Intelligent Search Engine with Semantic Web Technology

Stephen J.H. Yang[1] Angus F.M. Huang[2] Blue C.W. Lan [3] Irene Y.L. Chen[4]
Dept. of Computer Science and Information Engineering, National Central University, Taiwan [1,2,3]
Ching Yun University, Taiwan [4]
jhyang@csie.ncu.edu.tw[1] {945402002[2], 92542005[3]}@cc.ncu.edu.tw irene@cyu.edu.tw[4]

Abstract

According to the experiences of retrieving contents by search engines, people often get a great deal of information irrelevant to user's intentions. The main reason caused this problem is the lack of enough semantic descriptions of digital contents during analyzing, searching and matching processes. The purpose of this paper is to improve the searching efficiency, user's satisfactions and the practicability of a search engine. We apply exploit ontologyies theory to describe contents in a semantic manner. In our proposed example of digital library application, we utilized Digital Library Ontology to describe contents, domain knowledge and user profiles. Based on reasoning techniques, we developed an inference-based intelligent search engine to assist in literatures retrieval. Experiments show the proposed intelligent search engine can efficiently improve searching performances. To distinguish from traditional keywords search, our approach can provide better searching result, which is based on deduced needs from user. The responded literatures are verified to have better comprehension to user.

Keywords

Semantic Web, Ontology, Reasoner, Information Retrieval, Search Engine, User Profile, Web Ontology Language (OWL), Description Logic.

1. Introduction

The way of accessing information has constantly been changed, from the industrial revolution to the information revolution, and to the present knowledge revolution. How to find specific information from a massive like Internet becomes a difficult work. Information access is easier than before, but this convenience may cause "information overload". To make better information utilization, there are many researches discussed searching methods, like data indexing [Hammouda 2004], [Mili 1988], [Corby 1999] and keyword query [Bai 1998], [Xu 2004], [Oyama 2004].

Knowledge is a kind of well-organized information. Therefore, many facilities focus on knowledge engineering. To maximizing employees' ability, industry deploys knowledge management to increase the usage of knowledge. Knowledge management [Kemp 2001], including discrimination, gathering, exploitation, decomposition, storage and communication of knowledge, etc., called knowledge chain, is to incorporate the industry's knowledge resources. Correct information will bring success and the others will cause failure. The data mining research, likes [Clifton 1998], [Weiss 1999] and [Dorre 1999], is to find out implied but meaningful information.

People do not get right Web content with useful information that they really want. There are various Web content formats and lack of formal representation standards, which results in difficulties to formally describe Web content; there are no semantic relationships among Web content and lack of context and reasoning mechanism, which results in difficulties to retrieve right content. Wrong search results will misguide user's knowledge comprehension. Unreasonable search results will distort user's cognition. Efficient search mechanism will save user a lot of time. We think an intelligent search engine should not only help user do right things but also do things right.

The following points are considered as the main characters that a successful search engine should possess, i.e. Criterion of search rule, Understanding search contents, Possessing professional knowledge, Understanding users' request, Cognition for knowledge evolution.

If the search objective is new knowledge for the user, even search engine is devoted to find the correct subject which is still insufficiency, since users' ability on absorbing and recognizing also need to be considered. To assume the user who wants to learn the knowledge of *Petri nets*, if we can first collect the applications combined with *Petri net* and *workflow* by user's industrial management science background, this can lead user understand the new domain of knowledge rapidly.

Therefore, how can a search engine attracts user's attention and increase their confidence and satisfaction? Not only assisting user to get the document with accurate concept, but retrieved the document which is easy to be comprehended.

Hence, we anticipate extending the searching ability of search engine by users' background knowledge, which can enable search engine find the appropriate documents which conform uses requirements and also easy to be understand. From the Economic point of view, this research is aim to retrieve search result conform maximum unit efficiency, in another word, every result should assist user in understanding knowledge more efficiently, instead of retrieve abstruse sources of literatures and lead difficulty for users on comprehension.

Our goal is to assist users in searching for useful and reasonable contents on WWW. We apply additional and useful descriptions to web contents in order to improve computers understand the meaning within a document. Further, computer can search out the precise contents to users. We will meet some problems when we want do describe an E-document: 1. there are various and complex data type within contents, 2. there are no standard formats for documents, 3. the relations between documents cannot be indicated, 4. there are too much documents to describe manually.

In order to improve the efficiency that the user searches and takes information, an new technology claim that classifying information on internet and using metadata to describe information [Handschuh 2003], [Steinacker 2001]. Their goal is to make data has the capability that is self-describing. Therefore, computer can execute some operation automatically. A version from Tim Berners-Lee [Lee 2001], computers can support more automatic services by machine-readable meta-data, besides existing services on Internet, i.e. digital content management, knowledge sharing and information retrieval. Advantages of Semantic Web [Semantic Web] are decision support, information sharing, knowledge discovery, business development, administration, and automation. The services for the Semantic Web are reasoners, data stores, planners, schedulers, etc.

We use web ontology language (OWL) to improve the semantic description of E-documents and specialized knowledge according to the concept of Semantic Web. We tried to solve the problem that computer cannot understand the meaning within a content. Therefore, we use semantic content search mechanism to build an intelligent search engine. Further, we utilize user's background information to supply the more prefect search services, i.e. using web ontology language to describe the matters of documents, specialized knowledge and user background information, using Logic Reasoner to build an inference-based search engine, defining content correlation formula to calculate the correlation between document and user requirement.

Our Intelligent Search Engine will be explained by a searching behavior of laboratory digital library. And we utilized some experiments to verify the feasibility of inference-based search. In the limited of research time and practical environment, our scope and limited are: 1. The searching contents are bounded in our own digital library. 2. The searching contents' type is focus on PDF format presently. 3. We put forward this research for improving the efficiency of search engine. Therefore we do not consider the searching time for the moment.

In chapter 2, we review some related works about search engine and information retrieval. In chapter 3, we discuss the preliminaries of our study. In chapter 4, we present our system architecture and ontology building of intelligent content search engine. In chapter 5, we proceed to do some experiments. Later in 6, we conclude our study and discuss our contributions to the research.

2. Related Works

The purpose of research for information retrieval (IR) is to fulfill information need. Scholastic search engines are specifically used to search scientific or scholarly literatures and reports. We survey some information retrieval mechanisms. Finally we compare different approaches with ours.

The most important condition of search site is the search results arrangement. Since people do not pay much attention after carefully browsing the first forty results, therefore, the key technique of search site is to employ a mechanism which can arrange the most correct answer at the front part of the results.

The company of Google also develops a new academic search engine namely Google Scholar [Google Scholar], which works even better than Google on specialized information searching. The main purpose of Google Scholar is to enables scholarly literature searching, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all study areas. Furthermore, search efficiency can be enhance by author search, publication restrict, date restrict and some operators.

Take “Semantic Web” as an example, searching “Semantic Web” on Google , 12,300,000 results will be shown, but when searching the same word through Google Scholar, only 14,600 academic results are retrieved. And the preceding results are book, citation, PDF full text document etc. which is not general results as the result derived from Google.

Besides, SCIRUS [SCIRUS] is developed by Elsevier Science, which is designed especially for the researchers. The difference between SCIRUS and other search engines (Yahoo, Gais and Google) is which contain information of science. The main functions of SCIRUS are 1.To filter science irrelevant information 2.To record peer-reviewed document. 3. To narrow the search scope by utilizing author, journal, publish year when searching. 4. It can be used on search conference, abstract and patent information. In order to make search result more accurate, Some limitation are offered, i.e. publish date, information types, file formats, content sources, subject areas, number of results, results clustering , query rewriting.

Other well-known academic search site is Citeseer [CiteSeer], it was developed at the NEC Research Institute by Steve Lawrence, Lee Giles and Kurt Bollacker. They wanted to build a scientific literature digital library and search engine. Especially, the search engine also focuses on literature in computer and information. Sum up these information, we can understand the importance of this research by noticing the rising and flourishing of academic search sites

For common people, truly understanding a paragraph is not an easy task, still more for computers. Because that involve the knowledge and operation about character, word, syntax, semantics and pragmatics. In information retrieval research areas, there are often three types of techniques, i.e. text retrieval [Salton 1988], [Blair 1985], data retrieval [Aho 1979], [Finkel 1974] and knowledge retrieval [Martin 2000], [Staab 2001].We compared the difference between text retrieval, data retrieval and knowledge retrieval as Table 2.1. From these studies, we consider that knowledge retrieval is well for information retrieval. Therefore, we have an idea that is combine search engine with semantic web technologies to improve the efficiency of a scholastic search engine.

Table 2.1: Comparison between different information retrieval methods

	Features	Advantag es	Disadvantage s

Text Retrieval	1.Parse whole statements 2.Directly compare words	1.High Recall	1.Easily mislead semantics 2.Low Precision 3.Non-semantics
Data Retrieval	1.Structured data store 2.Specific schema query	1.High speed	1.Dependent on schema design 2.Low semantics
Knowledge Retrieval	1.Semantics match 2.Knowledge match	1.High Recall 2.High Precision	1.Dependent on knowledge construction

3. Preliminaries

Plenty of documents will be accumulated while the laboratory is constantly developing. We hope to establish efficient knowledge management and share mechanism to make researchers quickly obtain and employ plenty documents in the laboratory, and furthermore, to make the value of knowledge increase. Except structured information technology itself, human is still the main source of knowledge so that we hope knowledge can be constantly created and spread through people's shares.

3.1 Ontology

The functionalities of ontology are formalizing class hierarchies, constrained properties and relations between classes. Powers of Ontology are logic assertions, classification, formal class models, rules and trust. For information retrieval, ontology supplies an architecture description of the specific domain that lead us can comprehend certain knowledge from different viewpoints. Ontology is a conceptual specification; the most famous definition of ontology is contributed by Gruber [Gruber 1993]. Therefore, the features of ontology are utilized to extend search facilities of our search engine. Figure 3.1 shows, by means of descriptive features of ontology, such as Metadata Description, Class Hierarchy, Property Relation and Restriction Constraint; search facilities can be increased to Keyword Search, Taxonomy Search, Association Search and Inference Search respectively.

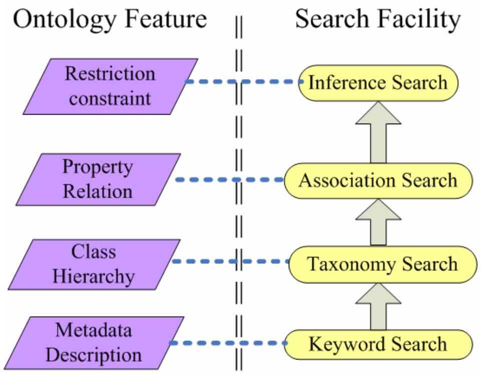

Figure 3.1: Extend search facility with ontology

3.2 Web Ontology Language

The Web Ontology Language (OWL) [OWL] is a semantic markup language utilized to publish and share ontology on the Web. OWL-DL can work on Description Logic-based inference system and contains two major features namely computational completeness and decidability. Computational completeness can guarantee that all results conform to users requirements can be calculated. Decidability can ensure that all computing can be completed in finite time.

In terms of techniques of OWL, knowledge can be operated widely by human via web ontology, which is the same principle as data can be operated through database. After the emergence of OWL, Ontology management system can be applied to certain specific domains, e.g. advanced search engines [Hubner 2004], collaboration context environments [Bernstein 2005], video retrieval [Wu 2004], smart device [Terziyan 2005], text retrieval, and business process integration [Yujie 2004].

The OWL axiom includes subClassOf, sameClassAs, disjointWith, sameIndividualAs, diferentFrom, samePropertAs, subPropertyOf, inverseOf, SymmetricProperty, TransitiveProperty, FunctionalProperty, and inverseFunctionalProperty. Accordingly it has become increasingly clear that how semantics can be added to web pages in order for computers to know how to handle them.

OWL can extend the features of property. The property will satisfy with some axioms that strengthen the description capability and inference capability of OWL [劉昕鵬 2003].

If the property P was marked as **Transitive**, the transitive property will satisfy with the axiom:

$$P(x , y) \text{ and } P(y , z) \rightarrow P(x , z)$$

If the property P was marked as **Symmetric**, the symmetric property will satisfy with the axiom:

$$P(x , y) \text{ iff } P(y , x)$$

If the property P was marked as **Functional**, the functional property will satisfy with the axiom:

$$P(x , y) \text{ and } P(x , z) \rightarrow y = z$$

If the property P1 was marked as **inverseOf** to a property P2, the inverseof property will satisfy with the axiom:

$$P1(x , y) \text{ iff } P2(y , x)$$

If the property P was marked as **InverseFunctional**, the inversefunctional property will satisfy with the axiom:

$$P(y , x) \text{ and } P(z , x) \rightarrow y =z$$

3.3 Description Logic System

We often embed a knowledge representation (KR) system into a larger environment. Other components can interact with the KR component. Description logics are a family of knowledge representation languages that can be used to represent the knowledge of an application domain in a structured and formally well-understood way. It includes some constructors, e.g. *conjunction* (Π), *negation* ($\neg$), *existential restriction* ($\exists$ R.C), *universal restriction* ($\forall$ R.C) and *number restriction* ($\geq$nR). We can describe some typical constructors by instances. For example, if we want do define the concept of "A woman that is married to a professor and has at least three children, all of whom are singers", this notion can be depicted as the following concept description:

Human
$\Pi \neg$ Male $\Pi \exists$ married.Professor Π ($\geq$ 3hasChild)
$\Pi \forall$ hasChild.Singer

First, we can understand that *conjunction* constructor (Π) is interpreted as set intersection and, *negation* constructor ($\neg$) as set complement. Second, the *existential restriction* constructor ($\exists$R.C) describes the set of individuals that have at least one specific kind of relationship with individuals that are members of a specific concept. Third, the *universal restriction* constructor ($\forall$R.C) describes the set of individuals that, for a given property, only have relationships with other individuals that are members of a specific concept. Finally, the *number restriction* constructor ($\geq$nR) describes the number of relationships that an individual may participate in for a given property.

4. Intelligent Content Search Engine

In the literature "The Semantic Web" [Lee 2001], the first diagram explains the purpose of Semantic Web. That is "I know what you mean". Further, we improve our Intelligent Content Search Engine with the Semantic Web technology to reach a bright realm. That is "I know what you need". We will explain the system architecture and search mechanisms in this chapter.

4.1 System Architecture

Our Semantic Content Search includes three main components as Figure 4.1, i.e. Content Register, Knowledge Reasoning and Content Search. The Content Register contains two components, i.e. Metadata Extraction and Property Generation. Metadata Extraction will extract the meaningful metadata from contents. We define thirteen metadata tags, i.e. *URI, Title, Creator, Description, Keyword, Publisher, Content Type, Content Format, Sub Title, Date, Language, Reference* and *Related Annotation*. This information of metadata will support the Description Logic Reasoner to infer the knowledge about content. After the Metadata Extraction, the Property Generation will generate new properties between classes within the DL-Ontology. The mechanism will extend the capacity of knowledge base. Because contents registered in Digital Library are more and more various.

Knowledge Reasoning is the core component of the architecture. It contains a Description Logic Reasoner and a Knowledge Base. They deal with the knowledge implied in content.

Content Search assists users in finding content. Requirement Inference will infer user's requirement from the user request and user profile. It will narrow the range of search accurately and make the search result more suitable for users. After Requirement Inference the Content Match will rank those contents covered within the concepts of user requirement.

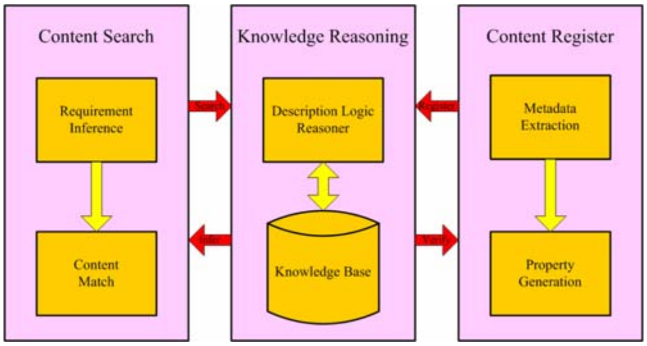

Figure 4.1: Semantic content search architecture includes three main parts, i.e. content search, knowledge reasoning and content register

4.2 Digital Library Ontology

Ontology is a connection between concept and concept. The most common ontology is dictionary, classified catalogue of library and organization chart of company. They can all regard as ontology.Have any relation between contents that can be as foundation and assistance for searching? What factors will influence user's search behavior and thinking model? That are important factors when we want do define an efficient Digital Library Ontology. In order to supply an efficient DL-Ontology, we propose a framework of Digital Library Ontology. As Figure 4.2, it includes three main elements, i.e. Content Element, User Profile and Subject Taxonomy. Purposes of the DL-Ontology are describing what documents were collected, and the knowledge field contain within the documents. For promoting the efficiency of information retrieval, we have increased the description of user's characteristic specially.

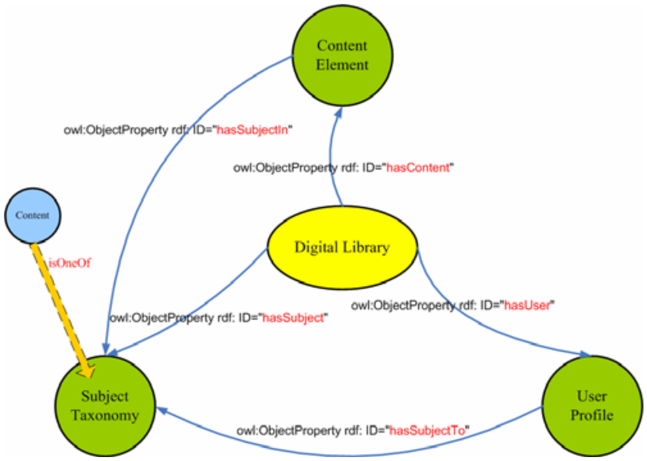
Content Element
Content
Digital Library
Subject Taxonomy
User Profile
owl:ObjectProperty rdf: ID="hasSubjectIn"
owl:ObjectProperty rdf: ID="hasContent"
isOneOf
owl:ObjectProperty rdf: ID="hasSubject"
owl:ObjectProperty rdf: ID="hasUser"
owl:ObjectProperty rdf: ID="hasSubjectTo"

Figure 4.2: The framework of Digital Library Ontology that includes Content Element, User Profile and Subject Taxonomy

Next we will explain the statement and meaning of the three main elements of DL-Ontology.

Content Element

Content Element describes two forms of content, i.e. Annotation and Core. Core means the metadata of original document published by the author. Annotation means some readers annotated information for a specific document. These elements are the author name, the title, the content type, etc.

User Profile

User Profile includes the descriptions of *Job* and *Interest* about users. Modeling user preference is one of the challenging issues in intelligent information systems. From various works [Jung 2005] and [Lise. 2004], we know that user's background knowledge will influence the judgment and comprehension among information retrieval behavior. Different users in their individual knowledge, intelligence, cognitive ability and judge experience will input various user queries. Therefore, we think it is well to define the User Profile into DL-Ontology. That can simulate the search context of user. Therefore, users will influence the result. Files which are highly-related to the user preferences get greater "feature value", and these highly-related files appear at the top of the search result list so that they are more noticeable.

We think that the job, interest, major and location will influence the user's behavior of searching. Therefore, we define these four elements within the User Profile. They are *User Job*, *User Interest*, *User Major* and *User Location*.

Subject Taxonomy

The Subject Taxonomy describes various research domains that the digital library covered. The main research domain categories can extend relevant subjects by the digital library manager. Taking our Intelligent Computer Unit as an example, in order to extend information share, research foundation and knowledge inheritance, we set up a laboratory named Digital Library, which is to collect relevant documents of research topic and to manage them through ICU Content Management System.

The research focus on nine subjects, which are Agent, Case based reasoning, Content Management, E-learning, Formal Methods, Recommendation System, Semantic Web, Web services and Workflow, and these nine subject taxonomies are set in the DL-Ontology. To aim with different research subjects, the relevant documents that were read and provided by researchers will be collected in ICU DL. And ICU CMS that contains these related documents will help researchers and promote the development of laboratory.

How can we use ontology after build it? To Semantic Web, it used for letting software agent describe and inquire the resource mainly. For example, the dictionaries and encyclopedias are our ontology. We will check the reference book possibly when we read the books or study course. Furthermore, we will usually make use of relation among the entries to understand the meaning of the characters. It is the same as agent to read the page marked with ontology.

4.3 Search Mechanism

The search mechanism referred to this research is shown in Figure4.3. Content Register mechanism extracts metadata from document and then selects the keywords from metadata. These keywords are view as an index and stored in Index-DB. In the meanwhile, document ID is registered as an individual of corresponding class within DL-Ontology through the same term of keywords.

When users input the user request, Requirement Inference mechanism will infer the concepts conform to users' requirement, i.e. Related Classes Sets (RCS) through the knowledge stored in DL-Ontology KB. Subsequently Content Match mechanism will gather the documents registered in the Related Classes Sets from the Index-DB. Then it used the corresponding document index terms and ontology-based content correlation formula to count the Content-Correlation between document and user requirement.

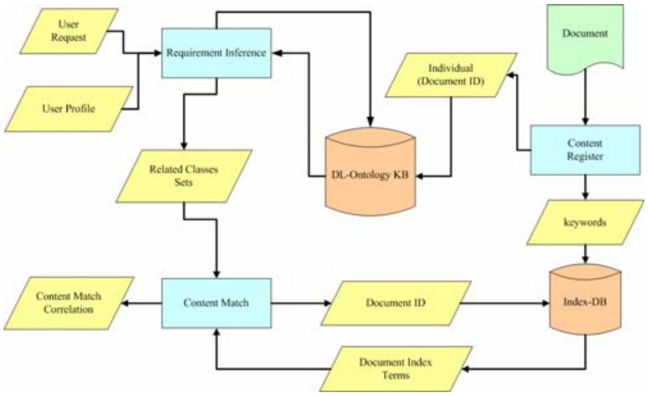

Figure 4.3: Flowchart of search mechanism

4.3.1 Content Register

In the Content Register, the operation processes are as follows:

1. Extracting the metadata from document

2. Selecting the key terms within the metadata.

3. Comparing these key terms with DL-Ontology classes. If they are the same as the classes within the Subject Taxonomy, we will build them as indexes for the document.

4. Then we stored these indexes in the Index-DB.

5. According the index terms, we registered the document as an individual into corresponding classes that they have the same name in DL-Ontology.

6. Adding the properties called *hasApplyTo* between the corresponding classes in DL-Ontology. They can expand the domain relation description in the knowledge base.

4.3.2 Content Search

Content Search includes two mechanisms：

1. Requirement Inference mechanism is used to infer the concepts conform to users' request. Therefore it can reduce the retrieval range exactly and speed the search.

2. Content Match mechanism will compute the content-correlation between user and documents that they belongs to the Related Classes Sets. It will rank these relevant documents by the content correlation.

4.3.2.1 Requirement Inference

We utilize the inference services of Description Logic System to infer the relevant concept about user requirement. The Requirement Inference processes are as follows:

1. Combining the User Request and User Profile as the User Query.

2. Inferring the classes that are conform to restrictions within knowledge base and related to user query.

2.1 The system only infers those classes that they have individuals.

3. Integrating the related classes as Related Classes Set (RCS)

The set means the related knowledge domains conform to user's background and requirement.

4.3.2.2 Content Match

The Content Match processes are as follows:

1. Calculate the correlation called *SetCorrel* between every Related Classes Set and user query such as definition 2

(1) Calculate the correlation between Inferred Class and Query Class, i.e. *ClassCorrel* such as definition 1

a. Multiply inferred class order by the property weight between inferred class and query class

b. The level of correlation of ontology's classes will differ as variance in properties, i.e. sameClassAs > subClassOf > objectProperty. We define these Property Weights as $W_{\text{sameClassAs}}$ =1, $W_{\text{subClassOf}}$ =0.6, $W_{\text{objectProperty}}$ =03

(2) Calculate the average about all ClassCorrel of the Related Classes Set

2. Sort the each related classes set by its correlation

3. Obtain each registered document of class of related classes set in sequence from higher correlation to lower; these documents belong to the concept sets that related to user requirement

4. Calculate each correlation called *ContCorrel* between document and user query, such as definition 3

(1) Select the repeated terms between Related Classes Set and document index

(2)Utilize these terms and formula of Definition 2 to obtain *ContCorrel*

5. Sort these documents by correlation as the result of #4

6. Examine the repeated terms between document indexes and RCS classes

7. Utilize the result of #6 to divide the searched documents into five content match types

Definition 1: Classes Correlation

$$ClassCorrel(ClassA, ClassB) = \frac{1}{\sum_{i=1}^{n}(W_{property} \times N_{Classi})}$$

Where

ClassCorrel : the correlation between two classes

$W_{property}$: the weight of property connect to the Class_i, and $W_{\text{sameClassAs}}$ =1, $W_{\text{subClassOf}}$ =0.6, $W_{\text{objectProperty}}$ =03

n : the number of classes between ClassA and ClassB

N_{Classi} : the order of Class_i

Definition 2: Related Classes Set Correlation

$$SetCorrel = \frac{\sum_{m=1}^{k}(ClassCorrel_m)}{k}$$

Where

SetCorrel : the correlation between the User Query and Related Classes Set

$ClassCorrel_m$: the correlation of class in the RCS

k : the amount of classes in a Related Classes Set

Definition 3: Content Correlation

$$ContCorrel = \frac{\sum_{i=1}^{p}(ClassCorrel_{MatchedClassi})}{p}$$

Where

$ContCorrel$: the correlation between the User Query and specific content

p : the number of matched class between Related Classes Set and Content Index

$ClassCorrel_{MatchedClassi}$: the correlation between the User Query and matched class

Example of Inference-based Search

We will explain how the Intelligent Search Engine fined a specific content. One student wants to find a paper about multi-agent and he majors in Operation Management. According to our mechanism, his User Request is Multi-Agent and User Major is Operation Management. The Intelligent Search Engine will find contents that are related to user's requirements and suitable for his background knowledge. Furthermore, users can understand the content easily. According the search scenario, we illustrate as follows:

User Request: Multi-Agent

User Major: Operation Management

Because there are these concept descriptions in the Description Logic Based knowledge base:

$$MiltiAgent \subseteq Agent$$
$$Service \cap (\exists hasApplyTo.Agent)$$
$$SOA \subseteq Service$$
$$WebService \subseteq SOA$$
$$Agent \subseteq Intelligent\ System$$
$$KnowledgeRepresenttion \cap (\exists hasApplyTo.IntelligentSystem)$$
$$RepresentationLanguage \subseteq KnowledgeRepresentation$$
$$Ontology \subseteq RepresentationLanguage$$
$$OpertationManagement \subseteq ProcessControl$$
$$ProcessControl \subseteq Workflow$$

Through the inference services supported by the Description Logic System, our Intelligent Search Engine will gather some Related Classes Set. One Related Classes Set RCS_1 is:

RCS_1= [Web Service, Ontology, Workflow, Multi-Agent, Operation Management]

According to the individuals registered in these classes of RCS1, we can get some documents from the digital library. In this example, one inferred document is that Paper407: Using Web Services and Workflow Ontology in Multi-Agent Systems [Korhonen 2002]. In Index-KB this paper was registered as [Web Service, Ontology, Workflow, Multi-Agent, Semantic Web, Transaction]

Through the comparison, the matched classes are Web Service, Ontology, Workflow and Multi-Agent. Next we will compute the Content Correlation between this paper and User Query step by step.

ClassCorrel(Web Service,Multi-Agent)=1/(5*0.6+4*0.6+3*0.3+2*0.6)=0.185

ClassCorrel(Ontology,Multi-Agent)=1/(6*0.6+5*0.6+4*0.3+3*0.6+2*0.6)=0.123

ClassCorrel(Workflow,Operation Management)=1/(3*0.6+2*0.6)=0.333

ClassCorrel(Multi-Agent,Multi-Agent)=1/1*1=1

ContCorrel(Paper407,UserQuery)=(0.185+0.123 +0.333+1)/4=0.4103

Therefore, we can get Content Correlation between Paper407 and User Query is 0.4103. Furthermore, because their equivalent classes are [Web Service, Ontology, Workflow, Multi-Agent] between Paper407= [Web Service, Ontology, Workflow, Multi-Agent, Semantic Web, Transaction] and RCS_1= [Web Service, Ontology, Workflow, Multi-Agent, Operation Management], their content match type is Intersection. Besides, users also can utilize the cluster selection of content match type to find out this paper.

5 Experiment Data and Result

A set of queries and their answers are prepared before search. Queries are sent to both Intelligent Search Engine (ISE) and Keyword Search in ICU-CMS (KSE). After two search engines respond, the results can be evaluated and compared through following three type experiments. We first collect all the publications from the nine major research topic literatures in the digital library. The number of collected documents is about 95.

5.1 Evaluation

Recall and precision are methods used to evaluate the performance of Intelligent Search

Engine system for this research. The definition of recall is: Recall=Number_Of_Document_Retrieved_In_Answer_Set/Number_Of_Document_In_Answer_Set. The definition of precision is: Precision = Number_Of_Document_Retrieved_In_AnswerSet /Number_Of_Document_Retrieved

Recall is often used to evaluate system ability on the retrieving answers for user, which means the rate of the amount of documents which answer the query occupied the correct answers. Precision is often used to evaluate how accurate of the answer is which provided by system, which means the rate of the documents belong to the answer of a query from retrieved documents.

We devised three experiments to evaluate the efficiency of Intelligent Search Engine system. First is efficiency of intelligent search without user profiles. Second is the efficiency of intelligent search with user profiles. And third is the effect of user profiles for intelligent search.

The Exp-1 is to input five different user requests into ISE and KSE respectively. And the ISE disable the User Profile. Then we prepared the specific answers for each query. The Exp-2 is that six different users input the same user request into ISE and KSE respectively. And the ISE enable the User Profile. Then these users prepare their individual and satisfiable answers for the query.The Exp-3 is that six different users input the five different requests into ISE with User Profile and ISE without User Profile respectively. And they prepare the specific answers for their own request individually.

5.2 Experiment Result

Experiment-1: ISE without User Profile V.S. Keyword Search

Figure 5.2 and Figure 5.3 illustrate the comparison of recall and precision between ISE without User Profile and Keyword Search. Experiment-1 shows the Intelligent Search Engine is more effective than Keyword Search even the user profile is not applied. The meaning of this experiment is that pure inference-based search is more effective than traditional keywords search.

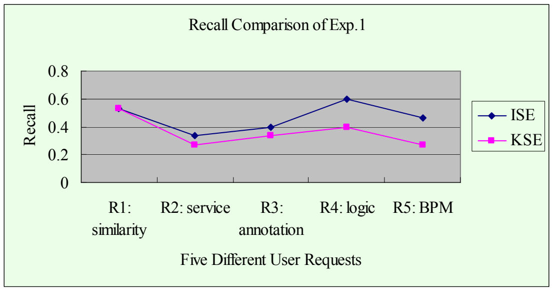

Figure 5.2: Recall comparison of Exp.1

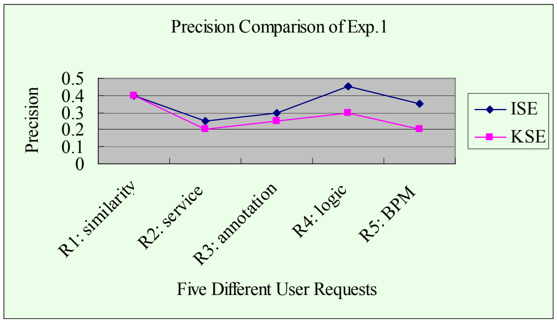

Figure 5.3: Precision comparison of Exp.1

Experiment-2: ISE with User Profile V.S. Keywords Search

Figure 5.4 and Figure 5.5 illustrate the comparison of recall and precision between ISE with User Profile and Keyword Search. Experiment-2 shows the Intelligent Search Engine is more effective than keyword search when user profile is applied.

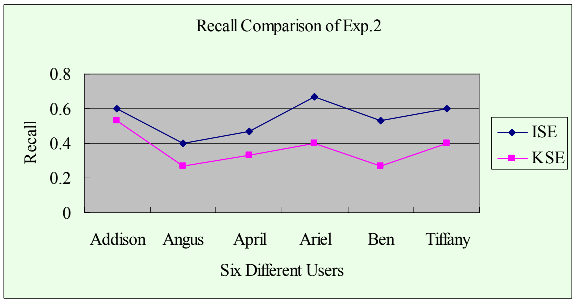

Figure 5.4: Recall comparison of Exp.2

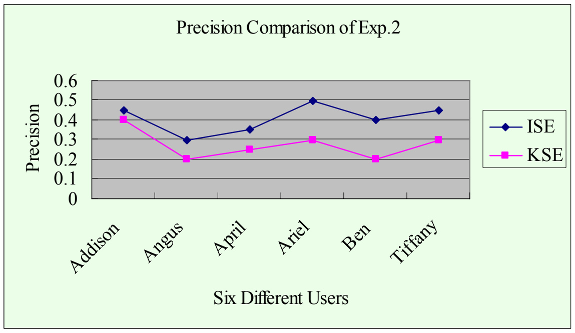

Figure 5.5: Precision comparison of Exp.2

Experiment-3: ISE with User Profile V.S. ISE without User Profile

Figure 5.6, and Figure 5.7 illustrate the comparison of recall and precision between ISE with User Profile and ISE without User Profile. Experiment-3 shows the User Profile can improve the efficiency of Intelligent Search Engine.

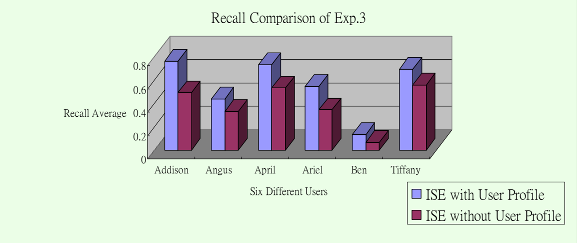

Figure 5.6: Recall comparison of Exp.3

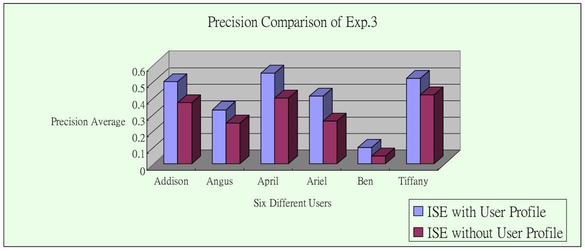

Figure 5.7: Precision comparison of Exp.3

6. Conclusions

The purpose of this research is to make use of Search Engine to combine Description Logic Inference System and Digital Library Ontology to complete Intelligent Search Engine. Major steps are following:

1. According to search engine mechanism, presenting demands and a formula evaluating present related technology of (meaning web) that can solve and promote the efficiency of search engine, and formulating the demands of wisdom search engine.

2. Sorting out the basic information of Digital Library Ontology by building up characteristic analyses of documents and users and the professional knowledge domains; DL-Ontology has to elaborate documents' contents, characteristic of users, category, nature and example of professional knowledge domains and the relationships among them.

3. According to the characteristics of ontology, proposing an operation mechanism of relative levels of concepts. This mechanism can properly operate the meaning relevance of concepts.

4. Making use of Description Logic Inference System to integrate Digital Library Ontology to proceed with the inference of user requirement.

5. Combing Content Search Mechanism and Knowledge Inference to accomplish the study of Intelligent Search Engine.

The Digital Library Ontology as it was presented depicts the contents and characteristics of documents and the construction of professional knowledge, and further integrates and classifies them. Also the application of user profile makes search engine understand users' knowledge background and fondness to search out what documents are users to demand. Furthermore, it can search out the documents fitting users, and easier to read and comprehend and not deviate from the topic. The research result reveals: 1. Applying users' characteristics into search engine can efficiently increase users' benefits of searching. 2. Applying the structure of domain knowledge into search engine can efficiently increase precision and recall of search mechanism. 3. After transforming document description into ontological format of machine-readable by means of OWL, the computer can automatically read and understand the information of documents without artificial help to reach the level of knowledge management. 4. The main purpose of this research system is to search for documents corresponding to users' demands. But academic research and documents are usually the integration of interdisciplinary and multiple-technology study. Therefore, only putting interdisciplinary relationship description of documents into the ontology and then operating ontology inference can enable search engine work more efficiently on search the documents of related domain studies. 5. The structure of domain knowledge built up by the experts can effectively extend the user request from the originally narrowed information to make system more precisely and widely search out the information in relevant category.

Discover via our intelligent Search engine that there are some directions that can be improved. We propose some plans as suggestions and reference to future study, i.e. Rich user profile description, factors that influence search behavior, Concept Visualization of User Interface, Granularity of Ontology. The Granularity of Ontology will influence reasoning. We found that the structure of domain knowledge and the description of characteristic will deeply influence the searching ability of system. We suggest that following related studies are supposed to elaborate the discussion of purposed knowledge so that the well-designed domain ontology could be contrived.

7. Acknowledgements

This work is supported by National Science Council, Taiwan under grants NSC 94-2524-S-008-001 and NSC 93-2524-S-009-004-EC3, and by Ministry of Education, Taiwan under grant A-91-E-FA06-4-4.

8. References

Chinese Reference

[劉昕鵬 2003] 劉昕鵬, “Ontology 理論研究和應用建模--<Ontology 研究綜述>、w3c Ontology 研究組文檔以及 Jena 編程應用總結”，28 March 2003

English Reference

[Aho 1979] A.V. Aho and J.D. Ullman, "Universality of Data Retrieval Languages," Sixth Symp. on Principles of Programming Languages, pp. 110-117,1979.

[Bai 1998] B.R. Bai, C.L. Chen, L.F. Chien and L.S. Lee, "Intelligent retrieval of dynamic networked information from mobile terminals using spoken natural language queries," *IEEE Transactions on Consumer Electronics*, Vol. 44, No. 1, pp. 62-72, Feb. 1998.

[Bernstein 2005] A. Bernstein, F. Provost, and S. Hill, "Intelligent Assistance for the Data Mining Process: An Ontology-based Approach," *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.4, pp. 503-518, April 2005.

[Blair 1985] D. C. Blair and M. E. Maron, "An evaluation of retrieval effectiveness for a full-text document retrieval system." Communications of the ACM, pp.289-299, 1985.

[Clifton 1998] C. Clifton and R. Steinheiser, "Data mining on text," *Computer Software and Applications Conference, COMPSAC 98. Proceedings.1998. The Twenty-Second Annual International*, pp. 630-635, 1998.

[Corby 1999] O. Corby and R. Dieng, "The WebCokace knowledge server,” *IEEE Internet Computing*, Vol. 3, No. 6, pp. 38-43, Nov./Dec. 1999.

[Dorre 1999] J. Dorre, P. Gerstl and R. Seiffert, "Text Mining: Finding Nuggets in Moutains of Textual Data," *KDD-99 SanDiego CA USA International*, pp. 632-635

[Finkel 1974] R.A. Finkel and J.L. Bentley, "Quad trees - a data structure for retrieval on composite keys", *Acta Informatics*, pp. 1-9, 1974.

[Gruber 1993] T.R. Gruber, “A translation approach to portable ontology specifications,” *Knowledge Acquisition*, Vol. 5, No 2, pp. 199-220, 1993

[Gruber 1993] T.R. Gruber, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing,” *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 1993.

[Hammouda 2004] K.M. Hammouda and M.S. Kamel, "Efficient phrase-based document indexing for Web document clustering," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 10, pp. 1279-1296, Oct. 2004.

[Handschuh 2003] S. Handschuh, R. Volz, and S. Staab, "Annotation for the Deep Web," *IEEE*

Intelligent Systems, Vol 18, No. 5, pp. 42-48, Sep/Oct 2003.

[Hubner 2004] S. Hubner, R. Spittel, U. Visser and T.J. Vogele, "Ontology-based search for interactive digital maps," *IEEE Intelligent Systems*, Vol.19, No.3, pp. 80-86, May/Jun, 2004.

[Jung 2005] S.Y. Jung, J.H. Hong and T.S. Kim, "A Statistical Model for User Preference," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 6, pp. 834-843, Jun. 2005.

[Kemp 2001] L.L. Kemp, K.E. Nidiffer, L.C. Rose, R. Small, and M.Stankosky, "Knowledge management, insights from the trenches," IEEE Software, Vol. 18, No. 6, pp. 66-68, Nov./Dec. 2001.

[Korhonen 2002] J. Korhonen, L. Pajunen and J. Puustijarvi, "UsingWeb services and workflow ontology in multi-agent systems," *Workshop on Ontologies for Multi-Agent Systems*, 2002.

[Lee 2001] T. Berners-Lee, J. Hendler, and O. Lassila. "The semantic web," *In Scientific American*, May 2001.

[Martin 2000] P. Martin, P. Eklund, "Knowledge Retrieval and the Word Wide Web." *In IEEE Intelligent Systems special issue Knowledge Management and Knowledge Distribution over the Internet*, 2000.

[Mili 1988] H. Mili and R. Rada, "Merging thesauri: principles and evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 2, pp. 204-220, Mar. 1988.

[Oyama 2004] S. Oyama, T. Kokubo and T. Ishida, "Domain-specific Web search with keyword spices," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 1, pp. 17-27, Jan. 2004.

[Salton 1988] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Information Processing and Management*, pp. 513-523, 1988.

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

[Steinacker 2001] A. Steinacker, A. Ghavam, and R. Steinmetz, "Metadata standards for Web-based resources," *IEEE Multimedia*, Vol 8, No. 1, pp. 70-76, Jan./Mar. 2001.

[Terziyan 2005] V. Terziyan, "Semantic Web Services for Smart Devices Based on Mobile Agents," *International Journal of Intelligent Information Technologies*, Idea Group, ISSN 1548-3657, April 2005.

[Weiss 1999] S.M. Weiss, C. Apte, F.J. Damerau, D.E. Johnson, F.J. Oles, T. Goetz, and T. Hampp, "Maximizing text-mining performance," *IEEE Intelligent Systems and Their Applications*, Vol. 14, No. 4, pp. 63-69, Jul./Aug. 1999.

[Wu 2004] Y. Wu, B.L. Tseng, and J.R. Smith, "Ontology-based multi-classification learning for video concept detection," *ICME '04. 2004 IEEE International Conference on Multimedia and Expo*, Vol.2, pp. 1003-1006, June, 2004.

[Xu 2004] C.Z. Xu and T.I. Ibrahim, "A keyword-based semantic prefetching approach in Internet news services," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 5, pp. 601-611, May 2004.

[Yujie 2004] M. Yujie, Z. Shensheng, and C. Jian, "Providing knowledge support in business process: a context based approach," *IEEE International Conference on Man and Cybernetics Systems*, Vol.3, pp.2143-2149, Oct. 2004.

URL Reference

[CiteSeer] CiteSeer http://citeseer.ist.psu.edu/

[Google Scholar] Google Scholar http://scholar.google.com/

[OWL] Web Ontology Language (OWL) http://www.w3.org/2004/OWL/

[SCIRUS] SCIRUS http://www.scirus.com/

[Semantic Web] Semantic Web http://www.w3.org/2001/sw/