A Knowledge Network Constructed by Integrating Classification, Thesaurus, and Metadata in Digital Library

Jun Wang[*]

[*] Information Management Department, Peking University, Peking, China. E-mail: junwang@pku.edu.cn

Available online 28 June 2003.

Abstract

For the digital libraries of China, the development of networked information resources, especially of metadata, is still a principal task. At the same time, many information resources have been accumulated and the exploitation of them is far from sufficient. The current chief utility — keyword search undermines seriously the potential value of information resources, especially the metadata, which contains valuable content description and indexing information. To solve this problem, we have devised a new paradigm — an integrated knowledge network. It is formed by merging the classification and thesaurus into a concept network and then distributing the metadata records into the concept nodes according to their subjects, just like the "shelving" of the metadata. We have built an experimental system, *VISION*, by incorporating a portion of the *Chinese Classification and Thesaurus* with the bibliographic data of the computing domain from Peking University Library. Such a knowledge network is not only a framework for metadata organizing, but also a structure for knowledge navigation, retrieval, and learning, and we believe it will be a catalyst for better knowledge management in digital libraries.

Article Outline

• Introduction
• A Review of the Classification and the Thesaurus in China
• The Knowledge Network
• Classification, Thesaurus, and Bibliographic Data
• The Construction of the Knowledge Network
• The Functions of KNICTM
• The Building of an Experiment: The VISION System
• The Ontology Design
• *VISION* on the Server Side: the Knowledge Network
• The *VISION* System on the Client Side: Knowledge Navigation and Retrieval
• Conclusion and Future Plans
• Acknowledgements
• Appendix A. VISION's Ontology in Ontolingua
• Appendix B. VISION's Ontology in RDFS (Protégé)
• References

Introduction

The library has been the center of the preservation, utilization, and distribution of information and knowledge of human beings for centuries. With the rising and the rapid development of the web, the

role and function of the library are questioned. The web is the most widespread and easily accessible information resource, and its popularity continues to grow unabated. In turn, digital libraries have found a home on the web. Digital libraries distinguish themselves from other web information resources by having a much greater capacity for knowledge management. To digitize information and transfer it onto the web is far from knowledge management. How then do digital libraries achieve knowledge management? This is the problem confronting all libraries worldwide, including those in China.

Let's examine the China digital libraries more closely. On the one hand, the current development of networked information resources, especially metadata, is critical for Chinese digital libraries. For example, the China Digital Library Project (CDLP) and the China Academic Library and Information System (CALIS), the two biggest digital library projects in China, both spent a large proportion of their funds on the development of information resources.

On the other hand, many networked information resources have been accumulated. For example, in just 4 years CALIS has imported 89 databases containing 7800 kinds of academic journals, and its hundreds of members have produced 1150 thousand records for the union catalogue, 1370 thousand records of indexes from over 5500 Chinese academic journals, and 70,000 records of abstracts of theses and dissertations.[1] Compared with the production of resources, the utilization of them is deficient. Keyword-based searching is applied everywhere, both for indexed databases and full-text web pages. In the keyword matching, the valuable content description and indexing of the metadata, i.e., the subject descriptors and the classification notations, are merely treated as common keywords to be matched with the user query. Without the support of the vocabulary control tools (e.g. classification and thesaurus), content analysis, descriptive term searches, and metadata indexes are underused. New retrieval paradigms are needed in order to exploit the potentials of the metadata resources.

With these problems in mind, we review the structure of the traditional library. In order to provide high-quality document services for users, the documents are described and indexed first, which produces catalogues. The documents include books, journals, cassettes, and other items from the library. Then the catalogues are arranged in various categories, and the documents are organized into a given hierarchy (i.e., shelving). All these are done according to the classification and thesaurus. What is absent in current digital library architecture is the classification and thesaurus — the vocabulary control and the knowledge organization tools which serve three purposes in a traditional library: The description, organization, and retrieval of document collections. The first corresponds to metadata production. The second, including catalogue ordering and book shelving, tailors the general knowledge of classification and thesauri into the specific domains of the library collections, and incarnates them in a concrete way, viz., the ordered catalogues and organized shelves of books in library. On this basis, the readers can browse and search to see what they want.

For more efficient and effective exploration, the networked information should be pre-arranged together with vigorous improvement of search techniques, before the applicable information is made available on the web. Could classification and thesauri that contain the condensed intelligence of generations of librarians be used in digital libraries to organize the networked information, especially metadata, to facilitate the information resources usability and catalyze the digital library into knowledge management?

Yes. In light of the structure of traditional libraries, we design and implement a new paradigm which incorporates the classification, the thesaurus, and the metadata. The classification and the thesaurus are merged into a concept network, and the metadata are distributed into the nodes of the concept network according to their subjects. The abstract concept node substantiated with the related metadata records becomes a knowledge node. A coherent and consistent knowledge network is thus formed. It is not only a framework for resource organization, but also a structure for knowledge navigation, retrieval, and learning. We have built an experiment system based on the *Chinese Classification and Thesaurus*,[2] which is the most comprehensive and authoritative in China, and we have incorporated

more than 5000 bibliographic records provided by Peking University Library. The results are indeed encouraging.

The next section presents a review on the classification and the thesaurus in China, and the section next presents the architecture of the knowledge network integrating classification, thesaurus, and metadata. In the penultimate section, the construction of our experimental system named *VISION* is explained step-by-step, and the final section is the conclusion and a look at the future.

A Review of the Classification and the Thesaurus in China

China, with its long literary tradition, has consistently used schemes of classification. Modern Chinese classification was greatly influenced by Dewey, although the *Dewey Decimal Classification* wasn't popular in China. All the classifications now in use were created after the founding of the People's Republic of China. The most perfect one among them is the *Book Classification of Chinese Libraries* (BCCL) which was published first in 1975 and has undergone four revisions. Others worthy of mention are the *Book Classification of the People's University Library in China,* which is now in its sixth edition, and the *Book Classification of the China Science Academy Library,* which is now in its third version. There are also dozens of domain-specified classifications, but none of them is well-known.

The first Chinese thesaurus was produced in 1964. The most famous comprehensive thesaurus in China is the *Chinese Thesaurus* (CT), compiled by more than 1000 people over a 6 year period (1974–1980). It was the biggest thesaurus at that time, containing 91,158 preferred terms and 17,410 non-preferred terms. Influenced by the faceted thesaurus, a huge project was started in 1986 to combine the BCCL and the CT into one — the *Chinese Classification and Thesaurus* (CCT). More than 40 institutions were involved, and it was finished in 1994, with a total of 14 millions words in six volumes. Currently, it has been used in all the public libraries and more than 90 percent of non-public libraries and information institutions of China.[3] Now there are more than 100 various thesauri in China. Among them, the *Military Thesaurus of China* is well-known for being the only computerized thesaurus in China.

The CCT is not designed to be used in a network environment. To apply it in a digital library has several inherent obstacles; most of them are common to other classifications and thesauri.[4] The CCT was complied almost 10 years ago. Although it has undergone several amendments, it cannot keep up with the myriad variations in dynamic networked information. Two parts of the CCT, the classification and the thesaurus, are relatively independent of each other and cannot be updated synchronously. The CCT is a comprehensive system. Its general knowledge cannot be tailored flexibly into the specific domain of a given information collection; the CCT is a professional tool to be used by librarians. Its infrastructure, such as syntaxes and rules, are too complicated for common users to master. The CCT is mainly meant for the description and organization of information resources. It is not so good at retrieval; CCT's design is more oriented for the management of hard-copy documents, and some of its usage criteria lose their sense in the digital environments, such as one-place shelving (i.e. there is just one unique shelf place for a given book). For these difficulties, there is seldom application of the CCT in the network environment.

The Knowledge Network

Classification, Thesaurus, and Bibliographic Data

It seems impossible to deploy traditional classifications and thesauri in digital libraries. But there exists one kind of networked information resource in which the classification and thesaurus are deployed thoroughly — the bibliographic data of the OPAC system. The bibliographic data is one of the most important resources the library (and only the library) owns. As collections of contemporary materials, they contain plenty of new professional terms in the title field, which supplement domain knowledge. More importantly, they are indexed strictly according to the classification and thesaurus.

Based on the subject indexing, the bibliographic records can be combined with the classification and thesaurus to complement each other. It's a kind of metadata "shelving". The knowledge structure of the classification and thesaurus provides a skeleton for the organization of the bibliographic data; the concrete bibliographic data provide substance for the frame work. Thus, a corpus of knowledge is formed where new terms can be extracted automatically from the bibliographic data to update the classification and thesaurus; the classification and thesaurus is customized to the specific domains of the OPAC resources. Such a knowledge network provides the user with a natural structure for navigation, searching, and learning. For convenience, we designate such knowledge network as "KNICTM" (Knowledge Network Integrating the Classification, Thesaurus, and Metadata).

The Construction of the Knowledge Network

The KNICTM is built in three steps:

• *Construction of the original concept node based on the classification and thesaurus.* First, the thesaurus is turned into an original concept network that consists of nodes and edges. A node is composed of the synonym set of a subject descriptor, including the descriptor and all the terms connected to it by the equivalent relationship in the thesaurus. If there is a hierarchical relationship between two terms, an "is-a" edge is set up between the corresponding concept nodes. The associative relationship is discussed in the next step. Then the classification scheme is embedded into this original concept network, and acts as a disciplined hierarchical backbone for the concept network (Fig. 1). The CCT is a reciprocal index between the BCCL and the CT rather than a faceted thesaurus, and there is no direct mapping between the categories of the BCCT and the concepts of the CCT. Therefore, the category nodes are created and the relationships between the category nodes and the concept nodes are set up.
• *Distribution of the bibliographic data to the concept network.* Secondly, the bibliographic data are distributed into the nodes of the original concept network according to their subjects. This is the key task of the KNICTM construction. Bound with the corresponding bibliographic data records, an abstract concept node becomes a knowledge node (Fig. 1) and the concept network turns into a knowledge network. The distribution of the bibliographic data records (BDR) in the original concept network is explained as follows: If a BDR contains only one subject descriptor, we take the BDR as one of the instances of the corresponding concept node and add the BDR into the node. If a BDR contains several descriptors, then we add the BDR into all the related concept nodes as instances of them. If a BDR contains a composite subject which is denoted by a coordination of descriptors, then we create a new concept node, and connect it to all the corresponding nodes of the coordinate descriptors with "related-to" edges. The newly-created concept node is called a "co-concept" node, which has the BDRs merely and no term for the moment. For example, a BDR with the title "*Internet Firewall Technologies*" is indexed with the string "Network-Security", for there is no "firewall" in the thesaurus. To add this BDR into the concept network, a co-concept node is created, and it is connected to the concept node of "Network" and the concept node of "Security" by 'related-to' edges. Because the associative relationship easily goes out of control in the thesaurus, we don't consider it in the first step. Only when the correlation of two concepts is supported by a BDR do we establish the 'related-to' relationship between them through a co-concept. Thus, the BDRs verify the associative relationship. In the next step, when there is a new extracted term with the meaning of the co-concept, the new term is added to the co-concept. In the above-given example, this is a 'firewall'. The KNICTM needs manual examination periodically to confirm the co-concepts created. When a preferred term is determined for the co-concept, the co-concept node turns into a common concept node.
• *Enhancement of the KNICTM.* The last and the most difficult task is to mine new terms from the metadata collection to enhance the KNICTM. The title of a scientific document usually summarizes its content and reveals its central topics. And, there exists a direct mapping between the keywords of the title and the subject indexing. Based on their semantic mapping, new terms can be extracted from the title and added into the concept network. There are three difficulties to accomplishing this. How to extract valuable terms out of the common terms in title? How to determine the position where the extracted terms should be inserted into the KNICTM? Prior to all these processes, the titles of natural

language must be segmented into word or phrase, the classical obstacle of Chinese language processing.

 (5K)

Figure 1. The knowledge network with a close-up of a knowledge node.

The construction of the KNICTM is similar to a tree. The classification hierarchy is the trunk and branches of the tree; the conceptual relationships of the thesaurus are the veins, and the metadata are the leaves. The basic construction unit of the KNICTM is the knowledge node. It is the concept node attached with references. Two kinds of edges join the knowledge nodes together: the hierarchical edges ("is-a"), which are the meridians of longitude of the KNICTM, and the associative edges ("related-to"), which are the parallels of latitude.

The Functions of KNICTM

• *A Framework for the organization of network resources.* The KNICTM provides a framework for the organization of the networked information resources, especially metadata. It is a network of knowledge with substantial data appended rather than a mere abstract concept network. As the instances of the concept, the metadata records inherit all the relationships among the concepts. The records of metadata which were isolated from each other become semantically connected now and are woven into an interconnected knowledge network.
• *An adaptive concept network based on the applied resources.* The classification and thesaurus represent general knowledge, and cannot fit a specific information collection perfectly. The KNICTM is an adaptive concept network capable of self-customizing, based on the scale and domain of the given collection. The nodes and edges supported by the metadata instances prove the usability of the correspondent concepts and relationships, and they are revealed to the user. The nodes and edges without the support of the metadata instances indicate that the correspondent concepts and relationships are unusable and may need updating. Furthermore, statistic and semantic techniques can be applied in mining new terms, concepts, and relationships from metadata collection to enrich concept networks automatically.
• *A structure for knowledge navigation and retrieval.* The keyword-based search undermines the value of the metadata seriously. The KNICTM provides a conceptual retrieval network and visual navigational ontology. First, the KNICTM can guide the user to clarify his information demands and express his query clearly. Second, because all the metadata have been arranged into the KNICTM, there is no need to dig into the metadata collection by keyword searching; it is only necessary to locate the knowledge node according to the user query, and follow the surrounding edges to reach other nodes to obtain the desired result. Third, now that all the metadata have been arranged into the structure of the KNICTM according to their subjects, the retrieval result is displayed in that structure, already ranked and classified.
• *A well-organized knowledge network to support knowledge learning.* The KNICTM consists of the knowledge node that is identified by a concept composed of a set of synonymous terms and supported by some metadata records. The knowledge nodes are organized into a hierarchy and clustered into topic areas through edges among them. A friendly interface like Cat-a-Core[5] can manifest the structure of the knowledge network. A user assisted by such an interface can learn the discipline structure of a domain, master the professional terms, understand the relationships among the subjects, and pick up the documents to study.
• *A digital library of knowledge management.* The most essential element of a library is information. In KNICTM, information resources are integrated into a knowledge structure described by both classification and thesaurus. The KNICTM can be easily extended to support other activities of a digital library, such as resource collecting and indexing. Thus, all the activities of a library, including indexing, organization, navigation, retrieval, and learning, center on this knowledge network. To

develop continuously, the KNICTM must move the digital library from information management to knowledge management.

The Building of an Experiment: The VISION System

We built an experimental system named *VISION*. In *VISION*, we combined the classification and indexing terms from the computing domain in the CCT with all the bibliographic records for Chinese computer science materials held by the Peking University Library, published between 1990 and 1999, totaling 5324 records. The *VISION* system has a client/server architecure; on the server side is the knowledge network supported by Oracle9i; on the client side is the user interface implemented in Java. We chose Oracle9i, for we need its powerful object-oriented feature such as nested table and variable array to support our complex objects. Java was selected in order to make it easy to transplant the system to the web.

We will first introduce the ontology of our system. Then we outline the architecture of the system on the server side and the function of the conceptual retrieval tool on the client side. Finally, we will introduce the work we have done in extracting new terms from the bibliographic data to enhance the knowledge network.

The Ontology Design

There are many objects in our system and their relationships are complex, so we used ontology tools such as Ontolingua[6] and Protégé[7] to design our system's ontology. Then, we converted this ontology into the database schema. Our ontology consists of seven classes: term, concept, co-concept, category, document, author, and publisher. Fig. 2 depicts their relationships, with the numbers indicating the cardinality of the links. Pieces of ontology in the syntax of Ontolingua and in the RDF schema rendered by Protégé are provided in Appendix A and Appendix B for further reference. We then convert the ontology into the relational schema of the database system and create the corresponding tables in Oracle.

VISION on the Server Side: the Knowledge Network

The original data set of the *VISION* system includes the e-text of the CCT and the bibliographic data of the computing domain. Both of these were provided by the Peking University Library. The characteristics of the original data have considerable influences on the system design and implementation. There are three steps to building the *VISION* system on the server side:

• *The e-text file of the CCT is processed to set up the fundamental structure of the VISION* system. A particular tool was developed to serve this purpose. The e-text of the CCT is read in, and all the entries (i.e. categories and terms) on computer science are processed. According to the structure, layout, and notation rules of these entries, the related records are created and appended into four tables, respectively: TERM, CONCEPT, CoCONCEPT, and CATEGORY. In this process, we get 2194 terms, including 1684 preferred terms and 278 non-preferred terms; and 1684 concepts are created. Some non-computing-domain terms are also captured, for they are the related terms of the computing domain terms. Every subject is treated as a concept, so the quantity of the preferred terms is equal to the quantity of the concepts.
• The bibliographic data are loaded in the database and distributed into the original concept network constructed in the preceding process. The bibliographic data are in CNMARC format. A particular tool is developed to decode the perplex CNMARC format and read each bibliographic record out, and the required fields are extracted (title, subject, author, etc.) to form a new record, which is appended into the table DOCUMENT. A total of 5053 document records have been created. Others are discarded for various reasons, for example, unrecognized titles or dual ISBN numbers. Such data processing requires a lot of time and energy. After the data loading, the records of the DOCUMENT table are connected with the records of the CONCEPT table, based on the semantic correspondence

between the subject indexing of the document records and the concept records. When necessary, new records are created in the CoCONCEPT table.

• *New terms are extracted from the DOCUMENT table and added to enhance the knowledge network of the VISION* system. Some of the problems have already been mentioned. This process is the focus of the ongoing second phase of the Vision project, so here we outline roughly what we have done to date.

The extraction: At present a statistic algorithm is applied to extract terms in the titles. First, the title is segmented into basic words and phrases by general segmentation tools, and then the co-occurrence frequency of the neighbors is counted. If the frequency is bigger than a given threshold, the combination is selected as a new-term candidate τ; then all the indexing terms are collected from the document records whose title field contain τ. Thus a set of indexing terms I is formed. The distribution of I is computed. If the distribution of I is convergent, then τ is a new term.

The insertion: The convergent point found above helps to determine the position where the new term should be inserted. Generally, the extracted new term is the narrower term of the subject represented by the convergent point. Now we are considering applying Lattice Theory[8] or Formal Concept Analysis [9] to this problem.

The *VISION* System on the Client Side: Knowledge Navigation and Retrieval

We implemented a knowledge retrieval system in Java to navigate and retrieve the *VISION* knowledge network. Fig. 3 is the snap-shot of the user interface. There are four physical areas in the interface: the query dialog, the concept network window, the information window, and the document window. Within the concept window, there are three basic ways to view the concept network: hierarchical tree, alphabetical list, and concept family. The hierarchical tree is similar to a faceted thesaurus. All the categories and concepts (identified by the preferred terms) are organized into an expandable conceptual tree. The alphabetical list is an index of all the terms in alphabetical (Chinese Pin Yin) order. The concepts can also be organized into concept families, that is, the term families of the thesaurus, and listed in alphabetical order by the top concepts. There is a fourth option, which is a hybrid of the hierarchical tree and the alphabetical list. When the user clicks on a concept, its detail information is displayed in the information window, including its term set, super-concept, sub-concepts, corresponding category, and the co-concepts around it; the documents connected with it are displayed in the document window. All the windows trigger each other and act in a chain, and all the objects in the windows are clickable. When a query is submitted, what is returned are not only the documents related, but also the concepts and co-concepts pertinent. The concept network window is triggered to highlight the position of the retrieved concepts in the concept hierarchy, and the relationships with other concepts are shown; the document window is triggered and the relevant documents are revealed. All these come to the user together as an integral knowledge unit.

Conclusion and Future Plans

Centuries of library work have proven that the organization of information is the basis for effective utilization of information resources, whether the library is traditional or high-tech. Because of the lack of organization, the potential of metadata as one of the most important networked resources is barely exploited. This article has presented an approach to organizing the metadata by both classification and thesaurus into an integrated knowledge network, establishing a new paradigm for the knowledge management in digital libraries. Our approach is distinguished from other ontology-driven and concept-based systems by incorporating the concepts and the relevant metadata records into integrated knowledge nodes that form the knowledge network. Our experiment also demonstrates that traditional resources such as bibliographic data still have indispensable value in spite of the ever-increasing networked resources.

The *VISION* system is entering its second phase of development. We seek to achieve the following:

• A better method for computing the appropriate position where the extracted term should be settled into the concept network. This is critical for the concept network to be self-sufficient. Now we consider applying an adjusted Formal Concept Analysis to this problem.
• A language for the concept query and manipulation in the concept network will simplify the operation of the concept network and add the automatic query expansion and contraction mechanism.
• A visualization interface such as Cat-a-Core[10] or Inxight Star Tree [11] can provide more friendly interaction with the user. A visualization interface that is structurally isomorphic to the concept network will support knowledge learning more powerfully.
When all the aspects of the system have been tested, the system will be transplanted to the web and incorporated with the current OPAC system of Peking University.
To integrate the classification, thesaurus, and metadata into a coherent knowledge network has promising applications in digital libraries. It could easily be expanded to support automatic classification and indexing in scientific domains. Enhanced by the bibliographic data, the knowledge network could absorb other metadata, such as index databases of journals, magazines, or newspapers. The web community also recognized the importance of the standardization and organization of the web information. XML, RDF, Dublin Core, and other specifications are preparing the web for the manageable web — the Semantic Web as envisioned by Tim Berners-Lee.[10] But how to construct it? Our paradigm provides one approach.

Acknowledgements

This research is an excerpt from my doctoral dissertation written at Peking University. I wish particularly to thank Zhu Xingguo, Deng Peng, and Zu Yong, all of whom assisted me in developing the Vision system. In addition, I would like to thank Professors Yang Dongqin and Tang Shiwei, as well as Dai Longji, President of the Peking University Library, and his staff.

References

1. Chen, L. (2001) The Report on the development of CALIS project at the end of first phrase (Chinese). http://www.sciencedirect.com/science?_ob=RedirectURL&_method=externObjLink&_locator=url&_cdi=6828&_plusSign=%2B&_targetURL=http%253A%252F%252Fwww.calis.edu.cn%252Frm%252Fcalis211.pdf

2. Liu, X.Sh., Zhang, Q.Y., *et al.* (Eds) (1994) *Chinese Classification and Thesaurus*, Beijing, Hua Yi Press. 1994.

3. Zhang, Q.Y., Liu, X.Sh. & Wang, D.B. (1996) Modern classification systems and thesauri in China. *62nd IFLA General Conference*, Booklet4, Beijing, August 25–31, 1996, pp. 31–37.

4. Zhang, Q.Y. (1998) Discussions on the information retrieval language of 21 Century. *Forum of Libraries* **21**(5).

5. Hearst, M.A. (1997) Cat-a-cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. *Proceedings of the Twentieth Annual International ACM SIGIR Conference*, Philadelphia, PA, July 1997.

6. Gruber, T.R. (1993) A translation approach to portable ontology specifications, *Knowledge Acquisition* **5**, pp. 199–220.

7. Noy, N., Fergerson, R. & Musen, R. (2000) The knowledge model of protege-2000: Combining interoperability and flexibility,

http://www.sciencedirect.com/science?_ob=RedirectURL&_method=externObjLink&_locator=url&_cdi=6828&_plusSign=%2B&_targetURL=http%253A%252F%252Fwww.smi.stanford.edu%252Fpubs%252FSMI_Reports%252FSMI-2000-0830.pdf

8. Pedersen, G.S. (1993) A browser for bibliographic information retrieval, based on an application of lattice theory. *ACM-SIG'93-6/93/Pittsburgh*, PA, USA.

9. Ganter, B. (1999). *Formal concept analysis*. Berlin, Heidelberg: Springer.

10. See footnote 5.

11. Inxight Software Inc. (2001), Inxight Software, http://www.sciencedirect.com/science?_ob=RedirectURL&_method=externObjLink&_locator=url&_cdi=6828&_plusSign=%2B&_targetURL=http%253A%252F%252Fwww.inxight.com