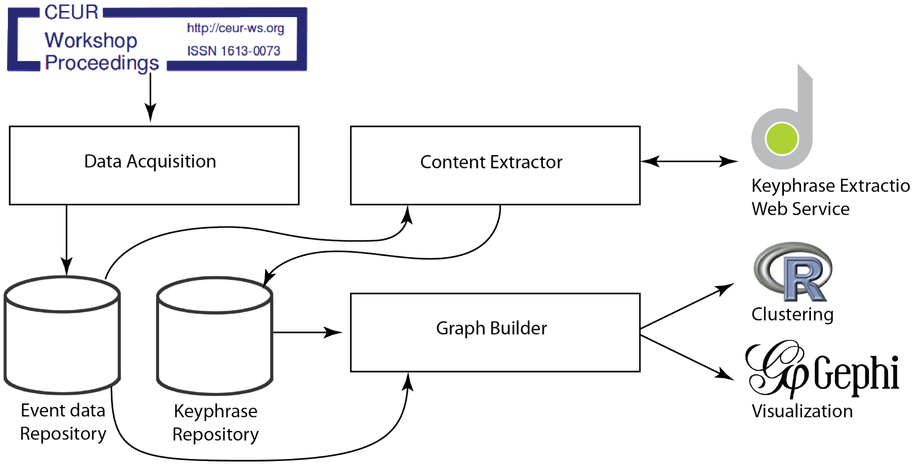

Fig. 1. System architecture overview.

The testbed system is made of three modules: the Data Acquisition (DA) module, the Content Extraction (CE) module, and the Graph Builder (GB) module, as shown in Figure 1. The DA module crawls the CEUR Website 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 one 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 were weighted proportionally to the number of authors shared. Communities of similar events in the graph are then identified with 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 with fine grain and to group events according to semantic similarities. For each CEUR event, all its 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).

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

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 CEUR 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 and compare the results of our two analysis on the 2014 CEUR volumes, namely the AG, that is the graph generated with SNA and the 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.

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

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

Table 1. Author Graph global statistics

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.

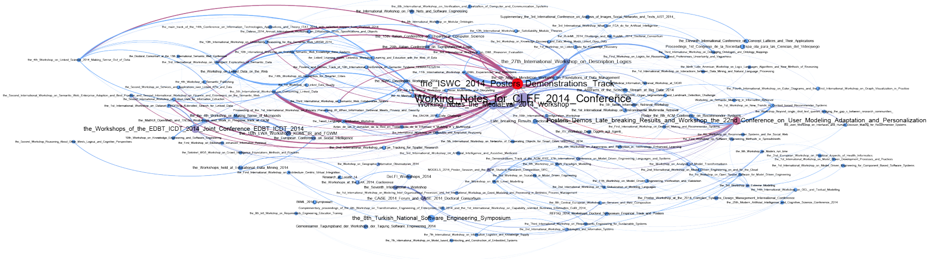
the_ISWC_2014_Posters_Demonstrations_Track
Working_Notes_for_CLEF_2014_Conference
Posters_Demos_Late_breaking_Results_and_Workshop_the_22nd_Conference_on_User_Modeling_Adaptation_and_Personalization
the_27th_International_Workshop_on_Description_Logics
the_Workshops_of_the_EDBT_ICDT_2014_Joint_Conference_EDBT_ICDT_2014
the_8th_Turkish_National_Software_Engineering_Symposium

Fig. 2. Overview of the Author Graph.

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 Figure 3, where all the isolated nodes are omitted.

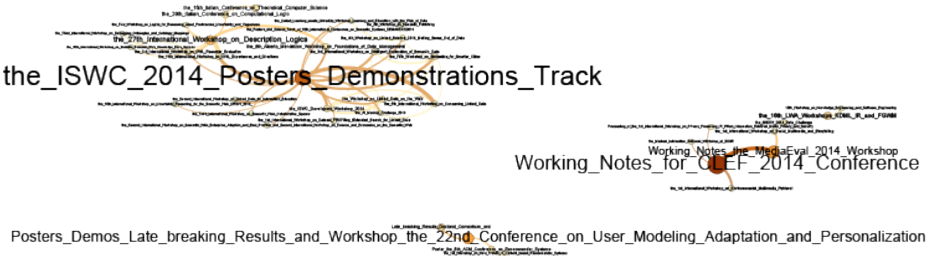
the_ISWC_2014_Posters_Demonstrations_Track
the_27th_International_Workshop_on_Description_Logics
Working_Notes_the_MediaEval_2014_Workshop
Working_Notes_for_CLEF_2014_Conference
Posters_Demos_Late_breaking_Results_and_Workshop_the_22nd_Conference_on_User_Modeling_Adaptation_and_Personalization

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. The generated TG, shown in Figure 4 is therefore extremely well connected and, considered as-is, it cannot provide any useful insight. Some nodes are clearly more central than others (the same color scheme of Figure 2 is used), but the network structure is so dense that the reasons of such properties are utterly unclear.

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

Table 2. Topic Graph global statistics

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 Figure 5 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 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.

the_ISWC_2014_Posters_Demonstrations_Track
Working_Notes_for_CLEF_2014_Conference
the_Workshops_of_the_EDBT_ICDT_2014_Joint_Conference_EDBT_ICDT_2014_
the_8th_Turkish_National_Software_Engineering_Symposium
Posters_Demos_Late_breaking_Results_and_Workshop_the_22nd_Conference_on_User_Modeling_Adaptation_and_Personalization
Working_Notes_the_MediaEval_2014_Workshop
the_27th_International_Workshop_on_Description_Logics
the_16th_LWA_Workshops_KDML_IR_and_FGWM
the_23th_International_Workshop_on_Concurrency_Specification_and_Programming
Gemeinsamer_Tagungsband_der_Workshops_der_Tagung_Software_Engineering_2014
the_CAiSE_2014_Forum_and_CAiSE_2014_Doctoral_Consortium
the_European_Conference_on_Social_Intelligence
the_29th_Italian_Conference_on_Computational_Logic
the_Workshops_held_at_Educational_Data_Mining_2014
the_International_Workshop_on_Petri_Nets_and_Software_Engineering
the_Seventh_International_i_Workshop
the_25th_Modern_Artificial_Intelligence_and_Cognitive_Science_Conference_2014
the_International_Workshop_on_Meta_learning_and_Algorithm_Selection
the_11th_International_Workshop_on_OWL_Experiences_and_Directions
the_ISWC_Developers_Workshop_2014
the_SNOW_2014_Data_Challenge
the_Tweet_Translation_Identification_Workshop
Actas_de_las_V_Jornadas_de_la_Red_en_Tratamiento_de_la_Informacion_Multilingue_y_Multimodal
CLEF_Workshops_2014

Fig. 4. Overview of the Topic Graph.

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.

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. With regard to the CEUR proceedings of 2014 events, our analysis led us to the following insights.

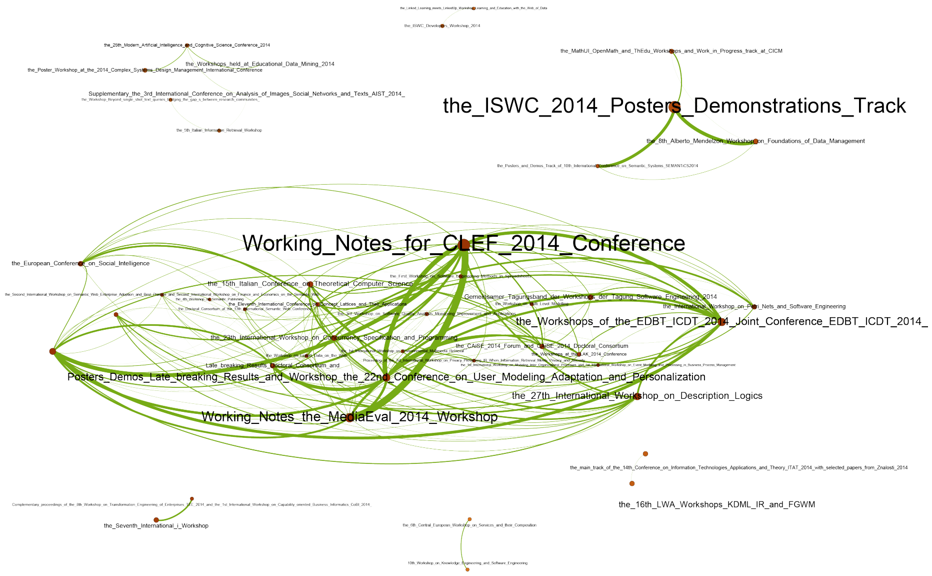
the_Linked_Learning_meets_LinkedUp_Workshop_Learning_and_Education_with_the_Web_of_Data
the_ISWC_Developers_Workshop_2014
the_25th_Modern_Artificial_Intelligence_and_Cognitive_Science_Conference_2014
the_Workshops_held_at_Educational_Data_Mining_2014
the_Poster_Workshop_at_the_2014_Complex_Systems_Design_Management_International_Conference
Supplementary_the_3rd_International_Conference_on_Analysis_of_Images_Social_Networks_and_Texts_AIST_2014_
the_Workshop_Beyond_single_shot_text_queries_bridging_the_gap_s_between_research_communities_
the_5th_Italian_Information_Retrieval_Workshop
the_MathUI_OpenMath_and_ThEdu_Workshops_and_Work_in_Progress_track_at_CICM
the_ISWC_2014_Posters_Demonstrations_Track
the_8th_Alberto_Mendelzon_Workshop_on_Foundations_of_Data_Management
the_Posters_and_Demos_Track_of_10th_International_Conference_on_Semantic_Systems_SEMANTiCS2014
Working_Notes_for_CLEF_2014_Conference
the_European_Conference_on_Social_Intelligence
the_First_Workshop_on_Software_Engineering_Methods_in_Spreadsheets
the_15th_Italian_Conference_on_Theoretical_Computer_Science
the_Second_International_Workshop_on_Semantic_Web_Enterprise_Adoption_and_Best_Practice_and_Second_International_Workshop_on_Finance_and_Economics_on_the_Semantic_Web
the_4th_Workshop_on_Semantic_Publishing
the_Eleventh_International_Conference_on_Concept_Lattices_and_Their_Applications
the_Doctoral_Consortium_at_the_13th_International_Semantic_Web_Conference
Gemeinsamer_Tagungsband_der_Workshops_der_Tagung_Software_Engineering_2014
the_Workshop_on_Multi_Level_Modelling
the_International_Workshop_on_Petri_Nets_and_Software_Engineering
the_3rd_Workshop_on_Software_Quality_Analysis_Monitoring_Improvement_and_Applications
the_Workshops_of_the_EDBT_ICDT_2014_Joint_Conference_EDBT_ICDT_2014_
the_23th_International_Workshop_on_Concurrency_Specification_and_Programming
the_CAiSE_2014_Forum_and_CAiSE_2014_Doctoral_Consortium
the_1st_International_Workshop_on_Environmental_Multimedia_Retrieval
the_Workshop_on_Linked_Data_on_the_Web
the_Workshops_at_the_LAK_2014_Conference
Proceeding_of_the_1st_International_Workshop_on_Privacy_Preserving_IR_When_Information_Retrieval_Meets_Privacy_and_Security
the_1st_International_Workshop_on_Modeling_Inter_Organizational_Processes_and_1st_International_Workshop_on_Event_Modeling_and_Processing_in_Business_Process_Management
Late_breaking_Results_Doctoral_Consortium_and
Posters_Demos_Late_breaking_Results_and_Workshop_the_22nd_Conference_on_User_Modeling_Adaptation_and_Personalization
the_27th_International_Workshop_on_Description_Logics
Working_Notes_the_MediaEval_2014_Workshop
the_main_track_of_the_14th_Conference_on_Information_Technologies_Applications_and_Theory_ITAT_2014_with_selected_papers_from_Znalosti_2014
the_16th_LWA_Workshops_KDML_IR_and_FGWM
Complementary_proceedings_of_the_8th_Workshop_on_Transformation_Engineering_of_Enterprises_TEE_2014_and_the_1st_International_Workshop_on_Capability_oriented_Business_Informatics_CoBI_2014_
the_Seventh_International_i_Workshop
the_6th_Central_European_Workshop_on_Services_and_their_Composition
10th_Workshop_on_Knowledge_Engineering_and_Software_Engineering

Fig. 5. Clusters obtained from the Topic Graph.

First, there exist some groups of events that serve as meeting point for some research communities and actually the Semantic Web community is the one that includes the largest share of events.

Secondly, there also exist some groups of events whose accepted contributions deal with a particular set of concepts and a large share of the analyzed events accepted papers dealing with user modelling, information access, information extraction and personalization. Moreover, these clusters have little overlap meaning that large parts of disjoint communities actually deal with similar topics while, on the other hands, parts of the a single community may deal with distinct topics. This result can be a sign that there exist, within the CEUR contributing authors, complementary communities which are now disjoint but could exchange knowledge and expertise to improve their research activities.

Finally, our content-based analysis allowed us to detect the most widespread buzzwords in the 2014 CEUR volumes, among which we find the words "Semantic web". About this topic, we find that:

- we have a relatively small Semantic Web related topic cluster detected in the TG;
- we have huge Semantic Web related cluster detected in the AG;
- the 2014 International Semantic Web Conference served as meeting place for researchers belonging to virtually any other community in CEUR;
- as just mentioned, "Semantic web" is a very widespread buzzword in our data source.

From these observations, in our opinion, it should arise the thorny question of what is the Semantic Web really about today.

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

Table 3. Most commonly extracted keyphrases ranked by their IDF

However, we won't proceed further in this 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. Anyway, we are just scratching the surface of the possibilities this method offers and our future work will be focused on further investigation on the research conference domain, possibly with more data than the CEUR volumes alone due to the fact that they correspond only to a small fraction of the research community.

Other specific kinds of analysis we are considering for future work concern the evolution over time (year after year) of topics, the discovery of useful detection indicators showing new research trends, and a more fine-grained analysis based on the scientific production of single authors participating to scientific events.

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. Physica A: Statistical Mechanics and its Applications 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., Kulikowski, J., Mroczek, T. (eds.) Human Computer Systems Interaction: Backgrounds and Applications 2, Advances in Intelligent and Soft Computing, vol. 98, pp. 311–326. Springer Berlin Heidelberg (2012)
4. De Nart, D., Tasso, C.: A domain independent double layered approach to keyphrase generation. In: WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies. 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. Proceedings of the National Academy of Sciences 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. pp. 50–57. SIGIR '99, ACM, New York, NY, USA (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 Semantics: Science, Services and Agents on the 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 (Jun 2001)
13. Newman, M.: Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132 (Jun 2001)
14. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98(2), 404–409 (2001)
15. Otte, E., Rousseau, R.: Social network analysis: a powerful strategy, also for the information sciences. Journal of Information Science 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.) Digital Libraries, Communications in Computer and Information Science, vol. 91, pp. 67–78. Springer Berlin Heidelberg (2010)
17. Sack, W.: Conversation map: a content-based usenet newsgroup browser. In: From Usenet to CoWebs, pp. 92–109. Springer (2003)
18. Scott, J.: Social Network Analysis: A Handbook. SAGE Publications (January 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 (1994)
21. Watts, D.: Small Worlds: the dynamics of networks between order and randomness. Princeton Univ Pr (1999)
22. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-worldnetworks. nature 393(6684), 440–442 (1998)