Making Sense of Massive Amounts of Scientific Publications: the Scientific Knowledge Miner Project

Francesco Ronzano, Ana Freire, Diego Saez-Trumper and Horacio Saggion

Department of Information and Communication Technologies
Universitat Pompeu Fabra
Carrer Tanger 122-140, Barcelona, Spain
first.last@upf.edu

Abstract. The World Wide Web has become the hugest repository ever for scientific publications and it continues to increase at an unprecedented rate. Nevertheless, this information overload makes the exploration of this content a very time-consuming task. In this landscape, the availability of text mining tools to characterize and explore distinctive features of the scientific literature is mandatory. We present the Scientific Knowledge Miner (SKM) Project, that aims to investigate new approaches and frameworks to facilitate the extraction of knowledge from scientific publications across different disciplines. More specifically, we will focus on citation characterization, recommendation and scientific document summarization.

Keywords: text mining, information extraction, recommender systems, indexing, crawling, online resources.

1 Introduction:

During the last decade the amount of scientific information available on-line increased at an unprecedented rate. Recent estimates reported that a new paper is published every 20 seconds [1]. PubMed[1], Elsevier' Scopus[2] and Thomson Reuther's ISI Web of Knowledge[3] respectively contain more than 24, 57 and 90 million papers. In this scenario, the exploration of scientific literature has turned into an extremely complex and time-consuming task. The availability of text mining tools able to extract, aggregate and turn scientific unstructured textual contents into well organized and interconnected knowledge is fundamental.

However, scientific publications are characterized by several structural (title, abstract, figures, citations...), linguistic and semantic peculiarities that make them difficult to analyze by relying on general purpose text mining tools. One of the special features of scientific papers is their network of citations, that are starting to be exploited in several context including opinion mining [2, 7] and scientific text summarization [3, 8]. Besides citations, the interpretation of the semantics of the actual textual contents of

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.scopus.com
[3] http://www.webofknowledge.com

scientific papers usually needs the availability of knowledge repositories with an adequate coverage of scientific concepts and relations that could not be found on global domain knowledge resources like WordNet, DBPedia, FreeBase or BabelNet.

Considering both the peculiar structural and semantic features of scientific publications and the huge amounts of papers that need to be taken into account when we mine scientific literature, customized information extraction, semantic indexing, search and content aggregation approaches are required in order to fully take advantage of the knowledge exposed by scientific articles.

In this context, we present the Scientific Knowledge Miner (SKM) Project. It aims at developing both knowledge resources and a complex scientific knowledge mining infrastructure that will be exploited to support fine-grained semantic analysis and large-scale studies of scientific document collections. In the context of the SKM Project, we are going to analyze publications by relying and extending the Dr. Inventor Scientific Text Mining Framework [9] (DRI Framework), a freely available Java-based library. The DRI Framework enables the automated analysis and characterization of several facets of publications including the identification of the scientific discourse category of sentences (Approach, Background, Future Work, etc.), the characterization of the purpose of citations and the annotation of Named Entities that occur inside the textual contents of a paper[4]. By performing fine-grained semantic analysis of articles and aggregating and merging this information across collections of papers, the scientific literature analysis supported by the SKM Project are characterized by a different, deeper level granularity when compared to platforms like CiteSeer and GoogleScholar: these platforms mainly aggregate scientific papers by extracting and normalizing a structured set of metadata, including titles, authors, citation counts, etc. In the SKM Project, the DRI Framework will be properly complemented by ad-hoc data normalization, indexing and content visualization infrastructures that will allow the integration of information across papers and the execution of large-scale experiments.

2 Overview of the SKM Project

The core objective of the SKM Project is the investigation of new approaches, the extension and development of software tools and the creation of new datasets that will facilitate the extraction of knowledge from scientific publications across different disciplines. In particular, we have identified three core research topics that we would like to explore thanks to the SKM Project:

1. The analysis, the characterization and the navigation of collections of research papers in order to test new, alternative metrics to evaluate their quality;
2. The investigation of new, state of the art multi-document summarization approaches, tailored to scientific publications;
3. The evaluation of new approaches to scientific content recommendation that relies on both the contents of a paper and its relations with other scientific results.

[4] We rely on Babelfy: http://babelfy.org/

To investigate these research topics, we will carry out the following activities:

- extension and improvement of the Dr. Inventor Text Mining Framework[5], a Java library that integrates several Document Engineering and Natural Language Processing tools customized to enable and ease the analysis of the textual contents of scientific publications, both in PDF and JATS XML format. To get more information on the framework, the interested reader can refer to [9]. In the SKM Project will extend the Framework by implementing semi-supervised or unsupervised methods for citation classification (polarity and purpose) and semantically aware relation extraction (e.g. causal inference), both features useful to support information extraction and automated semantic enrichment of scientific texts;
- enrichment with new features of the SUMMA document summarization Java library [10]. In particular, SUMMA will be able to support the summarization of scientific texts by relying on citation-based summarization approaches (both sentence and paper assessment based on peers opinion). We will also implement state of the art multi-document summarization customized to scientific papers, based on the extraction and aggregation of relevant sentences across publications in order to automatically create surveys;
- implementation of new methodologies for semantic enrichment, interlinking, indexing and navigation of corpora of scientific papers. We will develop Web crawling approaches specialized to repositories of scientific publications and model relevant structured Web contents (such as conference Websites) in order to complement, enrich or interlink the information mined from scientific publications. In the meanwhile, we will complement this activities by the definition of proper content indexing, normalization, search and aggregation methodologies and infrastructures to enable the aggregation and browsing of the information extracted from huge collections of scientific publications;
- creation and sharing of semantically enhanced scientific datasets to train and validate new information extraction approaches. To this purpose we will take advantage of Annote[6], the Web based collaborative annotation tool we developed in the context of Dr. Inventor to support annotators in carrying out complex annotation tasks such as rhetorical sentence classification or summarization.

3 SKM scientific publication mining infrastructure

In this section we introduce the high-level architecture of the infrastructure to crawl, process, index and visualize the contents of corpora of scientific publications in the context of the SKM Project (see Figure 1).

Our initial target collections of contents to analyze include open access Web sites of publishers, conferences as well as any kind of on-line repository of scientific publications. The crawler gathers papers and metadata (name of the conference, editors of a journal paper, etc.) from the input Web sites. The original paper (in PDF or XML) is stored in a repository together with its metadata. Then, the contents of each paper are

[5] `http://backingdata.org/dri/library/`

[6] `http://penggalian.org/annote/` - username: user, password: pswd

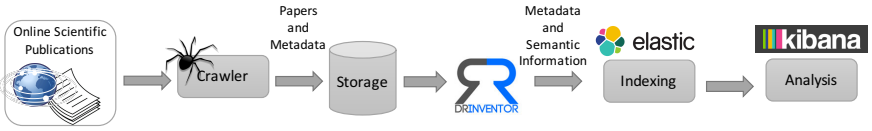

Fig. 1. Steps/components of our architecture

analyzed thanks to the DRI Framework. Both the metadata of a paper and the semantic information mined by the DRI Framework are properly indexed thanks to a mature open source engine: Elastic Search[7]. This platform is based on Lucene and has been designed to efficiently search across multiple documents, stored using the JSON format. Contents from different papers are linked by applying title and author normalization procedures. We will explore and analyze the collections of papers by directly querying Elastic Search by a graphical interface named Kibana. In Figure 2 we show some preliminary visualization of the information mined from a paper by the DRI Framework. These visualizations can be accesse on-line at: http://backingdata.org/dri/viz/.

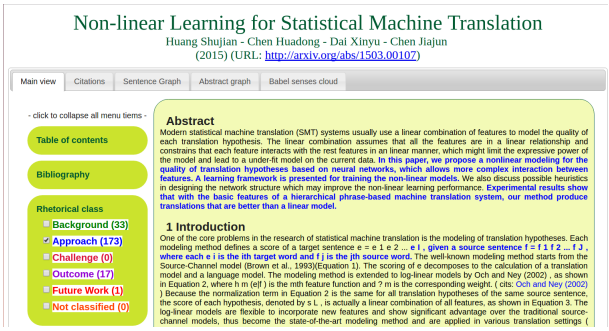

Fig. 2. Web based visualization of the information extracted from a paper thanks to the DRI Framework. In particular, we can see highlighted in bold the sentences of the paper classified as approach.

4 Scientific information analysis use cases

In this section we briefly present the three core use cases we are going to investigate in the context of the SKM Project. Even if our initial investigations will be focused on the exploration of these three application scenarios, the scientific publication mining infrastructure that constitutes the core of the SKM Project (see Section 3) can be easily adapted and thus exploited in any other context related to the analysis of large corpora of papers.

[7] https://www.elastic.co/products

4.1 Characterization of citations' purpose and polarity

The network of citations across papers constitutes one of the most characteristic traits of scientific publications: when a paper cites the work presented in another one the author explicitly identifies a relevant connection among both works. The count of the citation that a paper receives constitute the basis of the most common metrics exploited to evaluate the scientific production of papers, journals and researchers (i.e. h-index). The effectiveness of citation-based research evaluation metrics would benefit from the possibility to take into account not only the number of citations a paper receives but also the purpose and the polarity of each one of them. Several classification schemata and approaches have been proposed to characterize aspects related to the purpose and polarity of citations [2,12]. By relying on and extending the set of annotated citation included in the Dr. Inventor Multi-Layered Annotated Corpus of Scientific Papers [5][8], we aim at exploring new approaches to citation purpose and polarity classification, by placing special attention on their robustness across domains and on the limited availability of manually annotated data.

4.2 Scientific document summarization

Nowadays, the possibility to automatically identify the most relevant contents across a set of scientific publications is essential to deal with and perform screenings of the huge amount of articles currently available on-line. Several approaches to scientific papers summarization have been proposed [3, 8, 11]. Most of them extend general purpose document summarization methodologies by considering information facets that are characteristic of scientific publications. In particular, the sentences of the papers in which the article to summarize is cited provide valuable material to improve the quality of scientific summarization. Also the possibility to consider the rhetorical structure (background, approach, future work, etc.) of the different excerpts of the contents of a paper to summarize provides valuable information to generate summaries that include contents better balanced across the sections of a paper. In the SKM Project, we aim at investigating different strategies to improve content and graph-based summarization approaches by considering typed citation networks and by relying on the automated characterization of the rhetorical structure of scientific publications implemented by the DRI Framework.

4.3 Recommender system for citations

Citation recommendation is a complex task because of the difficulty in matching excerpts of the source paper to the contents of huge amounts of other candidate articles to be cited. Among the many approaches proposed, many of them rely on text classification as well as on question answering and query ranking [4, 6]. The goal of the SKM Project is to develop a recommender system for citations, that helps authors to find relevant articles by relying both on the semantic information extracted by the DRI Framework and on the data aggregated across corpora of papers crawled from the Web. In order to test our system, we will define a prediction task, where learning from the past, we will try to predict which citations a given article will contain.

[8] http://sempub.taln.upf.edu/dricorpus/

5 Conclusions

We introduced Scientific Knowledge Miner (SKM), a project that will facilitate the extraction of knowledge from scientific publications. We briefly described the SKM scientific publication mining infrastructure that will be exploited to analyze corpora of scientific papers, thus supporting large-scale investigations of scientific contents. We also presented our future venues of research by describing the three main application scenarios that we plan to investigate in the near future in the context of the SKM Project: characterization of the purpose and polarity of citation, summarization of scientific document and citation recommedation.

Acknowledgements. This work is supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502), by the European Project Dr. Inventor (FP7-ICT-2013.8.1 - Grant: 611383), the Catalonia Trade and Investment Agency (*Agència per la competitivitat de l'empresa*, ACCIÓ) and the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE).

References

1. The rise of open access. Science 342(6154), 58–59 (2013)
2. Abu-Jbara, A., Ezra, J., Radev, D.R.: Purpose and polarity of citation: Towards nlp-based bibliometrics. In: HLT-NAACL. pp. 596–606 (2013)
3. Abu-Jbara, A., Radev, D.: Coherent citation-based summarization of scientific papers. In: Proc. of 49th Annual Meeting of the ACL: Human Language Techologies. pp. 500–509. ACL (June 2011)
4. Balog, K., Ramampiaro, H., Takhirov, N., Nørvåg, K.: Multi-step classification approaches to cumulative citation recommendation. In: Proc. of the 10th Conference on Open Research Areas in Information Retrieval. pp. 121–128. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE (2013)
5. Fisas, B., Ronzano, F., Saggion, H.: A multi-layered annotated corpus of scientific papers. In: LREC Conference (2016)
6. He, Q., Kifer, D., Pei, J., Mitra, P., Giles, C.L.: Citation recommendation without author supervision. In: Proc. of the fourth ACM international conference on Web search and data mining. pp. 755–764. ACM (2011)
7. Nakov, P.I., Schwartz, A.S., Hearst, M.A.: Citances: Citation sentences for semantic analysis of bioscience text. In: In Proc. of the SIGIR'04 workshop on Search and Discovery in Bioinformatics (2004)
8. Ronzano, F., Saggion, H.: Taking advantage of citances: citation scope identification and citation-based summarization. In: Text Analytics Conference (2014)
9. Ronzano, F., Saggion, H.: Knowledge extraction and modeling from scientific publications. In: Semantics, Analytics, Visualisation: Enhancing Scholarly Data Workshop co-located with the 25th International World Wide Web Conference April 11, 2016 - Montreal, Canada (2016)
10. Saggion, H.: Summa: A robust and adaptable summarization tool. In: Traitement Automatique des Langues. vol. 49.2 (2008)
11. Teufel, S., Moens, M.: Summarizing scientific articles: experiments with relevance and rhetorical status. Computational linguistics 28(4), 409–445 (2002)
12. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proc. 2006 conference on empirical methods in NLP. pp. 103–110. ACL (2006)