Research Profiling Based on Semantic Mining

Zhi-xiong Zhang[1], Jian-hua Liu[1], Ying Ding[2], Na Hong[1]

[1]National Science Library, Chinese Academy of Science, China
{zhangzhx, liujh,hongn }@mail.las.ac.cn
[2]School of Library and Information Science, Indiana University, US
{dingying}@indiana.edu

Abstract. Web resources are a kind of useful information that could be used to dynamically reflect the changes and development of scientific researches. This paper presents a complete framework to generate research profiling based on scientific web resource using semantic mining techniques. It contains web resource crawling, semantic knowledge extraction, semantic mining and research profiling. At present, the authors take a pilot test in Artificial Intelligence area to show the validity of this proposed framework.

Key words: Research Profiling; Semantic Mining; Knowledge Extraction;

1 Introduction

Web resources are a kind of useful information that could be used to dynamically reflect changes and development of scientific researches. But they are unstructured, poor in semantic meaning, and sometimes unreliable and unstable [1]. It is a great challenge to depict the research development based on these web resources.

Nowadays, the authors are undertaking a project named Science Monitoring and Evaluation based on Scientific Web Resources (SMESWR), funded by National Key Technology R&D Program in the 11th Five Year Plan of China. The project's goal is to (semi) automatically monitor and profiling the development of scientific research. The core task of this project is research profiling [2] [3] [4] based on web resources using semantic mining techniques. Through semantic mining on web resources, people could form panoramic perspective of a specific research area, track evolution of one research topic or a research community, observe related objects in a research domain and gain interrelationship between research topics for a research area.

This paper just presents the profiling framework, a major outcome of SMESWR project, to generate research profiling based on web resources using semantic mining.

2. Overview of Framework

To achieve the goal, the overall proposed profiling framework (Figure 1) consists of three main layers: (1) Raw data Crawling layer. In this layer, we collect high relevant

scientific websites such as institutional websites, news websites, RSS feed in certain research fields and crawl regularly, and then clean the crawled web pages to obtain useful information from them such as title, main body text for further extraction and analysis. (2) Semantic knowledge extraction layer. For profiling, all the semantic knowledge we need contains research terms, research objects and their relations. All the knowledge components are organized in a pre-defined Research Ontology[1] referring to SWRC. To extract the instances of each knowledge component from the web resources automatically, we combine lexical-pattern approach and statistical approach together based on some matured NLP open source software such as GATE[2]. Then we transfer each extracted knowledge unit with its timestamps into a pre-defined knowledge repository. Take the structure "class, research object, harvest time" for example, one instance of this structure stored is "project, Science Monitoring and Evaluation based on scientific web resources, 2009-01-01".This computable knowledge repository will play an important role in future analysis. (3) Analysis and Visualization layer. Thanks to the semantic data stored in knowledge repository and a set of co-occurrence analysis, statistic analysis and semantic mining methods, we try to perform burst detection, hot topic detection, timeline tracking and relation mining to find semantic knowledge hidden behind the web resources. Then we could form panoramic perspective of a specific research area, detect research activities in certain research filed, track the evolution of one research topic or a research community, observe related objects within certain research domain and gain the inter-relationship between research topics for a whole research area and so on.

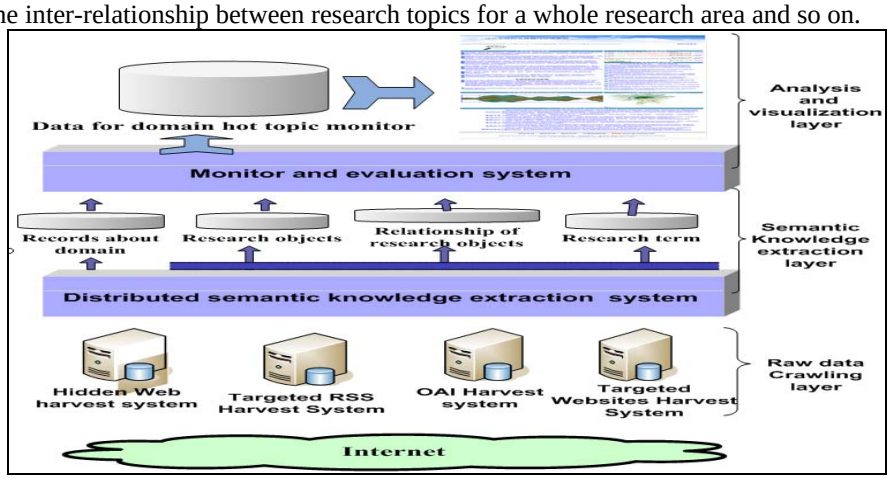

Figure1 SMESWR Framework

3. Implementation and Evaluation

At present, we choose "Artificial Intelligence" research area as test domain. Based on the proposed profiling framework, we identify the most important research objects,

[1] http://124.16.154.12/HotPortal/ResearchOntology.owl
[2] GATE Home. http://gate.ac.uk/: http://gate.ac.uk/

visualize whole structure of one research topic, use curve figures to intuitively illustrate the historical development of specific knowledge object, and predict the future development trends of a knowledge object; connect each knowledge object into relation networks in AI (Figure 2). More details could be found on the portal[1].

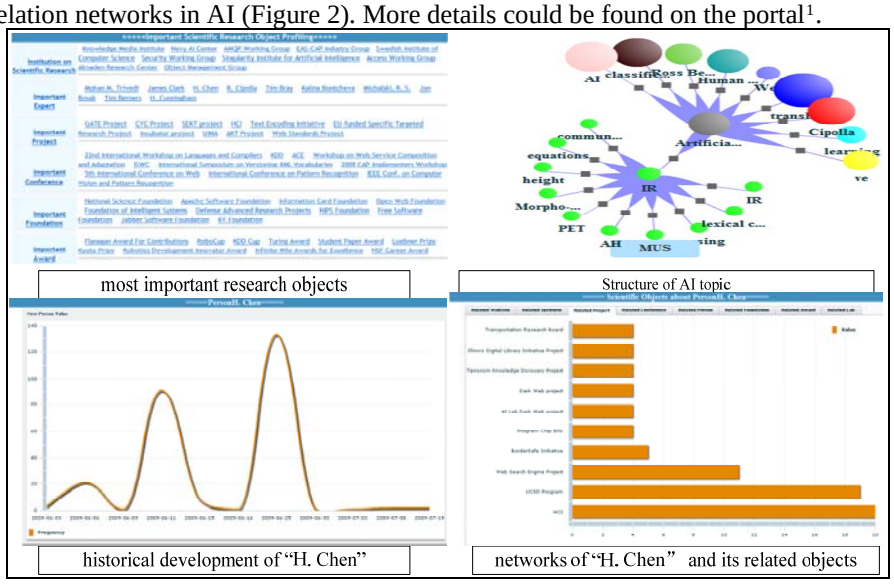

Figure 2 some profiling result of AI research filed

4. Conclusion

In the future, with the continuous data accumulation, we hope to cluster the research terms periodically, track timeline of topic and find topic changes. Besides, it is important to improve the extraction performance and test the scalability and efficiency of this method and link our data with other Linked Open Data sets.

5. REFERENCES

1. Zhang Zhixiong, Xu Jian, Liu Jianhua etc. Extraction Knowledge Objects in Scientific Web Resource for Research Profiling. In: Proceedings of International Conference on Machine Learning and Cybernetics, Baoding, July, 2009
2. Porter, A. L., A. Kongthon, et al. "Research profiling: Improving the literature review." Scientometrics, Vol.53, No.3 (2002) 351-370
3. Alan L. Porter, David J. Schoeneck, et al. Mining the Internet for Competitive Technical Intelligence. Competitive Intelligence Magazine, Vol. 10, No. 5 (2007) 25-28
4. Bragge, J., Sami R., et al. Enriching Literature Reviews with Computer-Assisted Research Extraction. Case: Profiling Group Support Systems Research. Proceedings of the 40th Annual Hawaii International Conference on System Sciences, Hawaii (2007)

[1] http://124.16.154.12/HotPortal_English/index.jsp