An Application of Topic Map-Based Ontology Generated from Wikipedia for Query Expansion

S. Eslami and E. Nazemi

***Abstract*—Topic maps are a Semantic Web technology for semantic annotation of resources to enhance the quality of search output. The main idea of this research is to present a query expansion method using topic maps-based ontology for query expansion process, furthermore this paper proposed a novel automatic approach to construct topic maps from Wikipedia XML corpus. Wikipedia is general purpose, freely available online, is containing up to date information so it is a suitable option for topic map development. The proposed model is implemented and then applied on a test collection. The results show that using topic map-based ontology in query expansion process improves search accuracy in keyword-based information retrieval.**

***Index Terms*—Ontology, information retrieval, semantic web, topic maps, query expansion.**

I. Introduction

Keyword-based information retrieval systems are widely used but have several problems related 1) to the quality of search results and 2) to the usage of systems [1] The low precision and recall of keyword-based search methods has many reasons [2]. For example, a keyword in a document does not necessarily mean that the document is relevant, and relevant documents may not contain the explicit word. Synonyms lower recall rate, homonyms lower precision rate, and semantic relations such as hyponymy, meronymy, and antonymy [3] are not taken into account. A prominent solution to these problems is to use ontology in keyword-based information retrieval systems. Some approaches utilize ontology for query expansion to find concepts that related to query concepts and removing the ambiguity of the query in such a way that system can understand user's intention [4].

One of the Semantic Web technologies is Topic maps [4]. They can be used as a means to organize and retrieve information in a more efficient and meaningful way [6]-[8].This paper describes a method which aims at improving search results quality and accuracy in keyword-based information retrieval systems. Its innovation lays in the ideas of:

- Using topic maps as knowledge base in query expansion process.
- Exploiting vast amounts of highly organized human knowledge encoded in Wikipedia to construct the topic maps-based ontology.
- Presenting an automatic approach for topic maps construction.

In our proposed approach, Wikipedia has been used as input in topic maps construction process and the resulted output used as knowledge base for query expansion process in a keyword-based information retrieval system. The rest of the paper is organized as follows: Section II will present the related work and the motivation for this paper. The proposed method is explained in Section III. Section IV shows the results of a prototype system implemented according to the proposed method and finally the paper is concluded in Section V.

Manuscript received January 8, 2013; revised May 20, 2013. This work was supported by Islamic Azad University of Qazvin (QIAU).

Saeedeh Eslami is now with the National Library and Archive of Iran Tehran, Iran (e-mail: s-eslami@nlai.ir, eslami.saeedeh@gmail.com, phone: +9881622440).

Eslam Nazemi was with Shahid Beheshti University (SBU), Tehran, Iran. He is now with the Electrical and Computer Engineering Faculty (e-mail: nazemi@sbu.ac.ir).

II. Related Works

In recent years, several researches try to apply otologies in information retrieval systems to overcome current search engines shortcomings. One of the solutions is query expansion methods which provide better query to improve retrieval process .Main objectives of query expansion is to overcome structural ambiguity and semantic ambiguity [9].They fills the gap between queries and documents by adding extra terms and reweighing terms in the original query [10].

Many search strategies use keyword based search with augmenting of ontology. These systems boost their performance using ontology-based query expansion which is oftentimes based on well-known WordNet ontology [10] or rely on knowledge resources such as thesauri [11] and inexistence ontologies [12], [13]. A considerable number of these attempts exploit Wikipedia. The idea to bring semantics into Wikipedia is not new; several studies on this topic have been carried out in the last few years. Wikitology is one this research using Wikipedia to make English Ontology [14]. Several different methods use Wikipedia for manipulating query such as [4]. [15] try to identify document topics using the Wikipedia category. [16] use Wikipedia for tagging a query with its category.

In recent years, various researches discussed methods to enhance topic map creation. Some solutions for topic map building exploit a specified classification. These solutions transferred the classification into a Topic Map. Ju ¨rgen Beier and Tom Tesche discussed a solution which transforms the MeSH classification of the National Library of Medicine with its approximately 19,000 entries into a topic map, thus, allowing an appropriate retrieval of medical information [17]. TOIR [7] exploit a semi-automatic created topic map to improve search quality.

Unfortunately, the current query expansion approaches use

inexistence ontologies which are too small and does not contain domain specific information [15]. As the accuracy of query expansion methods has been affected by ontology accuracy, thus a better and more general ontology will improve query expansion results. Furthermore, manual construction of topic maps are time consuming and most of methods restricted to utilize owl ontology representation language. On the other hand, Wikipedia is general purpose and very wide, containing up to date information making it a much better option. To solve these problems, we propose a method which build topic map based ontology automatically from Wikipedia and then use it in query expansion process.

III. Proposed Methodology

Our method consists of two main phases: Topic maps construction process, Query expansion process. The output of first phase is a topic maps-based ontology which is used as input of the second phase.

A. Topic Maps Construction Process

The topics of topic map model derived from Wikipedia. In our approach, Wikipedia is used because it is the largest knowledge repository on the Web and is available in dozens of languages. The largest Wikipedia corpus is its English version which is used in this research.

There are four Wikipedia features that are in particular attractive when building topic map-based ontology: internal links, redirects, disambiguations and categories [18], [19] describes these four features in detail. Each article has a title, assigned categories, and often refers to other articles. Some articles have more than one title; in this case, additional titles are implemented as special articles, or redirections, containing only a single link to the main article. Categories are organized hierarchically into sub and super categories. The category hierarchy is not a tree and some categories have multiple super-categories [15]. As depicted in

Fig. 1, first, we convert the Wikipedia corpus to appropriate form for topic map construction and second explain the constructing process.

1) Preperaing Wikipedia XML corpus

The original Wikipedia corpus in is not suitable directly for our approach. Before using it for topic maps development, it must be prepared for our purpose.

As mentioned, each article may have a number of titles, called redirections, and be associated to a number of categories. In our approach, for every article, only article titles, internal links of the article and the assigned categories are extracted. In this step, we obtain a list of concepts and their relations which will be used in topic map development. It doesn't utilize the text body of Wikipedia articles.

2) Generating topic maps

To construct topic maps, our approach exploits article's titles, categories of articles and the links between articles which obtained in previous step.

Each Wikipedia concept is represented as a topic in topic map model, thus titles, categories and links are considered as topic in topic map model. The text of each Wikipedia article contains hyperlinks to other articles in Wikipedia. We exploit two kinds of links: internal links and "see also" links. These two types of links are also mapped onto associations. The relation between each article and its categories are also mapped into an association in topic map model. We consider four association types for four kinds of relationship between two topics:

- *Part-of* association type specified the relation between categories and subcategories.
- *Assigned-to* association type is considered for the relation between each article and its assigned categories.
- *Has-a-relation-to* association type defined for the relation between internal links of an article and its title.
- *Has-SeeAlso* association type is considered for the relation between "see also" links of an article and its title.

Finally, the resulted topic map used as knowledge base for query expansion process. Fig. 2 shows small portion of the generated topic map.

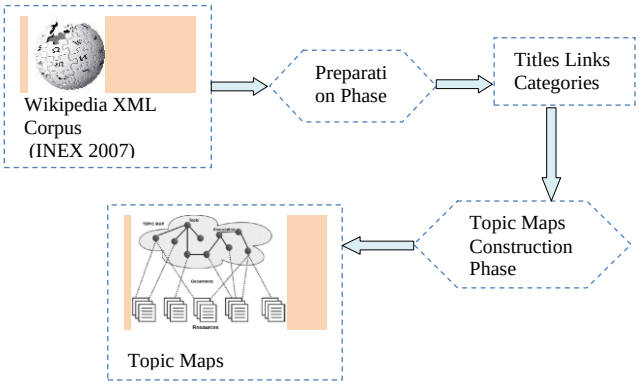


Fig. 1. Topic Maps construction using Wikipedia.

B. Query Expansion Process

It refers to the process that generates query expansion set for user's query keyword. According to the query expansion set, search program applies on index repository and search results with the characteristic of semantic information are returned to user. We proceed on the supposition that each query consists of n keywords, t_i denotes title topic in topic map, k_j shows keywords of the query, c_i indicated number of query keywords that exists in a title topic, tc_k indicates child topic of a each node up to two levels, $W_{j,i}$ denotes the weight of a keyword in a title topic and deg_i shows the emphasis degree of a topic on expansion process.

1) Step 1

Stopwords removed from user query and query keywords which are not present in any topics of topic maps are ignored.

2) Step 2

Find all title topics, t_i, that match query keywords. Each topic which has one matching keyword is selected. We save the number of matching keywords for each topic, c_i and use it for weighting in step3.

3) Step 3

For each title topic, t_i, and each query keyword, k_j, the keyword weight is calculated by considering each title topic in topic map-based ontology as father topic and selecting child topics, tc_k , up to two levels in topic map hierarchy. Child nodes are connected to each other by has-realtion-to,

has-Seealso association types. We consider a predefined weight for each of these relationships. Thus, the weight for each keyword is calculated as follows:

For tc_k such that $(k_j \in tc_k)$ and tc_i is a child node of t_i,

$$w_{j,i} = Maximum(c_i) \times \propto ,$$

Otherwise $w_{j,i} = 0$, we have $\propto$ as the relationship weight:

$\propto = 0{,}6$ for has-relation-to association type

$\propto = 0{,}7$ for *has-seealso* association type

4) Step 4

For each title topic, we calculate title importance by summing keywords weight according to following and consider it as the weight for that title topic:

$$deg_i = \sum_{j=1}^{n} w_{j,i}$$

5) Step 5

Sort title topics by weight and select topic with highest weight as candidate concepts for query expansion.

6) Step 6

We utilize a threshold value to omit concepts with lower weights from the candidate concepts and the remians are used for expansion process.

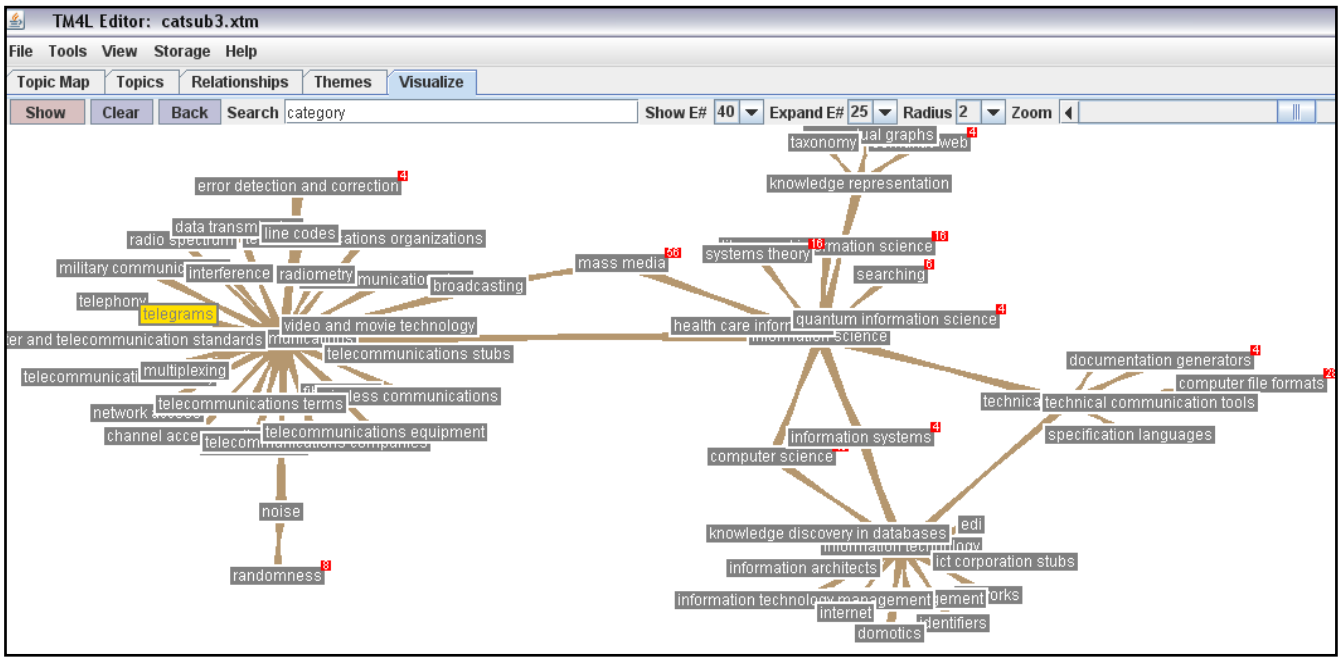


Fig. 2. A screenshot of small portion of the topic map generated as output of our method. It built from part of 'computer scinence' domain of wikipedia.

IV. Experimental Results

For experiments we used the Wikipedia XML corpus. The INEX community has produced a XML benchmark collections named INEX Wikipedia collection [20] which is available for the participants of INEX 2007.

This corpus is based on the English Wikipedia dump and contains about 659,388 articles. We exported the corpus into a SQL Server 2008 database and use it to construct the topic map-base ontology.

In order to validate the efficiency of our proposed method, we have constructed the topic maps-based ontology in computer science domain. The resulted topic maps contain approximately 2700 topics. The resulted topic map-based ontology validated using TM4L [21]. Then, 450 text documents have been collected for our experimental data set. We have implemented a prototype system which has keyword-based search then our query expansion method applied on system. Fig. 3 shows the experiment result with and without query expansion process. Precision measure used to evaluate the efficiency of search results. Precision is defined as proportion of retrieved related documents relative to all retrieved documents. It is showed that when the query is enriched by proposed query expansion approach, result precision will be higher.

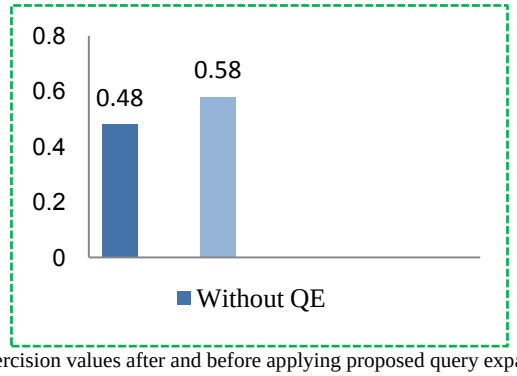


Fig. 3. Percision values after and before applying proposed query expansion approach using the gerentated topic-map based ontology.

V. Conclusions

In this research after an overview on approaches using ontology to improve information retrieval systems and approaches which exploits Wikipedia, a model for automatic construction of topic map-based ontology using Wikipedia XML corpus, is proposed. The prominent features of our approach are that it is an automatic method for constructing topic map-based ontology; it exploits Wikipedia, the largest up-to-date encyclopedia which is developed through a social process and kept current by the Wikipedia community. The

proposed method consists of two main phases: Topic maps construction process, Query expansion process. The output of first phase is a topic maps-based ontology which is used as input of the second phase. 0provides a summary of proposed method.

TABLE I: SUMMERY OF PROPOSED METHOD

Phase 1 : Topic Maps Construction Process	*Phase2: Query Expansion Process*
Step 1: Preparing Wikipedia XML Corpus	Step1:Remove stop words Step2: Find all topics that match query keywords. Step3: Set weight for each keyword.
Step 2 : Topic Maps Construction	Step4: calculate the weight for each title Step5: Sort topics by weight. Step6: apply threshold to omit less important topics.

In topic map constructing process, titles, categories and links are considered as topics in topic map model. Four association types are defined to express the relation between categories, internal links and article titles. Finally we proposed a query expansion method and applied it on the resulted topic map-based ontology. It is validated by using Google API and the results of experiments show that the precision of the retrieval system is improved by utilizing the proposed method. Our future work will consider the text of Wikipedia articles in topic map constructing process.

ACKNOWLEDGMENT

This works describes here is a research on topic map-based ontology for information retrieval at Islamic Azad university of Qazvin (QIAU).

REFERENCES

[1] E. Hyvonen, S. Saarela, and K. Viljanen, "Ontogator: Combining view-and ontology-based search with semantic browsing," in *Proc. Open Standards, XML, and the Public Sector, Kuopio*, Finland, October 30-31, 2003.

[2] F. V. Harmelen, I. Horrocks, J. Hendler, and D. L. Mcguinness, "The semantic web and its languages," *IEEE Intelligent Systems and Their Applications*, vol. 15, no. 6, pp. 67-73, 2000.

[3] C. Fellbaum, *WordNet. An electronic lexical database*, Cambridge, Massachusetts: The MIT Press, 2001.

[4] M. Farhoodi, M. Mahmoudi, A. M. Zare Bidoki, A. Yari, and M. Azadnia, "Query expansion using persian ontology derived from Wikipedia," *World Applied Sciences Journal*, vol. 7, no. 4, 2009.

[5] S. Pepper, "The TAO of topic maps," in *Proc. XML Europe*, 2004, [Online]. Available: http://www.ontopia.net/topicmaps/materials/tao.html

[6] *Document Description and Processing Languages–Topic Maps*, International Organization for Standardization ISO, ISO/IEC 13250: 2000, Geneva, 2000.

[7] R. Schweiger and J. Dudeck, "Improving information retrieval using XML and topic maps," in *Proc. Charitng the Topic Maps Research and Applications (TMRA)*, 2006, vol. 3873, pp. 253-262.

[8] M. Yi, "Information organization and retrieval using a topic maps-based ontology: Results of a task-based evaluation," *Journal of the American Society for Information Science and Technology*, vol. 59, no.12, pp. 1898-1911, 2008.

[9] H. H. Hoang and A. M. Tjoa, " The state of the art of ontology-based query systems: A comparison of existing approaches," in *Proc. IEEE International Conf. on Computing & Informatics*, Malaysia, 2006,

[10] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *Proc. the 17th annual international ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, Springer-Verlag Inc. 1994, pp. 61–69.

[11] Y. C. Wang, J. Vandendorpe, and M. Evens, "Relational thesauri in information retrieval," *Journal of the American Society for Information Science*, vol. 36, no. 1, pp.15–27, 1985.

[12] E. N. Efthimiadis, "Query Expansion," in *Martha, Annual Review of Information Systems and Technology (ARIST)*, E. Williams, Ed. 1996, vol. 31, pp. 121–187.

[13] G. Zou, B. Zhang, Y. Gan, and J. Zhang, "An ontology-based methodology for semantic expansion search," in *Proc. IEEE Conf. on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2008, vol. 5, pp. 453–457.

[14] Z. Syed, T. Finin, and A. Joshi, "Wikitology: Using Wikipedia as an ontology," in *Proc. the Second International Conf. on Weblogs and Social Media*, Seattle, Washington, USA, 2008.

[15] P. Schonhofen, "Identifying document topics using the Wikipedia category network," in *Proc. IEEE/WIC/ACM International Conf. on Web Intelligence*, 2006, pp. 456-462.

[16] M. Alemzadeh and F. Karray, "An efficient method for tagging a query with category labels using Wikipedia towards enhancing search engine results," in *Proc. IEEE/WIC/ACM International Conf. on Web Intelligence and Intelligent Agent Technology*, 2010, vol. 1, pp. 192-195.

[17] J. Beier and T. Tesche, "navigation and interaction in medical knowledge spaces using topic maps," *International Congress Series, Computer Assisted Radiology and Surgery*, Germany, vol. 1230, pp. 384-388, 2001.

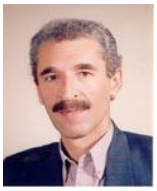
[18] M. Yi, "Information organization and retrieval using a topic maps-based ontology: results of a task-based evaluation," *Journal of the American Society for Information Science and Technology*, vol. 59 ,no.12, pp.1898-1911, 2008.

[19] C. Bøhn and K. Nørv˚ag, "Extracting named entities and synonyms from Wikipedia," in *Proc. IEEE International Conf. on Advanced Information Networking and Applications (AINA)*, 2010, pp. 1300-1307.

[20] L. Denoyer and P. Gallinari, "The Wikipedia XML Corpus," *INEX2006/SIGIR Forum/*, vol. 41, no. 1, pp. 64-69, 2006.

[21] TM4L. [Online]. Avaliable: http://compsci.wssu.edu/iis/nsdl/index.html

Saeedeh Eslami was born in 1983 in Tehran, Iran, She holds B.A. degree in Computer Software Engineering in 2006 and then graduated in M.A. in Computer Software Engineering in 2010. She is a fellow member of staff at National Library and Archive of Iran (NLAI) and has been working as a software specialist since 2005 and has participated prominently in NALI software projects. She is a software analyst and programmer of Software Architected and Design Group. She is a university lecturer and she teaches at Islamic Azad University and NLAI. Her research interest lies around semantic web with specific interest in developing linked data, ontologies and topic maps, free/libre open source software development and interoperability and so on. She has published several papers on such areas.



Eslam Nazemi was born in Sarab, Iran in 1954. He got the B.Sc. degree in Applied Mathematics and Operational Research from School of Planning and Computer Application, Tehran, Iran in 1977. He obtained the M.Sc. degree in both System Engineering and Economics in 1987 and 1996, and Ph.D. in Industrial engineering and Information technology in 2005, Iran. He was the faculty Member from 1978 in School of Planning and Computer application and then from 1986 to the present, he has been with the Electrical and Computer engineering Faculty at Shahid Beheshti Univversity (SBU), Tehran, Iran. He was the deputy of graduate and education affairs and now is the manager of informatics development of education in SBU. He is an associate professor of Computer Engineering Department. His main fields of research are self-software engineering, large scale software development, search engines, web mining, and self-adaptive software quality. He has authored and coauthored more the 90 papers in Journals and Conferences and has 10 books on mathematics, project management, software engineering, software quality and game theory.