Concept-Based Query Expansion for Information Retrieval within Digital Library

Thinn Mya Mya Swe
Computer University (Mandalay)
Myanmar
thinnmyamyaswe@gmail.com

Abstract

A digital library is a type of information retrieval (IR) system. The existing information retrieval methodologies generally have problems on keyword-searching. 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 query expansion for intelligent concept-based 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 is query expansion techniques using domain ontology and the other would be introducing a case based similarity measure for metadata information retrieval using Case Based Reasoning (CBR) approach. Results show improvements over classic method, query expansion using general purpose ontology and a number of other approaches.

1. 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 query expansion and traditional statistical information retrieval 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.

The rest of this paper is organized as follows. Section 2 presents review of the intelligent 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.

2. Review of the Intelligent Information Retrieval 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. 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.

In [2], it described geographical information retrieval with ontology of place that may be used to derive semantic distance measures for use in geographically-referenced information retrieval. The proposed ontology was characterised by a mix of qualitative and quantitative spatial data including topological relations and sparse coordinate data representing the spatial footprints of places. Places were classified according to their geographical categories and were linked to instances of non-geographical phenomena classified by conceptual hierarchies. An hierarchical distance measure is combined with Euclidean distance between place centroids to create a hybrid spatial distance measure.

In the approach [3], a query enrichment approach that used 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 was taken into account during the construction of the feature vectors. The ontology and its associated feature vectors were later used for post-processing of the results provided by the search engine.

In [4], it reported on how ontologies developed in the EU Semantic Web project SPIRIT were used to support retrieval of documents that were considered to be spatially relevant to users' queries. The query expansion techniques presented in this paper were based on

both a domain and a geographical ontology. An overview of ongoing research was presented in [5] based on the use of a concept network as the knowledge base for inducing a query expansion based on the concepts deduced from the original query terms. In this system, the quality of this conceptual query expansion depended on the quality of the concept network. Query terms were matched to those contained in the concept network, from which concepts were deduced and additional query terms were selected.

Although most concept-based IR systems used the WordNet as controlled vocabulary to expand query [4], [5] and [6], our proposed approach combined the advantages of concept-based approach with the benefits of statistical approaches based on IR techniques in this paper. Domain specific 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. 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 this model, a case base has to be created to represent document information (metadata). The system can prove how this approach enables various benefits for intelligent query processing and expansion. The system architecture is shown in figure 1.

The retrieval method which is used by traditional DL 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.

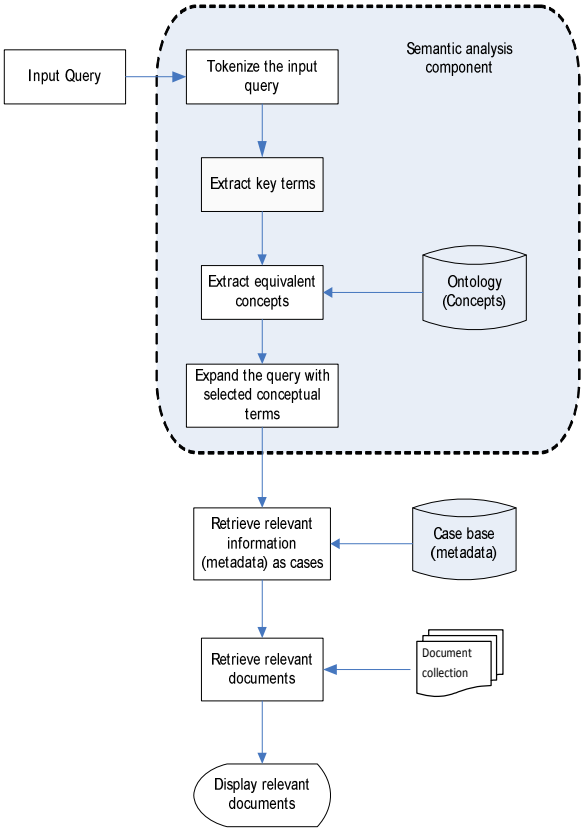

Figure.1 Architecture of the proposed system

3. 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. These processes are automatically carried out that is without any user intervention or feedback. This mechanism generates queries with appropriate and relevant concepts terms 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. 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. 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.

The aim of query expansion is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning 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 its equivalent domain concepts from the ontology along with the original terms.

4. Ontology Module

The main problem with traditional IR systems is that they typically retrieve information without an explicitly defined domain of interest to thie 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.

During searching and retrieval process, a novel and promising approach is concept-based search [7], [8], [9]. Ontology-based approach to 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.

A query expansion method was described in [10] which based on the expansion of geographical terms by means of WordNet synonyms and meronyms. This method was used for the participation to the GeoCLEF 2005 English monolingual task, while using the well-known Lucene search engine for indexing and retrieval. The obtained results show that the proposed method was not suitable for the GeoCLEF track, while WordNet can be used in a more effective way during the indexing phase, by adding synonyms and holonyms to the index terms.

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.

As regard with the construction of ontology Model, how the ontology of the categories of computer science domain [11] 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.

For ontology building, protégé of Standford University is used. It has a graphical user interface. In protégé, the process of constructing ontology includes building file, class, class hierarchy, and producing attribute, the effective value of attribute, and adding examples.

5. 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 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 this module, the processes are carried out as follow:

a new case is matched with all the other cases of the case base;
retrieve the most similar case (or cases) comparing the case to the library of past cases;
reuse the retrieved case to try to solve the current problem;
revise and adapt the proposed solution if necessary;
retain the final solution as part of a new case.

In the proposed approach, the only first two processes (retrieve and reuse) are applied in CBR component. The first three attributes (one or any two or all) from table 1 are used as case description in case retrieval. Case similarity measure is processed for “Author” and “Subject” attributes. The content of “Title” attribute is measured with the statistical IR methods based on the Apache Lucene search engine [12]. In Lucene, a combination of the Vector Space Model (VSM) of IR and the Boolean model was used to determine how relevant a given document is to a user’s query. In general, the idea behind the VSM is the more times a query term appears in a document relative to the number of times the term appears in all the documents in the collection, the more relevant that document is to the query. It used the Boolean model to first narrow down the documents that need to be scored based on the use of boolean logic in the query specification. Lucene also added some capabilities and refinements onto this model to support boolean and fuzzy searching, but it essentially remains a VSM based system at its heart. The cases are indexed with Lucene indexing mechanism.

In other words, it is important to ensure that high precision and high recall will be preserved during concept selection for documents or user requests. We propose 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 cases. The case attributes are the metadata element set. These attributes used in the case base are extracted from the Dublin Core Metadata Element Set and their descriptions are presented in table 1.

Table 1. Case Attributes and their Description

Case Attribute	Description
Title	A name given to the resource.
Author	An entity primarily responsible for making the resource.
Subject	The topic of the resource.
Abstract/Description	An account of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Publication	An entity responsible for making the resource available.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Type	The nature or genre of the resource.
Format	The file format, physical medium, or dimensions of the resource.

The main advantages of this search method are the good results and the applicability to non-structured texts. Our approach can overcome the lack of knowledge about the semantics of the texts with the use of domain ontology in conceptual query enrichment. So the proposed system combined the strength of the statistical IR algorithm with the benefits of ontology model to ensure high precision and recall in information retrieval within digital library.

6. Implementation and Evaluation

Proposed ontology in this paper is pre-tested for query expansion on 374 test collection. 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.

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 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 subcategories 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 its equivalent 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 2. And the extracted terms are added to the input string. However, this system has limitation that is two or more pair domain terms must be separated with underscore, "_", not space. For instance, when user wants to search "concurrency control system", the user must enter "concurrency_control system". There are so many pair domain terms such as data_mining,

machine_learning, artificial_intelligence, data_structure, knowledge_engineering, computer_algebra, computer_vision, error_detection, knowledge_representation, error_correction, cluster_analysis, memory_management, computational_number, computational_statistics, computational_physics, , pattern_matching and so on.

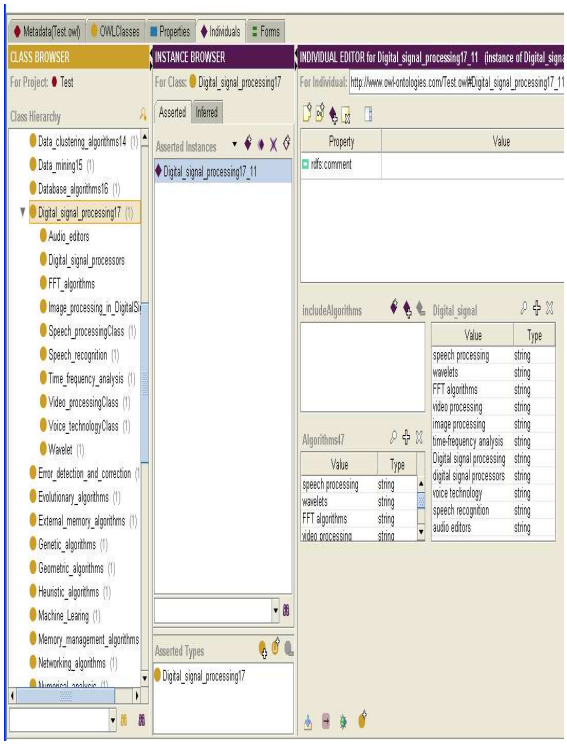

Figure.2 The partial ontology for computer science domain

Table 3 shows the comparison of precision and recall of the system retrieval with query expansion and without expansion. In the experiment, there are 374 total documents i.e 374 metadata cases. And the number of total relevant documents in this collection is 236. The original input string is searched against with "Title" field. The expanded string is found in "Abstract/Description" field. The local similarity values from the respective fields (Title and Abstract) are used to calculate the global similarity or average similarity value for each case. And finally, the top most similar cases which have the largest average similarity values are returned to the user.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|}$$

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|}$$

Table 2. Conceptual Terms Extraction using SPARQL Language

```
PREFIX rdfs: <http://www.owl-ontologies.com/Test.owl#>
SELECT ?Digital_signal
WHERE {
?Digital_signal_processing rdfs:Digital_signal
?Digital_signal.
        }
```

7. Conclusion

Our experiments have shown that already created domain-specific ontology can be effective for query expansion. Nowadays, semantics retrieval technology based on ontology has been the popular research direction. It brings hope to solve problems of lack semantics correlativity in traditional retrieval technology. 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

[1] Greenstein, Daniel I., Thorin, Suzanne Elizabeth. The Digital Library: A Biography. Digital Library Federation (2002) ISBN 1933645180. Accessed June 25, 2007

[2] Geographical Information Retrieval with Ontologies of Place

[3] Tomassen, S.L., Gulla, J.A., Strasunskas, D.: Document Space Adapted Ontology: Application in Query Enrichment. 11th International Conference on Applications of Natural Language to Information Systems. Springer, Klagenfurt, Austria (2006)
[4] Gaihua Fu , Christopher B. Jones , Alia I. Abdelmoty.: Ontology-based Spatial Query Expansion in Information Retrieval. Lecture Notes in Computer Science, Volume 3761, On the Move to Meaningful Internet Systems: ODBASE: OTM Confederated International Conferences, Vol. 3761 / 2005
[5] F. A. Grootjen , Th. P. Van Der Weide.: Conceptual Query Expansion. Data & Knowledge Engineering, Volume 56, (2004) 174-193
[6] Davide Buscaldi, Paolo Rosso and Emilio Sanchis Arnal.: Using the WordNet Ontology in the GeoCLEF Geographical Information Retrieval Task. Lecture Notes in Computer Science in Accessing Multilingual Information Repositories.
[7] Grootjen, F.A., van der Weide, T.P.: Conceptual query expansion. Data & Knowledge Engineering 56 (2006) 174-193
[8] Qiu, Y., Frei, H.-P.:Concept based query expansion. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM Press, Pittsburgh, Pennsylvania, USA (1993) 160-169
[9] Chang, Y., Ounis, I., Kim, M.: Query reformulation using automatically generated query concepts from a document space. Information Processing and Management 42 (2006) 453-468
[10] Buscaldi D., Rosso P., and Arnal E. S.: "A WordNet-based Query Expansion method for Geographical Information Retrieval", 2005.
[11] http://en.wikipedia.org/wiki/Category:Computer_science
[12] E. Hatcher and O. Gospodnetic. Lucene in Action (In Action series). ManningPublications Co., Greenwich, CT, USA, 2004

Table 3. Comparison of Precision and Recall based on the Preliminary Test

Input string	Type of Retrieval	Total retrieved documents	Retrieved documents		Precision	Recall
			Relevant	Irrelevant		
a	Keyword-based retrieval	64	49	15	0.7656	0.2076
a	Concept-based retrieval	245	219	26	0.8939	0.9280