Domain Knowledge Navigator—Topic Map for Practical Usages

Jiangning Wu Qiaofeng Liu Xiaohuan Wang

Institute of Systems Engineering
Dalian University of Technology, Dalian, 116024, China
jnwu@dlut.edu.cn, qfliu@student.dlut.edu.cn, xhwang@student.dlut.edu.cn

Abstract

This paper starts from a literature survey on Topic Map in the latest seven years in terms of the quantity of articles published in different databases, the trend of quantitative change, the distribution of affiliations and countries. The purpose of the survey is to show the research and development trend of Topic Map. Then a Topic Map in the domain of Customers' Service of Mobile Communication Corporations (MCCs) of China is constructed for knowledge navigation and information retrieval. Its building process contains topic selection, occurrence appending, and association analysis, while the implementing process involves domain knowledge navigation and information retrieval. Experiments have been made to compare the precision and recall of two kinds of information retrieval systems, i.e., keyword-based IR system and topic map-based IR system. From the application we can see that Topic Map is efficient in knowledge navigation. In the end, the paper addresses the future directions related to Topic Maps.

Keywords: Topic Maps, Knowledge Navigation, Information retrieval, MCC

1 Introduction

Topic Maps (TMs) are highly flexible and powerful standards for the organization and representation of knowledge and tools for providing access to knowledge. So we can say that TMs are able to convey knowledge of information resources in a semantic way. Accordingly, they are being embraced by a wide number of organizations and companies [1]. Considering the theoretical and pragmatic values of TMs, many related research papers have been published in recent years. So a literature survey in the years of 2000 to 2006 is first made for the following purposes: to outline the latest seven-year research results about TMs based on the collected papers from statistical point of view, and to review some of the current and potential application areas for TMs.

For practical use of TMs, this paper focuses on the domain of Customers' Service of Mobile Communication Corporations (MCCs) in China. It was reported that the number of mobile phone users has been more than 461 millions till the year of 2006. The rapidly increasing amount of mobile phone users and types of services provided by network operators leads to the accumulating of complaining information. How to use this information to enhance the quality of customers' service is a big issue to be concerned at present. To handle this kind of problem, the paper presents an approach to construct a TM for navigating domain knowledge and retrieving the more meaningful information to managers and operators in call centers of MCCs, which includes domain TM construction, a semantic topic expansion algorithm and VSM-based similarity calculation. The experimental results show the better performance of TM-based information retrieval system. Such approach is helpful for the enhancement of competition capability of MCCs.

2 Topic Map—Its TAO and IFS

There are various definitions about TMs, but their essence is the same, that is, TM is a tool for representation of model-based data on the Web for enhanced access. In comparison with resource description framework (RDF), TMs are developed separately from the resources they refer to [2].

2.1 The TAO

The TAO of TMs refers to topic, association, and occurrence respectively. A *topic*, in its most ge-

neric sense, can be any "thing" whatsoever — a person, an entity, a concept, really anything — regardless of whether it exists or has any other specific characteristics, about which anything whatsoever may be asserted by any means whatsoever. A topic may be linked to one or more information resources that are deemed to be relevant to the topic in some way. Such resources are called *occurrences* of the topic. A topic *association* represents relationships between topics and information resources relevant to them [3].

2.2 The IFS

The IFS of TMs refers to subject identity, facet, and scope respectively. Sometimes the same subject is represented by more than one topic, especially when two topic maps are being merged. In such a situation it is necessary to have some ways to establishing the identity between seemingly disparate topics. *Subject identity* is considered to be one of good ways and can be established by reifying a particular topic. When two topics have the same subject identity, they are considered to be "about" the same thing, and must therefore be merged. Sometimes it is convenient to be able to assign metadata to the information resources that constitute the occurrences of a topic from within the topic map. To provide this capability, the standard includes the concept of the facet. *Scope* specifies the extent of the validity of a topic characteristic assignment. It establishes the context in which a name or an occurrence is assigned to a given topic, and the context in which topics are related through associations [4].

3 Statistical Analyses on Related Literature

TMs have a long and complicated history. Many authors date their beginning back to the Davenport group, which in 1991 started a process to create a standard SGML DTD for software documentation [5]. This group quite quickly spun off an offshoot called CApH (Conventions for the Application of HyTime), one of whose tasks was to design an application for computerized back-of-book indexes. These indexes were intended to have one novel feature: it should be possible to merge them automatically. The ideas behind this application were what eventually TMs became.

TM was accepted by ISO's SGML working group as a new work item in 1996. ISO then spent another four years working on the standard before it was approved as ISO/IEC 13250: 2000 in January 2000. TM then had the form of an SGML architecture based on HyTime. Work was later done by an informal organization known as TopicMaps.Org, which produced the XTM (XML Topic Maps) syntax for TMs [5].

Regardless of the exact origins, the recent and rapidly growing literature on TMs among individuals addresses various questions, such as How to develop TMs? How to represent and merge TMs? How to maintain TMs? How to evaluate TMs? What have the others done with TMs? and so on. This survey does not bog down in a lot of technical details about the answers to the above questions due to the space limitation; instead it just gives a general view of the quantity and distribution of related literature.

A broad literature search for articles concerning TMs has been undertaken. A list of resources searched can be found in Table 1. The dates of survey range from 2000 to 2006. Keyword searched includes the term "Topic Maps". The relevant articles are then categorized based on the published year. To the end, there are total 112 articles and dissertations retrieved both in English and in Chinese.

The quantity of published articles on TMs distributed in the years of 2000 to 2006 corresponding to each database is listed in Table 1 below.

According to the above numbers, we plot seven curves corresponding to different databases, which illustrate the trend of quantity changes of published articles year-by-year, as shown in Figure 1. It gives us an animated graphical view, from which we can see that since the year of 2005, there has been an expansive growth in literature on TMs. This phenomenon may result from the TMRA (The international conference series on Topic Maps Research and Applications, and TMRA 2005 was the first TMRA conference).

Table 1 Quantity of published articles on TMs from 2000 to 2006

Yr/DB	Elsevier	Springer	EI	SCI	ProQuest	Dissertation	CNKI	Proceedings
2000	0	0	0	0	0	0	0	0
2001	1	1	2	0	0	0	0	1
2002	0	1	2	1	0	0	0	2
2003	2	1	5	2	0	0	0	3
2004	1	2	6	1	2	0	2	4
2005	1	2	14	5	0	1	4	12
2006	4	11	24	4	0	3	7	20
Total	9	18	53	13	2	4	13	42

Note: CNKI is the abbreviation of China National Knowledge Infrastructure. In the table, it is possible that the same paper has been retrieved several times from different databases.

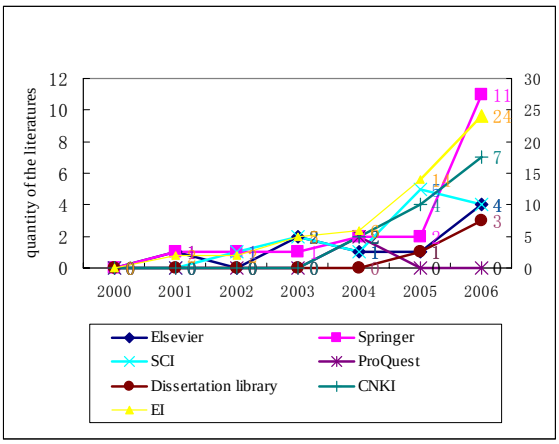

Figure 1 Curves corresponding to the quantity of articles about TMs published yearly

Besides providing the literature quantity display, Figure 2 shows the distribution of the quantity of articles about TMs for different databases.

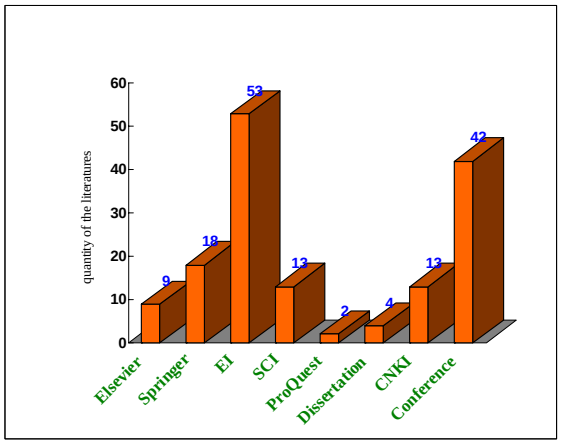

Figure 2 Distribution of the quantity of articles about TMs for different databases

For the purpose of comparing with research results on TMs between different countries, we plot a curve which represents the quantity of published articles in different countries for the years of 2000 to 2006, as shown in Figure 3. The articles we use are all in English. From Figure 3, we can see that there is a great gap between western and eastern countries in this research area although the survey is not complete in terms of the involved languages.

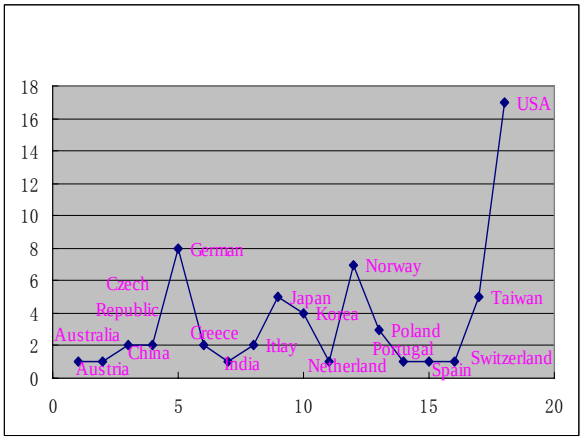

Figure 3 Quantity of literature published in different countries, 2000-2006

We also investigate the contributors to the retrieved articles on TMs, and group them into three categories: universities, institutes and corporations. Figure 4 shows the percentage of each group working on TMs, from which we can see that academic researchers at universities are the main contributors to TM articles, and there are also some institutes and corporations showing concerns, such as Morpheus Software of Netherlands [6], Ontopia of Norway [7], [8], [9], Computing Associates [12] and Siemens [10] of United States, etc.

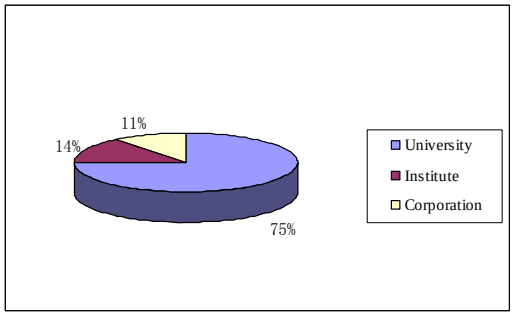

Figure 4 Distribution of affiliations of researchers working on TMs

TMs provide a general, powerful, and user-oriented way to convey knowledge of information resources under consideration in any specific domain. Many research works have been done from different ways. In the survey, we divide 64 articles retrieved from different databases into 3 categories depending on their contents: theory, application, and general review. Figure 5 shows the distribution of each category.

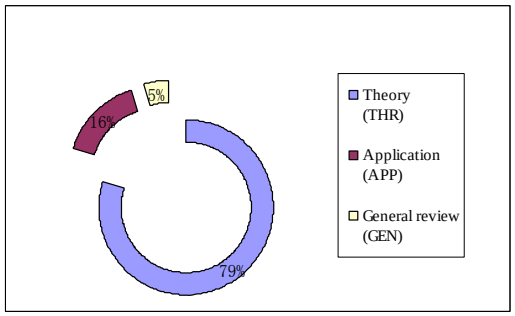

Figure 5 Categories of literature based on the contents

Comparing with TMs' theoretical values, many organizations and companies pay more attention to their practical uses.

From the survey, we find that most early works about TMs focus on providing introductory materials [11], [12].

Few of them are devoted to applications of TMs.

The earliest two works that addressed application issues were reported in [13] and [14]. Thomas Luckeneder, etc. [13] organized TMs in three layers: meta layer, schema layer and instance layer and established a well structured architecture to build world wide Internet directory services by cooperating with one of the largest Internet sites in Austria. Mahabal, Ashish, etc. [14] from CA used TMs as a tool of doing astronomy and for the first time TMs were implemented within the VO framework. Their works were published both in 2001, which are the earlier applications of TMs.

4 Application to Mobile Communication Corporations (MCCs) in China

In the study, we construct a TM in the domain of Customers' Service of Mobile Communication Corporations (MCCs) of China for knowledge navigation and information retrieval. The data is collected from a MCC of a certain city as well as the official website of MCC of a certain province in China. The test data are 500 pieces of customers' complaining documents in the form of Excel.

4.1 TM Construction

According to the TM structure, there are three main phases involved in the TM construction process: *topic selection*, *occurrence appending*, and *association analysis*. In *topic selection* phase, topics are selected in the following ways: word segmentation, domain thesaurus construction, conceptual clustering, topic selection, etc. Figure 6 shows the process of topic and topic type selection [15].

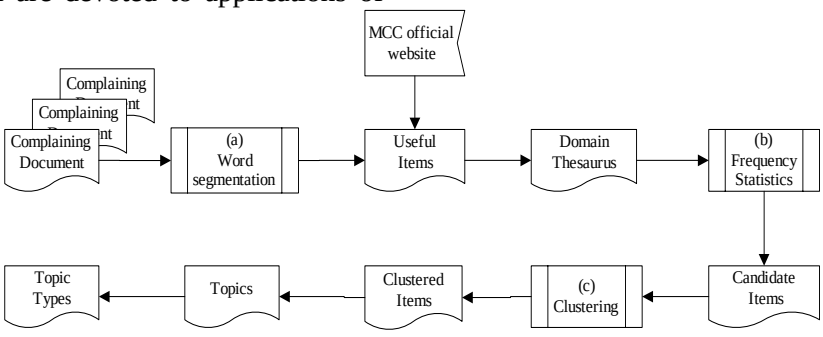

Figure 6 Process of topic and topic type selection

(a): 500 pieces of complaining documents are segmented based on the algorithm in Ref. [16], and 5228 segmented items are obtained, of which 420 items are selected as useful items.
(b): 210 items are selected after frequency statistics, with the principle of choosing the items with 2-6 characters and appearing above 5 times.
(c): 110 items are selected after conceptual clustering process based on the algorithm in Ref.[17], 89 items are chosen among the above items and named as topics, which are quite relative to the given domain and can describe the domain well.
In *occurrence appending* phase, occurrences are appended in the following steps:

Step 1: Map the multi-dimension space namely knowledge level of domain TM into the one-dimension space, see Figure 7.
Step 2: Construct an M×N Topic-Document Matrix, where the number of complaining documents is M and the number of topics is N.
Step 3: Appending occurrences back to the TM.

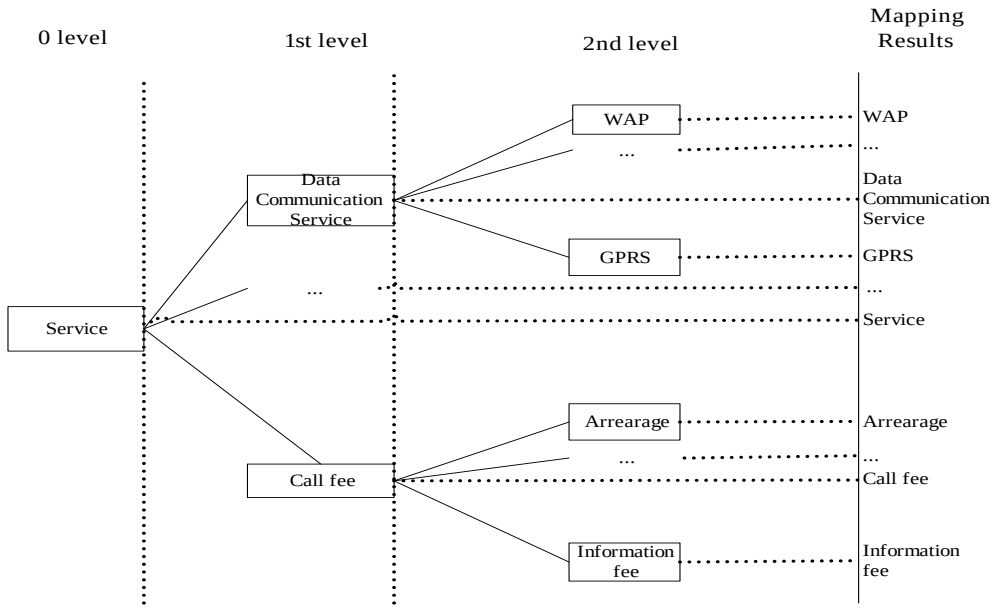

Figure 7 A part of domain TM and the mapping results

In *association analysis* phase, relations between topics and topic types are analyzed. These relations are associations in the domain TM. From 500 pieces of complaining documents, 3 kinds of associations and 6 kinds of association roles are extracted manually.

Up to now, the whole TM in the domain of Customers' Services for MCCs has been constructed.

4.2 Function Implementation

There are two main usages of TM in the knowledge navigation system: knowledge navigation and information retrieval.

4.2.1 Knowledge Navigation

In this paper, the TM is constructed based on the complaining documents of MCC. The purpose is by using of TM the customer service managers and the call center operators can have efficient works. By the link of the information/resources level, they can scan the service state conveniently and settle the customers' complaining documents problems timely.

Figure 8 gives the flowchart of the knowledge navigation.

Input: Target topics, target associations, target occurrence types
Output: Expanded sub-TM
Step 1. Give the object *topic* T_k, get a topic set $\mathbf{T}=\{t_1,t_2,\ldots t_k,\ldots t_i\}$,the element in **T** has "contain" association with T_k;
Step 2. Give the object *association type* A_k, get a new topic set $\mathbf{T}^*=\{t_1,t_2,\ldots t_k,\ldots t_j\}$, the element in $\mathbf{T}^*$ has the A_k with **T**;
Step 3. Consider the element in $\mathbf{T}^*$ as T_k, repeat Step1 and Step2 until there is no new topic appears.
Step 4. Give *occurrence type* O_k, get all occurrences belonging to this type. Then the expanded sub-TM is obtained.

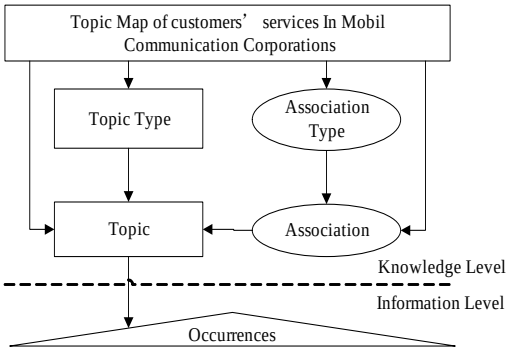

Figure 8 Knowledge navigation of TM

4.2.2 Information Retrieval

In a TM, the topics are linked by associations which contain certain semantic information that makes information retrieval more powerful than based on keywords. So a semantic topic expansion algorithm is proposed.

The process of topic expansion based on the above algorithm is given in Figure 9, which is created by *TM4J*. (http://compsci.wssu.edu/iis/nsdl/download.html)

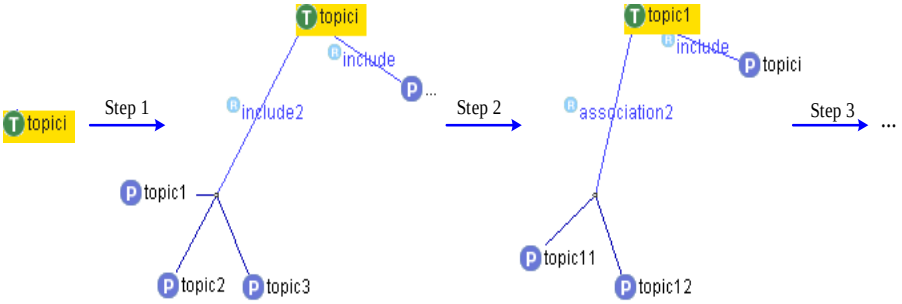

Figure 9 Process of topic expansion

When the number of the occurrences is large, the above algorithm isn't efficient, so we purpose a similarity calculating formula, which is measured by the cosine measure based on VSM. Similarities between complaining documents and the given query should be calculated as following:

Each complaining document can be represented as:

$\boldsymbol{D}_s=\{w_{s1}, w_{s11}, \ldots w_{s1i}, \ldots, w_{s1jk}, \ldots, w_{si}, \ldots, w_{sij}, \ldots, w_{sijk}, \ldots\}$

The query can be represented as:

$\boldsymbol{Q}=\{w_{q1}, w_{q11}, \ldots w_{q1i}, \ldots, w_{q1jk}, \ldots, w_{qi}, \ldots, w_{qij}, \ldots, w_{qijk}, \ldots\}$

The similarity between $\boldsymbol{D}_s$ and $\boldsymbol{Q}$ is given by Equation:

Then a threshold is set to limit the relevant result outputs.

$$sim(\boldsymbol{D}_s,\boldsymbol{Q}) = \cos\theta = \frac{w_{s1}w_{q1}+\ldots+w_{si}w_{qi}+\ldots+w_{sij}w_{qij}+\ldots+w_{sijk}w_{qijk}}{\sqrt{w_{s1}^{\ 2}+\ldots+w_{sijk}^{\ 2}}\sqrt{w_{q1}^{\ 2}+\ldots+w_{qijk}^{\ 2}}} \qquad (1)$$

4.3 Experiment Evaluation

The performance of the retrieval system is evaluated by two criterions: precision and recall. The formulas are given following:

$$\text{Precision} = \frac{\text{corret\_identification}}{\text{output\_documents}} \times 100\% \quad (2)$$

$$\text{Recall} = \frac{\text{correct\_identification}}{\text{all\_documents}} \times 100\% \quad (3)$$

We choose 10 queries to conduct an experiment to test the performance of the developed IR system and evaluate the precision and recall.

Table 2 Testing results on TM-based information retrieval system

	Output_Documents	Correct_Identification	All_Documents	Precision	Recall
1	47	44	71	93.60%	61.97%
2	47	47	65	100.00%	72.31%
3	5	5	14	100.00%	28.57%
4	5	5	6	100.00%	83.00%
5	10	7	8	70.00%	87.50%
6	9	7	9	77.78%	77.78%
7	22	16	23	72.73%	69.57%
8	2	2	3	100.00%	66.67%
9	62	20	27	32.26%	74.07%
10	3	3	4	100.00%	75.00%
Average				84.64%	69.68%

Table 3 Testing results on Keyword-based information retrieval system

	Output_Documents	Correct_Identification	All_Documents	Precision	Recall
1	110	51	71	46.36%	71.83%
2	76	43	65	56.58%	66.15%
3	16	10	14	62.50%	71.43%
4	6	6	6	100.00%	100.00%
5	2	2	8	100.00%	25.00%
6	10	5	21	50.00%	23.81%
7	29	13	23	44.83%	56.52%
8	2	2	3	100.00%	66.67%
9	9	8	27	88.89%	29.63%
10	5	4	4	80.00%	100.00%
Average				72.92%	61.10%

For TM-based information retrieval system, the average precision and recall are 84.64% and 69.68% respectively while the Keyword-based are 72.92% and 61.10% respectively, which means the performance of TM-based IR system is better than that of keyword-based IR system.

5 Conclusion and Future Direction

This paper mainly contains two aspects; first, we summarize the current state of TMs through the survey of large numbers of articles; second, we construct a TM in the domain of MCCs of China for knowledge navigation and information retrieval.

In the former part, we address core conception of TMs, chart the current state of TMs research, represent the main applications of TMs, and construct a TM for knowledge navigation and information retrieval in the domain of Customers' Service of MCCs. From the literature survey on TM, we can see there's a fast growth of the amount of articles about TM research since 2000; however, there's still a great gap between western and eastern in this research area; and the main contributors to TM articles are the academic researchers at universities. As to the later part, a TM-based IR system is constructed, it can navigate knowledge efficiently in MCCs and the competition capability of MCCs can be enhanced to some extent.

TMs have been developing quickly for the last 2 or 3 years and we believe that they will be a bright future. However, further study of TMs is still needed. The future direction of efforts can be pointed to the following aspects:

- Retrieve multimedia information such as images and audio/video clips, etc. with TMs

technologies

- Improve a powerful semantic-based information retrieval algorithm
- Construct TMs automatically or semi-automatically
- Perform the organizational support for modeling large, complex, and urgent problems

Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 70431001 and 70620140115.

References

[1] Schweiger R., Dudeck, Joachim. Improving Information Retrieval Using XML and Topic Maps. Artificial Intelligence Vol. 3873, 2005. 253-262

[2] Tian H.Y, Wu J.N., Yang G.F., Topic Map and Its Application to Document Retrieval. In Proceedings of The First International Congress of the International Federation for Systems Research, Kobe, Japan, November 14-17, 2005. 338-344.

[3] Pepper S. The TAO of Topic Maps-Finding the Way in the Age of Infoglut. Available at http://www.ontopia.net/topicmaps/materials /tao. html, 2001.

[4] Pepper S., Grønmo G.O., Towards a General Theory of Scope. Available at http://www. ontopia.net/topicmaps/materials/scope.html, 2002.

[5] Garshol L.M. Topic Maps, RDF, DAML, OIL, A comparison. Available at http://www. ontopia.net/topicmaps/ materials/tmrdfoildaml. html, 2003.

[6] Hopmans G., Kruijsen P.P., Oud L., et al. Hopmans, Gabriel, etc. Topic Maps for European administrative nomenclature. Artificial Intelligence Vol.3873,2005.177- 182

[7] Garshol L.M., Bogachev D., Maicher L., et al. TM/XML - Topic maps fragments in XML. Artificial Intelligence Vol.3873, 2005. 210-230

[8] Garshol L.M. Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. Available at http://www.ontopia.net/ topicmaps/materials/tm-vs-thesauri.html, 2004.

[9] Garshol L.M., Maicher L., Park J. TMRAP-Topic maps remote access protocol. Artificial Intelligence Vol. 3873, 2005. 53-68

[10] Biezunski M., Newcomb S.R., Liu P. XML topic maps: Finding aids for the Web. IEEE Multimedia, 2001. 104-108

[11] Pascal A., Patrice O., Mendez. A Formal Model for Topic Maps. Computer Science. Vol. 2342, 2002. 69-83

[12] Adam C. Just for Me Topic Maps and Ontology. Computer Science, Leipzig, Germany, 2005. 145-159

[13] Luckeneder T., Steiner K., Wöß W. Integration of Topic Maps and Databases. Computer Science. Vol. 2113/2001, 2004. 744

[14] Ashish M., Robert B. Topic Maps as a Virtual Observatory tool. Astronomical Data Analysis.San Diego,UnitedStates, 2001. 161-172

[15] Jiangning WU, Xiaohuan WANG, A Knowledge Navigation Method for the Domain of Customers' Services of Mobile Communication Corporations in China, Proceedings of International Conference on Intelligent Computing, Qingdao, China, August 21-24, 2007.

[16] Jiang S.H. Segmentation Algorithm for Chinese Text Based on Length Descending and String Frequency Statistics, Vol. 25, No. 1, (2006) 74-79 (in Chinese)

[17] Jiangning Wu, Haiyang Tian, Guangfei Yang, A Multilayer Topic-Map-Based Model Used for Document Resources Organization, Lecture Notes in Control and Information Sciences, in Proceedings of International Conference on Intelligent Computing, ICIC 2006, August 16-19, 2006, Kunming, China, 344: 753-758, Springer-Verlag Berlin Heidelberg.