DynamicView: Distribution, Evolution and Visualization of Research Areas in Computer Science

Zhiqiang Gao, Yuzhong Qu, Yuqing Zhai, and Jianming Deng

Department of Computer Science and Engineering, Southeast University, China
{zqgao, yzqu, yqzhai, jmdeng}@seu.edu.cn

Abstract. It is tedious and error-prone to query search engines manually in order to accumulate a large body of factual information. Search engines retrieve and rank potentially relevant documents for human perusal, but do not extract facts, or fuse information from multiple documents. This paper introduces *DynamicView*, a Semantic Web application for researchers to query, browse and visualize distribution and evolution of research areas in computer science. Present and historical web pages of top 20 universities in USA and China are analyzed, and research areas of faculties in computer science are extracted automatically by segmentation based algorithm. Different ontologies of ACM and MST classification systems are combined by SKOS vocabularies, and the classification of research areas is learned from the ACM Digital Library. Query results including numbers of researchers and their locations are visualized in SVG map and animation. Interestingly, great differences of hot topics do exist between the two countries, and the number of researchers in certain areas changed greatly from the year 2000 to 2005.

1 Introduction

The web is increasingly becoming the primary source of research areas to modern researchers. With millions of pages available from thousands of web sites, finding the distribution and evolution of research areas in different countries and regions is a problematic task. Imagine that a young Chinese researcher, who has just received his PH. D degree in *artificial intelligence*, is planning his future research. He may want to know how many people are doing relative researches, such as *machine learning*, *multi-agent system*, *knowledge representation*, etc. He may also want to examine history and prognosticate tendency of *machine learning*. Additionally, if he intends to be a visiting scholar in USA in the near future, finding differences of hot topics between the two countries is rather helpful. However, browsing web sites, extracting related information and analyzing this information is too time consuming for individuals. Therefore, we develop *DynamicView* to tackle this challenge.

In the following of the paper, we begin with the introduction of major components of *DynamicView* in Section 2. In Section 3, we describe key services. Related works are discussed briefly in Section 4. Lastly, conclusions and ongoing works are summarized in Section 5.

Y. Gil et al. (Eds.): ISWC 2005, LNCS 3729, pp. 1054–1058, 2005.

2 Major Components

Crawler. Hub pages (faculty lists) are found by human intervention, and the *Crawler* searches and stores the homepage of each faculty by link analysis. Top 20 universities in USA are chosen mainly according to the ranking of US News [1], but not the exactly same. Top 20 universities in China are selected in accordance with the ranking of Ministry of Education of China [2], with a few universities excluded whose web pages could not be accessed (May, 2005). Historical web pages of top 20 universities in USA are downloaded from the Web Archive [3].

Extraction Engine. English pages are processed automatically, while Chinese ones by hand due to its complexity. Extraction results of research areas, names of researchers and universities are stored into relational databases. Web pages of top 10 universities are browsed manually and 65 *cue phrases* indicating *start* positions for information extraction are obtained, such as *research areas*, *research interests*, etc. *End* positions may be character '.', html tag <p>, end of file, or the position where the window size exceeds 300. Meanwhile, 1274 *pattern phrases* used for KMP algorithm are obtained. Combining *cue phrases* and KMP algorithm with segmentation of pages for each faculty, the average performance of our algorithm reaches 68.00% recall and 73.11% precision.

Ontology Learner. The ACM digital library [4] is utilized to learn classification of research areas. Each research area is input as a keyword, and top 60 papers returned with primary and additional classifications are used as training samples. Three cases of classification distribution may occur. 1) If one peak exists, the peak classification is the answer. 2) If more than one peak exist, and they belong to the same super classification, the super classification is the answer. 3) If more than one peak exists but they belong to different super classifications, or there is no peak, the classification is specified by human interaction. For each research area, the following relations are defined in SKOS (Simple Knowledge Organisation System) [5] vocabularies: *skos:prefLabelENG*, *skos:prefLableCHN*, *skos:altLabelENG*, *skos:altLabelCHN*, *skos:narrowerACM*, *skos:narrowerMST*, *skos:broaderACM* and *skos:broaderMST*. ENG means the label is expressed in English, and CHN in Chinese. ACM refers to the ACM Computing Classification System (1998)[6], with MST to classification and code of disciplines GB/T 13745/92 by Ministry of Science and Technology, China.

Query Processor. Users may query by country (USA or China), ontology (ACM or MST), hot topics and history. Note, users have to install SVG Viewers [7] to see SVG (Scalable Vector Graphics)[8] maps and animation.

[1] http://www.usnews.com
[2] http://www.cdgdc.edu.cn/zhxx/index.jsp
[3] http://www. archive.org
[4] http://portal.acm.org/dl.cfm
[5] http://www.w3.org/TR/swbp-skos-core-guide/
[6] http://www.acm.org/class/1998/ccs98.html
[7] http://www.adobe.com/svg/viewer/install/main.html
[8] http://www.w3.org/Graphics/SVG/

3 Key Services

Distribution of researchers in different countries and areas based on different ontologies. Given as an example, the number of researchers in *artificial intelligence* is shown in Fig.1, which are 327 (USA, ACM), 315 (USA, MST), 125 (China, ACM) and 169 (China, MST), respectively.

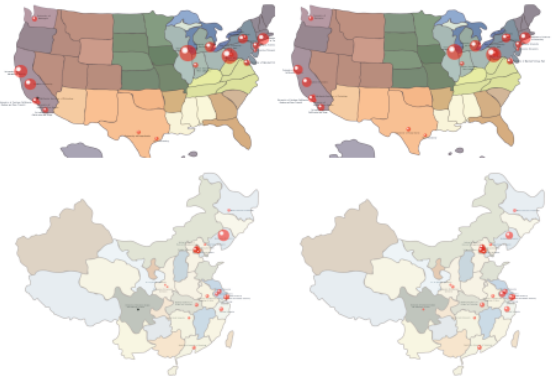

Fig. 1. Distribution of researchers in *artificial intelligence* according to ACM (left) and MST (right) ontologies in USA (top) and China (bottom)

Distribution of hot topics in different countries. Top 10 hot topics are deduced from original research areas with synonym relations, as depicted in Fig. 2 (top). Grey color refers to USA and pink color China. Surprisingly, the 1st

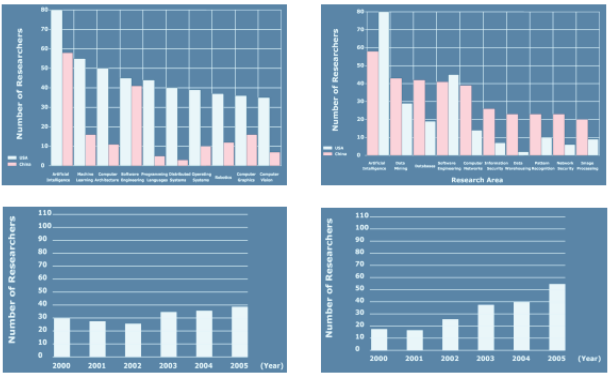

Fig. 2. Distribution of researchers in top 10 hot topics (top) in USA (left) and China (right), as well as evolution of hot topics (bottom) of *operating system* (left) and *machine learning* (right)

and 4th hot topics in two countries are the same: *artificial intelligence* and *software engineering*. But the other 8 hot topics are totally different. It seems that researchers in USA prefer theory and foundation, including *machine learning*, *computer architecture*, *programming languages*, *distributed systems* , *operating systems*, *robotics*, *computer graphics* and *computer vision*. By contrast, Chinese researchers emphasize application, including *data mining*, *databases*, *computer networks* , *information security*, *data warehousing*, *pattern recognition*, *network security* and *image processing*.

Evolution of hot topics from the year 2000 to 2005. The number of researchers in some research areas does not change significantly such as *operating systems*. Meanwhile, The number of researchers in other areas such as *machine learning* has increased nearly 2 times, as demonstrated in Fig. 2 (bottom).

4 Related Works

CS AKTive Space [1] provides a way to explore the UK computer science research domain across multiple dimensions for multiple stakeholders, from funding agencies to individual researchers. *Flink* system [2] extracts, aggregates and visualizes online social networks from a number of information sources including web pages, emails, publication archives and FOAF profiles. However, *D*ynamicView faces a much larger challenge, namely to extract and demonstrate research areas of computer science in different countries, languages, ontologies and over time. To the best of our knowledge there is no well-established approach for this task.

5 Conclusions and Ongoing Works

*D*ynamicView is designed for researchers of computer science to visualize the distribution and evolution of research areas in top 20 universities of USA and China. Research areas are extracted automatically by segmentation based algorithm with the performance of 68.00% recall and 73.11% precision. In order to combine different ontologies of ACM and MST, SKOS vocabularies are extended. Classifications of research areas are learned from the ACM Digital Library by analyzing the peaks of classification distribution. Except for *artificial intelligence* and *software engineering*, the other 8 hot topics in the two countries are different. The number of researchers in some areas such as *machine learning* has changed greatly in the past 6 years. In the near future, we will design a link grammar based information extraction algorithm to detect new areas[9].

Acknowledgments

This work is supported in part by National Key Basic Research and Development Program of China under Grant 2003CB317004, the NSF of Jiangsu Province, China, under Grant BK2003001, Hwa-Ying Culture and Education Foundation as well as Ministry of Education of China under Grant 6809001001.

[9] http://xobjects.seu.edu.cn/DynamicView/index.html

References

1. Nigel R. Shadbolt, Nicholas Gibbins, Hugh Glaser, et al.: Walking Through CS AKTive Space: A demonstration of an integrated Semantic Web Application. Journal of Web Semantics, volume 1, issue 4. 2004
2. Peter Mika: Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. Journal of Web Semantics, volume 3, issue 2. 2005