Semantic-Powered Research Profiling

Zhixiong Zhang
National Science Library
Chinese Academy of Science
Beijing, China
+86-10-82629426
zhangzhx@mail.las.ac.cn

Ying Ding
Indiana University
1320 E 10th
Bloomington, IN
+1-812-855-5388
dingying@indiana.edu

Na Hong
National Science Library
Chinese Academy of Science
Beijing, China
+86-10-82629426
hongn@mail.las.ac.cn

ABSTRACT

Research profiling is a widely-adopted method to monitor research development and rank research performance. This paper describes a novel infrastructure to generate semantic-powered research profiling for research fields, organizations and individuals. It crawls related websites and news feeds, extracts research terms, research objects and relations from these resources and uses the proposed Research Ontology to model them into RDF triple stores to facilitate semantic queries and semantic mining on novelty detection, hot topic detection, dynamics of research, and topic clustering. The authors implement a research profiling experiment in Artificial Intelligence area to show the effectiveness of the research profiling based on semantic mining.

Categories and Subject Descriptors

I.2.4 [**Knowledge Representation Formalisms and Methods**]: Semantic Networks

General Terms

Measurement, Design

Keywords

Research Profiling, Semantic mining, Knowledge Object Extraction, Visualization

1. INTRODUCTION

Research profiling is a widely-adopted method to monitor research development and rank research performance within a certain research field [1] [2]. It gathers related research materials via automatic crawling on organizational websites, related news feeds, personal websites, online journals and related databases, extracts valuable data by automatic information extraction tools, and creates evaluation metrics based on co-occurrence analysis and other research policy indicators. Through the daily monitoring the selected research field, it can obtain a "big picture" on the research activity, understand the research community, gain insight into how innovation is progressing, and map (graphically represent) topical interrelationships for a whole research field.

Related work on research profiling include [1],[2],[3],and [4]. Most of them either have limited coverage of crawling or shallow statistical analysis. For example, Alan Porter and his team have tried to mine the Internet for competitive technical intelligence (CTI). They tried to bring Research Profiling into Web resources mining to discover competitive intelligence in commercial area through Google Soap Search API [4]. But it focuses on statistics of search results rather than deep analysis of text contents.

This demo reports one of the major outcomes of a project named Science Monitoring and Evaluation based on Scientific Web Resources (SMESWR), which is funded by National Key Technology R&D Program in the 11th Five Year Plan of China. The main goal of this project is to form a comprehensive methodology on automatically (or semi-automatically) extracting intelligence from web resources especially for scientific research analysis. Developing technologies to detect scientific research activities, to monitor the progress of one research field, to track the evolution of one research topic or a research community, and to profile the key research unit is the heart of the project.

2. SMESWR Infrastructure

This framework consists of five major components: *(1) Web resource collecting.* We collect important institutional websites, news websites and newsgroups of related research fields, RSS from related website, OAI repository of one institution, personal homepages and blogs of one researcher from the Internet. *(2) Semantic knowledge extraction.* We named these extracted research terms and research objects as knowledge objects, and model them based on the proposed Research Ontology. The main classes of this ontology include *Research Activity*, *Research Outcome*, *Research Organization and Person*, *Research Facilities* and *Basic Concepts*. These classes can be classified further into more specific subclasses. For example, *Research Activity* includes *Project, Conference, Lecture, Research Award, Experiment, Investigate* and *Training*. In addition to hierarchical structure of the concepts, the Research Ontology also describes relationships between the research objects. For example, we could define *has_attendees* relation between *Research Activity* and *Person*, *supports* relation between *Foundation* and *Project*, etc. We utilize this ontology with other data extraction algorithms to achieve more refined extraction results. *(3) Knowledge repository construction.* Knowledge repository is composed of a series of extracted knowledge objects with timestamps, for example, an instance of structure "class, research object, harvest time" is "research project, Science Monitoring and Evaluation based on scientific web resources, 2009-01-01"; Based on this computable data structure, knowledge repository is more clear and effective for future analysis. *(4) Semantic mining.* By using a set of co-occurrence analysis and semantic mining methods, we perform semantic mining based on the data stored in knowledge repository, try to form panoramic perspective of a specific research field, perform novelty detection, hot topic detection, timeline tracking and topic clustering to discovery knowledge hidden behind the web resources. *(5) Research profiling.* In this process, we perform visualization analysis to profile the targeted research field. We detect the scientific research activities in one research field, figure out the key components in the area, depict the relation between those components, monitor the progress of one research field, track the evolution of one research topic or a research community.

Based on the Research Ontology mentioned above, we construct a knowledge repository containing all the classes, relationships and their instances. This semantic repository is built upon RDF triple stores and stored in MS SQL Server. It provides information retrieval, inference, and statistics interfaces to enable intelligent semantic queries and reasoning.

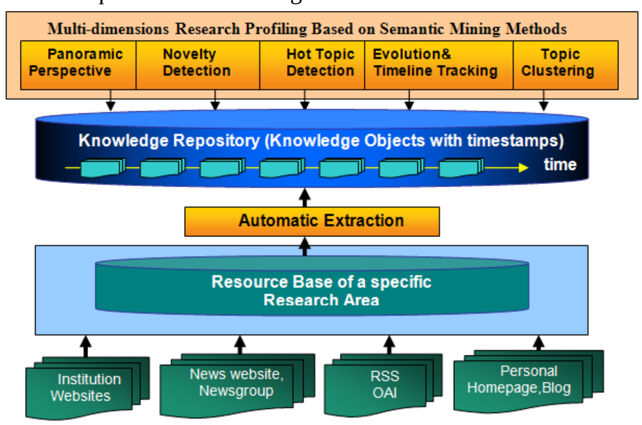

Figure 1.SMESWR Infrastructure

3. Implementation and Evaluation

At present, we choose "Artificial Intelligence" domain as our test domain. We have obtained 444 web site seed and 166 RSS harvest seeds. After the filter steps, we collected 89705 unique webpages and 18591 news articles. Based on these data, deep data mining has been conducted: through analyzing and scoring changes of each research object along the time axis, we identified the most important research objects in AI domain which provides the "big picture" of this field; tracking timeline, we use curve figures to illustrate the historical development of a certain knowledge object, and predict its future development trends; by analyzing the relationships among extracted knowledge objects, we interlink the research profiles of different objects, such as, relevant websites, projects, conferences, research personnel, funds, awards and laboratories related to a certain knowledge objects; and via clustering the top ranked research terms (e.g., top 2000 terms) along the timelines, the hot topics in AI have been visualized (Figure 2), more results displayed on the web portal*.

Comparing with related work, Arnetminer [5] mainly crawled academic personal websites and provides integrated overview of one researcher, but it does not conduct hot topic detection. CiteSeer [6] and GoogleScholar [7] are two of the largest collection of academic articles and provide basic citation data and analysis, but it cannot monitor one specific organization or research field. Furthermore, all of them do not crawl news articles and take them into consideration for research evaluation.

* http://124.16.154.12/

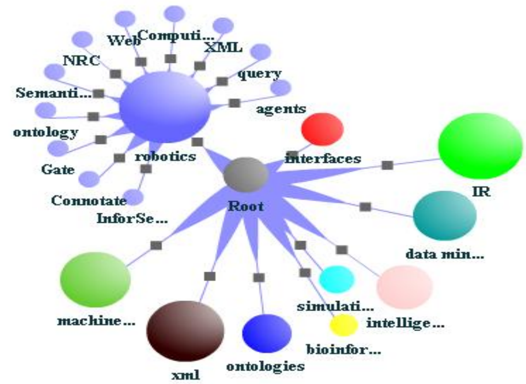

Figure 2. Hot Topic Detection

4. CONCLUSION

This paper describes a novel infrastructure to generate semantic-powered research profiling for research fields, organizations and individuals. It crawls related websites and news feeds, and on the basis of Research Ontology, they were modeled into RDF triple stores to facilitate semantic queries and semantic mining on novelty detection, hot topic detection, dynamics of research, and topic clustering based on the large-scale database. In the future, with the continuous accumulation of data, we hope to cluster the research terms periodically and tracking timeline of topic, and find topic changes through comparing every clustering result. Another important and hard-working task is to try to refine the extraction result to improve the performance. We also need to test the scalability and efficiency of our approaches and link our data with other Linked Open Data sets.

5. REFERENCES

[1] Bollen, J., Rodriguez, M. A., & Van De Sompel, H. (2007). MESUR: Usage-based metrics of scholarly impact. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (Vancouver, BC, Canada, June 18-23, 2007), 474.

[2] Carr, L., Bechhofer, S., Goble, C., & Hall, W. (2001). Conceptual linking: Ontology-based open hypermedia. In Proceedings of the 10th International WWW Conference (Hong Kong, China, May 01 - 05, 2001), 334-342.

[3] Handschuh, S., Staab, S., & Ciravegna, F. (2002). S-CREAM | Semi-automatic CREAtion of Metadata. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02) (Siguenza, Spain, October 1-4, 2002).

[4] Alan L. Porter, David J. Schoeneck, et al. 2007. Mining the Internet for Competitive Technical Intelligence. Competitive Intelligence Magazine 10 (May 5), 25-28

[5] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, & Zhong Su 82008). ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008) (LAS VEGAS,US, Aug. 24-27) 990-998.

[6] CiteSeer: http://citeseer.ist.psu.edu/

[7] GoogleScholar: http://scholar.google.com