International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:4, 2009

Semi-Automatic Trend Detection in Scholarly Repository Using Semantic Approach

Fereshteh Mahdavi, Maizatul Akmar Ismail, and Noorhidawati Abdullah

***Abstract*—**Currently WWW is the first solution for scholars in finding information. But, analyzing and interpreting this volume of information will lead to researchers overload in pursuing their research.

Trend detection in scientific publication retrieval systems helps scholars to find relevant, new and popular special areas by visualizing the trend of input topic.

However, there are few researches on trend detection in scientific corpora while their proposed models do not appear to be suitable. Previous works lack of an appropriate representation scheme for research topics.

This paper describes a method that combines Semantic Web and ontology to support advance search functions such as trend detection in the context of scholarly Semantic Web system (SSWeb).

***Keywords*—**Trend, Semi-Automatic Trend Detection, Ontology, Semantic Trend Detection.

I. Introduction

THE term trend describes the continuing directional change in the value of an indicator [1]. Trend is a foundation of technical analysis in business such as in financial markets to show traders the variation in market prices. It can be used in fashion forecasting to support designer, manufacturer and retail businesses or in science such as in biology, chemistry and weather forecast.

Trend is typically classified by their changes as upward, downward and sideways. An upward trend means increasing in volume whereas downward trend indicates decreasing in volume and sideways trends is used when changes are only rather small. For instance, upward trend in scientific corpora means systematic and extended rise in the relevant and popular topics of the given documents over some period of time. Trend detection system takes a large collection of textual data as input and identifies topics that are emerging in a trial period.

The main two reasons for trend detection are to extract the useful information on a particular time series and to make it possible to forecast future events [2]. Classical trend detection methods employ some text mining techniques for detecting topics then monitors these topics over times and defines whether these topics are emerging or not.

Authors are with the Department of Information Science of University Malaya (UM), Kuala Lumpur, Malaysia (e-mail: fdotnet@yahoo.com, maizatul@um.edu.my, noorhidawati@um.edu.my).

Current trend detection methods fall into two categories: fully-automatic and semi-automatic. In fully-automatic approach, a list of emerging topics is developed afterwards researcher pursues these topics and the evidence to determine which truly emerging trends are.

Semi-automatic approach on the other hand, requires user to input a topic then it provides the user with the evidence whether the input topic is truly emerging [3]

Most of the existing trend detection systems focus on keyword matching, statistical techniques and link analysis. Currently there are emerging efforts to employ Semantic Web technologies to provide enhance information search and retrieval mechanisms [3].

The Semantic Web has been desired as an extension of the current Web, which makes a well-defined meaning of information to enable a better connection between computers and people to work together. Ontology is used to visualize knowledge on the Semantic Web. Recently, ontology is a formal representation of a set of concepts within a domain into a machine-readable format that is also understandable by humans, consisting of entities, attributes and relationships [4].

This research implements trend detection system in scholarly repository to improve traditional search (free-text search) using Semantic Web, ontology and identifying the concepts in the search query.

II. Classical Web-Based Information Retrieval Methods

Some examples of trend detection using classical information retrieval or text mining methods, are Technology Opportunities Analysis System (TOAS), Constructive collaborative Inquiry-based Multimedia E-Learning (CIMEL), ThemeRiver, Envision, TimeMines, Hierarchical Distributed Dynamic Indexing (HDDI), PatentMiner, Emerging Topic Tracking System (ETTS), Sequence-Based Self-Organizing Map (SBSOM) have been developed.

TOAS is a semi-automatic trend detection system developed for technology opportunities analysis. This system extracts query relevant documents and provides analysis of data by using word counts, data information, citation information and publication information to track activity in a subject area. TOA facilitates the analysis of the data available within the documents by showing lists of frequently occurring keywords or lists of author affiliations, countries, or states [5][6].

CIMEL is a multi-media framework which uses a semi-automatic trend detection method to enhance computer

science education. The students input the topic which they want to see the main topic area of their research and new conferences and workshops. The database of this system can be any Web resources [5] [7].

ThemeRiver is a fully-automatic method which summarizes the topics area and symbolizes a "river" of information for the changes in them over times. This system generates automatically a list of possible topics, called theme words, of which a subset is manually chosen. Then, it counts number of documents containing a particular theme word.

The Envision system focus more on visualization of textual data by enabling users to see trends in multimedia digital library of computer science literature, with full-text searching and full-content retrieval.

TimeMines is a fully- automated system which shows a ranked list of topics in corpus. Users can figure out how vital a topic is in the area. This system begins processing with the default model that assumes the distribution of a feature depends only a base rate of occurrence that does not vary with time. [5]

HDDI aims to identify features and methods to improve the automatic detection of emerging trends by generating clusters based on Semantic similarity of textual data. It uses neural network for classifying topics as emerging or non-emerging.

Most of the above systems use traditional information extraction techniques to extract features from the corpus but PatentMiner employs a sequential pattern matching technique that is often used in data mining systems [5] [7] [8].

ETTS is an intelligent software application for detecting and tracking the emerging topic (hot topic) from a particular information area on the Web. It uses a new (Term Frequency * Proportional Document Frequency) TF*PDF algorithm to detect the prominent topics in the changes. It crawls the Web, collects the changes and journalizes a summary of popular topics to the user[5][9].

SBSOM discover three aspects of a topic which are hotness, period and their relations within the map. SBSOM shows hotness within certain times, relations among topics, and period of topics [5] [10].

III. Semantic Web-Based Information Retrieval

As to date there is no evidence in the literature for trend detection system employing Semantic Web technology. However, a successful development of the Semantic Web applications depends on availability and adoption of ontologies and Semantic data [11]. Several works which apply ontology-based information retrieval in various domains are Swoogle ontology search engine [12], TAP generic Semantic search framework [13], and Semantic annotation platform in Knowledge and Information Management (KIM) [14].

Another research which is closely related to this study is done by Elsevier (the leading scientific publisher). Elsevier solves physical and syntactic heterogeneity problems by translating a large amount of its content to an XML format that allows cross-journal querying.

Although Semantic problem remains largely unsolved, searching through journals for articles containing user's query is not a complicated task but the point is among vast amount of journals providing satisfactory result is doubtful because there are extensive homonym and synonym problems. Elsevier overcome this free-text search problem by Elsevier's life science thesaurus named "EMTREE". EMTREE groups relevant terms in a specific area and provides a controlled vocabulary for indexing information as well.

As part of its experiments, Elsevier has sponsored the Drug Ontology Project for Elsevier (DOPE) project, where the EMTREE thesaurus was used to index approximately 10 million medical abstracts from MedLine as well as approximately 500,000 full text articles from Elsevier's own collections.[15]

IV. Semi-Automatic Trend Detection In Ssweb

This research employs semi-automatic trend detection method based on Semantic Web and ontology. It is implements Sesame Resource Description Framework (RDF) storage. The search and browse interface is constructed using Perl. Fig. 1 describes the system's process as follows:

Step 1: Retrieving Synonyms In Query Statements By Using WordNet:

This approach relies on WordNet lexical ontology. The system has an interface to the WordNet lexical reference system, which is responsible for retrieving synonyms in query statements.

Step2: Tokenized Abstract, Topic and Keyword:

In this step, system tries to collect tokens from abstracts, topic and keywords from Sesame RDF storage.

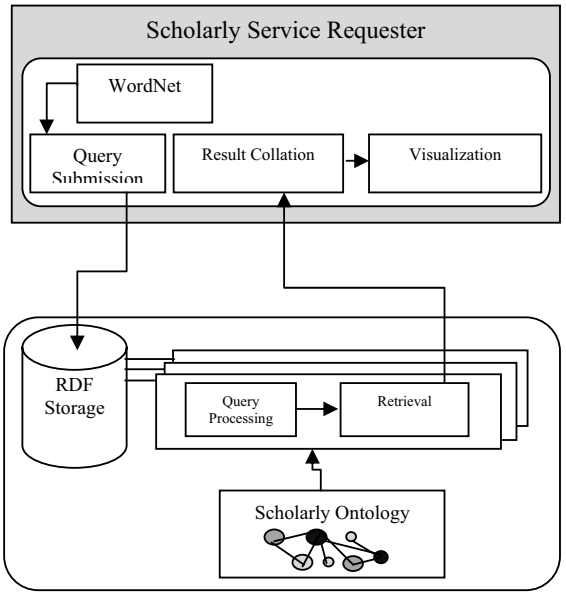

Fig. 1 Scholarly Information Retrieval Process

Step 3: Free-text Search Based on Query and Synonyms of Query:

In this step, documents which contain query statement and synonyms will be retrieved. The system keeps the number of these documents and represents it by histograms in special period time.

Step 4: Searching Ontology (try to find match category):

This step compares concepts in main ontology (sub classes of area) and query statements based on synonyms and other variations of keyword. If the search term is not an exact match for any of the concepts in the ontology the synonym matching is performed (See Fig. 2)

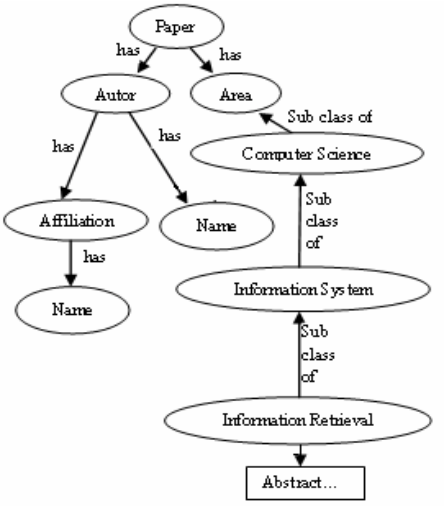

Fig. 2 Classes and subclasses of the Scholarly ontology

Fig. 2 represents a part of scholarly ontology that is used in this system. The entire ontology is much larger. This ontology indicates a paper with Author and Area. This figure specifies there is only one area named "Computer Science" which is super class of "Information System" and "Information Retrieval". Moreover, "Information Retrieval" class has a value which is abstract of the paper. On the other hand, each paper has authors and each author has name and affiliation as well.

Step 5: Search The Articles Attached To Nearest Super Category:

Finally, system will count the number of articles attached to selected concept. Repeated documents are ignored.

V. Conclusion

This paper presents the semi-automatic trend detection method for supporting retrieval of scholarly information through the Semantic Web. It also describes the system architecture of scholarly Semantic Web (SSWeb). Although the focus has been on scholarly documents, the basic principles describe in this paper can be adapted to other document types.

The future focus will be on other factors such as citation analysis and measure of growth in interest which also has influence on trend detection.

Acknowledgment

Fereshteh Mahdavi would like to thank Associate Prof. Dr Sameem Abdul Kareem as project leader. This research is funded by ministry of Higher Education under the Fundamental Research Grant #P096/2007l.

References

[1] Urquhart, S., Larsen, D. (1998) Monitoring For Policy-Relevant Regional Trends Over Time. Ecological Applications, 8.

[2] Box, G. (1976) Time Series Analysis: Forecasting And Control (2nd ed.), Holden-Day, San Francisco

[3] Aleman-Meza, B., Halaschek-Wiener C., Sahoo S (2005) Template Based Semantic Similarity for Security Application.

[4] Ismail M.A, Yaacob, M., Abdul Kareem, S. (2008) Semantic Support Environment for Research Activity. Journal of US-CHINA Education Review, 5, 36-51.

[5] Hoang, L. M. (2006) Emerging Trend Detection from Scientific Online Documents. Japan Advance Institute Of Science and Technology.

[6] Kontostathis, A., Galitsky, L., Pottenger, W., Roy, S. (2003) A Survey of Emerging Trend Detection in Textual Data Mining.

[7] Roy, S., Gevry, D. , Pottenger, W. (2002) Methodologies For Trend Detection In Textual Data Mining.

[8] Lent, B. A., Srikant, R. (1997) Discovering Trends In Text Databases. Third International Conference on Knowledge Discovery and Data Mining. California.

[9] Bun, K. K. (2005) Topic Trend Detection and Mining in World Wide Web. Japanese Society for Artificial Intelligence.

[10] Fukui, K., Saito, K., Kimura, M. , Numao, M (2004) SBSOM: Self-Organizing Map For Visualizing Structure In The Time Series Of Hot Topics. Joint Workshop of Vietnamese Society of AI, SIGKBS-JSAI, ICS-IPSJ, and IEICE-SIGAI on Active Mining.

[11] Shadbolt, N., Hall, W., Lee, B. (2006) The Semantic Web Revisited. IEEE Intelligent Systems.

[12] Ding, L., Finin, T., Joshi, A., (2004) Swoogle: A Search And Metadata Engine For The Semantic Web. 13th ACM International Conference On Information And Knowledge Management.

[13] Guha, R., McCool R.(2003), Semantic Web Testbed. Journal of Web Semantics.

[14] Kiryakov, A., Popov, B., Terziev, I. (2005) Semantic Annotation, Indexing, and Retrieval, Elsevier's Journal of Web Semantics.

[15] Harmelen, F., Antoniou, G. (2008) A Semantic Web Primer.