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

Wang Jun
Information Management Dept.
Peking Uni., Beijing, China
junwang@pku.edu.cn

Abstract

For the digital libraries of China, the development of networked information resources, especially of metadata, is still a principal task. In the same time, a huge amount of information resources has been accumulated and the exploitation of them is far from sufficient. The current chief utility—keyword search undermines seriously the potential values of them, especially the metadata, for the metadata contain valuable content description and indexing information. To solve this problem, we devise a new paradigm—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 experiment system Vision by incorporating a portion of the Chinese Classification and Thesaurus with the bibliographic data of 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 the digital libraries to come to knowledge management.

Keywords

Knowledge Management Conceptual Retrieval Metadata Classification Thesaurus

1. Introduction

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 developing of the Web, the role and function of library are questioned. People are absorbed almost completely into the Web, the most widespread and easily accessible information warehouse human beings ever owned. Digital libraries, as the reply to this questioning, have bloomed in the Web. And they wave the banner of knowledge management to distinguish themselves from the various Web information facilities [Dong 00]. To digitalize the information and transfer them onto the Web is far from the knowledge management. How do digital libraries achieve knowledge management then? This is the problem confronting all the libraries around the world, including those in China.

Let's examine the China digital libraries more closely. On the one hand, currently the development of networked information resources, especially that of metadata, is fundamental to and still a principal task for China digital libraries. For example, 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, huge amounts of networked information resources have been accumulated. For example, just in 4 years the CALIS has imported 89 databases containing 7,800 kinds of academic journals, and its hundreds of members have produced 1,150 thousand records of the union catalog, 1,370 thousand records of the content indexes of 5,500 Chinese academic journals, and 70 thousand records of the abstracts of the theses and dissertations [Chen 02]. Compared with the production of resources, the utilization of them is far from enough [Liu 00]. The keyword-based searching is applied everywhere no matter the resources are indexed databases or full-text Web pages. In the keyword matching, the valuable content description and indexing of the metadata,

such as 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), the intelligent labors of content analyzing, describing and indexing in metadata production are wasted seriously. New retrieval paradigms are needed in order to exploit the potentials of the metadata resources sufficiently.

With these problems in mind, let's retrospect the structure of traditional library again. In order to provide high quality document services for users, the documents are described and indexed first, which produces catalogs. The documents include books, journals, cassettes and other articles in library. Then the catalogs are arranged in various orders, and the documents are organized into a given hierarchy (i.e. shelving). All these are done under the instruction of 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 traditional library: the description, organization and retrieval of document collections. The first is corresponding to the metadata production of today. The second, including catalog ordering and book shelving, tailors the general knowledge of classification and thesaurus into the specific domains of the library collections, and incarnates them in a concrete knowledge corpus—the ordered catalogs and organized shelves of books in library. On this basis, reader could browse and search to pick up what he wants.

For more efficient and effective exploration, the networked information should be pre-arranged besides the vigorous improvement on the search techniques before it is made use of on the Web. Could classification and thesaurus which contain the condensed intelligence of generations of librarians be used in digital library to organize the networked information, especially metadata, to facilitate their usability and catalyze the digital library into a knowledge management environment?

Yes. In the light of the structure of traditional library, 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 resources organizing but also a structure for knowledge navigation, retrieval and learning. We have built an experiment system based on the Chinese Classification and Thesaurus, which is the most comprehensive and authoritative in China, incorporated with more than 5,000 bibliographic data of computing domain from Peking University Library. And the result is encouraging.

The next section presents a review on the classification and the thesaurus in China, and in section 3 presents the architecture of the knowledge network as the integration of classification, thesaurus and metadata. The construction of our experiment system named Vision is explained step by step in section 4. The last section is the conclusion and the future works.

2. A Review of the Classification and the Thesaurus in China

China has a long tradition of using classification due to her abundant ancient books. The classification of modern China was influenced by Dewey greatly though his classification wasn't popular in China. All the classifications now in use were created after the foundation of the People 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 4 revisions. Others worthy of mentioning are the Book Classification of the People's University Library in China which is in 6th version now and the Book Classification of the China Science Academy Library which is in 3rd version now. There are also dozens of domain-specified classifications, but none of them is well-known. [Liu 96]

The first thesaurus of China occurred in 1964. The most famous comprehensive thesaurus in China is the Chinese Thesaurus (CT) which was compiled by more than one thousand people who spent 6 years (1974-1980) on it. It was the biggest thesaurus at that time, containing 91,158 preferred terms and 17,410 non-preferred terms. Influenced by faceted thesaurus, a huge project was started in 1986 to combine the BCCL and the CT into an incorporated one—the Chinese Classification and Thesaurus (CCT). More than 40 institutions were involved, and it was finished in 1994, with the total amount of 14 millions words in 6 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 [Zhang 96]. Now there are more than one hundred 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 the network environment. To apply it in digital library has several inherent obstacles, most of them are common to other classifications and thesauri [zhang 98]: the CCT was complied almost 10 years ago. Although it has undergone several amendments, it can not keep up with the rapid varying of the dynamic networked information; the 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 user to master; the CCT is mainly meant for the description and organization of information resources. It is not so good at retrieval aiding; the CCT is design for the management of hard-copy documents, and some of its usage criteria lose their sense in the network and the digital environment, 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.

3. The Knowledge Network

3.1 Classification, Thesaurus and Bibliographic Data

It seems impossible to deploy traditional classification and thesaurus in digital libraries. But there exists one kind of networked information resources in which the classification and thesaurus are deployed thoroughly—the bibliographic data of OPAC system. The bibliographic data is one of the most important resources library (and only library) owns. They are updated timely and representing the specific domain knowledge effectively. They are a particularly collection of living materials containing plenty of new professional terms in the title field which is indexed in the controlled vocabularies of the classification and thesaurus. Based on this mapping, the classification and thesaurus and the bibliographic data could be combined 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 restore blood and flesh to the skeleton. Thus a living knowledge corpus is formed where new terms could 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 the convenience of writing, we call such knowledge network as KNICTM (Knowledge Network Integrated of Classification, Thesaurus and Metadata).

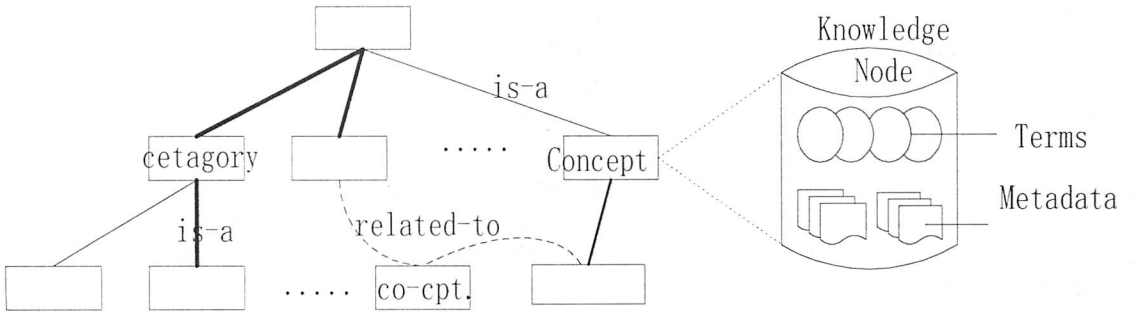

Fig. 1 the knowledge network with a knowledge node zoomed in

3.2 The Construction of the Knowledge Network

We incorporate the portion of the computing domain in the CCT with all the bibliographic data of the computer science of Peking University Library between 1990-1999 which amounts to more than 5,000, and built an experiment system named as Vision.

The KNICTM is built in 3 steps:

- *Construction of the original concept node based on the classification and thesaurus.* First, the thesaurus is turned into a original concept network which 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 (Use/Use For) in thesaurus. If there is a hierarchical relationship (Broader Term/Narrower Term) between two terms, an "is-a" edge is set up between the corresponding concept nodes. The associative relationship (Related Term) is discussed in the next step. Then the classification scheme is embedded into this original concept network. The hierarchy system of the classification is transferred to the concept nodes corresponding with the categories of the classification, and acts as a discipline hierarchical backbone to the concept network (fig.1). In the case of the CCT, for it is a reciprocal index between the BCCL and the CT rather than a faceted thesaurus, there is no direct mapping between the categories of the BCCT and the concepts of the CT. 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 arranged into the nodes of the original concept network according to their subjects. This is the key task of the KNICTM construction. Supported by the records of bibliographic data, the abstract concept node becomes the knowledge node where the abstract concept composed of terms is bound to the metadata records closely (fig.1). And the concept network turns into a knowledge network formed by the integration of classification, thesaurus and metadata. The distribution of the bibliographic data in the original concept network is explained as follows: we follow Gruber to call a record of bibliographic data "reference", for it identifies a specific document in the library [Gruber 94]. If a reference contains only one subject descriptor, we take the reference as one of the instances of the corresponding concept node and add the reference into the node. If a reference contains several descriptors, then we add the reference into all the related concept nodes as instances of them. If a reference contains a composite subject which is described 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 call "co-concept" node which has the references merely and no term for the moment. For example, a reference with title "Internet Firewall Technologies" is indexed with the string "Network-Security" for there is no "firewall" in thesaurus. To add this reference 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. For the associative relationship easily becomes out of control in thesaurus, we don't consider it in the first step. Only when the correlation of two concepts is supported by a reference, we establish the "related-to" relationship between them through co-concept. Thus the references function as the verifications of the associative relationship. In the next step, when there occurs a new extracted term with the meaning of the co-concept, the new term is added to the co-concept. In the above example, this is "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 indexing of the subject descriptors and the classification notation. Based on this mapping, statistic and semantic techniques could be applied to extract new terms from the title and add them into the concept network. There are three difficulties to accomplish 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? There are three possibilities. If there is a corresponding concept (including co-concept) node in the KNICTM, the new term is added to the synonym set of the concept node; if there is no such node, a co-concept node is created; preceding to all these processing, the sentences of title must be separated into word or phrase. This is the classical problem for Chinese—segmentation.

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 bring leaves to it. 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.

3.3 The Advantages of the 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 a interconnected knowledge network.

- *An Adaptive Concept Network Based on the Applied Resources.* The classification and thesaurus are the representation of the general knowledge, and can not 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 in metadata collection to enrich concept network automatically. This is a technique to enrich the controlled vocabulary with the keywords in title.

- *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 user to clarify his information demand and express his query clearly. Secondly, because all the metadata have been arranged into the KNICTM, there is no need to dig into the metadata collection by keyword matching to find what the user needs, it is only necessary to locate the knowledge node according to the user query, and follow the surrounding edges to reach other nodes to pick up what he wants. Thirdly, 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 is made of the knowledge node which 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 discipline hierarchy and clustered into topic areas through edges among them. A friendly interface like cat-a-core [Hearst 97] manifests the organization of the knowledge nodes. A user facilitated 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 elements of a library are the information resources and the classification and thesauri, which are the information organization and retrieval tools. In KNICTM, these elements have been integrated into a coherent and consistent knowledge network. And the KNICTM could be easily extended to support other activities of digital library, such as resources collecting and indexing. Thus all the activities of the digital library, including indexing, organization, navigation, retrieval and learning, are centering on this knowledge network. To develop continuously, the KNICTM will bring the digital library onto the platform of knowledge management from that of information management.

4. The Construction of the Knowledge Network

We have built a experiment system named Vision. What we introduce here is the issues of implementing the Vision system in the first phase. The Vision has a client/server architecure; at the server side is the knowledge network supported by Oracle9i; at 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 is selected in order to make it easy to transplant the system to the Web. We first introduce the ontology of our system. After that we depict the whole data flow of the system creating. Then we outline the architecture of the system at the server side and the function of the conceptual retrieval tool at the client side. At last we introduce the work we have done in extracting new terms from the bibliographic data to enhance the knowledge network.

4.1 The Ontology Design

There are many objects in our system and their relationships are complex, and so we used ontology tools such as Ontolingua and Protégé to design the ontology of our system. Then we convert this ontology into the database schema. Our ontology consists of seven classes: term, concept, co-concept, category, document, author, and publisher. Their naming manifests their meanings, but the relationships among them are entangled and perplexing. A simple diagram will depict their relationships with the numbers indicating the cardinality of the links (fig.1). And pieces of ontology in the syntax of Ontolingua and in the RDF schema rendered by protégé are provided in appendix 1 and 2 for further reference. We then convert the ontology into the relational schema of the database system and creat the corresponding tables in Oracle.

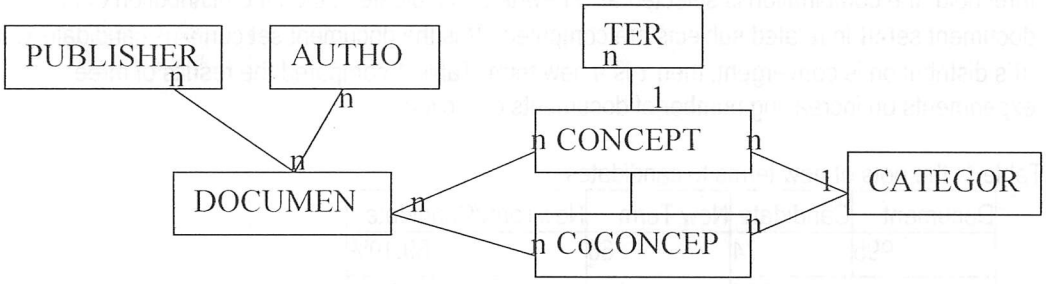

Fig.2 the objects and the relationships in the Vision system

4.2 The Vision at the Server Side: the Knowledge Network

The original data set to build the Vision have two parts: the e-text of the CCT and the bibliographic data of computing domain. Both of them are 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 build the Vision as the server:

- *The e-text file of the CCT is processed to set up the fundamental structure of the Vision.* 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 2,194 terms, including 1,684 preferred terms and 278 non-preferred terms; and 1,684 concepts are created. Some non-computing-domain terms are also caught 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 organized into the original concept network constructed in the preceding process.* The bibliographic data are within the period of 1990-1999 in computer science and amount to 5,314. They are of 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 (e.g. title, subject, author etc.) to form a new record which is appended into the table DOCUMENT. Totally 5,053 document records are created. Others are discarded for various reasons, for example, unrecognized title, two ISBN numbers, etc. Such data processing spends 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 correspondence between the subjects of the document record and the meaning of the concept record. And when it is necessary, new record is created in CoCONCEPT table. These processes accomplish the task of organizing the metadata into the concept network to form the knowledge network. This has been explained in section 3 in detail.

- *New terms are extracted from the DOCUMENT table and added to enhance the knowledge network of the Vision.* It is depicted in the dashed line square in figure X. Some of the problems have been mentioned in section 4. It is the aim of the ongoing second phase of the Vision project. So we just outline roughly what we have done now. The principal obstacles to achieve this are: to extract new terms from the titles and to insert them into the knowledge network.
  1) The extraction: At present the statistic algorithm is applied to extract terms in titles. First, the title is segmented into basic words and phrases by general segmentation tool, and then the co-occurrence frequency of the neighbors are counted. If the frequency is bigger than a given

threshold, the combination is selected as a new-term candidate τ; then the distribution of the document set $\Re$ in related subjects are computed, $\Re$ is the document set contains candidate τ. If $\Re$'s distribution is convergent, then τ is a new term. Table 1 compared the results of three experiments on increasing number of documents collection.

Table 1: the riots of new terms to candidates

Document	Candidate	New Term	NewTerm/Candidate
995	74	66	89.19%
2310	234	219	93.59 %
3453	368	344	93.47%
4623	512	477	93.16%

2) The Insertion: The convergent point of the former computing will help to determine the position where the new term should be inserted. And we are considering applying the theory of lattice [Pedersen 93] and Formal Concept Analysis [Ganter 99] in the Vision.

4.3 The Vision at the Client Side: Knowledge Navigation and Retrieval

We implement a knowledge retrieval system in Java to navigate and retrieve the Vision knowledge network. Figure 3 is the snap-shot of the user interface. There are four areas: the query dialog, the concept network window, the information window and the documents window. Basically there are 3 ways to view the concept network: hierarchical tree, alphabetical list, concept family and hybrid tree. The hierarchical tree is similar to a faceted thesaurus. All the categories and concepts (identified by the preferred terms) are organized into a expandable conceptual tree; the alphabetic list is an index of all the terms in alphabetic order of Chinese Pin-Yin; the concepts can also be organized into concept families (i.e. term families of the thesaurus) and listed in alphabetic order of the top concepts (i.e. top terms); the hybrid tree is a hybrid of the hierarchical tree and the alphabetic list.

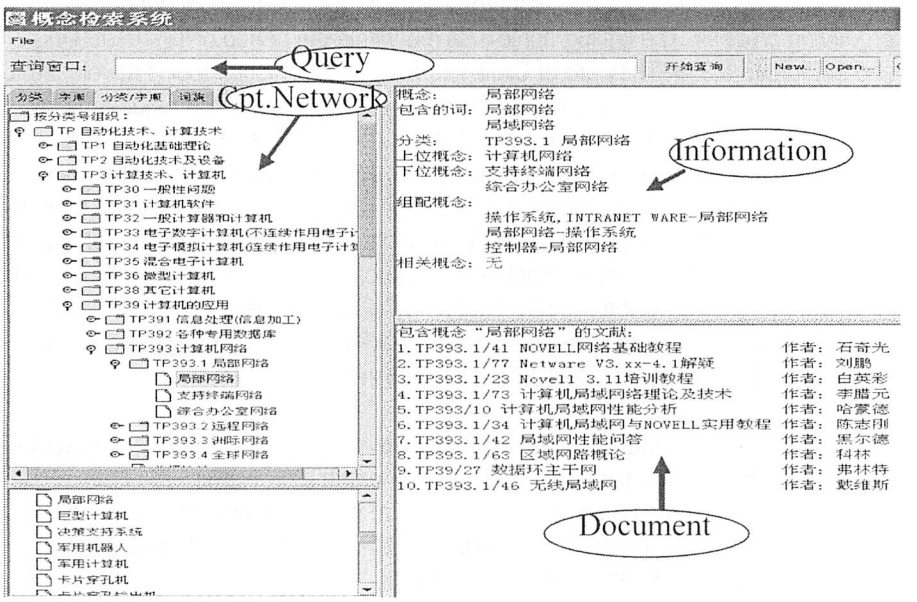

Fig.3 the interface of the Vision system

When the user clicks 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 is not the documents containing the querying words, but the concepts and co-concepts pertinent to it. The concept network window is triggered to highlight the position of the retrieved concepts in 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.

5. Conclusion and Future Works

Centuries of library work has proved that the organization of information is the basis for the sufficient utilization of information resources. It's just the value of library. So is digital library. Just for the lacking of organization, the potentials of the metadata as one of the most important networked resources are not exploited sufficiently. This article presents an approach for organizing the metadata by the classification and thesaurus into an integrated knowledge network and set up a new paradigm for the knowledge management in digital library. Our approach distincts itself from other ontology-driven and concept-based systems by incorporating the concept and the relevant metadata records into an integrated knowledge node which forms the knowledge network. We have built an experiment system Vision based on the Chinese Classification and Thesaurus and the bibliographic data of the Peking University Library. Our experiment also demonstrates that the traditional resources such as bibliographic data still have indispensable values worth of further exploring in spite of the continuous increasing of various digital resources.

The Vision is on the way to the second developing phase now. We are endeavoring to achieve the following aims:

- There is still no perfect method to compute the extension and intension of a term and determine its position in the concept network. This is critical for the concept network to be a self-sufficient system. Now we consider applying an adjusted Formal Concept Analysis in the Vision.
- A language for the concept query and manipulation in the concept network will simplify the operation on the concept network and add the automatic query expansion and contraction mechanism.
- A visualization interface, such as Cat-a-Core [Hearst 97] or Inxight-Star-Tree [Inxight 02], is planned to provide more friendly interaction with the user. The visualization interface is isomorphic to the concept network in structure; it 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.

To integrate the classification, thesaurus and metadata into a coherent knowledge network has promising application in digital library. 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 database 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 of seven layers [Tim 00]. But how to construct it? Our paradigm provides an optional approach.

Acknowledgement
This research is a portion of my doctorate dissertation. I wish to thank Prof. Yang Dongqin and Prof. Tang Shiwei. I am particularly grateful to Dai Lingji, president of the Peking University Library, Shen Yunyun, director of the catalog branch of PKU Library, Li Shun Min, director of the automation branch of PKU Library.

I thank very much Zu Yong and Deng Peng who assisted me in the developing of the Vision. I also thank Prof. Zhang Han and Prof. Ma Zhanghua who gave me much valuable advice. My special thank should also go to my fiacee, Liu Hao, without whose love, the work would not be as good as it is.

References

Chen, L. (2002), The Report on the development of the CALIS at the end of the 1st phrase. Internal Technical Report for the Government.

Dong Xiaoying, (2000). *The Information Management and Services in Network Environment*, Beijing, The Translation and Publishing Press.

Ganter, B., (1999). Formal concept analysis, Springer-Verlag Berlin Heidelberg 1999

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies,* special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation

Gruber, T. R. (1994). Introduction to the bibliographic data ontology, http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data.txt.html

Hearst, M. A., (1997). Cat-a-Cone: An Interactive 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

Liu Jia, (2000). *The Metadata of Internet* Doctorate Dissertation, Beijing: Peking Uni.

Liu, X.S. (1996). The classification and thesaurus of the contemporary China, *62nd IFLA GENERAL CONFERENCE. Booklet4*, [HHU31-37].

Ma, Z. H., (1999). The introduction to the document classification and thesauri, the Beijing Library Press, 1999.

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

Shiri, Ali Asghar; Revie, Crawford; Chowdhury, Gobinda. (2002.2) "Thesaurus-enhanced Search Interfaces". *Journal of Information Science.* Vol.28, Number 2, 2002.

Chinese Book Classification and Thesaurus (3rd ed.), (1990). Beijing: Hua Yi Press.

Tim Berners-lee, (2000.12.16). Semantic Web, *xml2000 conference,* 2000.12.16, Retrieved 2002.4 from: http://www.w3.org/2000/Talks/1206-xml2k-tbl/

Zhang, Q.Y., (1996). Introduction to Chinese classification and thesauri, *The works on the information retrieval language by Zhang Qiyu*, Books and Documents Publisher.

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

Zhang Qiyu, (2001,3). The Developing Trends of the Language of Network Information Retrieval, *The Journal of Library(Chinese),* NO.4, 2001

Appendix I: VISION's Ontology in Ontolingua

```
(define-class TERM (?x)
  :def (and (ctms-thing) ?x))
  :axiom-def (subclass-partition
                TERM (setof preferred-term non-preferred-term)))
(define-class CONCEPT (?x)
  :def (and (ctms-thing) ?x)
        (has-one ?x cpt.identifier))
(define-function CPT.IDENTIFIER (?cpt) :-> ?identifier
  :def (and (concept ?cpt)
        (preferred-term ?identifier)))
(define-relation CPT.TERMS (?cpt ?terms)
  :def (and (concept ?cpt)
        (term ?terms))))
(define-relation CPT.CATEGORY (?cpt ?ctg)
  :def (and (concept ?cpt)          (category ?ctg))))
(define-relation CPT.BROADER-CONCEPT (?cpt ?b-cpt)
  :def (and (concept ?cpt)
        (concept ?b-cpt))))
(define-relation CPT.NARROWER-CONCEPT (?cpt ?n-cpt)
  :def (and (concept ?cpt)
        (concept ?n-cpt))))
(define-relation CPT.DOC-CONTAINED(?cpt ?docs)
  :def (and (concept ?cpt)
        (document ?docs))))
```

Appendex II: VISION's Ontology in RDFS (Protégé)

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE rdf:RDF [
        <!ENTITY rdf 'http://www.w3.org/1999/02/22-rdf-syntax-ns#'>
        <!ENTITY thesaurus 'http://protege.stanford.edu/thesaurus#'>
        <!ENTITY rdfs 'http://www.w3.org/TR/1999/PR-rdf-schema-19990303#'>
]>
<rdf:RDF xmlns:rdf="&rdf;"
        xmlns:thesaurus="&thesaurus;"
        xmlns:rdfs="&rdfs;">
<rdfs:Class rdf:about="&thesaurus;*ConceptFamily"
        rdfs:label="*ConceptFamily">
        <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;?NewTerm"
        rdfs:label="?NewTerm">
        <rdfs:subClassOf rdf:resource="&thesaurus;Term"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Book"
        rdfs:label="Book">
        <rdfs:subClassOf rdf:resource="&thesaurus;Document"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Category"
        rdfs:label="Category">
        <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Category_Code"
        rdfs:label="Category_Code">
        <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Concept"
        rdfs:label="Concept">
        <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdfs:Class rdf:about="&thesaurus;Conference"
        rdfs:label="Conference">
        <rdfs:subClassOf rdf:resource="&thesaurus;Event"/>
</rdfs:Class>
```

Discussion Points

1. What does the author mean by classification?

2. Does the research involve creation or implementation of a classification scheme?

3. How does the researcher use classification to improve the automated approach?

4. How do these methods compare to current human-generated approaches to classification?

5. How does the reported research expand our understanding of classification?

6. Does the research suggest an improvement over human-generated classification?

7. What do you think are the most important lessons learned in this research?

8. What do you think are the best practices reported in this research?

9. What would you recommend to the researcher as the next step in this approach?

10. Is there other related research that you would recommend the researcher become acquainted with?