Knowledge discovery and enrichment from scholarly data for expert finding

Stella Zevio[1] [0000−0003−0877−3633]

LIPN - CNRS UMR 7030 - Université Paris XIII
99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
zevio@lipn.univ-paris13.fr

Abstract. With the generalisation of the digitalization of scientific publications, scholarly data is now tackled with a big data perspective. In this context, interest about applications like expert finding, research recommandation systems or collaborators discovery grows and challenges related to knowledge discovery on large-scale scholarly data arise. The Unified Knowledge Platform (Plateforme de Connaissances Unifiées) project aims at developing an open-source platform valuing scholarly as well as business data. In that respect, we propose an approach and a methodology for discovering knowledge and enrich it from workings documents, more precisely scientific publications, in the particular use case of expert finding. Compared to the state of the art, the originality of our approach lies in the combination of text mining as well as graph mining methods, more specifically graph abstraction. We present an experiment with already published results on the 9-years acts of a French workshop on semantic information retrieval. In this experiment, we managed to obtain a graph mapping researchers who participated in the workshop. In this graph, researchers are linked together by co-publication relationships and described by their topics of publication. We were also able to detect dense communities of researchers with the help of graph abstraction. Based on these results and in the light of the state of the art, we discuss further research tracks.

Keywords: Scholarly data · Graph mining · Knowledge discovery · Expert finding.

Early Stage PhD - EKAW 2018 Doctoral Consortium

1 Introduction

An expertise is ”an individual’s skill, knowledge, aptitude or behaviour” [6]. The task of expert finding consists in assessing individuals’ expertises (*i.e.* constructing their expert profile [5]). This task has various applications in industries such as finding employable and appropriate candidates or assigning an expert to a task or a project for example. In academia, expert finding is also useful for assigning a researcher to a program committee or a project expertise, or setting up research projects, to name a few. According to the claim that an author of a text is an expert of its content, text appears like a solid source of knowledge for expert finding. More precisely, we focus on working documents (*e.g* CV, project reports, *etc.*). Such documents contain crucial information about an individual's expertises. In academia, they mainly consist in scholarly data.

The PCU[1] (Plateforme de Connaissances Unifiée, *i.e* Unified Knowledge Platform) project's aim is to propose an open source industrial platform valuing business (and scholarly, to an extent) data. With the recent explosion of digitization of academic and technical documents, scholarly data has known such a rapid growth [18] that we now talk about big scholarly data. In that respect, interest in big scholarly data platforms has emerged [17, 12]. In this context, our aim is to discover knowledge from text (*i.e* scientific publications) for expert finding, represent it automatically into graphs and enrich knowledge with the hypothesis that new knowledge will emerge from the graph structure. Our research question is the following: how precise and accurate knowledge can become thanks to the knowledge enrichment process, and what methods are providing the best results on working documents, or more specifically, scientific publications?

To answer this question, we will enrich PCU with a semantic platform with the aim of supporting experiments that will test our hypothesis and enable us to answer this research question. This research question lies in the area of knowledge extraction and at the interface between knowledge discovery, natural language processing and artificial intelligence, more precisely machine learning (unsupervised methods) and graph mining.

We will present a state of the art in section 2, define our approach in section 3, define our methodology in section 4, present results that we have reached so far in section 5 reflect on the relevance of our approach and discuss further work and research tracks in section 6.

2 State of the art

According to the literature, expert finding is closely related to the problem of expert profiling [3, 11], which implies identification of expertises and their assignation to appropriate individuals owning them [5]. Initially, expert finding systems were based on people assessing their own expertises by selecting predefined keywords [2], and the use of manually generated heuristics was predominant [19]. For the sake of the automation of expert finding, textual sources of knowledge were harvested [2]. With the explosion of online data stored in digital libraries, scholarly data [18, 10] became a solid source of knowledge for expert finding. In the rest of this paper, we will consider scholarly data, more precisely scientific publications, as a relevant textual source of expert finding concerning scholars.

From scientific publications, classical methods of expert finding described in the literature are based on information extraction methods [10] such as metadata (title, authors, abstract, date of publication, *etc.*) as well as citations and

[1] Plateforme de Connaissances Unifiée : https://www.smile.eu/fr/publications/smile-lab/pcu-plateforme-connaissances-unifiees

author information (co-authorship, authors' affiliations) extraction. Open-source systems for metadata, citations and author information extraction exist [16] but still need improvement according to their error analysis. To identify underlying expertises within scientific publications, concept extraction methods are used [4] by means of classical keyphrase extraction algorithms for example. Such algorithms can be domain-dependant, thus rely on a model trained on an annotated corpus [9] or take advantage of knowledge of the domain by investigating relations between expertise topics extracted [5, 8]. Concerning the computer science domain, an ontology has recently been released [13].

Some scholarly data platforms have already been developed, such as Rexplore [12]. Representation of knowledge extracted from text through a graph is quite common. Rexplore takes advantage of this representation by providing a semantic network of fine-grained research areas, linked by semantic relations. As described in the literature, researchers have mostly used graph and machine-learning techniques for expert finding [1]. As suggested [1], issues related to the identication and ranking of experts [2] can be avoided by "combining content-based expertise indicators and social relationships". Combining machine learning algorithms with graph mining methods for expert finding in order to discover knowledge and enrich it is a challenging research question. Inspired from the analysis of social networks, graph mining techniques have been applied to the detection of frequent k-communities [14]. With the claim that researchers and expertises are represented through an attributed graph, detecting strongly connected k-communities would be interesting to investigate for expert finding. Moreover, the application of the recent hub-authority core theory [15] is also a promising investigation trail for directed citation graphs. As far as we know, no scholarly data platform takes advantage of the recent advances of graph mining techniques, even if graph representation is common.

3 Proposed approach

In the light of the state of the art, we propose an original approach for expert finding consisting in combining text mining, more precisely machine learning algorithms applied on text (*i.e* scientific publications), with graph mining methods. The graph mining methods are applied on the graph representing knowledge extracted from the text. The machine learning algorithms considered are keyphrases [9] and semantic relationships [8] extraction algorithms. From the text, the keyphrase extraction algorithm is initially applied, in order to extract fine-grained topics or thematics of publication within scholarly data, more precisely on full-text scientific publications. The algorithm considered is based on a model trained on an annotated corpus thus it is language-dependent and applied to the domain of computer science. It is based on a conditional random fields model trained with keyphrase candidated filtered with part-of-speech tag sequences.

Then, the semantic relationships extraction algorithm is applied on the output of the keyphrase extraction algorithm. It is based on semantic similarity between extracted keyphrases belonging to the same sentence. The semantic similarity is measured thanks to most frequent patterns and clustering methods. The algorithm is unsupervised, which enables the automatic extraction of already known as well as brand new relationships between extracted keyphrases, such as "is-a" between "information retrieval" and "task". It has been tested on the ACL corpus [7]. The sequential application of these two algorithms enables us to collect the knowledge required for building a graph representing knowledge extracted from the text.

The originality of our approach described in Figure 1 lies in the combination of these algorithms with graph abstraction [14]. From the representation of a text as a graph, the idea is to focus on strongly connected vertices applying a topological constraint, for example by searching for the k-core (*i.e* the largest subgraph verifying a topological constraint such as "each vertex in the subgraph has a k degree"). Several representations are possible with attributed graphs, one of them being that vertices represent researchers and are labeled by topics of publication. In Figure 1, experts on topic 1 are obtained by removing R5 (who does not have 2-degree), then R4 for the same reason after R5's removal. Our expectation considering strongly connected vertices is that it may bring out new and interesting knowledge from the graph that is itself built from the representation of text. In Figure 1, all experts on topic 1 obtained by 2-core abstraction on the graph are also experts on topic t2, which is a new knowledge. This hypothesis has been raised from the state of the art, more precisely from network analysis for social sciences [15] and our claim is that interesting knowledge could emerge for expert finding from the generalisation of such graph mining methods.

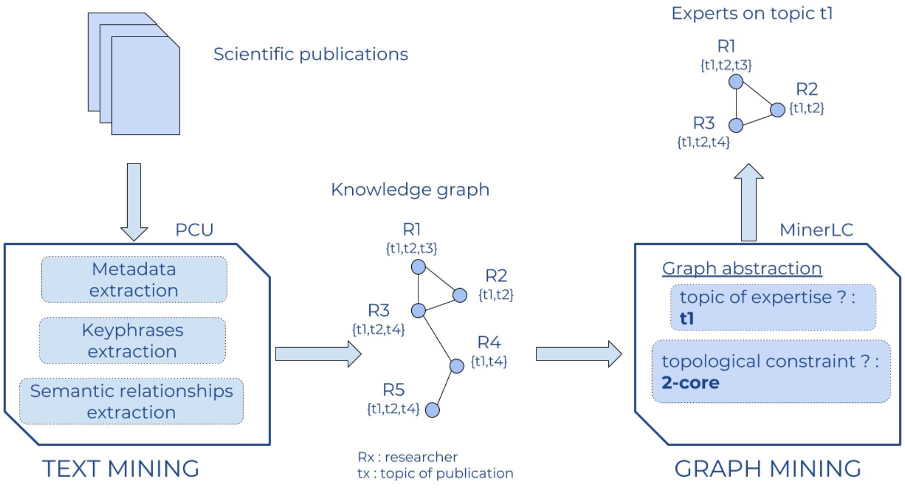

Fig. 1. Our approach combining data and graph mining for expert finding from scientific publications

4 Methodology

To support our experiments and meet PCU project's purpose, we have developed an open source semantic platform making the machine learning algorithms [9, 8] that we selected available as a workflow. We also developed a Mathematica workflow as a complementary tool for our experiments, mostly for data manipulation and displaying as well as for the automation of the connection between the machine learning algorithms outputs and the input of MinerLC (*i.e* the software for applying graph abstractions). We selected the following datasets : ACL corpus written in English and the 9-years scientific acts of the Recherche d'Information SEmantique (RISE) workshops[2], mostly written in French. We plan on building a larger dataset, by collecting the scientific publications of the members of our lab. Our experiments consist in applying the machine learning algorithms on the full-text of the scientific publications. We also collect structured metadata such as titles, authors or keywords thanks to CERMINE [16].

From the knowledge discovered, we build a graph G = (V,E) (V being the vertices, *i.e* the objects considered and described by items, E the edges, *i.e* the relations between the vertices) for each dataset. Our first experiments consisted in describing researchers (*i.e* vertices, objects) by topics of publication (*i.e* items) and linking them together by relationships of co-publications (*i.e* edges). On this graph, 2-core abstraction is applied in order to identify communities of strongly connected researchers who published on a common set of topic of publication with at least two other researchers that also belong to the community. Further expriments will emerge, consisting in identifying the best way to represent knowledge from scientific publications, being describing researchers or publications with topics of publication, dates of publication, locations of laboratories or keywords, for example. The nature of the semantic relationships between objects described could also be different, like co-citation relationships for example. Also, more interesting topological constraints should be applied on the graph during graph abstraction, such as ”each vertex in the subgraph belongs to a star” for example.

To evaluate our work, no gold standard exists as far as we know. We should evaluate the costs of building a gold standard with manually annotated corpus for supporting our experiments as well as the possibility of conducting an evaluation campaign in the domain of scholarly data. As we aim at enriching knowledge extracted with the use of graph abstractions, an error analysis as well as a comparison of knowledge extracted with and without (baseline) the application of graph abstraction should be proposed. Indeed, we should be able to detect communities straight from the graph, but such communities should be narrower after the application of graph abstraction. As a matter of fact, our hypothesis consisting in new knowledge emerging from graph abstraction would be verified, as we would be able to detect core scientific publications or core researchers of a domain with more precision.

[2] RISE : https://sites.google.com/site/frenchsemanticir/documents

5 Results

We presented the preliminary results obtained on our experiment on the 9-years scientific acts of Recherche d'Information SEmantique (RISE) workshops (from 2009 to 2017) during the 10th edition in 2018 [20]. During this experiment, we managed to obtain an attributed graph mapping the researchers of the workshops, described by their topics of publication and linked by coauthor relationships. As scientific publications in RISE are mainly written in French and the keyphrase extraction algorithm is language-dependant and trained for English, the topics of publications were simply extracted from the keywords given by the authors. We obtained a graph by applying text mining methods on the scientific acts of the workshops. We were able to detect dense communities of researchers based on graph abstraction applied on this graph.

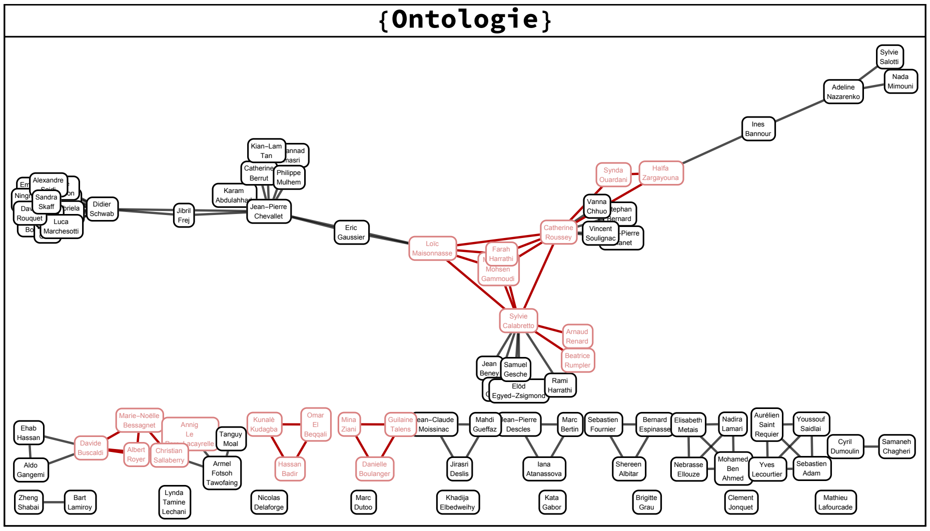

Fig. 2. Community of researchers who participated in RISE workshops (2009-2017). In red : community of researchers obtained by the application of a graph abstraction with topological constraint of 2-degree on the topic of "ontologie" (*i.e* ontology, in French)

For example, we queried researchers who published on the subject of "ontology" (*i.e* "ontologie" in French). We obtained a community of researchers based on authors who published on this particular topic. With the application of 2-core graph abstraction on this community, we managed to remove researchers from the community because of their lack of co-publication relationships with the others within the workshops. Thus, we managed to identify a narrower community of researchers strongly connected to each other, according to a degree 2, whose members would be the core experts of the domain. The results are showed in Figure 2.

6 Discussion

As our preliminary results seem to imply, our approach looks quite promising. We already managed to obtain finer-grained communities of researchers according to a given topic of publication, which seems to validate the relevance of our approach. We run into difficulties related to considerations such as the size of the graph obtained and the lacking of the quality of semantic relationships describing the objects (*i.e* vertices) of the graph. Indeed, the graph we obtained is quite small (less than 50 vertices) and lacks items describing the objects and relationships between objects. Such a graph is interesting for experiments and ease of manipulation for showing examples, but we should consider large-scale data or at least larger corpus. Also, resources for French-language are quite limited.

Recommandations for future work would be supporting the semantic interoperability of our graphs and opening to the semantic web by integrating the Computer Science Ontology to PCU. We should enrich our vertices' descriptions with the concepts of the ontology recognized in the full-text or abstract of the scientific publications thanks to semantic annotation, with the idea of a bottom-up enrichment by generalization. For supporting multilingual processing, a translation of the Computer Science Ontology in French as well as a training of the keyphrase extraction algorithm in French would be useful, including in the aim of meeting PCU's purpose, but costs of translation should be evaluated. We should also consider conducting further experiments such as described in the section 4, among other things describing objects with more than topics of publication (dates of publication or locations of laboratories for example) and finding the best parameters for graph abstraction for expert finding.

Acknowledgements

This thesis is supervised by Professor Thierry Charnois, Dr. Guillaume Santini and Dr. Haïfa Zargayouna. It is financed by the FUI project PCU (Plateforme de Connaissances Unifiées).

References

1. Al-Taie, M.Z., Kadry, S., Obasa, A.I.: Understanding Expert Finding Systems: Domains and Techniques. Social Network Analysis and Mining **8**(1), 57 (2018)
2. Angelova, M., Boeva, V., Tsiporkova, E.: Advanced Data-driven Techniques for Mining Expertise. In: 30th Annual Workshop of the Swedish Artificial Intelligence Society SAIS 2017, May 15–16, 2017, Karlskrona, Sweden. pp. 45–52. No. 137, Linköping University Electronic Press (2017)
3. Balog, K., De Rijke, M., et al.: Determining Expert Profiles (With an Application to Expert Finding). In: International Joint Conference on Artificial Intelligence. vol. 7, pp. 2657–2662 (2007)
4. Bordea, G.: Concept Extraction Applied to the Task of Expert Finding. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications. pp. 451–456. Springer Berlin Heidelberg (2010)

5. Bordea, G.: Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining. Ph.D. thesis (2013)
6. Draganidis, F., Mentzas, G.: Competency Based Management: a Review of Systems and Approaches. Information Management & Computer Security **14**(1), 51–64 (2006)
7. Gábor, K., Tellier, I., Charnois, T., Zargayouna, H., Buscaldi, D.: Détection et classification non supervisées de relations sémantiques dans des articles scientifiques. In: JEP-TALN-RECITAL 2016. Actes de la conférence conjointe JEP-TALN-RECITAL 2016, vol. 2. Paris, France (2016)
8. Gábor, K., Zargayouna, H., Buscaldi, D., Tellier, I., Charnois, T.: Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. In: LREC 2016. Proceedings of the LREC 2016 Conference (2016)
9. Hernandez, S.D., Buscaldi, D., Charnois, T.: LIPN at SemEval-2017 Task 10: Filtering Candidate Keyphrases from Scientific Publications with Part-of-Speech Tag Sequences to Train a Sequence Labeling Model. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 995–999 (2017)
10. Khan, S., Liu, X., Shakil, K.A., Alam, M.: A Survey on Scholarly Data: From Big Data Perspective. Information Processing & Management **53**(4), 923–944 (2017)
11. Lin, S., Hong, W., Wang, D., Li, T.: A Survey on Expert Finding Techniques. Journal of Intelligent Information Systems **49**(2), 255–279 (Oct 2017)
12. Osborne, F., Motta, E., Mulholland, P.: Exploring Scholarly Data With Rexplore. In: International Semantic Web Conference. pp. 460–477. Springer (2013)
13. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas (2018)
14. Soldano, H., Santini, G., Bouthinon, D.: Local Knowledge Discovery in Attributed Graphs. In: International Conference on Tools with Artificial Intelligence (ICTAI). pp. 250–257 (2015)
15. Soldano, H., Santini, G., Bouthinon, D., Lazega, E.: Hub-Authority Cores and Attributed Directed Network Mining. In: Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on. pp. 1120–1127. IEEE (2017)
16. Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P.J., Bolikowski, Ł.: CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature. International Journal on Document Analysis and Recognition (IJDAR) **18**(4), 317–335 (2015)
17. Wu, Z., Wu, J., Khabsa, M., Williams, K., Chen, H.H., Huang, W., Tuarob, S., Choudhury, S.R., Ororbia, A., Mitra, P., Giles, C.L.: Towards Building a Scholarly Big Data Platform: Challenges, Lessons and Opportunities. In: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 117–126. IEEE Press (2014)
18. Xia, F., Wang, W., Bekele, T.M., Liu, H.: Big Scholarly Data: A Survey. IEEE Transactions on Big Data **3**(1), 18–35 (2017)
19. Yimam-Seid, D., Kobsa, A.: Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing and Electronic Commerce **13**(1), 1–24 (2003)
20. Zevio, S., Zargayouna, H., Santini, G., Charnois, T.: Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers Recherche d'Information SEmantique (RISE). In: Actes de la dixième édition de l'atelier Recherche d'Information SEmantique (RISE). pp. 6–13 (2018)