INTELLIGENT RETRIEVAL
Open Journal

*Review*

# Intelligent Information Retrieval within Digital Library using Domain Ontology

**Thinn M. M. Swe**
*Computer University, Mandalay, Myanmar*

**Corresponding author*
**Thinn M. M. Swe**
Computer University, Mandalay, Myanmar

***Article information***
**Received:** February 26th, 2024; **Revised:** March 18th, 2024; **Accepted:** April 7th, 2024; **Published:** April 29th, 2024

***Cite this article***
Swe TMM. Intelligent information retrieval within digital library using domain ontology. 2024; 2(1). doi: https://doi.org/10.70705/ppp.ir.2024.v02.i01.pp30-34

**ABSTRACT**

One kind of information retrieval system is a digital library. Keyword searches are often problematic for the current information retrieval technologies. To address this, we put forth a model that makes use of a metadata case base and a concept-based approach (ontology). The user's query is used to discover domain ideas, which are then expanded upon in this model. By suggesting a conceptual question expansion for intelligent concept-based retrieval, the technology hopes to help make search results from digital libraries more relevant. The idea of ontology, with its wealth of semantics and common concepts, has to be imported. By shifting the focus from keyword matching to semantics matching, domain specific ontology may help enhance information retrieval from a strictly keyword-based approach to one that is more grounded on lay knowledge or concepts. Two methods exist for retrieving metadata: one uses domain ontology for query expansion techniques, while the other introduces a case-based similarity measure using the Case Based Reasoning (CBR) method. In comparison to other methods, including the conventional technique and query expansion utilizing general purpose ontology, the results demonstrate significant gains.

**Keywords**
Digital library; Domain ontology; Information access; Intelligent information retrieval; Query expansion.

## INTRODUCTION

A digital library (DL) is one that stores its contents digitally, rather than in print, microform, or any other physical medium, and makes those collections available to computers [1]. It is possible to save the digital information locally or to view it remotely via computer networks. Many online libraries have developed out of more conventional library models with an emphasis on expanding access to their collection of books and other scholarly materials. Digital information processing, distribution, storage, search, and analysis are now all part of digital library research and development, and many firms now maintain their own digital libraries. There are no time limits or location restrictions with digital libraries, unlike traditional ones. So, it's safe to say that digital libraries are necessities for modern knowledge workers. Innovative computer science solutions have always found digital libraries to be an attractive playground. They thereby rose to the status of a leading field of study.

The effective retrieval of information from digital libraries is the major topic of this study. We use domain ontology, a regulated vocabulary, to extend the input query string. These days, users have challenges while trying to manage and share the massive amounts of information stored in DLs. This paper lays forth a technique and technical foundation for semantic retrieval-based document recommendation systems. It is common practice to get data by comparing query phrases with terms found in documents. Searches based on keywords are used in the conventional approach. Only documents that contain the keywords entered by the user will be returned. Despite this, many publications provide the necessary semantic information even when these keywords are absent. Building a search mechanism is the biggest obstacle to effective and user-friendly retrieval. This system combines concept-based query expansion with classic statistical information retrieval techniques, which have shown to be very effective in the area of information retrieval (IR), to assist end users in effectively retrieving documents that are relevant to their information requirements. The process of an intelligent retrieval system based on ontology is highlighted in order to ensure the delivery of minimum irrelevant information (high accuracy) while also ensuring that significant information is not forgotten (high recall). In order to assist users in defining their information requirements and developing semantic representations of documents, a growing number of information retrieval systems use ontologies. Integrating these semantic techniques with standard search technologies is a particular problem in this context.

A collection of ideas and the connections between them that provide a high-level overview of a topic of potential applications is called an ontology. Finding suitable ideas that characterize and identify documents and language used in user requests is the main difficulty with translating words to meaning. Since its inception in the late 90s, one of the reasons for the Semantic Web's existence has been the use of ontology to circumvent the shortcomings of keyword-based search. Although some progress has been made in this area in recent years, the majority of these efforts either rely on boolean retrieval models or only partially leverage the expressive power of ontology-based knowledge representations. Consequently, they do not have the proper ranking model to handle massive amounts of data.
Here is how the remainder of this paper is structured. Section 2 provides an overview of the AI-powered database search engine. In Section 3, the ontology model is described. Structure of the case as outlined in Section 4. Section 5 elaborates on the part dealing with semantic analysis. Section 6 delves into the implementation and discusses the early test findings. Lastly, section 7 concludes this study.

## 2. REVIEW OF THE INTELLIGENT INFORMATION RETRIEVAL SYSTEM

In order to address the issue of low-quality retrieval in digital libraries, we provided the benefits and related uses of ontology in the semantic retrieval domains of digital libraries. The proposal of a conceptual framework for semantic retrieval also helps to enhance the relevance of results obtained from digital libraries. If we wanted to fix the problem of conventional retrieval technology's lack of semantic connection, we could use semantic retrieval technology. It would greatly increase retrieval quality. This paper lays forth a strategy and technical framework for semantic retrieval and case-based metadata that may provide a list of relevant publications to the user. Users may input questions in natural language and have them processed. The query's conceptual representation is compared to a database of representations in order to get the one that is most similar. The user may start the search by entering a pertinent text, a query in a natural language, or a Boolean expression. Once a relevant document is located, the user is able to peruse similar papers.

In [2], the process of retrieving geographical information using an ontology of places was detailed, which might lead to the development of semantic distance metrics for use in GIR. Topological relations and sparse coordinate data characterizing the geographical imprints of places were among the quantitative and qualitative spatial data that made up the suggested ontology. Geographical categories were used to classify places, while conceptual hierarchies were used to classify examples of non-geographical phenomena. A hybrid spatial distance measure is formed by combining a hierarchical distance measure with the standard geometric distance between two points.

To make the searches more relevant to the user's tastes and the documentation collection's features, the method [3] suggested a query enrichment strategy that made use of contextually enhanced ontologies. To adapt these ideas to the particular document collection and language in use, the plan is to link each ontology concept (both classes and instances) to a feature vector ($fv$). When building the feature vectors, we considered the ontology's structure. Afterwards, the search engine's output was post-processed using the ontology and the feature vectors that went along with it.

According to the findings published in [4], the retrieval of content deemed geographically relevant to users' queries was aided by ontologies created in the EU Semantic Web project SPIRIT. The methods for query expansion discussed in this article used a geographical ontology in addition to a domain ontology. In [5], the state-of-the-art research was reviewed using a concept network as the foundational knowledge to induce an extension of the query using the concepts inferred from the initial query words. In this setup, the idea network's quality dictated how well this conceptual query extension worked. Concepts were inferred and further query phrases were chosen by matching them to those in the concept network.

Despite the fact that the majority of concept-based IR systems relied on WordNet as a controlled vocabulary to enlarge queries [4, 5], and 6, our suggested method in this article integrated the strengths of both the concept-based and statistical methods grounded on IR techniques. When expanding queries, domain-specific ontology is used as a regulated vocabulary. By the same token, the underlying premise is that when a person writes a search query, they are also articulating an issue that needs fixing. As a problem description, a request for information retrieval is processed by the case based reasoning component. A solid group of links to pertinent material in relation to the search query—a search result—would be an excellent option in this situation. The creation of a case base to describe document information (metadata) is necessary for this paradigm. The system may demonstrate how this method allows for several advantages in intelligent query processing and expansion. Figure 1 depicts the system architecture.

Traditional DL's keyword-based retrieval strategy is too focused on mathematical research, ignoring the implications of semantics and mining the semantics of the keywords themselves. The bag-of-words paradigm states that relevant documents will not be retrieved unless they include all of the keywords in the query. In computer science, expanding a user's query with more keywords to better results is called query expansion.

## 3. 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.

A fresh and encouraging strategy for the retrieval and searching process is concept-based search [7], [8], [9]. An ontology-based meth-

od for IR is introduced. With this method, the user doesn't have to worry about understanding the papers' structure; instead, they can concentrate on searching conceptually. Finding excellent ideas is a challenge with this strategy. By multiplying the input words with the applicable domain ideas, domain ontology is helpful for query expansion. Specifically, the system relies on a schema for representing domain notions via theory of things. Ontology is a tool for building domain-specific concepts and relations that reflect ideas about a certain text.

Extending geographical terms using WordNet synonyms and meronyms was the basis of a query expansion strategy disclosed in [10]. Using the popular Lucene search engine for indexing and retrieval, this strategy was used for the participation in the GeoCLEF 2005 English monolingual challenge. Evidence from the GeoCLEF track indicates that the suggested approach did not work, but that WordNet may be more useful when indexing phrases by include holonyms and synonyms.

Extracting semantic ideas from keywords and document indexing are two major challenges when using an ontology-based methodology. In order to solve the first challenge, we need to understand both the language used in user queries and the right ideas to describe and identify documents. It is critical to avoid associating and matching unnecessary ideas and to avoid discarding relevant ones in this process.
Regarding the building of the ontology model, the presentation is made of the development of the ontology of the categories of the computer science domain [11]. There are twenty-two subcategories within this domain. The field ontology is built using the professional field's (Computer Science) definitions of concepts and property relationships in mind. This building model makes use of the Seven-Steps Method, which was created by the American Physic Institute at Stanford University.

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.

## 4. 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.
The suggested method relies on the CBR component alone for the first two steps, retrieval and reuse. For case retrieval, the first three qualities from table 1 are used as the case description. This might be one, two, or all three. The "Author" and "Subject" properties undergo case similarity measurement processing. The statistical IR techniques that are based on the Apache Lucene search engine are used to assess the content of the "Title" property [12]. To find out how relevant a page is to a user's query, Lucene uses a mix of the Boolean model and the Vector Space Model (VSM) of IR. The basic premise of the VSM is that a document is more relevant to a query if its frequency of occurrence is higher than the frequency of occurrence of the word in the whole collection. Prior to scoring any documents, it employed the Boolean model to filter them according to the query specifications' usage of boolean logic. Although Lucene improved and enhanced this model to accommodate fuzzy and boolean searches, at its core it is still a VSM based system. The Lucene indexing technology is used to index the cases.

To rephrase, when selecting concepts for documents or user requests, it is critical to guarantee that high recall and accuracy will be maintained. An alternate method of solution description is suggested. If your search does not provide the best possible results, you may enhance it by running a "improved" query that will help you solve your issue more effectively.

The CBR method uses cases to describe metadata, which is data about the content stored in a digital library. Metadata element set consisting of case attributes. You can find explanations of these properties in the table that was retrieved from the Dublin Core Metadata Element Set. They are utilized in the case base.

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.

## 5. SEMANTIC ANALYSIS COMPONENT

The implementation semantics retrieval function relies on reasoning based on semantic analysis. Simply said, it's the process of examining the meaning of user-submitted search phrases. Improving the user interface by extending the word semantic categorization framework and obtaining data appropriately. Computer science publications and user inquiries include ideas that need to be identified. It is necessary to do conceptual matching between the retrieved ideas. Finding precise concept matching is simple at this point; what's crucial is matching the remaining relevant ideas using the knowledge store. Concepts and the interactions between them are detailed in the knowledge repository. At this point, it is essential to have a database of information that covers all the bases in terms of ideas and connections relevant to the application domain.

To begin, the query entered by the user must be tokenized. Afterwards, the domain phrases that are most important are retrieved from the tokenized words. Additionally, the ontology-related ideas are built upon for the only domain words. Here, a significant innovation is that it eliminates superfluous ideas while letting pertinent ones link to documents and take part in query development. Without any input from the user, these procedures are executed automatically. By using information contained in ontology form, this system is able to construct queries that include suitable and pertinent concept words. An integral part of this component is query extension, which involves adding more words or phrases to the initial query in order to enhance retrieval performance. The query may be expanded in three ways: manually, interactively, or automatically. User participation is essential for both manual and interactive query expansion. Intelligent search

The method of expanding a query involves adding more words or phrases to enhance retrieval performance automatically, without requiring the user to intervene. There is a need for query expansion techniques that do not involve the user since there are instances where the user does not offer enough information.
The goal of query expansion is to narrow the gap between the query and the relevant documents by including terms that are comparable in meaning. But there are certain risks that come with expanding queries. When expanding a user's initial inquiry, the most important thing to consider is which expansion phrases to utilize. A thesaurus is a collection of words and phrases that have common meaning; it is a common component in information retrieval systems. Another issue associated with query expansion is query drift, which occurs when the query moves in a direction opposite to the user's purpose. When the question is vague, this occurs often. A search for "windows" might refer to either the Windows operating system from Microsoft or to real windows in homes. The suggested method avoided this issue by using domain ontology to glean domain ideas for use as a thesaurus rather than as synonymous terms. There is no expansion for every tokenized term. As part of the query expansion procedure, any words found in the domain ontology are substituted with their corresponding domain concepts stored in the ontology.

## 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.

As demonstrated in table 2, the domain ontology is queried using SPARQL to retrieve conceptual terms such as "digital signal processing, speech processing, wavelets, FFT algorithms, video processing, image processing, time-frequency analysis, digital signal processors, voice technology speech recognition, audio editors" that are equivalent to the key term "digital signal" in the text "digital signal processing" that is input string "a". The input string is then updated to include the extracted phrases. Nevertheless, there is a restriction to this scheme that states that underscores, not spaces, must be used to separate two or more pair domain words. The correct way to search for "concurrency control system" is "concurrency_control system," for example. Data mining, machine learning, AI, data structure, knowledge engineering, computer vision, computational algebra, computational statistics, cluster analysis, memory management, pattern matching, computational physics, computational number, and many more pair domain terms are available.

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.

## 7. CONCLUSION

We found that domain-specific ontologies that have previously been constructed work well for query expansion in our trials. Research on ontology-based semantics retrieval technologies has recently been trending. It offers optimism for resolving issues with conventional retrieval methods' lack of semantic correlation. We investigate the possibility of enhancing search results by using ontology princi-

ples. Here, we utilize the query words to find ontology conceptual concepts that match. The ideas from ontology are modified to fit the language of the domain. After putting our query expansion approach through its paces, we saw a big improvement in recall and a little improvement in accuracy.

## REFERENCES

[1] Thorin, Suzanne Elizabeth; Greenstein, Daniel I. A Biography of the Digital Library. In 2002, the Digital Library Federation published an ISBN called 1933645180. Retrieved on June 25, 2007.

[2] Ontologies of Place for Geographical Information Retrieval3] "Document Space Adapted Ontology: Application in Query Enrichment" by S.L. Tomassen, J.A. Gulla, and D. Strasunskas. The Eleventh International Conference on Natural Language Applications to Information Systems. Heidelberg, Germany: Springer (2006)

Spatial query expansion in information retrieval based on ontologies [4] Gaihua Fu, Christopher B. Jones, Alia I. Abdelmoty. Title: On the Move to Meaningful Internet Systems: ODBASE: OTM Confederated International Conferences, Volume: 3761 / 2005 - Lecture Notes in Computer Science

Reference: [5] F. A. Grootjen and Th. P. Van Der Weide: Expanding Conceptual Query. Volume 56, Issue 4, Pages 174–193 of Data & Knowledge Engineering, 2004.

[6] WordNet Ontology Utilization in the GeoCLEF Geographical Information Retrieval Challenge by Emilio Sanchis Arnal, Paolo Rosso, and Davide Buscaldi. Lecture Notes in Computer Science on Accessing Multilingual Information Repositories.

Conceptual query expansion (Grootjen, F.A., & van der Weide, T.P., 2007). Article 174–193 in Data & Knowledge Engineering 56 (2006)

Concept-based query expansion (Qiu and Frei, 2008). This book is the result of the research presented at the sixteenth annual international ACM SIGIR conference on information retrieval. The ACM Press, and Pages 160–169, Pittsburgh, Pennsylvania, USA (1993)

[9] Chang, Y., Ounis, I., and Kim, M.: Reformulating queries using notions derived automatically from a document space. Journal of Information Processing and Management, 42(2006), 453–468

10 "A WordNet-based Query Expansion method for Geographical Information Retrieval" by Buscaldi, Rosso, and Arnal, 2005.

[11] Wikipedia article: "Computer science"

In reference 12, there is work by E. Hatcher and O. Gospodnetic. This is the Lucene in Action series. Publish by Manning Publications in 2004.