Application and Evaluation of Topic Maps for the Cultural Resource Data — Experiment with the Graphic Data of "Genji Monogatari (the Tale of Genji)" —

Project Leader	Mari Nagase	Department of Information Shizuoka University
Collaborator	Motomu Naito	Research & Development Synergy Incubate Inc.

Abstract

The objective of this research is to explore and study the new paradigm for effectively utilizing and evaluating, from users' perspective, not from creators', diverse cultural resource data that are archived at libraries, museums, research institutes, universities, and corporations today. The research starts with applying the Topic Map paradigm to relatively small data. The data size is then gradually scaled up in order to examine the applicability of the Topic Map technology to large-scale database, where it is expected to be most effective. For the small data, we use 227 pieces of graphic data from "Genji Monogatari (the Tale of Genji)" and related hypertext data in Japanese and English. As extraction of topics, grouping and association to other graphic data and topics can be done with ease on graphic data, we expect that our experiment will serve as a model which users may follow when they attempt to build Topic Maps on the same graphic data later on. We are still in the early phase of the research and expect many things to be done in order to attain our ultimate goal. In this paper, we put forth our view of anticipated future developments in the research method as well as in the research itself and invite advice and/or criticism from readers. In the later phase of the research we plan to apply the Topic Map system to large-scale database, evaluate its usefulness and effectiveness and discuss issues found in the process. We hope that eventually the knowledge and theory developed through this research will help manage, analyze, and evaluate massive and diverse cultural resource data that have been accumulated in our society over hundreds of years and also help bring users' point of view to the forefront in the world of computer.

1 Introduction

Whereas every conventional paper book has its index, typically at the end of the book, electronic documents are rarely equipped with such arrangements that help users. What is worse is that neither common rules nor protocols exist for preparing indexes for digital

documents. Indeed, with today's high-speed search engines, one can easily retrieve, "precisely" in an alphabetical order, digital documents containing the texts he/she specifies, out of billions placed on the internet. Using portal sites like Yahoo, one can also obtain a long list of URLs related to the topic of his/her choice. The retrieved documents or search results, however, often do users practically more harm than good. In some cases, it takes the user too much time to go through all the "hits". In others, the "hits" are simply irrelevant documents containing relevant keywords.

Today we input a large amount of data on paper documents into the computer everyday with a lofty goal of accumulation and sharing of our knowledge and wisdom. Unfortunately, however, the database is often of little value to its users, because it is edited and organized with its creators' point of view only. As Steve Pepper, one of the proponents of the Topic Map paradigm, pointed out by saying "A book without an index is just like a country without a map", an electronic document without a index is useless when it comes to building a solid knowledge base. The Topic Map system was invented to solve the problem with users in mind and subsequently formalized by the ISO as an international standard.

In this research we attempt to examine the usefulness and effectiveness of the Topic Maps system by applying it to "Genji Monogatari (the Tale of Genji)", a masterpiece of Japanese classic literature and evaluating the results. We also attempt to develop and evaluate a prototype of the Topic Map authoring tool.

We hope that, with the authoring tool, anyone who wants to build Topic Maps on his/her own will be able to do so without knowing its syntaxes or being an Topic Maps expert.

2 Graphic Data from Genji Monogatari (the Tale of Genji)

With the support of the Universite de Charles-de-Gaulle-Lille III, the French government, the Japanese government and private funds, the graphic database had developed by one of the author of this paper from 1996 to 1998 in France. And now it can be seen at UNESCO site. (http://webworld.unesco.org/genji/)

The Tale of Genji was written by Murasaki-shikibu about 1000 years ago in the Heian era(794-1192.) The graphic data of 227 pictures we use this research were scanned into the computer from "Ehon Genji Monogatari" an illustrated book for Genji Monogatari. The pictures were created with woodcut in 1650 by Shunshou Yamamoto, one of the finest Japanese lacquer artists in the Edo era(1603-1867.) The book was used for the processing, instead of the original woodcuts that are archived at the Library at Tsurumi University and the Office of Japanese Literature at Tokyo University, largely because the originals were in the unfavorable conditions and their availability for observation was restricted. Another factor was that the book featured a five-to-six line caption for each picture, allowing readers to follow the story with ease.

The 227 pictures can be grouped under the following six themes.

1. Waka (Japanese poems) composition (including letter writing, writing practice, and

painting): 38 pictures

2. Music performance: 9 pictures
3. Party and Play (including hunting, boating, Japanese dolls, go (a kind of board game), sugoroku (Japanese backgammon), and dancing): 35 pictures
4. Love (including secret love and peeping): 41 pictures
5. Traveling (carriage, horse and boat): 16 pictures
6. Others (nature, talk, ghost, death, prayer, religious service, religious event, birth, visit, walk, hiking, dream, education, etc.)

The sum of 1. Waka, 2. Music performance, 3. Party, and 4. Love totals 123, accounting for 54.4% of all the pictures. Many, however, fall in multiple categories, so the percentage is not necessarily accurate.

Figure 1-4 shown below are select scenes from some of the category. The picture numbers in the caption of figures represent the volume number and the serial number of the picture in the original database.

Figure 1: Picture 2-8

Figure 2: Picture 4-18

Figure 3: Picture 33-124

Figure 4: Picture 3-14

3 Research Process

3.1 Topic Map Application to Genji Monogatari

In this research we applied the Topic Map system to the graphic data from Genji Monogatari as described in Chapter 2 in order to acquire know-how on Topic Map building and evaluate the effectiveness of the Topic Map. In building the Topic Map, we followed the process developed by Steve Pepper as shown below as a guidance.

1. Ontology Definition
   (a) Domain Definition
   (b) Domain Analysis
   (c) Topic Definition
   (d) Association Definition
   (e) Occurrence Definition
2. Tool Selection
3. 'Legacy' Data Handling
4. Topic Map Building
5. Ontology Review
6. Repeat 2) to 5) to refine the Topic Map

As a first step of the process, we first defined the subject domain as whatever covered in the illustrated book for Genji Monogatari.

We then analyzed the domain and extracted the candidates of topics, associations, and occurrences. Many subjects are included in the pictures. So we can extract various topic candidates from various point of view.

We utilized a regular editor for Topic Map building purposes and Ontopia's Omnigator for Topic Map display purposes. Omnigator was selected because it was readily available for free and also supported Japanese language.

Initially, we built the Topic Map manually using the text editor and made improvements by switching back and forth between Omnigator and the editor. In the process of developing the Topic Map, we also reviewed the ontology from time to time and gave refinements as deemed necessary.

3.2 Authoring Tool

Our initial approach to Topic Map building was just-do-it, using samples as a model and the published standard as a reference. In the meantime we actively participated in tutorial sessions to learn more about Topic Maps and gain skills in Topic Map building. We had lots of valuable advice from Ontopia.

In addition to how to construct the ontology, major problems we found while developing the Topic Map were as follows.

- Moderate or better understanding of the syntax was necessary.
- Typos and syntax errors were inevitable.
- It was impracticable to input manually large volume of data.
- Topic map's simplicity, flexibility, and versatility made it difficult to decide what was the 'proper' representation of a given Topic Map.

We worked on developing the prototype of the authoring tool that would solve the above problems. At the time of this writing, we can input types and instances of the major Topic Map elements (topic, association, and occurrence) through the interface generated using JSP and XSLT. We can also create XTM syntax-based Topic Maps based on the input data, which is viewable with Ontopia's Topic Map navigation tool, Omnigator.

4 Evaluation

The Topic Map system allows users to create data maps outside and independent of data resources and navigate freely between various points of view. We have an impression that the Topic Map technology has a significant potential. As for cultural resource data, Topic Maps are expected to benefit its users in many ways, including:

1. To help organize the cultural resource data from various point of user's view and navigate it through the view

2. To help expand ways to utilize it
3. To help transfer creator's idea, view, experience etc. to another person
4. To help build effective cultural resource databases without being affected by the database editor's point of view
5. To help mine wisdom out of data
6. To help provide new research subjects

On the other hand, we see the need for further development and improvement with Topic Map system and Topic Map tools. Indeed, a pile of issues exists because of it's high potential. We will show "the ten issues." We will discuss such issues and possible solutions in the following.

1. Improve Topic Maps' knowledge representation capabilities

   Though their have the possibility of losing simplicity of Topic Maps.

   (a) Allow topics to have attributes and methods

   If topics had attributes and methods, just as objects do in the world of the object-oriented technology, they would have more power to represent themselves better. For example, if such attributes as "size", "color", "singing voice", "habitat" and "diet" are attached to a topic "bird", the information that the topic carries becomes richer.

   (b) Increase common association types

   Currently, only "SuperClass-SubClass" and "Class-Instance" are available as common association types. It would be beneficial to users to add more association types, such as "Whole-Part" that represent common relationships between topics.

2. Establish a general method for developing ontologies

   A topic can be anything. It can be something real and thus very intuitive such as "cranes", "birds", etc. It can be very abstract in nature as with "ideal", "happiness", "wisdom", etc. A general method to reify and organize these subjects and relationships between them is needed.

3. Provide Topic Map users with inference capabilities

   If a user is able to find out new or hidden relationships between topics by inferring them from those expressly defined and contexts, then he/she can make better use of existing Topic Maps.

4. Enable users to view Topic Maps flexibly

   Topic maps can easily go beyond human comprehension if the ontology gets overly complex or loaded with massive data. A tool which enable users to view Topic Maps in a graphical format, zoom in for parts and layers, and zoom out for the big picture needs to be developed.

5. Develop a tool with automatic count capabilities

   It is beneficial to Topic Map users to be able to evaluate Topic Map elements (topics, associations, etc) and know the relative importance of each element. One way to do so is to count movements between Topic Map elements as users navigate on it. A Topic Map tool with auto-count capabilities needs to be developed.

6. Enable handling of large data

   It is practically impossible to handle large volumes of data manually whether it's inputting, editing, archiving, retrieving, or displaying. Computer software and devices that enable handling of such large data need to be developed.

7. Enable auto-generation of Topic Maps

   If we were able to generate Topic Maps automatically from existing data, not only would it save significant amounts of time and work but help expand the horizon of Topic Maps application. Auto-generation of Topic Maps is expected to be relatively easy with data in the XML format.

8. Enable dynamic response to changes

   It is nothing unusual for Topic Map elements to change after the Topic Map is built on a certain knowledge domain. For instance, web sites and their contents change constantly with additions, updates and deletions. Topic maps need to respond dynamically to the changes in order for them to stay useful and relevant.

9. Standardize topics

   Standardization of topics is critical for efficient and effective exchange, distribution and integration of Topic Maps. The Topic Map Published Subjects Technical Committee at OASIS has been working on this issue.

10. Establish a Topic Map evaluation method

    We all think more or less differently. 100 people may possibly create 100 different Topic Maps on the same domain of knowledge. Because of this, it is extremely difficult to decide whether a Topic Map is proper and appropriate. A standardized Topic Map evaluation method needs to be established.

5 Outlook

One of the major new trends that have been brought by growing networks is personalization. Today one can have a glimpse of the trend toward personalization in many areas, particularly in business. One-to-one marketing and on-demand publishing are just a few such examples.

Meanwhile, with information technology more advanced and sophisticated, knowledge sharing over the network has been one of the hot topics in our society attracting much public interest and discussions. On-line libraries and on-line museums, the classic case of knowledge sharing over the network, are expected to be available soon.

In our opinion, however, those systems that archive and manage massive information resources for the public should be the first to apply the concept of personalization. To supply information tailored to individual needs and to personalize databases from users' point of view are the key to the effective use of information.

In addition to the traditional database building with administrators' logic, the mapping of electronic documents from users' perspective will be critical in the future. The goal should be the restructuring and reorganization of information resources on the semantic and metadata level, with the use of new paradigms for information resource management such as Topic Maps.

This restructuring, however, calls for published subjects to be developed for each individual domain of knowledge and then evaluated for validity. Expertise in knowledge information, information technology and artificial intelligence is necessary for successful development. Involvement of experts in each domain of knowledge and study of information user behavior are also indispensable.

The next step will be to explore ways to restructure information resources on the semantic and metadata level. The ultimate goal will be to complete the standardization of topics for each domain of knowledge and make the restructured information resources available to users as semantic networks. This research is a small step toward the final destination.

References

[1] Benedicte Le Grand, Michel Soto, "Information Management - Topic Map Visualization", Conference Paper at ACM CIKM'99, 1999.

[2] Steve Pepper, "The TAO of Topic Maps." Conference Paper at ACM CIKM'99, 1999.

[3] International Standard Organization, "ISO/IEC 13250 Information Technology - SGML Applications - Topic Maps", ISO/IEC 13250:2000, 2000.

[4] Charles F. Goldfarb, "SGML Handbook", Tokyo ASCII, 1993.

[5] John F. Sowa, "Knowledge Representation Logical, Philosophical, and Computational Foundations" Brooks/Cole, ISBN 0-534-94965-7.

[6] TopicMaps.org, "XML Topic Maps (XTM) 1.0", 2001.

[7] Topicmaps.net, "Topicmaps.net's Processing Model for XTM 1.0, Version 1.0.1", 2001.

[8] Steve Pepper, "Modelling Topic Maps", Tutorial Handout XML 2001 Orlando Florida, 2001.

[9] ISO/IEC JTC1 SC34, "Topic maps, roadmap for further work", ISO/IEC JTC1 SC34 N0278, 2001.

[10] Michel Biezunski, Steven R. Newcomb, "The Topic Maps Technical Workshop", KT 2001 Austin Texas, 2001.