Topic Map and Its Application to Document Retrieval

Haiyan Tian, Jiangning Wu and Guangfei Yang
Institute of Systems Engineering, Dalian University of Technology, Dalian, 116024, China
E-mail: peanutthy@student.dlut.edu.cn , jnwu@dlut.edu.cn

ABSTRACT

As computer and network technologies develop, information on line becomes flooded. Under this condition searching requested documents is really a difficult task. People are always seeking for an efficient searching tool and Topic Map is one of them. Topic Map is a new tool proposed by the International Organization for Standardization (ISO) to solve problems about knowledge representation and knowledge management. It has been widely used in many knowledge fields and in this paper we mainly make research on its application to information (document) retrieval. We firstly address some issues related to Topic Map: topic, association, occurrence as well as identity, facet and scope. By means of the current technologies for Topic Map developing, such as TMCL, TMQL, XML, SGML and so on, we propose a multi-layer Topic Map-based document retrieval model (TMDRM). TMRDM is based on Topic Map's richly cross-linked structure and capabilities of topics used to group together objects that relate to a single abstract concept. This model helps people narrow the search scope step by step in order to facilitate them to proper documents.

Keywords: Topic Map, document retrieval, TMDRM

1. INTRODUCTION

As more and more text becomes available on-line as part of the world wide web of electronic information, finding information about a specific topic becomes harder and harder. The ambiguities found in most languages mean that few terms have a single meaning. Whilst some words have the same meaning in a number of languages, many meanings have different expressions in different languages. Traditional full-text searching typically fails to distinguish between the different meanings of a word. It also cannot distinguish between the uses of the same word in different languages. Because of this, full-text searching often provides too many "hits" for users to have time to fin d the information they need from the morass of irrelevant information[1].

Besides full-text searching, there also exist some other types of search engines on the WWW, e.g., the catalogue-based approach, but such category-based search engines often make it difficult for experienced users to find information related to specific disciplines, or to specialist areas within disciplines.

That is why new methodologies are called for and Topic Map appears in response. Topic map provides an approach that marries the best of several worlds, including those of traditional indexing, library science and knowledge representation, with advanced techniques of linking and addressing[2]. It is a new ISO standard for describing knowledge structures and associating them with information resources. As such they constitute an enabling technology for knowledge management. Dubbed "the GPS of the information universe", topic map is also destined to provide powerful new ways of navigating large and interconnected corpora.

In the next part of this paper, we mainly focus on some concepts about Topic Map including topic, association, occurrence and so on. Then we discuss the critical technologies for Topic Map development and its application. Based on the discussions, we propose a multi-layer TMDRM model to be used for document retrieval.

2. SOME ISSUES RELATED TO TOPIC MAP

2. 1. What is a Topic Map?

There are various of definitions about Topic Map, but their essence is the same, that is, Topic Map is a tool for representation of model-based data on the web for enhanced access. Topic maps are based on topics, associations and occurrences. In comparison with resource description framework (RDF), topic maps are developed separately from the documents they refer to.

The basic concepts of a Topic Map are topic, association, and occurrence — the TAO of topic map. In the following two subsections we will introduce them

in detail as well as the additional concepts of identity, facets and scope — the IFS of topic map.

2. 2. The TAO of Topic Map — Topic, Association and Occurrence

2. 2. 1. Topic

A topic, in its most generic 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[3].

In fact, this is almost word for word how the topic map standard defines subject, the term used for the real world "thing" that the topic itself stands in for. Strictly speaking, the term "topic" refers to the object or node in the Topic Map that represents the subject being referred to. However, there is (or should be) a one-to-one relationship between topics and subjects, with every topic representing a single subject and every subject being represented by just one topic. To a certain degree, therefore, the two terms can be used interchangeably.

A topic may have one or more of the three characteristics: topic name, topic type and association, thereinto association will be introduced in Section 2.2.3.

Normally topics have explicit names, since that makes them easier to talk about. However, topics don't always have names. A simple cross reference, such as "see page 97", is considered to be a link to a topic that has no (explicit) name. Topic's name can be divided into three kinds: base name (required), display name (optional) and sort name (optional). The standard provides the facility to assign multiple base names to a single topic, and to provide variants of each base name for use in specific processing contexts. In the original ISO standard variants were limited to display name and sort name.

Topics can be categorized according to their kinds. In a topic map, any given topic is an instance of zero or more topic types. This corresponds to the categorization inherent in the use of multiple indexes in a book (index of names, index of works, index of places, etc.), and to the use of typographic and other conventions to distinguish different types of topics.

2. 2. 2. Occurrence

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.

An occurrence could be a monograph devoted to a particular topic, for example, or an article about the topic in an encyclopaedia; it could be a picture or video depicting the topic, a simple mention of the topic in the context of something else, a commentary on the topic (if the topic were a law, say), or any of a host of other forms in which an information resource might have some relevance to the subject in question[3].

Occurrences, similar to topics, may be of any number of different types. Such distinctions are supported in the standard by the concepts of occurrence role.

2. 2. 3. Association

Up to now, all the constructs that have been discussed have had to do with topics as the basic organizing principle for information. The concepts of "topic", "topic type", "topic name", "occurrence" and "occurrence role" allow us to organize our information resources according to topic (or subject), and to create simple indexes, but not enough.

The really interesting thing, however, is to be able to describe relationships between topics, and for this the Topic Map standard provides a construct called topic association to describe relationships between topics.

Just as topics and occurrences can be grouped according to their individual types, so associations between topics can also be grouped according to their types that are called association types.

Each topic that participates in an association plays a role in that association called the association role.

2. 3. The IFS of Topic Map — Identity, Facet and Scope

2. 3. 1. Identity

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.

When the subject is an addressable information resource (an “addressable subject”), its identity can be established directly through its address. However most subjects, such as Puccini, Italy, or the concept of opera, are not directly addressable. This problem can be solved through the use of subject indicators. A subject indicator is “a resource that is intended by the topic map author to provide a positive, unambiguous indication of the identity of a subject.” Because it is a resource, a subject indicator has an address (usually a URI) that can be used as a “subject identifier”.

2. 3. 2. Facet

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.

Facets basically provide a mechanism for assigning property-value pairs to information resources. A facet is simply a property; its values are called facet values. Facets are typically used for supplying the kind of metadata that might otherwise have been provided by SGML or XML attributes, or by a document management system. This could include properties such as “language”, “security”, “applicability”, “user level”, “online/offline”, etc.

Once such properties have been assigned, they can be used to create query filters producing restricted subsets of resources, for example those whose language is “Italian” and user level is “secondary school student”.

2. 3. 3. Scope

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[5]. Every characteristic has a scope, which may be specified either explicitly, as a set of topics, or implicitly, in which case it is known as the unconstrained scope. Assignments made in the unconstrained scope are always valid.

Scope is considered to establish a namespace for the base names of topics. This leads to the constraint, imposed by the topic map paradigm, called the topic naming constraint, that any topics having the same base name in the same scope implicitly refer to the same subject and therefore should be merged. With the exception of this constraint, the interpretation of a characteristic's scope and its effect on processing is left to the application and is in no way constrained by this specification.

Scope is defined in terms of themes, and a theme is defined as “a member of the set of topics used to specify a scope”. In other words, a theme is a topic that is used to limit the validity of a set of assignments.

In fact, the well-designed, consistent and imaginative use of scope in topic maps is not only intent to remove ambiguity. It can also be used as the navigator, for example by dynamically altering the view on a topic map based on the user profile and the way in which the map is used.

3. KEY TECHNOLOGIES FOR TOPIC MAP DEVELOPMENT

Technologies used for developing the Topic Map are shown in Figure 1.

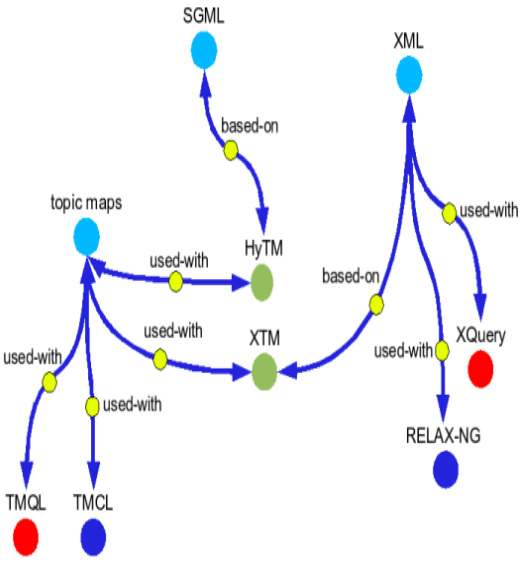

Figure 1. Key technologies for developing the Topic Map

3. 1. TMCL (Topic Map Constraint Language)

TMCL is a formal language for defining schemas and constraints on topic map models. Specifically, TMCL is to constrain topic map models as defined by the Data Model for Topic Maps. The constraint language will provide a formal constraint language, related to operational semantics, and the syntax [9].

Topic Map Constraint Language provides a means to express constraints on topic maps conforming to ISO/IEC 13250:2000. These constraints will be over

and above the constraints currently defined in the Topic Map Data Model.

TMCL is designed to allow users to constrain any aspects of the topic map data model. TMCL adopts Topic Map Query Language (TMQL) as a means to express both the topic map constructs to be constrained and topic map structures that must exist in order for the constraint to be met. TMCL defines TMCL-Schema and TMCL-Rule. TMCL-Schema provides a type-based model of constraints. TMCL-Schema is defined in terms of a more abstract model TMCL-Model. TMCL-Rule provides a generalized model of constraint based on TMQL. For each language a model, semantics and syntax are defined [10].

Both TMCL-Rule\Schema define sets of constraints. In general these constraints consist of terms that identify parts of the Topic Map to be constrained and terms that define the predicate or truth that must hold for the Topic Map to be considered to be consistent.

3. 2. TMQL (Topic Map Query Language)

TMQL is an XML-based extension of Structured Query Language (SQL), a query language developed for use in meeting the specialized data access requirements of Topic Maps. Two types of data access for Topic Maps are: information retrieval (IR), which is focused on separate search instances of a single user looking for specific information; and information filtering, which is a query process that builds up a sort of user profile, filtering information to construct a selection of data relevant to a particular user [11].

TMQL is intended to be easier to learn by developers, most of whom are likely to already be familiar with SQL. However, SQL was created to be used on the data in a relational database, which has a well-defined pre-existing structure; TMQL must be able to retrieve information from a vast and constantly changing body of information. A relational database may be expressed in terms of a simplified Topic Map, but TMQL will need to work with a much more complex data repository. A typical SQL select query (used to retrieve data from a table in a database) could be rewritten to retrieve data from a Topic Map.

3. 3. SGML (Standard Generalized Markup Language)

The interchangeable form of Topic Maps are formally expressed in SGML. SGML is considered here a syntax, which can be understood, parsed, validated and interchanged on many systems [12]. However, it's different from classical SGML in the sense that even if it's editable by hand in an SGML editor. It's not very comfortable to do so, because it's almost like dealing with assembly code with a text editor.

Philosophically speaking, Topic Maps are a SGML application, because they are describing in a structural form which is applied to the semantics of the information. Such application is actually a tagging mechanism.

3. 4. XML (Extensible Markup Language)

Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML. It is a kind of descriptive language based on Topic Map criterion, which mainly defines document type definition (DTD) files used to describe Topic Map and supplies syntaxes and models to describe structured information. Such syntaxes can define topics, association between topics, and so on.

The linking part of XML, called XLL (Extensible Linking Language) is made of two parts: XPointer and XLink. XPointer is linked to a specific location and XLink is designed to establish links among all documents or resources. XLink supplies two kinds of links in XLL: simple links and extended links. Simple links define a kind of unitary links from initial resources to terminative resources. Extended links apply the same concept of independent linking as HyTime does. They contain the attributes necessary to fully describe Topic Map information so that XLL extended links are a good candidate to represent Topic Map interchange information.

3. 5. Topic Merging

The ISO standard provides some guidelines and constraints to be used when merging maps. These include the Topic Naming constraint and the concept of identity [13].

An important problem needed to be dealt with during Topic Map Merging is topic merging. This also occurs after description of information resource and merging information resource to several basic facets in Topic Map. As we know that there must be some overlaps among topics, so it is necessary to merge various topics especially for the resource that has latent repetition caused by the people's subjectivity on descriptive point of view and categorization. Concretely speaking, the reasons lie in: every topic may have different origins; every topic may be produced by different technologies; every topic may use different syntaxes.

The principles for Topic Map Merging are:

- Unity of topics. Topics standing for the same subject should be merged into one single topic;
- After two topics are merged, topics' characteristics should also be unified.

3. 6. Topic Map Engine

Topic Map Engine is the core of the Topic Map. It provides a comprehensive Application Programming Interface (API) to allow programmers to create and modify Topic Map structures. The engine can be used to manage topic maps which are maintained in-memory or which are persistently stored either in the Ozone object-oriented database, or in a relational database using the Hibernate O-R mapping. Current Topic Map Engines are TM4J, Omnigator, etc.

The Omnigator is a free generic application built on top of the Ontopia Navigator Framework that allows users to load and browse any conforming Topic Map, including their own. Users can view Topic Map data using a Web-based text-oriented interface or with an highly intuitive graphical visualization interface. Designed primarily as a teaching aid to help newcomers understand the Topic Map concepts, it is also an extremely useful tool for debugging Topic Maps and for building demo applications.

4. APPLICATION TO DOCUMENT RETRIEVAL

Topic Maps can be used in many research fields, such as knowledge representation, knowledge repository, knowledge navigation, information retrieval, knowledge inference, and so on.

The traditional method for document retrieval is full-text retrieval, for example, like Google, when you enter key words included in the information you need, the system will search for texts that include the key words in full text area, and results will be those texts. The retrieved results are numerous and some are void so that they waste people a lot of time to acquire proper and useful information.

There are some precise ways to increasing retrieval efficiency. In most documents retrieval systems, we can choose searching areas, like "*Title*", "*Abstract*", "*Key words*", "*All fields*", etc., and when we enter key words, the system will search for matched information in the area you have chosen. Such method is quicker and more precise but returned results are still too numerous.

Considering the facts above, we intend to construct a model based on Topic Map for document retrieval named TMDRM.

4. 1. Basic Idea

Basic idea for this TMDRM lies in the following two aspects:

1.Topic Map provides the casual browser of knowledge in it, with a richly cross-linked structure, see Figure 2.

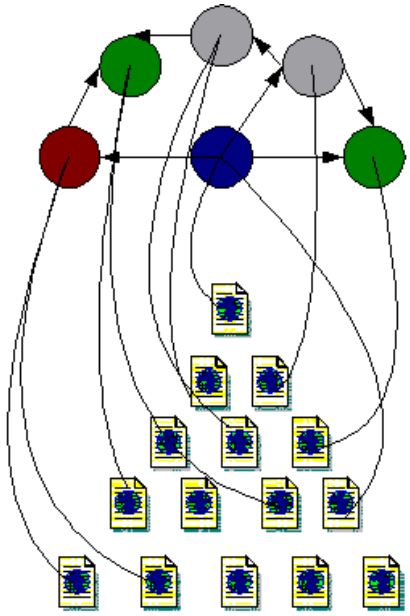

Figure 2. Associative browsing with a Topic Map

Topic occurrences create "sibling" relationships between objects. A single object may be an occurrence of one or more topics, each of which may have many other occurrences. When a user finds/browses to a given object, this sibling relationship enables them to rapidly determine where there are other objects regarding the same topic as the current one. Topic associations create "lateral" relationships between subjects - allowing a user to see what other concepts are related to the subject of current interest and to easily browse to them. Associative browsing allows an interested data consumer to wander in a guided manner. A user might also find associative browsing useful in increasing the chance of serendipitous discovery of relevant information.

2. Topics can be used to group together objects that relate to a single abstract concept. Each object may be defined as an occurrence of the topic. Occurrences may be assigned a role, defining the relationship with the parent topic. These typed relationships mean that a user may first query on a concept and then rapidly narrow the size of the results set by occurrence role.

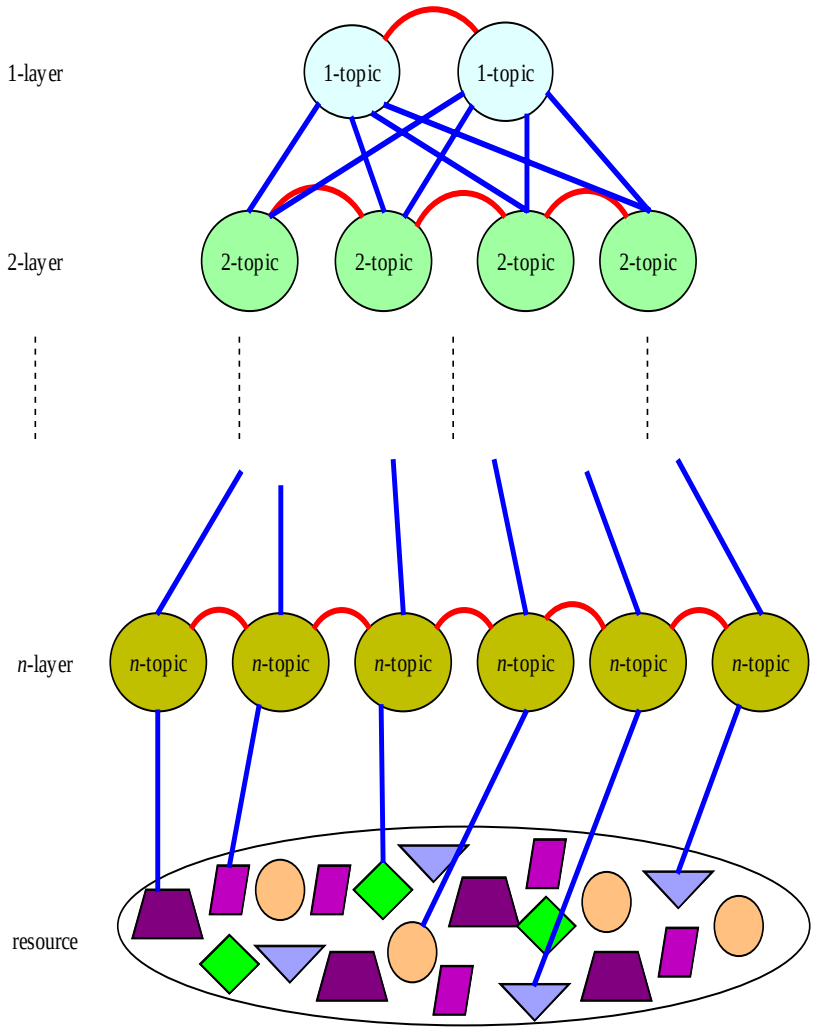

Figure 3. Structure of TMDRM

4. 2. A Multi-layer TMDRM for Document Retrieval

In order to realize document navigation, it is needed to exhibit Topic Map's data model in the Web, design electronic index for "Topics" and link all the topics according to their associations. Because the topic area is defined upon information area, all topics make sense in topic areas. When we search for something, the searching system is operated first in topic areas, its search scope is narrower than in information area and the results are a series of related topics; then we can easily find information we need through the links between topics and their occurrences. It makes "precise retrieval" come true and also improves retrieval efficiency.

Our TMDRM consists of *n* layers whose framework is shown in Figure 3

In the first layer, the topics are concerned with some broad research fields and their corresponding branches, such as "*knowledge management*", "*knowledge representation*", "*ontology*", etc., in which the associations show the relationships between two topics, and topics' occurrences are those topics in the second layer.

From 2-layer to *(n-1)*-layer, topics in each layer are subtopics or small categories of topics corresponding to their upper layers. By using the proposed multi-layer retrieval model, we can narrow the search scope step by step.

In the last layer, topics are names of documents related to their corresponding parent topics in *(n-1)*-layer. Associations between two documents are similarities that can be calculated by VSM method for example. Occurrences are documents themselves or details that can be found in "resource" by links.

When we search for some documents in some research fields, such as "*Knowledge Representation*", first we can search for the topic "*knowledge representation*" among the first-layer topics. Then we can get the subtopics linked with the parent topic "*knowledge representation*" and this leads us to the second layer. Go on in turn and at last we can get to the last layer where we can see the similarities between two documents so that we can seek for another document by means of the other topics in the same layer. In fact the proposed multi-layer model is like a navigation map that guides us for correct direction and useful information (documents).

5. CONCLUSION REMARKS

In this paper we mainly introduce three main concepts about Topic Map: topic, association, occurrence. Afterward, we discuss the critical technologies for Topic Map development and application. By using the current technologies mentioned, a multi-layer Topic Map-based document retrieval model is proposed in order to guide people to proper documents. It is only a primary model and is being improved and realized in our study.

REFERENCES

[1] Topic Navigation Maps - An Overview. By Martin Bryan, The SGML Centre, Churchdown, Gloucestershire, United Kingdom. http://www.isgmlug.org/n3-4/n3-4-15.htm

[2] What's in a topic map? http://www.webreference.com/xml/column77/2.html

[3] The TAO of Topic Maps-Finding the Way in the Age of Infoglut, Steve Pepper, Chief Strategy Officer, http://www.ontopia.net/topicmaps/materials/tao.html

[4] XML Topic Maps (XTM) 1.0-TopicMaps.Org Specification, Members of the TopicMaps.Org Authoring Group, http://www.topicmaps.org/xtm/1.0/#desc-subject-identity

[5] Towards a General Theory of Scope, Steve Pepper, Geir Ove Grønmo, http://www.ontopia.net/topicmaps/materials/scope.htm

[6] Information Retrieval with Conceptual Graph Matching, Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh, http://ccc.inaoep.mx/~mmontesg/publicaciones/2000/IRwithCG_New-dexa00.pdf

[7] Taking RDF and Topic Maps seriously - what happens when you drink the Kool Aid, Kent Fitch, http://ausweb.scu.edu.au/aw02/papers/refereed/fitch2/paper.html

[8] Semantic Networks, John F. Sowa, http://www.jfsowa.com/pubs/semnet.htm

[9] The Topic Map Constraint Language, http://www.isotopicmaps.org/tmcl/

[10] Topic Map Constraint Language , Graham Moore, Mary Nishikawa, Dmitry Bogachev , http://www.jtc1sc34.org/repository/0549.htm

[11] Topic Map Query Language, http://searchwebservices.techtarget.com/sDefinition/0,,sid26_gci520929,00.html

[12] The SGML Standardization Framework and the Introduction of XML, Walter Fierz, MD; Rolf Grutter, DVM, http://www.jmir.org/2000/2/e12/

[13] Topic Map technology - the state of the art, Graham Moore, http://www.gca.org/papers/xmleurope2000/papers/s22-04.html

[14] Topic Maps for repositories, Kal Ahmed, http://www.gca.org/papers/xmleurope2000/papers/s29-04.html

[15] Research of "Topic Maps" standard and its application, Peiyun Zhang, Jiang Wu, Yun Jia, Journal of Anhui University(Natural Sciences), Vol. 28, No. 3, May 2004, pp. 19-22 (in Chinese)

[16] Research of Knowledge Representation technology based on Topic Map, Feng Dai, Journal of South-Central University for Nationalities(Natural Sciences), Vol. 23, No. 1, March 2004, pp. 84-87(in Chinese)

[17] Towards knowledge organization with Topic Maps, Alexander Sigel M.A. Bonn Germany http://www.gca.org/papers/xmleurope2000/papers/s22-02.html

[18] Making topic maps more colorful, Hans Holger Rath, Rimpar Germany, http://www.gca.org/papers/xmleurope2000/papers/s29-01.htm