http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (1): 132-138, 2010
ISSN 1812-5638

Distributed Knowledge Integration Based on Intelligent Topic Map

Huimin Lu and Boqin Feng
School of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an, 710049, China

Abstract: We propose a novel concept of Intelligent Topic Map, which extends the conventional topic map in structure and enhances the reasoning functions. With the Intelligent Topic Map as infrastructure, a mechanism of distributed knowledge integration is designed. The structure is divided into three layers: local Intelligent Topic Map layer, similarity measure layer and global Intelligent Topic Map layer. It provides a uniform query interface to a multitude of knowledge sources and lays the foundation for high-quality knowledge services. Moreover, we propose a new similarity measure algorithm based on comprehensive information theory and merging rules for knowledge integration. The experimental results show that our method is feasible and it has the significance of reference and value of further study for the distributed knowledge integration.

Key words: Knowledge integration, topic map, intelligent topic map, knowledge management

INTRODUCTION

Along with the up-rising of knowledge economy, massive amounts of knowledge which are often geographically distributed and owned by different organization are being mined (Zhang *et al*., 2008). There is a need to provide solutions that integrate knowledge from different sources and make them available for application queries. Knowledge Integration (KI) plays the role of giving a common representation for the different information sources handled it and offers users a global view of the information sources that can be accessed (Seng and Kong, 2009). The KI is a complicated task because it requires creating a common data model, finding semantic correspondences between two entities, satisfying the merge requirements and generating the duplicate free entities, etc. However, previous works find the semantic correspondences between entities rather than entity merging. They do not consider defining merge problems and providing solutions to those problems.

Topic map is a new ISO standard (ISO/IEC 13250) (ISO/IEC JTC 1/SC34 N323, 2002; ISO/IEC, 2008) for describing knowledge structures and associating them with information resources. It absorbs the ideas contained in the semantic web. The semantic organization and joining between the physical resource entities and the abstract concepts are implemented.

Previously, many methods used an object model to deal with the integration problem of distributed information sources (Tomasic *et al*., 1998; Carey *et al*., 1995). Such object models are represented in different forms. After XML standardization, many researches choose XML as the underlying data model (Baru *et al*., 1999). XML has been the W3C standard document format for exchanging information on the Web. It is the lowest common denominator for integration tasking. However, while XML can indeed establish interoperability between different information sources on the Web, its main limitation is that it copes only with structural heterogeneity and it can barely handle semantic heterogeneity (Seng and Kong, 2009). So, ontology is employed to tackle not only structure but also semantic interoperability in information integration. An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes, attributes and relationships. Ontology plays a key role in providing a shared terminology and supporting for the semantic representation and integration process. For example, FCA-Merge (Stumme and Madche, 2001) used the FCA (Formal Concept Analysis) method to merge between ontologies sharing a set of instances. Ontology reconciliation techniques such as merging, alignment and integration were focused on providing a formal description (Silva *et al*., 2007).

The method based on topic map uses XML as the common data model and adds ontology to enable the information integration. Merging between topic maps, which combines two topic maps to create a new ontology based on their semantic correspondences. XTM (XML

Corresponding Author: Huimin Lu, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, China

Topic Maps) 1.0 specification (Pepper and Moore, 2001) describes how to merge between entities of topic maps to produce an integrated entity. However, the merging method proposed by topic maps standard community processes integration only between equivalent entities. The method cannot merge between entities which have different structures but have semantic correspondences. In order to overcome the above shortcomings, many researches proposed merging approach to find correspondences between ontologies based on the syntactic or semantic characteristics and constraints of the Topic Maps (Lu *et al*., 2008; Korthaus *et al*., 2009).

In this study, we propose a novel concept of Intelligent Topic Map (ITM), Lu and Feng (2009) construct the KI framework based on ITM. We define a detailed process for ITM merging. First, local ITMs for local knowledge resources are generated. Next, the similarities of local ITMs are computed and then the topic pairs and the knowledge element pairs which have high similarity are found respectively. Finally, the global ITM is generated by merging local ITMs according to specific rules.

INTELLIGENT TOPIC MAP

The structure of conventional topic map composed of Topics, Associations and Occurrences (TAO) (Pepper, 2001), which is shown in Fig. 1.

Topics define the concepts. Associations define the relationships between the topics and could represent arbitrary number of roles among arbitrary number of topics. Occurrences link the information resources (e.g., documents) with topics. Topic maps are dubbed the GPS (Global Positioning System) of the information universe. Topic maps are also destined to provide powerful new ways of navigating large and interconnected corpora, but the conventional topic maps can not describe the relationships between knowledge elements. Moreover, as the knowledge resources becoming mass, the only topic level is difficult to locate the knowledge points and can not provide users with efficient knowledge navigation. Conventional topic map is a graphical index but lack of knowledge reasoning abilities and we unable to acquire implicit knowledge.

Extended topic map in structure: In our framework of the ITM, we define a clustering level above the topic level. Furthermore, a knowledge element level is inserted above the resource level. The structure of ITM is shown in Fig. 2.

The ITM establishes a novel multi-resource knowledge organization which depicts the hierarchical relationship cluster-topic-knowledge element-occurrence.

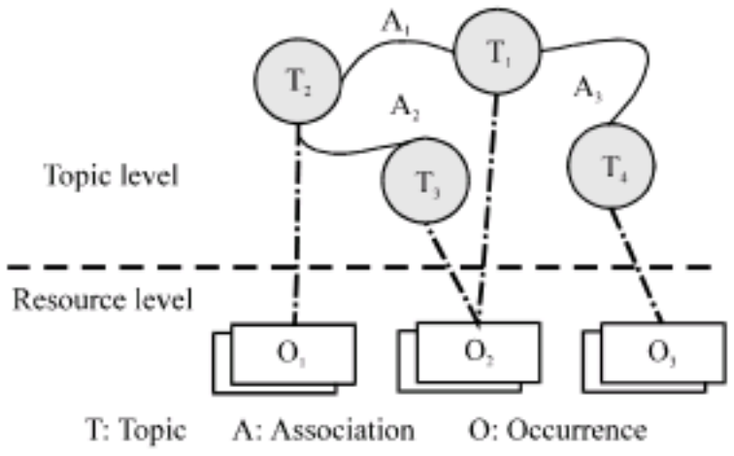

Fig. 1: The structure of conventional topic map

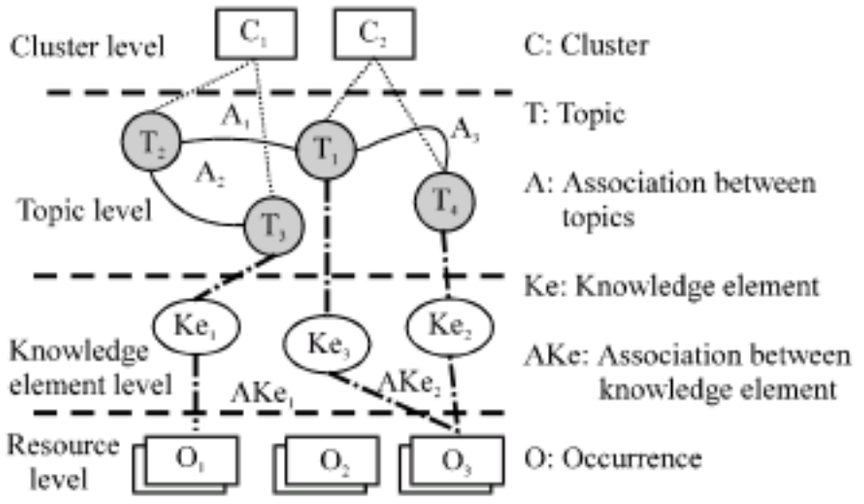

Fig. 2: The structure of extended topic map

ITM organizes knowledge from four levels: cluster level, topic level, knowledge element level and resource level. It constructs multi-granularity knowledge representation architecture which includes clusters, topics, knowledge elements, associations and occurrences. Knowledge elements allow users to access to more detailed knowledge information and provide knowledge elements navigation. Each cluster contains several closely related topics. Clusters provide the effective navigation and browsing mechanism for users after processing the topics by clustering analysis. Clustering analysis is the assignment of a set of topics into subsets (called clusters) so that topics in the same cluster are similar in some sense. The expression of multi-level, multi-granularity and inner relevant characteristics of knowledge is improved.

Knowledge reasoning: The implicit knowledge is acquired by the reasoning based on the custom rules or internal rules. Knowledge reasoning mainly includes Relationship Type Reasoning, Association Reasoning, Knowledge Architecture Reasoning and Order Reasoning, etc. In this study, we discuss the knowledge architecture reasoning which is related to knowledge visualization display. Knowledge architecture reasoning can obtain the level and class structure of the knowledge. Knowledge architecture reasoning mainly implements the following function. Given knowledge node, knowledge architecture

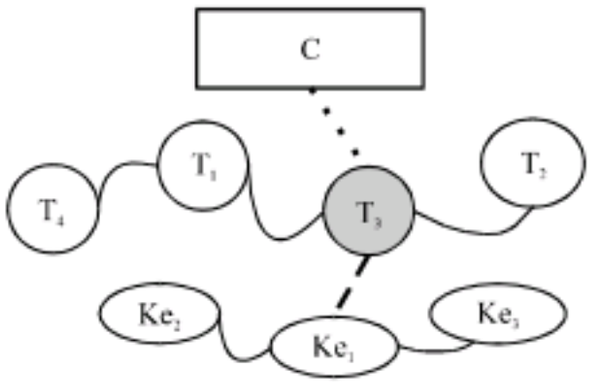

Fig. 3: Topic as the user interest node

reasoning return the cluster, all the topics and knowledge elements associated with the node within a certain knowledge radius. In the ITM, if there is a concept sequence C_p, C_1, C_2,..., C_m, C_q and there are association-between (C_p, C_1), (C_1, C_2),..., (C_m, C_q), then we said there exist a knowledge path between concept C_p and C_q. Association-between (C_i, C_j) denotes that the concept C_i is directly related to the concept C_j. the knowledge radius is the number of concepts traversed in a knowledge path, i.e. the length of the path. Reasoning results is shown in Fig. 3.

ITM extends the conventional topic map in structure and enhances the reasoning functions.

DISTRIBUTED KNOWLEDGE INTEGRATION

Distributed knowledge integration based on ITM is realized by merging local ITMs. Local ITM merging is divided into three layers: local ITM layer, similarity measure layer and global ITM layer. The structure is shown in Fig. 4.

Local ITM layer: The distributed knowledge resource is managed by a knowledge logical organization model, which is based on ITMs. Local ITMs for local knowledge resources are generated. We perform extraction of the elements of ITM to obtain the topics, the knowledge elements, the relationships between topics and the relationships between knowledge elements. Topics and knowledge elements extraction is the scope of information extraction, but the relationships between topics and the relationships between knowledge elements are acquired based on semantic understanding. And then we cluster the topics to get the clusters. After local ITM elements are extracted, local ITM will be generated.

Similarity measure layer: Similarity computing is the prerequisite and basis for ITMs merging. Many researchers have done a lot of work in this area. Subject Identity Measure (SIM) (Maicher and Witschel, 2004) was used to measure the similarity between topics based on their name similarity and occurrence similarity. TM-MAP

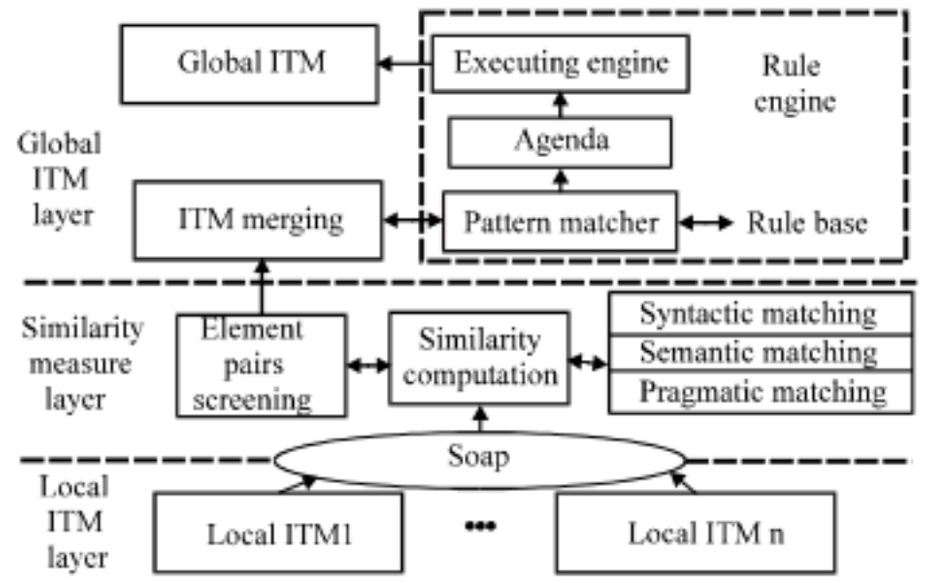

Fig. 4: The structure of knowledge integration based on ITM

(Kim *et al*., 2007) was a multi-strategic matching technique, which measured four facets of similarity: name-based similarity, property-based similarity, hierarchy-based similarity and association-based similarity. Topic and Occurrence-oriented Merging (TOM) (Wu *et al*., 2006) can be implemented when two topics may be merged, after establishing that there is topic name similarity and occurrence data/resource similarity and so on.

However, most of them can only operate at syntactic level. They are statistical methods, which are purely based on the character composition of two words. The semantic similarity and pragmatic relevance are not being considered. We propose the similarity measure method based on comprehensive information theory. The process used in the similarity algorithm consists of syntactic matching, semantic matching and pragmatic matching. The algorithm is summarized as follows:

SIM (w1, w2)
{
Csim1: $SIM_{syntactic}$ (w1, w2)
// Th is the threshold of syntactic similarity
If $SIM_{syntactic}$ (w1, w2) > Th goto Csim3
Csim2: $SIM_{semantic}$ (w1, w2)
Csim3: $SIM_{pragmatic}$ (w1, w2)

$$\text{If } SIM_{syntactic}(w1, w2) > Th \text{ then}$$

$$SIM(w1,w2) = \beta \times SIM_{syntactic}(w1,w2) + (1-\beta) \times SIM_{pragmatic}(w1,w2)$$

else

$$SIM(w1,w2) = \beta 1 \times SIM_{syntactic}(w1,w2) + \beta 2 \times SIM_{semantic}(w1,w2) + \beta 3 \times SIM_{pragmatic}(w1,w2)$$

}

Syntactic matching: Syntactic matching is used to compute the syntactic similarity by analyzing the character composition of topics or knowledge elements.

When linking a pair of topics (or knowledge elements), the syntactic similarity $SIM_{syntactic}$ (w1, w2) is defined as follows:

$$SIM_{syntactic}(w1,w2)=\frac{2c}{|w1|+|w2|} \quad (1)$$

The c denotes the number of characters of the largest common substring contained in two words. |w1| and |w2| denotes the number of characters of a pair of topics (or knowledge elements).

Semantic matching: Semantic matching analyses the static semantic similarity with aspect to synonyms. A pair of topics (or knowledge elements) is given. It is assumed that topics (or knowledge elements) are words and ES is the set of sense similarity valueES = $\{sv_1, sv_2,..., sv_{m\times n}\}$. ES is divided into four intervals: A: [0.0, 0.1) B: [0.1, 0.2) C: [0.2, 0.8) D: [0.8, 1.0). We analyze the contribution of these four intervals in words similarity cognitive ambiguity and certainty. semantic similarity is defined as follows:

$$SIM_{semantic}(w1,w2)=\begin{cases}Max(ES_D)-\dfrac{(0.2\times|ES_A|-\sum ES_A)+\sum ES_B}{|ES_A|+|ES_B|}\\ Max(ES_C),\ ES_A=ES_B=ES_D=\varnothing\end{cases} \quad (2)$$

Sense similarity is defined as follows:

$$SIM_{sense}=\beta_1 SIM_{MP}+\beta_2 SIM_{OP}+\beta_3 SIM_{RP}+\beta_4 SIM_{SP},\quad \beta_1+\beta_2+\beta_3+\beta_4=1 \quad (3)$$

where, β_1 is weight and SIM_{MP} is mainly sememe similarity, SIM_{OP} is other sememe similarity, SIM_{RP} is relative sememe similarity and SIM_{SP} is symbol sememe similarity.

Pragmatic matching: Pragmatic matching computes dynamic semantic similarity, which resolves the problem of polysemy. It considers the pragmatic relevance in linguistic context. When linking a pair of topics, the pragmatic similarity $SIM_{pragmatic}$ (Ta, Tb) is defined as follows:

$$SIM_{pragmatic}(Ta,Tb)=wSIM_{pt}(C_{Ta},C_{Tb})+(1-w)SIM_{pk}(C_{Ka},C_{Kb}) \quad (4)$$

where, SIM_{pt} (C_{Ta}, C_{Tb}) is the similarity between set C_{Ta} and C_{Tb}. C_{Ta} is the set of all topics which are directly related to the topic T_a. C_{Tb} is the set of all topics which are directly related to the topic T_b. SIM_{pt} (C_{Ta}, C_{Tb}) is defined as follows:

$$SIM_{pt}(C_{Ta},C_{Tb})=\frac{\sum\sum SIM_{syntax}(\varphi_i,\varphi_j)}{|C_{Ta}|\times|C_{Tb}|} \quad (5)$$

where, SIM_{pk} (C_{Ka}, C_{Kb}) is the similarity between set C_{Ka} and C_{Kb}. C_{Ka} is the set of all knowledge elements which are directly related to the topic T_a. C_{Kb} is the set of all knowledge elements which are directly related to the topic T_b. SIM_{pk} (C_{Ka}, C_{Kb}) can be calculated by the same as Eq. 5.

The measure based on comprehensive information not only considers the similarity of character composition of two words, but also considers the meaning and relevance in linguistic context. It solves the problem of synonym and polysemy and improves the veracity of similarity measuring.

Global ITM layer: This layer implements knowledge integration by merging local ITMs. Topic maps merging describe the process of integrating two topic maps into a new topic map. Merging operation ITMM (Intelligent Topic Map Merging) is defined as the following expression:

$$ITMM:(ITM_A\times ITM_B)\rightarrow ITM_C \quad (6)$$

We propose the method of ITM merging based on rule engine. The merging rules are described by rule description language based on topic map. The rule descriptions are saved in the rule documents and the documents are loaded and parsed by rule engine. Topic-merging rule and association-merging rule are defined as follows:

- **Rule 1:** Topic-merging rule. If topic T_1 in ITM_a has high similarity with T_2 in ITM_b, the two topics must be merged into a single topic (T_1 or T_2) in intelligent topic map ITM_c. The same is true for knowledge element
- **Rule 2:** Association-merging rule. Consider that in the ITM_a, a relationship Ra (T_{a1}, T_{a2}) exists between two topics of T_{a1} and T_{a2}. Also, consider that in ITM_b, a relationship R_b (T_{b1}, T_{b2}) exists between two topics of T_{b1} and T_{b2}. A merged topic T_{c1} of T_{a1} and T_{b1} has two relationships R_a (T_{c1}, T_{a2}) and R_b (T_{c1}, T_{b2}). The same is true for the relationship of knowledge element. The rule is depicted in Fig. 5.

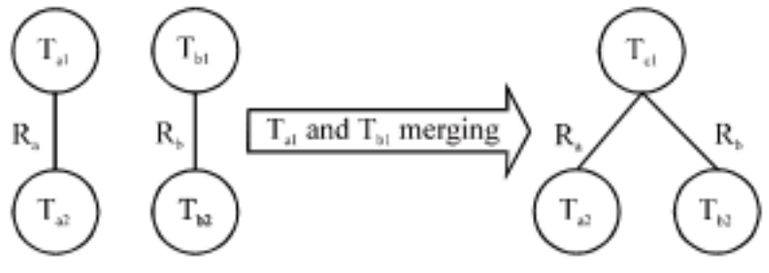

Fig. 5: The rule of association merging

The process of ITM merging based on rule engine is divided into four steps:

Step 1: Merging rules added

```
RuleBase rb = RuleBaseFactory.getRuleBase();
TopicMap ruleName = null;
String ruleNamePath = "ruleName.xml";
rule1 = Transformer.File2Obj (ruleNamePath);
rb.addEtm (ruleNamePath);
```

Step 2: The high similarity element pairs are loaded into Working Memory

```
WorkingMemory workingMemory = new WorkingMemory();
TopicMap tm = merger.getEtm();
workingMemory.insert(tm);
rb.setWorkingMemory (workingMemory);
```

Step 3: Configure executing engine

```
Agenda agenda = new Agenda();
HashMap < String, ExecEngine> execMap = new HashMap < String, ExecEngine>();
execMap.put("ruleName", new ruleNameExecutor());
agenda.setExecMap(execMap);
rb.setAgenda(agenda);
```

Step 4: Executing the rules in Agenda, until to all the rules in Agenda are finished.

```
rb.matchAllRules();
rb.getAgenda().execute();
```

SYSTEM FRAMEWORK

The system framework of distributed knowledge integration based on ITM mainly includes two modules: background processing and foreground processing, as shown in Fig. 6.

Background processing: Background processing module includes the following functions:

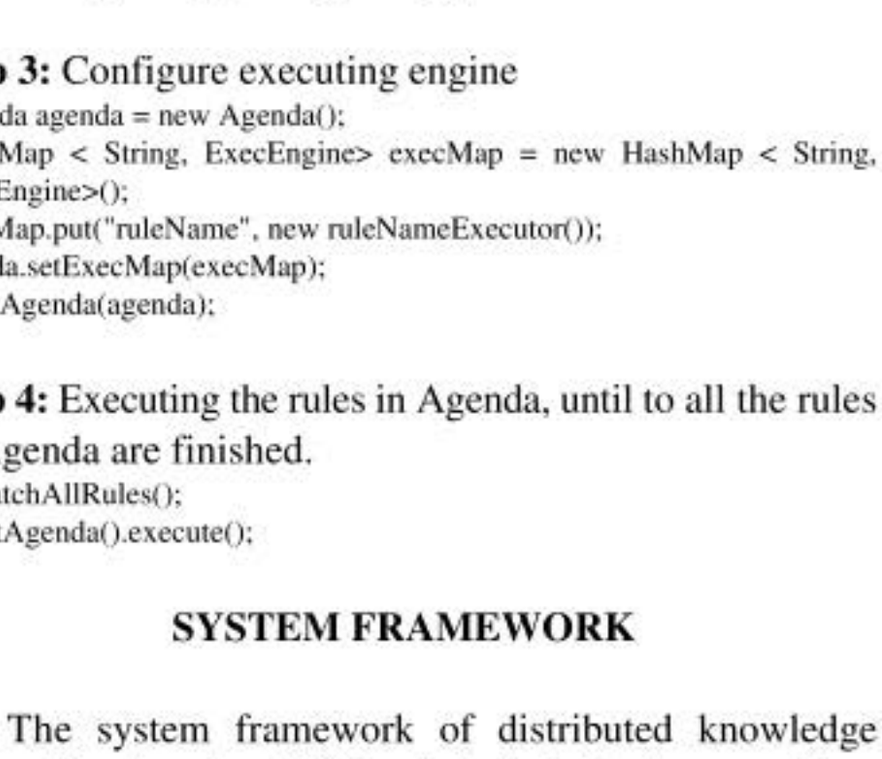

Fig. 6: The system framework of distributed knowledge integration based on ITM

- **Knowledge storage:** Since knowledge resources are massive, knowledge distributed storage are built up. Local ITMs for local knowledge resources are generated and local autonomy management is carried out
- **Knowledge processing:** Knowledge discovery extracts and defines the new patterns from the original data set, which includes data preparation, data mining and outcome interpretation and evaluation and realizes knowledge reasoning
- **Knowledge organization and management:** It implements the centralized management of local ITMs merging, global ITM, metadata, indexing and the interest models of users. In the mean time, this step also accomplishes the optimizing process of user query

Foreground processing: The mainly function of foreground processing is implements interactive learning with user. Portal is an interface between end users and web service. Here, it represents the new knowledge service for web accessing. It provides hot authoritative resource discovery, knowledge navigation, knowledge recommendation, knowledge visualization display and so on.

The system framework of distributed knowledge integration based on ITM implements distributed storage, centralized management and dynamic data integration of knowledge resources. It constructs the global views during the information integration and provides good knowledge services for users.

EXPERIMENTAL RESULTS

We evaluate the effectiveness of the proposed ITM-based distributed knowledge integration system model from similarity algorithm, because similarity algorithm is the foundation of ITM merging and its performance is directly related to the quality of merging. It is applied in a part of the knowledge domain of computer network (Table 1).

In this study, we use performance measurement of information retrieval F (F-measure). We get true-positive set (TP) which includes correctly identified matches, false-positive set (FP) includes false matches and false-negative set (FN) which includes missed matches. We can

Table 1: The experimental data of merging

Local ITM	TN	KeN	ATN	AKeN	ATKeN
Partial computer network	194	217	88	197	232
Data link layer	100	104	112	113	104
Network layer	100	201	82	200	112
Physical layer	100	100	100	94	44

TN: Topic No., KeN: Knowledge element No., ATN: Association No. between topics, AKeN: Association No. between knowledge elements, ATKeN: Association No. between topics and knowledge elements

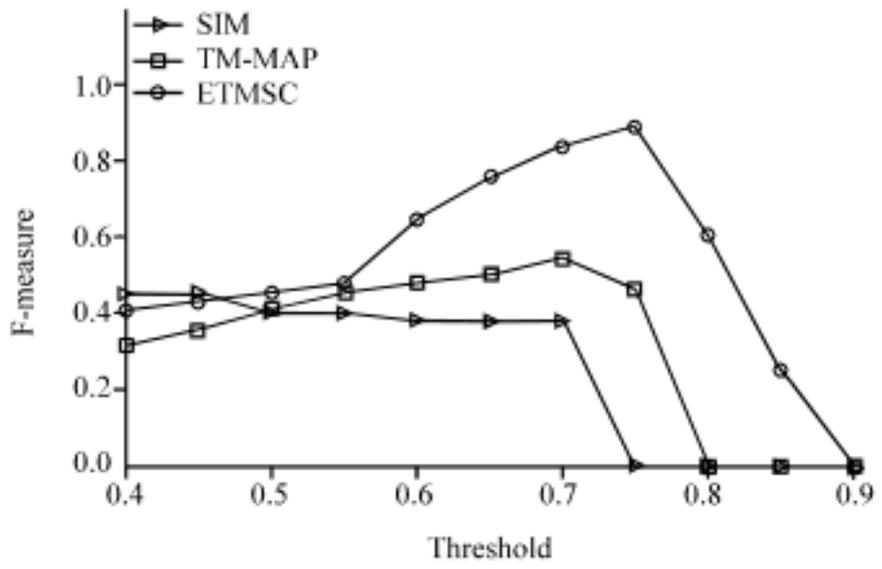

Fig. 7: F-measures of SIM, TM-MAP, ETMSC similarity algorithms

$$F = 2 \times \frac{TP}{FP + FN} \quad (7)$$

measure match quality of automatic matching process by evaluating following expression.

We compare our method named ETMSC with other topic maps similarity algorithms which are called SIM and TM-MAP (TOM is similar as SIM, which measures the topic similarity and occurrence similarity on syntactic level). Figure 7 shows the experimental result that F-measures of similarity algorithms each other.

The experimental results indicate that ETMSC has higher F-measure than SIM and TM-MAP, when the threshold is more than 0.55. Because SIM measures the similarity between topics based on their names similarity and occurrence similarity, it does not consider the external structures of topic maps, such as hierarchy and association. TM-MAP measured four facets of similarity: name-based similarity, property-based similarity, hierarchy-based similarity and association-based similarity. However, our similarity algorithm ETMSC comprehensively considers syntactic matching, semantic matching and pragmatic matching. Compared with the traditional algorithms which purely based on the syntactic or semantic similarity, the F-measure of ETMSC is improved by 9.2-11.1%.

Compared with the traditional algorithms which purely based on the syntactic or semantic similarity, the accuracy of ETMSC is improved, but it has great relationship with threshold selection. Threshold selection is relative to the importance degree of similarity and the different levels of fusion demand. Threshold selection should give full consideration to achieve the desired objectives and outcomes of the time and effort. Before determining the threshold, the test must be carried out in a certain amount of target data. Threshold should be selected in the average F-value of best-case scenario.

CONCLUSION AND FUTURE WORK

This study has presented a mechanism of the distributed knowledge integration based on a new concept of ITM. Using ITM in distributed knowledge integration field is a novel direction and presents a new way for distributed knowledge management. We concisely summarize the main advantages of our proposed framework as follows: (1) ITM not only express the multi-level, multi-granularity of knowledge, but also fully reflect the association between the knowledge and the information resources related to the knowledge. It contains rich knowledge and information. (2) Distributed knowledge integration is easy to realize by merging local ITMs into a global ITM and (3) Graphic display based on ITM is more perceivable, it can provide visual knowledge navigation.

The dynamic updating mechanism, automatic adaptation and real system deployment of ITMs are the essential future work. We hope that, from our initiative framework, the standards could be made and the real system will be widely deployed in the future.

ACKNOWLEDGMENTS

This research is sponsored by the National High-Tech Research and Development Plan of China under Grant No. 2008AA01Z131; The National Natural Science Foundation of China under Grant No. 60803162. This study is also partially supported by National High-Tech Research and Development Plan of China under Grant No. 2008AA01Z136.

REFERENCES

Baru, C., A. Gupta, B. Ludäscher, R. Marciano, Y. Papakonstantinou, P. Velikhov and V. Chu, 1999. XML-based information mediation with MIX. ACM SIGMOD Record, 28: 597-599.

Carey, M.J., L.M. Hass, P.M. Schwarz, M. Arya and W.F. Cody *et al*., 1995. Towards heterogeneous multimedia information systems: The garlic approach. Proceedings of the 5th International Workshop on Research Issues in Data Engineering-Distributed Object Management, Mar. 6-7, Taipei, Taiwan, pp: 124-131.

ISO/IEC JTC 1/SC34 N323, 2002. Guide to the topic map standards. International Organization for Standardization. http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0323.htm.

ISO/IEC, 2008. Information technology topic maps part 2: Data model. http://www.isotopicmaps.org/sam/sam-model/data-model.pdf.

Kim, J.M., H. Shin and H.J. Kim, 2007. Schema and constraints-based matching and merging of topic maps. Inform. Process. Manage., 43: 930-945.

Korthaus, A., M. Aleksy and S. Henke, 2009. A distributed knowledge management infrastructure based on a Topic Map grid. Int. J. High Performance Comput. Network., 6: 66-80.

Lu, H., B. Feng, Y. Zhao, Q. Zheng and J. Liu, 2008. A new model for distributed knowledge organization management. Proceedings of the 7th International Conference on Grid and Cooperative Computing, Oct. 24-26, Shenzhen, China, pp: 261-265.

Lu, H. and B. Feng, 2009. An intelligent topic map-based approach to detecting and resolving conflicts for multi-resource knowledge fusion. Inform. Technol. J., 8: 1242-1248.

Maicher, L. and H.F. Witschel, 2004. Merging of distributed topic maps based on the subject identity measure (SIM) approach. Proceedings of the LIT 2004, Sept. 29-Oct. 1, Leipzig, Germany, pp: 1-11.

Pepper, S., 2001. The TAO of topic maps. http://www.gca.org/papers/xmleurope2000/.

Pepper, S. and G. Moore, 2001. XML topic maps (XTM) 1.0. TopicMaps Org.

Seng, J.L. and I.L. Kong, 2009. A schema and ontology-aided intelligent information integration. Expert Syst. Appl., 36: 10538-10550.

Silva, P.A., C.M.F.A. Ribeiro and U. Schiel, 2007. Formalizing ontology reconciliation techniques as a basis for meaningful mediation in service related tasks. Proceedings of the ACM 1st Ph.D. Workshop in Conference on Information and Knowledge Management, Nov. 9, Lisbon, Portugal, pp: 147-154.

Stumme, G. and A. Madche, 2001. FCA-merge: Bottom up merging of ontologies. Proceedings of the 17th International Joint Conference on Artificial Intelligence, Aug. 4-10, Seattle, Washington, USA., pp: 225-234.

Tomasic, A., L. Raschid and P. Valduriez, 1998. Scaling access to distributed heterogeneous data sources with DISCO. IEEE Trans. Knowledge Data Eng., 10: 808-823.

Wu, X., L. Zhou, L. Zhang and Q. Ding, 2006. TOM algorithm in distributed topic maps merging. Eng. J. Wuhan Univ., 39: 131-136.

Zhang, C., X. Yang and S. Du, 2008. A distributed knowledge model for knowledge management system. Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing, Oct. 12-14, Dalian, China, pp: 1-4.