Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2004 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 2004

OBDMS----A Document Management System

Zhonglei Yuan
Fudan University

Xiao-Bai Li
Fudan University

Jin Chen
Fudan University

OBDMS——A Document Management System Based On CBR and Ontology Techniques

Zhonglei Yuan
Management School, Fudan University
Shanghai 200433
Yzl_1981@sina.com

Xiaobei Li
Management School, Fudan University
Shanghai 200433
Luckypage_li@hotmail.com

Jin Chen
Management School, Fudan University
Shanghai 200433
Jchen924@sohu.com

Abstract

Document Management Systems (DMSs) are playing an increasingly critical role in organizational knowledge management. However, the most often used DMSs, such as windows Resources Manager, cannot settle the issues like inefficient query and redundancy because of the lack of document content and relationship management. In this paper, we present a prototyping system, based on improved CBR and two-layer ontology techniques, which can deal with the issues above and support effective document retrieval and management.

Keywords: Document Management System, Ontology, CBR

1. Introduction

With the development of organizations, the competition among them grows increasingly fierce. And knowledge Management is becoming the important basis for organizations to gain advantages in competition. Knowledge, which contains experiences and values of organizational members, mostly comes from data of daily operations. When data has been explained and analyzed by members, it turns into information, which, at last, can be generalized to knowledge that can be shared by others. For a research team, document management is the key for the organization-wide data management.

The current DMSs (Document Management System) mostly adopt CBDM (Category-based Document Management) or KBDM (Jie Lu 2004) (Keyword-based Document Management) techniques. CBDM, (such as Windows Resources Manager (WRM),) classifies and stores documents by directory, while KBDM organizes documents via abstracted keywords, for example, a library system.

Commercial DMSs (A.Maedche 2003) implemented for business use, mostly adopt the document classification and complex keyword query technique mentioned above. However, with the increase of document size and quantity, the document directory structure becomes more complicated. Meanwhile, it is inevitable that document classification may overlap. Thus, current DMSs have trouble in retrieving and managing documents, to be specific, storing missing, term missing, or term mismatching (Jie Lu 2004). If there isn't any proper technique that can be applied to handle these problems, the efficiency of document retrieval, reuse, and replenishment (Chengyao Wang 2003) would be influenced when documents are updated dynamically.) The newly-rising CBR technique (A.Aamodt 1994) has the capability of attribute-value relationship management, and allows adding attribute dynamically. At the

same time, two-layer ontology technique (Jie Lu 2004) can be used in keyword query. The research of these two techniques provides new ways to improve DMS. Basing on these two techniques, we developed the prototype system, Ontology-based Document Management System. (OBDMS) This paper will introduce this system in detail.

OBDMS mainly integrates three particular functions:

I. By document relationship management, OBDMS can display the structure of knowledge snippets, and maintain the Knowledge Evolution Network (KEN) (Limin Lin 2003).

II. By Dynamic Attribute Management based on CBR technique, OBDMS can identify and classify documents more clearly, and cope with the problem of document overlapping.

III. By keyword ontology management, OBDMS can improve the efficiency of keyword query, and decrease the possibility of term missing which may be encounter by other retrieving techniques.

The remainder of the paper restates the basics about two techniques in Section II. The profile of OBDMS is introduced in Section III and the specific functions of OBDMS are further illustrated in Section III. System evaluation description and results analysis are presented in Section IV and Section V where the new system is compared with the currently-accepted keyword-based document retrieval approach. Finally, we discuss conclusions and further research in Section VI.

2. Theory Foundations

2.1 CBR in Document Management

CBR (Roger C. Schank 1982) is a decision support technique proposed by Schank and Abelson (Ian Watson 1997). It makes decision by retrieving and adjusting current solutions of similar problems, and at the same time, stores the new problem and its solution for future reasoning use. The research of CBR has gone through two stages. The first one was in late 1980s, when CBR was gradually enriched and systematized in theory after ten-year's development. The technique then contained case presentation, case index, case storage, and similar arithmetic (F. Crestani 2003) of case retrieval. The intercrossing with other AI techniques (For instance, Neural Network Technique, genetic arithmetic, and Rule-Based-Reasoning) also made contribution to its development. At the same time, CBR began being taken into practice for business use, and successfully applied to customer supporting, medical diagnoses, law consulting, and failure maintenances etc. This process resulted in the first upsurge of CBR, and made CBR an important embranchment of AI. The second stage, which was brought by the research of Knowledge Management, began at the late 20th century. The common process concept of CBR and KM made scholars on CBR to think over the relationship between them. In management principles, CBR and KM was similar, because both of them emphasized that Rule/Universal knowledge is hard to gain, and both of them paid attention to the innovation and development management of practical knowledge.

On the bases of the results in the second stage, CBR technique can be applied to document management. In the process of collecting and arranging documents, we pay more attention to extracting and recording attributes that can describe the content of documents. It is different from the way we used before, in which we just recorded the managing information, for instance, created date, subject, and author. As a result, DMSs that use CBR technique can clearly present the relationship of documents in content, and index the documents according

to their content features. Thus the efficiency and precision of query has been greatly improved.

2.2 Ontology in Document Management

2.2.1 Integrating Case-Based Reasoning with Feature-Adding and Relationship-Managing Technique

With the development of research, the research team will generate more and more documents including the referred articles, comments made by members, and assumptions to be recorded at any time anywhere during the research process. All of them should be well managed for future use. When the quantity of documents is gradually increasing, document classification is imperative under the situation. But in the process of building up and drilling down classification hierarchy there will be overlap of sub-attributes. In order to manage the document classification better, OBDMS has adopted "Integrating Case-Based Reasoning with feature-adding and relationship-managing technique" (Chengyao Wang 2003).

The core of "Integrating Case-Based Reasoning with feature-adding and relationship-managing technique" is an ontology (S. Staab et al. 2001) (William S. 1999), which records values of the feature that reflects the document classification. Each document has a relevant node in this ontology, which represents its classification in the domain. By designing the structure of ontology, and combining two tables:" attribute value relationship table" and "attribute value mapping table", this technique solves the overlapping problems of document classification mentioned above. More details can be found in" Integrating Case-Based Reasoning with feature-adding and relationship-managing technique for the research team knowledge management" written by Wang Chengyao, etc.

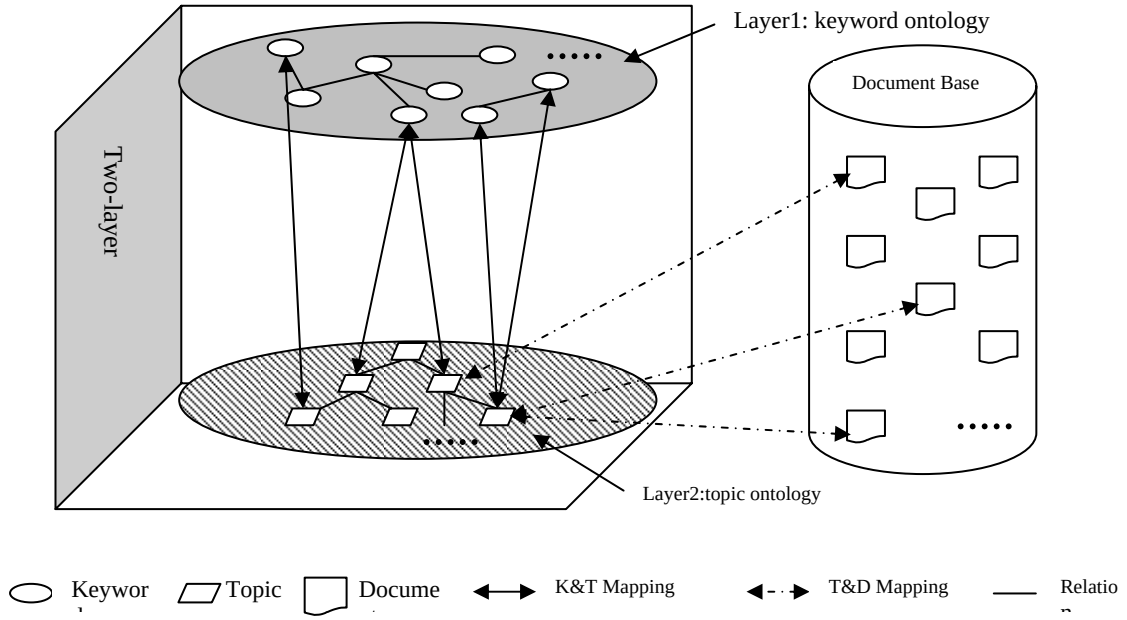

Figure 2.1 Two-Layer Ontology Frameworks

2.2.2 Keyword Query Based on Two-Layer Ontology

The most important application for DMSs is document retrieval. An efficient system always does well in retrieving. But current DMSs have many disadvantages in this practice.

Therefore, we have designed an efficient DMS searching engine according to “keyword query technique based on Two-layer ontology” proposed by the paper ” Keyword-based Document Retrieval Approach Under Two-layer Ontology Framework”. The search mechanism of this engine will be illustrated in Part 3. Now we will introduce this technique in general.

On the basis of the ontology mentioned in 2.2.1, this technique creates and maintains a keyword ontology. And it sets up the mapping between these two ontologies, thus converting keyword query into more normative and definite query about document classification, and improving the retrieving efficiency too. The technique uses two ontologies--one for keywords and the other for classification, and it decomposes a query into three phases: from the keyword ontology to the topic ontology (ontology of document classification), and finally to the object document domain. It is the two ontologies and the mapping between them that play the key role in this technique. Therefore, it is named as “keyword query technique based on Two-layer ontology”.

3. OBDMS

3.1 General Description of OBDMS

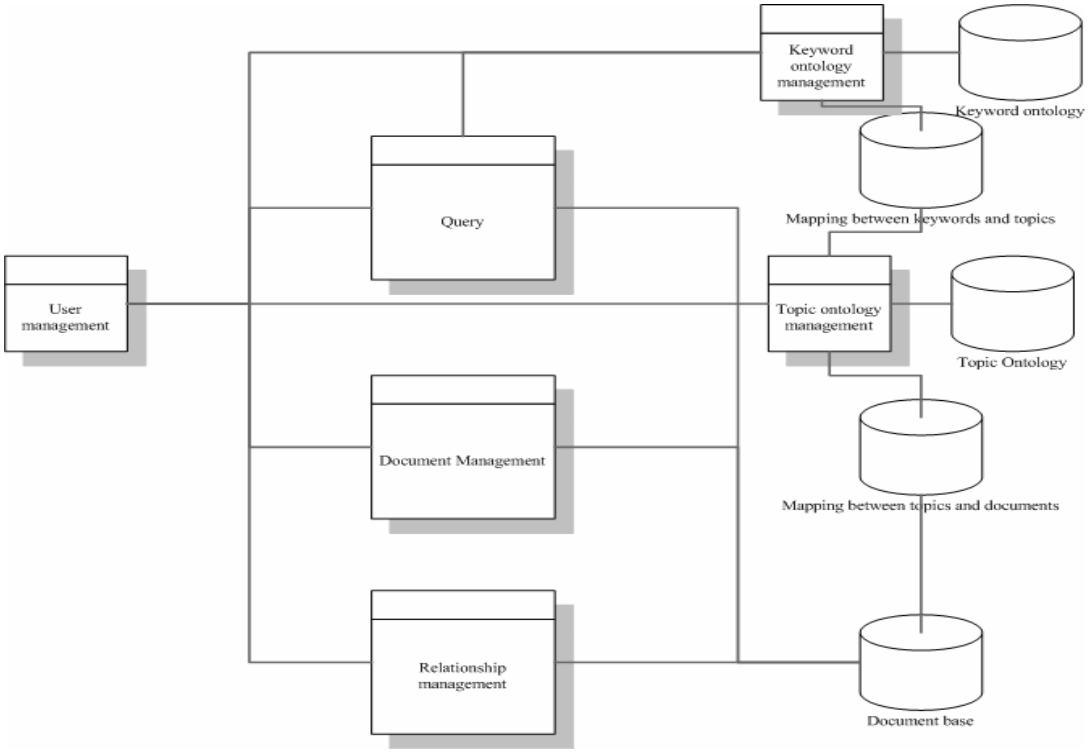

Figure 3.1 System Deployment Architechure

3.1.1 Keyword Ontology Management Module

Users can execute operations relevant to keyword ontology in this module, such as adding/deleting nodes, building up new relations, defining / modifying similarity degrees etc.

3.1.2 Topic Ontology Management Module

Users can execute operations relevant to topic ontology in this module such as adding/deleting nodes, establishing new relations, drilling down documents structure etc

3.1.3 Complex Query Module

In this module, queries submitted to the system through an interactive interface are translated into query statements that are readable by the system database. At the same time, a relevant analysis is processed according to the topic ontology aiming at getting the set of all matching topics. During a query based on keywords, related mappings are automatically executed in line with the two-layer ontology framework and the query results are then generated.

3.1.4 Document Management Module

New documents are added into database in this module, by defining the basic information (such as title, author and press etc), keywords, and relevant classification/dynamic features. The document maintenance, such as document modification, attributes editing, is also implemented in this module.

3.1.5 Relationship Management Module

In this module, document relationships according to Knowledge Evolution Network are established. When users browse a document, the links to its related documents are available automatically as well.

3.1.6 User Management Module

In this module, 5 types of users, which enjoy different operational authorities, are defined.

3.2 Query Process

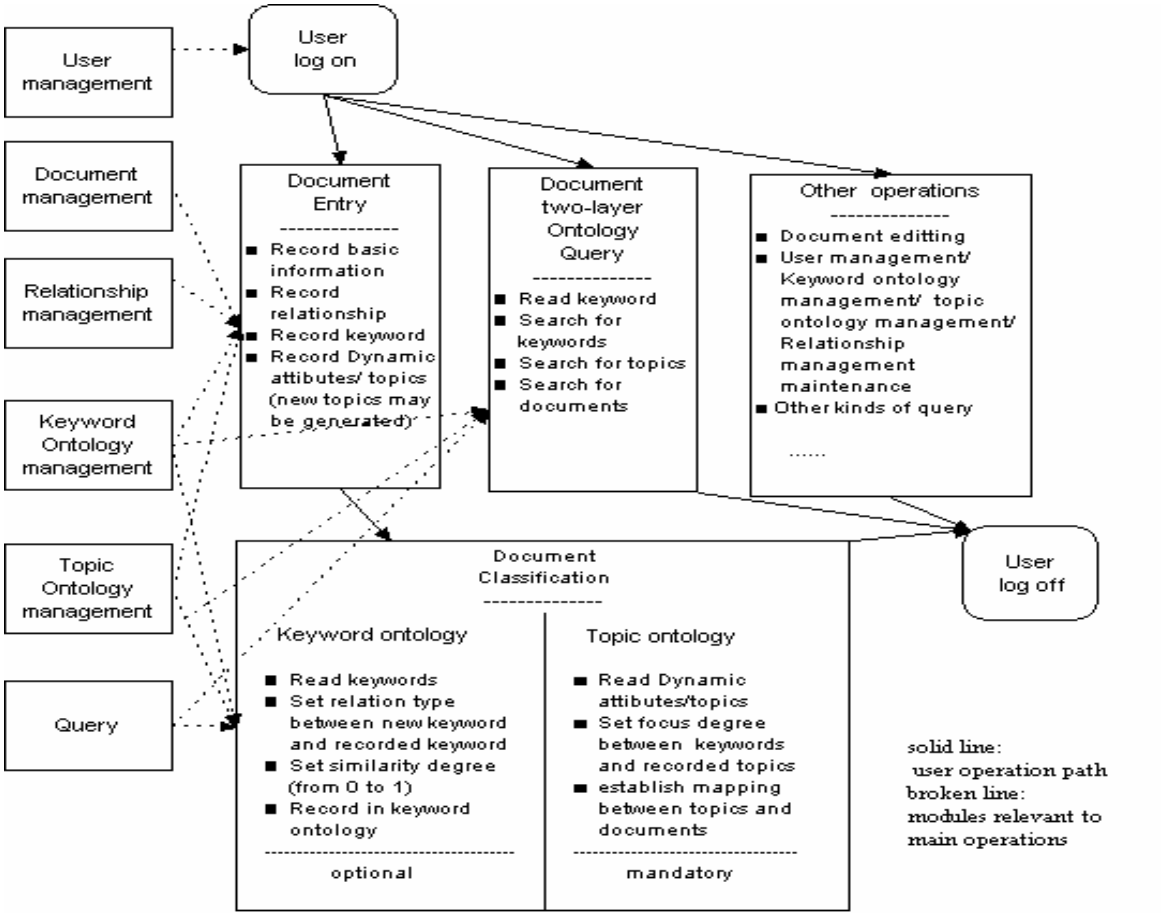

Figure 3.2 System Process

3.2.1 Documents Classification based on CBR Technique

Using the dynamic documents storage method based on the CBR technique, OBDMS classifies documents, which are stored together in physical medium, according to specific attributes (dynamic features). In this way, all the documents of a research team can be stored in OBDMS database without defining their physical location. Whereas, users can find that all the files are clearly listed in their proper classification path via OBDMS interfaces. In other words, they can understand the knowledge infrastructure of the team, the current classification hierarchy, and the exact classification of a document with the help of OBDMS.

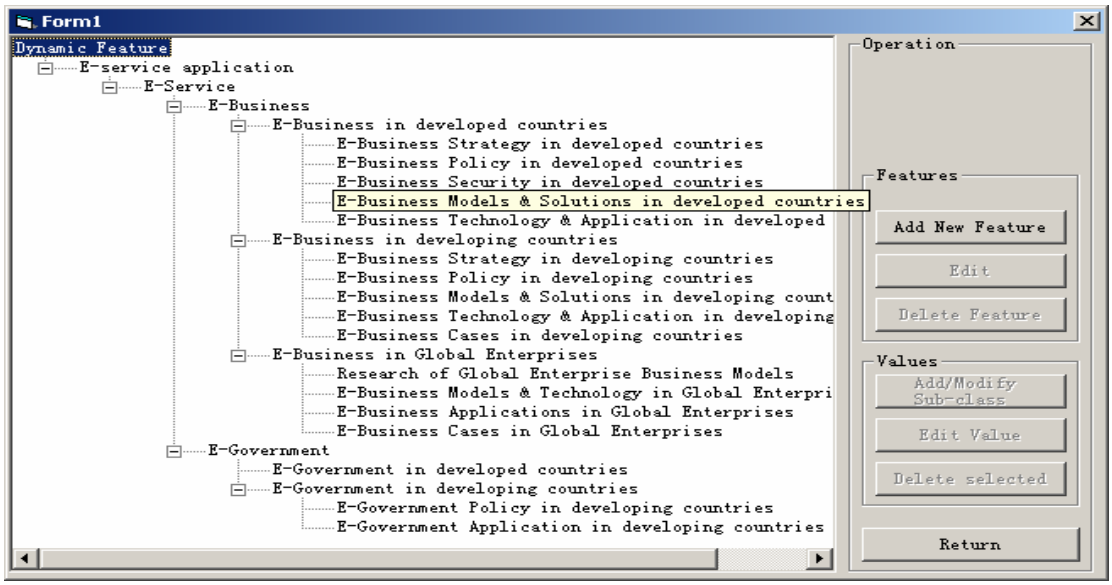

Figure 3.3 "E－service Application" study group's document classification hierarchy

We call this dynamic document storage method used in OBDMS "Dynamic Feature Management"(DFM). There is an example of a "E－service Application" study group's document classification hierarchy as Figure3.3 shows:

Users can set up or modify the dynamic feature structure and edit the value of dynamic

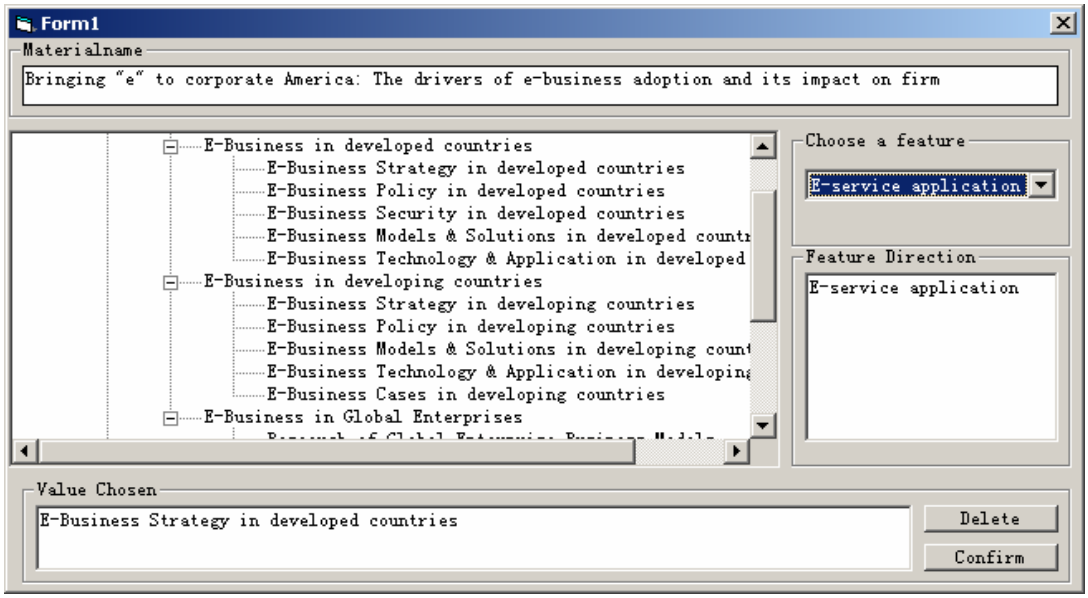

Figure 3.4 System Process

features for a specific document in OBDMS. The interface of OBDMS that allows users to set up or modify the structure is shown in Figure3.3. And in the interface as is shown in Figure3.4, users can define the dynamic feature value and classify a document easily.

3.2.2 Document Relationship Management

There is a Knowledge-Reference-Net (KRN) among the documents of a research team. And OBDMS has the ability of managing the relationship between knowledge snippets, that is allowing users defining, editing or deleting the relationship between any two snippets. The relationship of reference can be viewed as a kind of filiations. Sometimes it is extraordinarily needed to find where the knowledge snippet (document) stored in database derives from and what other snippets are derived from it. In the main interface of the OBDMS, users can easily find these two kinds of related records when they are checking a specific document.

3.2.3 Keywords Query Using Ontology Technique

One prominent capability which makes OBDMS exceeds other document management systems is the keywords query based on the "two-layer ontology" technique. As is shown in Figure 3.5, users can input keyword "strategy" directly, if they are not sure about how the research team expresses the same concept. Unlike other searching engine, OBDMS does not search the documents immediately.

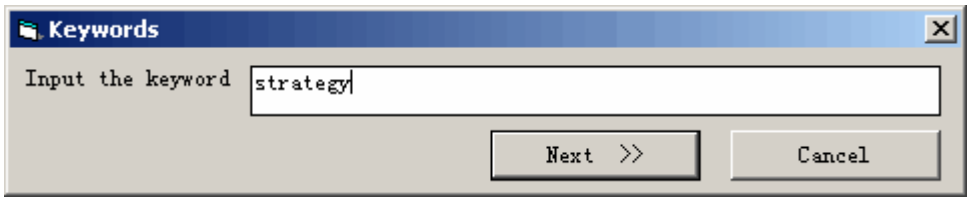

Figure 3.5 Keyword Input Interface

Contrarily in the interface shown in Figure3.6, it lists all keywords recorded in the keywords ontology that are related to the user-defined keyword "strategy", together with the similarity

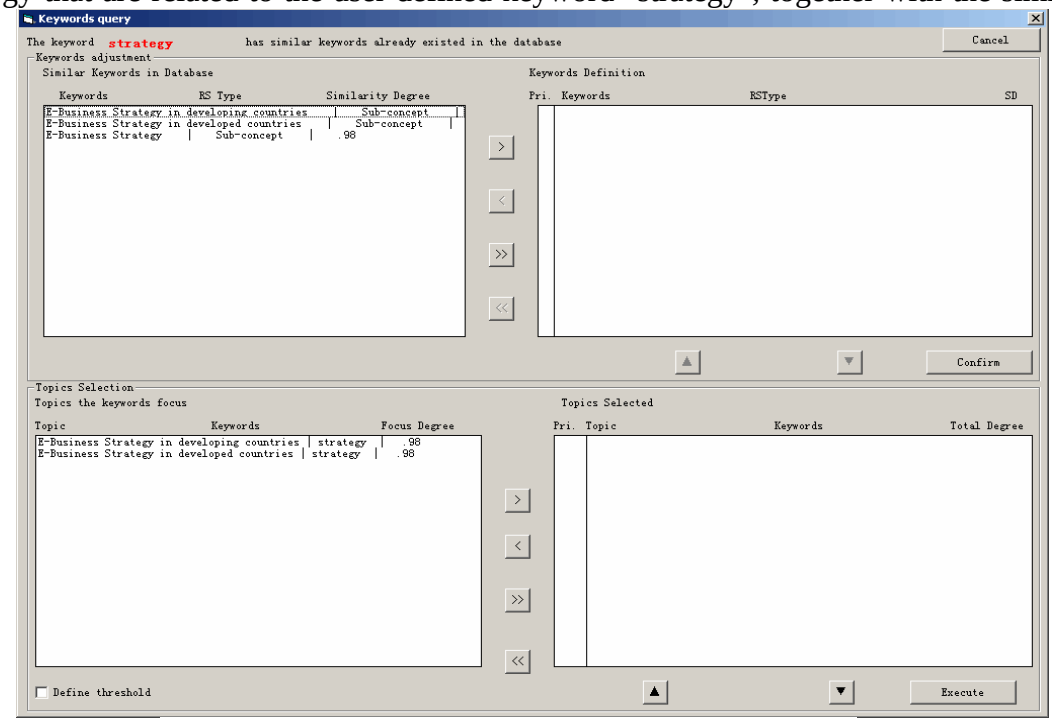

Figure 3.6 Keyword Query List

degree between each two of them. For example, “E－Business Strategy in developing countries”/“E－Business Strategy in developed countries”/“E－Business Strategy”. In the frame left below, OBDMS lists the topics that have direct relation with "Strategy”:” Strategy of E-Business in developing country”/“Strategy of E-Business in developed country”, and also the focus degree.

Next, users can select any related keywords listed in the frame left above according to their judgments to enhance the accuracy of this single search. Suppose only one related keyword “E－Business Strategy in developed countries” is selected. This item is added to the frame right above at once. If the button "confirm" is pressed, OBDMS winkles the unwanted one "Strategy of E-Business in developing country".

And at the same time, topics that are related to “E－Business Strategy in developed countries” are added to the frame left below too. Users then decide which topics listed in frame left below are the final ones they are interested in. After they press "Execute", OBDMS makes the final search in the database that stores documents, and returns the result, which is “E-strategies remain on the back burner in Hungary”, as is shown in Figure 3.7.

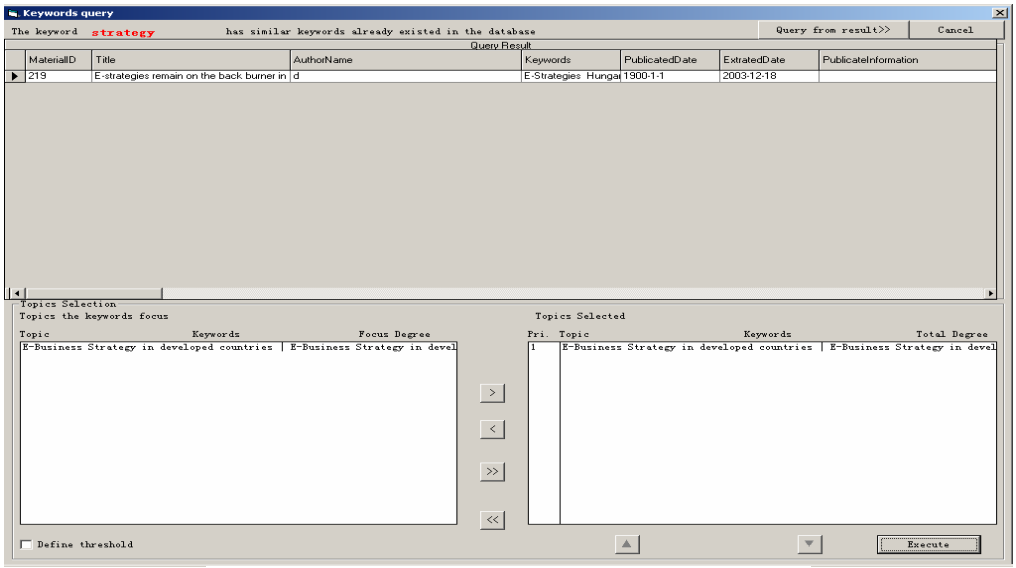

Figure 3.7 Keyword Query Result

4. Evaluation of OBDMS

4.1 The Evaluation of query based on two-layer ontology framework（two-layer ontology query VS common query based on keyword）

OBDMS employs the dynamic feature-adding technique to classify documents logically, thus improving the focusing relationships (Jie Lu 2004) between documents, and finally making queries easier. Besides this technique, OBDMS adopts two-layer ontology framework as well, which enhances the query effectiveness, decreases the possibility of the term mismatching/term missing that may be brought about by other kinds of query. This kind of efficiency improvements can be best illustrated in two aspects: 1) classification-oriented query improvements 2) differently-expressed topic query improvements.

First, OBDMS improves the efficiency of classification-oriented queries, which can be

measured by recall rate (W.Meng 1997), by generalizing semantics and semantic relations with the help of ontology. The concerned topics (classifications) will be positioned by the system automatically according to the topic ontology. For example, in a single search, a user concerns the query results of "E-Business Strategy in developed countries". In the concept layer, "developed countries" is a general concept and meanwhile it is a specific description to "E-Business Strategy". If we employ common query technique based on keywords, the results may be 0. But in fact, such documents do exist in the document base, so the recall rate = 0. Without two-layer ontology framework, the terms, such as "America", "Japan", which are instances of "developed countries", could not be considered. In OBDMS, it is easy to locate relevant topic through the dynamic features/attributes and the recall rate can be improved a lot.

Second, OBDMS improves the efficiency of differently-expressed topic query, which can also be measured by recall rate. Under two-layer ontology framework, we realize query intelligence to some extent. For example, when a user concerns "E—Business Strategy in developed countries", the submitted topics by him/her can be "E-Business Strategy in developed country"、"Strategy of E-Business in developed countries"、"Strategy of E-Business in developed country",etc. In a traditional keyword query, different query results may be generated with different expression of terms and the recall rate is hard to reach 1. However, in OBDMS, the semi-intelligent query module will point all relevant expressions to the topic "E—Business Strategy in developed countries" and then list all alternatives for further query. Thus, the recall rate is always improved. This framework is very beneficial and convenient to those who are not familiar with the research team.

4.2 The Evaluation of documents classification (CBR-integrated technique and document relationship management)

OBDMS employs the Relationship Management Module to improve the performance of document managing that has complicated KRN. This kind of efficiency improvements can be measured by precision rate (W.Meng 1997) and average time. The precision rate for OBDMS and WRM (Windows Resources Manager) with different set of documents is listed in Figure 4.1, and average time in Figure 4.2.

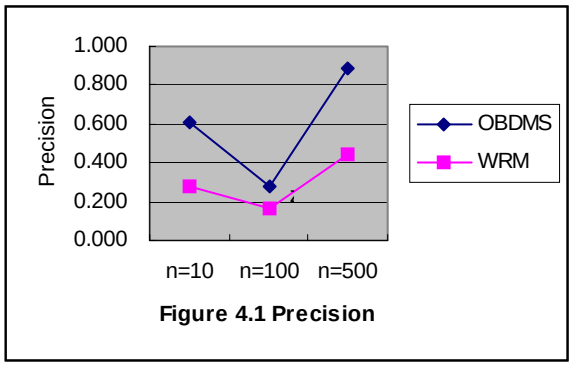

Figure 4.1 Precision

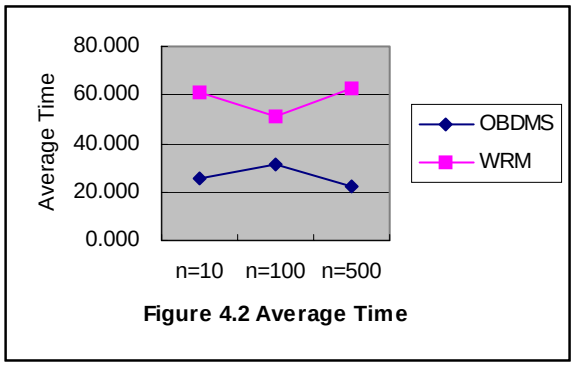

Figure 4.2 Average Time

In our experiment, OBDMS and WRM were compared. We mimicked three sets of documents with size of 10, 100, and 500, and kept the distribution of documents in almost the same complexity by setting the key parameters similar with each other. It is not difficult to find out that OBDMS has better performance than WRM.

OBDMS records all reference relationships between any two knowledge snippets in a research team. Users can easily find other related references when retrieving a special

document. On the other hand, WRM keeps no record of the reference relationship between documents. Therefore when file name is not precise enough for identifying the content of document, or there are too much documents around, the retrieval could be non-efficient.

5. Conclusion and Further Research

OBDMS proposed in this paper, enjoys a significant improvement in query efficiency compared with other DMSs as the result of adopting ontology technique and CBR technique for documents classification. However, well-defined keyword ontology, mapping between two ontologies and mapping between topics and documents are obviously the prerequisites to such efficiency improvement.

OBDMS could be particularly useful for the industries in which the team documents are organized as a topic-oriented collection, such as consulting industry. The knowledge structure of this industry has the features of case-oriented (topic-oriented) and reference relationship -- first the enterprise practice data, then the comments based on the cases, and then the generalizations based on the comments and cases. In particular, the two-layer ontology technique can well solve topic-based team document retrieval problems by using ontology to describe keyword space and case space and mapping them.

During our research, we come to realize that how to define a complete keyword ontology is critical enough to be one of the decisive factors for system efficiency and effectiveness. The essence of keyword ontology is the vocabulary of the domain the documents cover or the research team works in. However, issues on the creation, maintenance and update for the vocabulary remain unsolved.

At the same time, mapping between keywords and topics/dynamic attributes is also a factor that is not to be sneezed at. In our OBDMS, the mapping have to be built up and maintained manually, while the ideal state is that system can readjust the mapping between keywords and topics automatically according to the query history, such readjustments refer to adding/deleting mapping, modifying mapping focus degree (Jie Lu 2004), etc.

6. References

A.Aamodt and E.Plaza, “Relating Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches”, AI Communications, 7(1), 1994

A.Maedche, B.Motik, L.Stojanovic, R.Studer, R.Volz, “Ontologies for enterprise knowledge management,” Intelligent Systems, IEEE, 2003, vol.18, no.2, pp. 26 –33.

Chengyao Wang, Limin Lin etc, “Integrating Case-Based Reasoning with feature-adding and relationship-managing technique for the research team knowledge management”, Proceedings of the 9th Cross-Strait Conference on Information management Development Strategy, 2003

F. Crestani, “Combination of similarity measures for effective spoken document retrieval”, Journal of Information science, 2003, vol. 29, no. 2, pp. 87.

Ian Watson. Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann Publishers, Inc., 1997

Jie Lu, Limin Lin and Guangquan Zhang, Keyword-based Document Retrieval Approach Under Two-layer Ontology Framework, Proceedings of the 2nd International Conference on Information Technology & Application (ICITA 2004)

Limin Lin, “Enhancing the Quality of E-service in Consulting Industry using Case-Reference-Net CBR Technique”, 2003

Roger C. Schank, "Dynamic Memory: A Theory of Reminding and Learning in Computers and People", Cambridge University Press, New York, NY, 1982.

S. Staab et al., “Knowledge Processes and Ontologies,” IEEE Intelligent Systems, vol. 16, no. 1, Jan./Feb. 2001, pp. 26–34.

W.Meng and C.T.Yu. Principles of Database Query Processing for Advanced Applications. San Francisco: Morgan Kaufmann, 1997

William S. , Austin, T. 1999. Ontologies. IEEE Intelligent Systems, 1999 Jan/Feb: 18-19