Abstract – A digital library is a type of information retrieval system. The existing information retrieval methodologies generally have problems on keyword-search problem. We proposed a model to solve the problem by using concept-based approach (ontology) and metadata case base. This model consists of identifying domain concepts in user’s query and applying expansion to them. The system aims at contributing to an improved relevance of results retrieved from digital libraries by proposing a conceptual framework for semantic retrieval. We need to import the concept of ontology, making use of its advantage of abundant semantics and standard concept. Domain specific ontology can be used to improve information retrieval from traditional level based on keyword to the lay based on knowledge (or concept) and change the process of retrieval from traditional keyword matching to semantics matching. One approach in this context is using query expansion techniques and the other would be introducing a case based similarity measure for metadata information retrieval using CBR approach. Results show improvements over classic method, query expansion using general purpose ontology and a number of other approaches.

Index Terms – intelligent information retrieval, domain ontology, digital library

I. INTRODUCTION

A digital library (DL) is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers [1]. The digital content may be stored locally, or accessed remotely via computer networks. Many digital libraries have evolved from traditional libraries and concentrated on making their information sources available to a wider audience. Today, many companies maintain their own digital libraries, and research and development for digital libraries now includes processing, dissemination, storage, search and analysis of all types of digital information. In contrast to physical libraries, digital libraries enable concurrent access at any time without physical boundaries. As such, digital libraries can be regarded as indispensable tools for today’s knowledge workers. Digital libraries have always been an appealing playground for innovative computer science solutions. So they became a prominent research area.

In this paper, we focus on digital library within the efficient information retrieval using domain ontology as a controlled vocabulary to expand the input query string. Nowadays, user faces problems of management and sharing of huge amount of documents saved in the DLs. The work proposes methodology and technological framework allowing the user to be provided with a set of relevant documents based on semantic retrieval. Typically, information is retrieved by matching terms in documents with those of a query. The traditional solution employs keyword-based search. The only documents retrieved are those containing user specified keywords. But many documents convey desired semantic information without containing these keywords. The key problem in achieving efficient and user friendly retrieval is the development of a search mechanism. To help end users efficiently retrieve documents relevant to their information needs, this system provides concept-based (ontology) query expansion and traditional statistical information retrieval (IR) algorithms that has given such good results in the IR field. To guarantee delivery of minimal irrelevant information (high precision) while insuring relevant information is not overlooked (high recall), the process of intelligent retrieval system based on the ontology is particularly presented. An increasing number of recent information retrieval systems make use of ontologies to help the users clarify their information needs and come up with semantic representations of documents. A particular concern here is the integration of these semantic approaches with traditional search technology.

An ontology is a collection of concepts and their interrelationships, which provide an abstract view of an application domain. With regard to converting words to meaning the key issue is to identify appropriate concepts that both describe and identify documents, as well as language employed in user requests. The use of ontology to overcome the limitations of keyword-based search has been put forward as one of the motivations of the Semantic Web since its emergence in the late 90’s. While there have been contributions in this direction in the last few years, most achievements so far either make partial use of the full expressive power of an ontology-based knowledge representation, or are based on boolean retrieval models, and therefore lack an appropriate ranking model needed for scaling up to massive information sources.

In the approach [2], a query enrichment approach that uses contextually enriched ontologies was proposed to bring the queries closer to the user’s preferences and the characteristics of the document collection. The idea is to associate every concept (classes and instances) of the ontology with a feature vector (fv) to tailor these concepts to the specific document collection and terminology used. The structure of the ontology is taken into account during the construction of the feature vectors. The ontology and its associated feature vectors are
later used for post-processing of the results provided by the search engine.

The rest of this paper is organized as follows. Section 2 presents overview of the proposed information retrieval system. Section 3 describes the semantic analysis component. Section 4 discusses the ontology model. Case base module described in section 5. And implementation and preliminary test results are discussed in section 6. Finally, this paper is concluded in section 7.

II. OVERVIEW OF THE PROPOSED SYSTEM

Aim at the problem of poor retrieval quality in digital library, the advantage and correlative application of the ontology in digital library’s semantics retrieval fields was introduced. And contributing to an improved relevance of results retrieved from digital libraries by proposing a conceptual framework for semantic retrieval. Semantics retrieval technology would improve retrieval quality extremely, and would be the preferred method to solving the lack of semantic relation in traditional retrieval technology. The work proposes methodology and technological framework allowing the user to be provided with a set of relevant documents based on semantic retrieval and case-based metadata. It also concentrates on formalizing information demand of the user by processing data information to make it useful and provide readers with knowledge services.

In this paper, a novel approach that combined the advantages of concept-based approach with the benefits of statistical approaches based on IR techniques. Most concept-based IR systems use the Wordnet as controlled vocabulary to expand query [6], [7] and [9]. Domain ontology is used as controlled vocabulary for query expansion. And the basic assumption is that a user composing a search query simultaneously is describing a problem he or she seeks to solve. The case based reasoning component handles an information retrieval request as a description of a problem being part of a case. The system can prove how this approach enables various benefits for intelligent query processing and expansion. A good solution for such a case would be a good search result, i.e. a set of links to relevant information with respect to the search query. For the proposed model, a case base is created to represent document information (metadata) contained in the Digital Library. The detailed process flow of the system is shown in figure 1.

The user is able to enter natural language queries which, in turn, are analyzed. The conceptual representation of the query is matched against the database of conceptual representations to select the closest match. It allows the user to start the search with a relevant document or a natural language or Boolean query. It allows the user to browse related documents once a relevant document is found. The retrieval method which is used by traditional digital library based on keyword, it is too unilaterally concerning research of arithmetic to ignore consequence of semantics and mining of semantics of keyword itself.

Under the bag of words model, if a relevant document does not contain the terms that are in the query, then that document will not be retrieved. Query expansion is the process of augmenting the user’s query with additional terms in order to improve results in computer science. For example the query “digital signal processing” from the computer science domain, the concept terms extracted from the domain ontology might be automatically added to the input query string so that more documents that contain these additional terms along with the original terms get higher relevancy.

![Fig. 1 The overview architecture of the information retrieval system](image)

we developed a query expansion method that exploits these relationships. First of all, the query is tagged with POS labels. After this step, the query expansion is done in accordance to the following algorithm:

1. Select from the query the next word (w) tagged as proper noun.
2. Check in WordNet if w has the {country, state, land} synset among its hypernyms; if not, return to 1, else add to the query all the synonyms, with the exception of stop-words and the word w, if present; then go to 3.
3. Retrieve the meronyms of w and add to the query all the words in the synset containing the word capital in its gloss or synset, except the word capital itself. If there are more words in the query, return to 1, else end.

III. SEMANTIC ANALYSIS COMPONENT

Semantic analysis reasoning is the key of implementation semantics retrieval function. It is just that analyzing the
semantics of search terms which is submitted by users. Expanding the classification structure of words semantics then retrieving accordingly data to user interface. We need to identify concepts in information resources (Computer Science documents) and user queries. We need to do conceptual matching between extracted concepts. At this stage it is easy to find exact concept matching but the important part is to match remaining relevant concepts with the help of knowledge repository that is used. The knowledge repository gives information about concepts and their relationships with other concepts. So this stage requires a knowledge repository that does not miss any concepts and any relationships in the application domain.

Firstly, it needs to tokenize the user input query. And then the key domain terms form the tokenized words are extracted. And the only domain terms are expanded with the relevant concepts from the ontology. In this case, an important novelty is that prunes irrelevant concepts and allows relevant ones to associate with documents and participate in query generation. An automatic query expansion mechanism that deals with user requests expressed in natural language has also been proposed. This mechanism generates queries with appropriate and relevant expansion through knowledge encoded in ontology form.

In this component, query expansion that is the process of supplementing additional terms or phrases to the original query plays as an important role in order to improve the retrieval performance. There are three different ways of expanding the query: Manual, Interactive and Automatic. Manual and Interactive query expansion requires users involvement. Sometime user may not be able to provide sufficient information for query expansion, therefore query expansion methods are needed which do not require user’s involvement. Automatic query expansion is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance without user’s intervention.

The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. However, the query expansion has some inherent dangers. The central problem of query expansion is the selection of the expansion terms based on which user’s original query is expanded. Thesaurus has frequently been incorporated in information retrieval system for identifying the synonymous expressions and linguistic entities that are semantically similar. A phenomenon named query drift, that is moving the query in a direction away from the user’s intention, is also related to problem of query expansion. This happens frequently when the query is ambiguous. For example the query “windows” might be about actual windows in houses or the Microsoft Windows operating system. In order to solve this problem, the proposed system used domain ontology to extract the domain concepts as a thesaurus but not as the synonymous expressions. Every tokenized word is not expanded. The query expansion is processed that the terms included in the domain ontology are replaced with the extracted domain concepts from the ontology along with the original terms.

IV. ONTOLOGY MODULE

The main problem with traditional information retrieval (IR) systems is that they typically retrieve information without an explicitly defined domain of interest to the user. Consequently, the system presents a lot of information that is of no relevance to the user. The research presented in this paper examines how ontologies can be efficiently utilized for traditional vector-space IR systems. The ontologies are adapted to the document space within multi-disciplinary domains where different terminology is used. The objective is to enhance the user-experience by improvement of search result quality for large-scale search systems. An approach to concept-based retrieval in Digital Libraries (DLs) is proposed.

During searching and retrieval process, a novel and promising approach is concept-based search [3], [4], [5]. Ontology-based approach to information retrieval (IR) is presented. With this approach, the burden of knowing how the documents are written is taken off the user and hence the user can focus on searching on a conceptual level instead. One problem with this approach is to find good concepts. Domain ontology is useful for query expansion by proliferating the input words with the relevant domain concepts. The system is based on a domain concepts representation schema in form of ontology. With the use of ontology, concepts and relations representing concepts about a particular document in domain specific terms are built.

There are two key problems in using an ontology-based model: one is the extraction of the semantic concepts from the keywords and the other is the document indexing. With regard to the first problem, the key issue is to identify appropriate concepts that describe and identify documents on the one hand, and on the other, the language employed in user requests. In this it is important to make sure that the irrelevant concepts will not be associated and matched, and that relevant concepts will not be discarded.

From the point of construction of ontology model, how the ontology of the categories of computer science domain [10] developed is presented. This domain has 22 subcategories. Concept and property relationship in professional field are defined and field ontology is constructed, according to the professional field (Computer Science). In this construction model, it is used of Seven-Steps Method developed by American Stanford University Physic Institute.

Step1: Confirm the professional field and category of ontology;
Step2: Seeing about possibility of reusing existing ontology;
Step3: List important terms in ontology;
Step4: Define classes and grading system of classes;
Step5: Define property of classes;
Step6: Define aspects of property;
Step7: Create instances.

V. CASE-BASE MODULE

A concept-based search approach based on Case-based reasoning and specific domain ontology is presented. A case is defined by a set of metadata associated with the relevant document. A case base is created to represent the document information within digital library. It is used to retrieve the
information of relevant documents and for contextualizing the search process. This work aims at improving ontology-based information retrieval by the integration of the traditional information retrieval process, the use of domain ontology and the Case-Based Reasoning (CBR) process. In fact, the proposed approach uses the ontology for concept-based query expansion and a combine approach of case-based similarity and textual similarity is used to retrieve metadata information of the related documents and to provide end users with alternative documents recommendations.

In other words, it is important to insure that high precision and high recall will be preserved during concept selection for documents or user requests. We propose to include an alternative way of describing a solution: Given a search query that does not result in the optimal set of available information, a good solution is an “improved” query, i.e. performing this query would deliver better support for solving the given problem.

The CBR approach represents metadata (information about the document in digital library) as a set of cases, a case base. The case attributes are the metadata element set. These attributes used in the case base are extracted from the Dublin Core Metadata Element Set.

Title
Author
Subject
Abstract/Description
Publication

Fig. 2 The partial ontology for computer science domain with Protégé 3.4 editor

To verify the concept-based intelligent IR technique, some experiments were carried out. In this section, we will demonstrate how concept-based query expansion techniques make improved search results to get more relevant information and reduce irrelevance. We also report on the experiments which were carried out to retrieve most relevant documents.

Building complete domain ontology and metadata case base for the computer science domain in digital library is an enormous undertaking, or would need very advanced semi-automatic knowledge extraction techniques that are not available yet in current state of the art.

Query expansion techniques are implemented using Java embedded with SPARQL language for domain terms extraction via Jena Ontology API. The domain ontology contains the terms that are in the categories and subcategories of computer science. There are 22 subcategories of the computer science domain [10] encoded as classes such as Algorithms, Artificial_Intelligence, Computational_Science, Computer_Architecture, and so on. In this case, these subcategories consist of several subcategories included in domain ontology as subclasses, for example, “Algorithms” subcategory contains 47 subclasses encoded as subclasses in ontology as shown in figure 2.

For example, in input string “a” that contains “digital_signal processing”, in which the term “digital_signal”
is the key term in the domain and so the conceptual terms ("digital signal processing, speech processing, wavelets, FFT algorithms, video processing, image processing, time-frequency analysis, digital signal processors, voice technology speech recognition, audio editors") are extracted from the domain ontology using SPARQL query language as shown in table 1. And the extracted terms are added to the input string. Table 2 shows the comparison of precision and recall of the system retrieval with query expansion and without expansion.

VII. CONCLUSION

Concept-based access to information promises important benefits over keyword-based access. One of these benefits is the ability to take advantage of semantic relationships among concepts in finding relevant documents. Another benefit is the elimination of irrelevant documents by identifying conceptual mismatches. Specific domain ontology has been applied as expansion of queries with related terms in the effective operation of an IR system. We explore the idea of using the concepts in ontology to improve search results. In our approach, the query terms are used to match conceptual terms in the ontology. The ontology concepts are adapted to the domain terminology. Our query expansion method was tested and demonstrated that a small improvement could be obtained in precision, but in recall higher increase gained.

REFERENCES


TABLE I. CONCEPT TERMS EXTRACTION SPARQL LANGUAGE

<table>
<thead>
<tr>
<th>Prefix rdfs: <a href="http://www.owl-ontologies.com/Test.owl#">http://www.owl-ontologies.com/Test.owl#</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT ?Digital_signal</td>
</tr>
<tr>
<td>WHERE {?Digital_signal_processing rdfs:Digital_signal ?Digital_signal.}</td>
</tr>
</tbody>
</table>

<p>| TABLE II. PRECISION AND RECALL OF THE PROPOSED SYSTEM BASED ON THE PRELIMINARY TEST |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>Input string</th>
<th>Query generation approach</th>
<th>Total retrieved documents</th>
<th>Retrieved documents</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Not expand</td>
<td>20</td>
<td>19</td>
<td>1</td>
<td>0.95</td>
</tr>
<tr>
<td>a</td>
<td>Expand</td>
<td>74</td>
<td>72</td>
<td>2</td>
<td>0.97</td>
</tr>
</tbody>
</table>