An Ontology Based Intelligent Information Retrieval System for Domain Documents

Priyadharshini.G[1], Saravana Balaji.B[2]
Sri Ramakrishna Engineering College,
TamilNadu, India.
[1]dharshini.cse19@gmail.com
[2]saravanabalaji.b@gmail.com

***Abstract-* An ontology based intelligent information retrieval system is aimed at improving the retrieval performance. Information retrieval is the process of retrieving the accurate information for the keyword which is entered by end-users. The Proposed system uses keyword-based semantic retrieval approach. Keyword-based interface provide the most comfortable and relaxed way of querying for the end-user. It includes semantic indexing, word sense disambiguation and query expansion method. This system uses the state of art technologies such as Protégé, WordNet, etc. The accuracy of information retrieval is improved by Ontology based Intelligent Information Retrieval while compared with traditional retrieval system.**

***Keywords-* Ontology, Information Retrieval, Semantic Web, Word Sense Disambiguation, OWL.**

1. INTRODUCTION

The Web is expanding greatly, from its first generation until now it is entering the third generation. But the current situation of the Web is still lacking in establishing links between the resources [1], which make the Web 3.0 is still at its early stage. Initially, Web is developed as a global document repository with a very easy way to access, publish and link documents [2]. And web content is intended for direct human processing. In its current form, machine–based approaches are impossible, unless the content is transformed into machine-readable format [3].

The main drawback of conventional retrieval system is the result retrieved is not concerned about the user search's intent. Mostly it based just on the keyword representation of user queries. The retrieval result is normally high in the recall but low in precision [4]. This means the retrieval system returns hundreds of links for users to check which link pages is relevant and fulfill their need [5]. At the same time, there a lot of evidence showed that most users will only afford to click and examine the first top 10 links from the search results [6]. It means on the first round of the search activity ended with frustration until the user re-query for several times.

Therefore the current challenge for information retrieval system is to put more semantic value into its structure. In simple words, make the machine, understand the content of documents and also, understand the user query so that it will be able to link them in a better way. In addition, the machine should also understand the concept of the application field. But the question is how to make the machine / computer understand in all those aspects? There comes the ontology and the RDF (Resource Description Framework) and linked open data to overcome this problem.

A.Semantic Web

Tim Berners Lee [7] the inventor of WWW has coined the idea of Semantic Web. Because of only human can interpret the content of document, but not machine, so Berners Lee has suggest enriching the Web with machine-process able information which supports the user in his tasks. Meaning that, trying to make the machine able to interpret the meaning of the content in Web by itself. Therefore building the Semantic Web is not an easy task. Thus there is a list of steps show the direction where Semantic Web is heading to [8]:

i. Providing common syntax for machine understandable
ii. Establishing common vocabularies
iii. Agreeing on a logical language
iv. Using the language for exchanging proofs.

The Semantic Web Layer Cake Architecture which is also proposed by Berners Lee. Fig. 1 shows stacks of the components.

B. Classic Information Retrieval

Information Retrieval (IR) is a broad area in Computer Science, where it focuses on fulfilling the user need in finding information of their interest. This area has evolved a lot in step with the growth of the web technology. Information retrieval is described as the task of retrieving relevant document from the collection based on the user request. Research in this field is greatly done by a lot of researchers since its first emergence in 1950s. However, the basic process in IR framework is hardly changing. The main processes are text operation, indexing, query operations, searching and ranking.

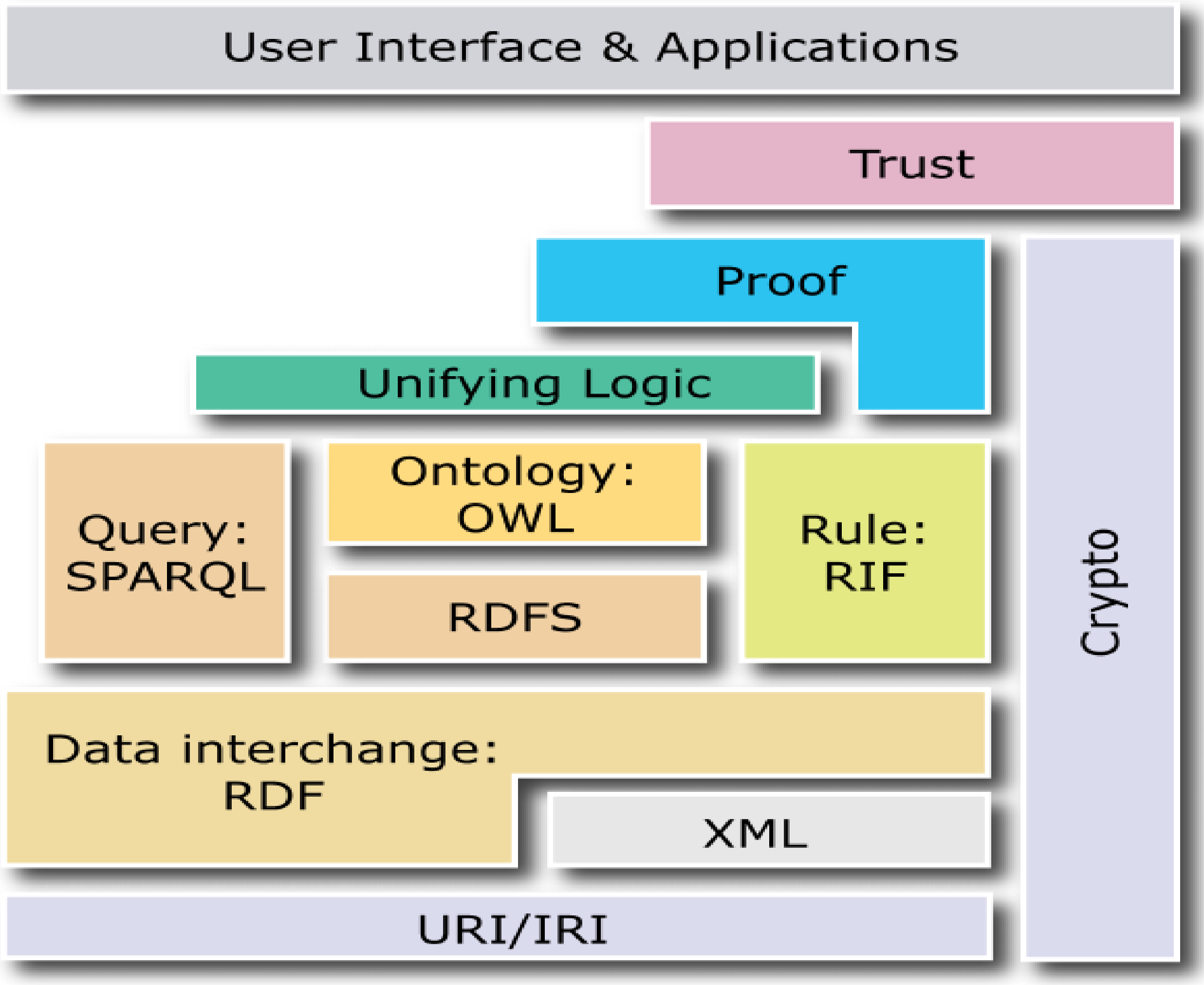

Fig.1 Semantic Web Architecture

Basically there are three main models being used in IR, which is the Boolean model, Vector Space Model (VSM) and probabilistic model. The main difference between these three models is at its index terms. In Boolean, the index terms don't have any weight, while the other two assign weights at its index terms. The performance of IR can be evaluated using precision and recall metric, the most commonly used in the IR field.

IR is also a foundation for various search engines available today such as Google, Bing and many more. In fact it is also a foundation for the semantic search which is actively being researched currently in the Semantic Web era.

C. Semantic Search

The most activity done by user on the Web is searching [10], and searching on the Web also has become a daily activity for many people today [11] [12]. That's why many efforts are continuing to improve the search capabilities. Semantic search is an application of Semantic Web. Semantic search is about search by meaning. Current search techniques are not smart enough to extract the meaning of data; hence it ends by giving irrelevant result to a user's query. In fact the search should permits complex query and can do reasoning to retrieve relevant information [13].

Currently there are four approaches can be applied to semantic search [14]. As seen in Fig. 2, one of the approaches is using contextual analysis, where it emphasizes on how to disambiguate queries. Another approach is reasoning. This type of approach can infer additional information from existing facts in the system. The third approach is to apply natural language understanding, which aim to identify the entity in a sentence. Last approach is ontology, where it can enrich the retrieval of specific domain related. This approach is the most used by many researchers to develop the semantic retrieval system. And many semantic search engine mix and match between those four approaches in various ways to give the best search experience to their user.

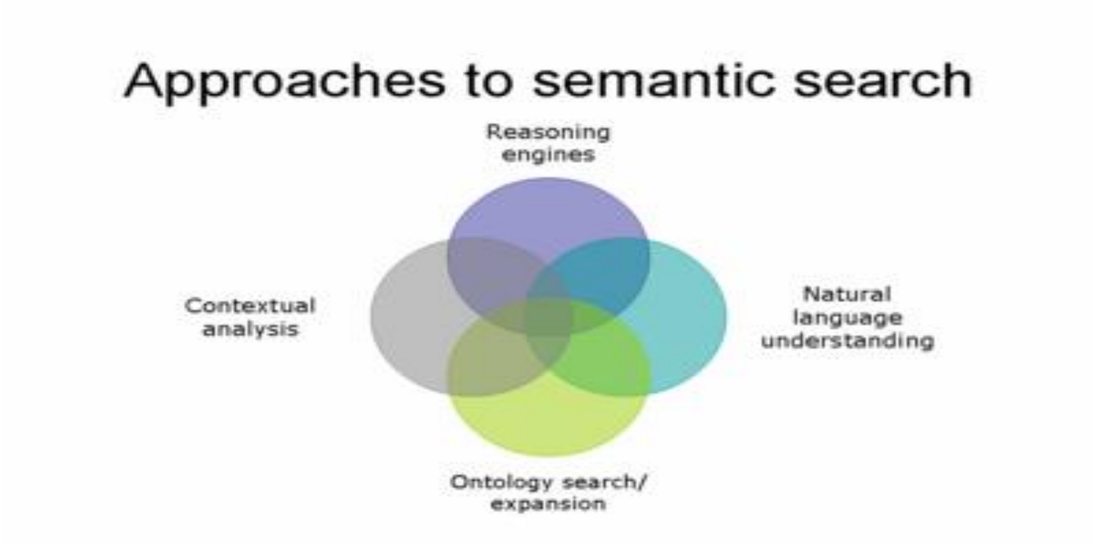
Approaches to semantic search
Reasoning engines
Contextual analysis
Natural language understanding
Ontology search/ expansion

Fig.2 Approaches to Semantic Search

D. Ontology

In the past, the term ontology was quite limited in range of philosophy alone, but now it has a special role in the field of semantic technology. The study of ontology is getting prevailing in the community of computer science.

Information retrieval can benefit a lot from the ontology presence. The usage of ontology in Information Retrieval is getting more attention lately. There are many definitions of ontology being published, but the most popular cited is the definition given by Gruber; ontology is an explicit specification of a conceptualization [15]. In other words ontology is a formal specification of a concept in a specific domain. Some people said ontology is a way to define one concept together with its relations. A concept is an abstract that can easily represent the semantic or hidden knowledge. Simple ontology normally has a set of relationships, but a more complex ontology has rules and the constraint used to manage the relationship.

A vast amount of data on the web is in structured, semi-structured and unstructured form, so there is a need to standardize these data in a formal way. Ontology can be the mechanism to solve this problem [1]. Other than that, there are several other reasons for using ontology such as for knowledge sharing, logic inference, and knowledge reuse [18]. Ontology is said as the backbone of the semantic web, because it may provide machine process able semantics of data and resources can be linked together.

Basic activities related to the ontology are; define the class / concept, arrange the concept in a taxonomy hierarchy (superclass - subclass) and define the relationship together of the value permitted. Ontology development is a complex and mostly it is a domain oriented process. To support this activity, many ontology development tools have been developed by researchers, such as Protégé (Protégé), TopBraid Composer (TopQuadrant), Ontolingua [16] and many more. Among those tools, Protégé is the most popular being used by many people and it is also a domain independent tool [17].

Ontologies are not just presenting information, but they are designed to be used in applications that need to process information and also to do reasoning. They allow greater machine interpretability by providing additional vocabulary along with formal semantics value. There are many aspects of ontology is actively being studied such as ontology development, learning, matching, mapping, alignment and population [19].

However ontology development is not an easy task because it is a collaborative approach. It needs several parties to involve such as the knowledge domain expert, the web developer and the software engineer. They must know how to model the knowledge ontologically. There are also other issue arises in ontology development such as ontology is really domain dependent, ontology alignment is difficult to manage and till now, construction of ontology is still done manually or semi-automatic. So, it's a big challenge to make the Semantic Web success, since ontology is one of the most important elements in it.

2. RELATED WORKS

In this section, a general survey is conducted of the recent works related to information retrieval in semantic web using ontology.

A. D. Vallet, M. Fernández, and P. Castells, "An Ontology-Based Information Retrieval Model," [20]

Authors [20] claimed their approach can be seen as an evolution of the classic vector space model, where they have replaced the keyword based index to semantic knowledge base. Document annotation and weighting processed is done semi-automatic. Instances from the ontology created are used to annotate related documents, and then weight is assigned to the term using TFIDF. For ranking they adapt the vector-based model and takes advantage of ontology. Then they combine the semantic search result with keyword search result and assign combination weights to discover the incompleteness of the ontology knowledge base. From the research they discover, better recall when querying for instances. Better decision by using structured semantic queries. Better precision by using query weights. Better recall by using class hierarchy and rules. Better precision by reducing polysemic ambiguities using instance label and classification of concepts and documents. The model proposed by this author is clear and easy to understand and apply.

B. A. Karthikeyan, "An Novel Approach Using Semantic Information Retrieval For Tamil Documents," [21]

In the paper [21], author proposed an ontology-based retrieval system for Tamil documents. The author creates domain ontology of banking for the Tamil language. This author used the same model suggested by [20] and they apply it to Tamil document. For document processing (Tamil document) they use the normal process in text processing which are tokenizing, special character removal and stop word removal. For retrieval part, they use SAX parser to get list of the instances from the ontology and then match them with the user query. The result produce from their research has shown better precision and recall. So this has proof that model suggested by [21] is suitable for other language such as Tamil.

C. A.Bouramoul, M.K. Kholladi and B.L. Doan "PRESY: A Context Based Query Reformulation Tool for Information Retrieval on the Web" [22]

The author [22] aimed to develop an animal search finder system using semantic approach. In their work they include user profile at the query processing module to reformulate user query, and the result shows more relevant document being retrieved, in fact the user satisfaction has also increase remarkably. Extension to that, the author uses the ontology at query reformulation and document indexing in his second proposal. The study is based on their initial hypothesis that a document can be viewed as a set of concept. The author has made a good effort to enrich query reformulation and document semantic annotation using domain ontology. It has proven that ontology really help in increasing the precision and recall of the retrieval.

D. S. M. Patil and D. M. Jadhav, "Semantic Information Retrieval Using Ontology and SPARQL for Cricket," [23]

Authors in [23] develop a simple information retrieval system for a cricket match domain. They compare between traditional search, extracted search and SPARQL search. Firstly they gathered cricket information by crawling on the Internet. Then the document is processed to extract information and it is stored in the domain ontology. The authors have put some rules so that Pallet reasoner produced inference. The retrieval is done from ontology using SPARQL. From the project, they founded that SPARQL able to answer complex query and SPARQL result are better than traditional search (vector space model using Lucene toolkit) results. The challenge with SPARQL is the user must really know the structure of the related ontology before they can run the query.

E. M. Supiah, "Ontology –Based Semantic Search For Documents Related To Chilli Crop," [24]

In study [24], the author explore ontology based semantic search for agriculture focusing in chili crop. From the literature review, it is shown that recently agriculture domain is also actively involved in semantic technology for improving their management and production. As an example, FOA (Food and Agriculture Organization) has developed AGROVOC; a comprehensive thesaurus encompasses the field of fishery, forestry, food safety and others. In the research, the author created chili ontology adapted from Agropedia (http://agropedia.iitk.ac.in/) knowledge model. There are several crop knowledge models available at Agropedia such as banana, rice, mango, potato, tomato and many more. Agropedia is a digital content organization in the agricultural domain, supported by FOA. The semantic annotation which links the instances in the ontology with related document has been created separately. For evaluation, the proposed model has been compared to searching without using ontology. The result showed that the precision of ontology based search has increased 32.3% compared to search without ontology. In this study, ontology has been used only at the document annotation part, and the result showed quite a good improvement in precision.

3. ARCHITECTURE OF AN ONTOLOGY BASED INTELLIGENT INFORMATION RETRIEVAL SYSTEM FOR DOMAIN DOCUMENTS

The overall process of the proposed system is begin by

1. Collecting data mining domain documents.
2. Creating ontology (OWL file) for each document.
3. Apache Lucene Indexer takes the OWL files as input to create index for those documents.
4. This system given to the users, while the user enters the keywords a query expansion is automatically generated to provide a suggestion to the user.
5. User entered keywords are given as an input to a Word Sense Disambiguation algorithm (IJSAW) to find the sense of the word.
6. The sensed word is given as input to the Lucene Index to find the relevant files.
7. Then those files are ranked based on the frequency of terms present in the document.
8. Ordered documents are displayed to user as result with some recommendations.

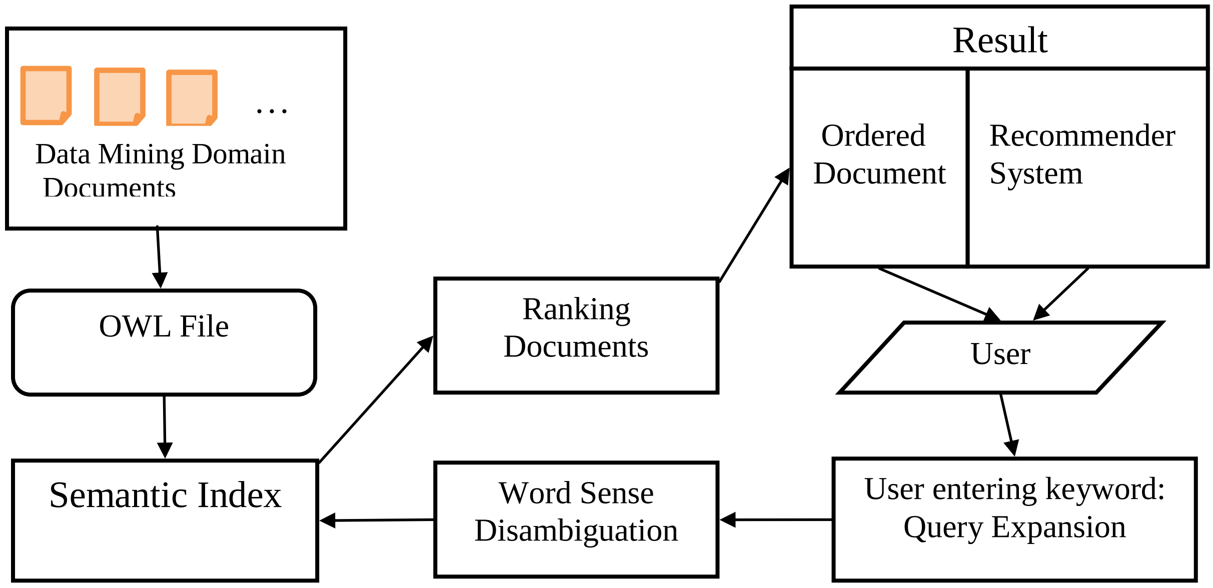
Data Mining Domain Documents
...
OWL File
Semantic Index
Ranking Documents
Word Sense Disambiguation
Result
Ordered Document
Recommender System
User
User entering keyword: Query Expansion

Fig. 3 Architecture of An ontology Based Intelligent Information Retrieval System for Domain Documents

4. IMPLEMENTATION

Implementation of the proposed system consist five of parts:

1. Ontology Creation
2. Semantic Index
3. Query Expansion
4. Word Sense Disambiguation
5. Ranking

A. Ontology Creation

Ontology is created using Protégé tool. For each document a separate OWL file is created.

B. Semantic Index

The indexing mechanism is built upon Apache Lucene version 4.0, a scalable and high performance indexer and searcher, which is essentially designed for free-text search. Semantic retrieval is achieved by implementing a custom ranking for Lucene indices so that documents containing ontological information get higher rate.

Index Structure

The structure of the semantic index has utmost importance in the retrieval performance. A Lucene index is constructed using Apache Lucene such that each entry represents a soccer event. Each event has its own properties associated with it, such as subjects and objects. This information is also included with each event.

Searching

In the traditional keyword search, the indexed documents usually contain nothing but raw text associated with that document. Lucene can easily handle such indices and its default ranking gives usually good results. In order to take the advantages of ontology-aided index structure, have slightly modified the default querying mechanism of Lucene. First of all, boosted the ranking of fields containing the extracted and inferred information to stress the importance of them. Second, these fields are re-ranked according to their importance.

C. Query Expansion

Assume that *Z* is a pool of term-candidates for query expansion. The formulas below present the method to select terms to expand queries. *N* is a number of original query terms, and *j* is an index of them. Values for the WordNet component should be above zero in order to choose related terms.

$$weight(w_i) = \begin{cases} 1, if \sum_{j=0}^{N} \frac{score(w_1, w_2)}{N} > t_1 \\ 0, otherwise \end{cases}$$

$$score(w_1, w_2) = \begin{cases} 1, if \mu_{WNP} > 0; C_\xi > 0; \mu_{EWC} > t_2 \\ 0, otherwise \end{cases}$$

This approach can be interpreted as follows: A term is selected from the list of term-candidates, if the similarity score between this term and the majority of original query terms is higher than a given threshold. The term-candidate should have non-zero values for components.

D. Word Sense Disambiguation

In IR, both terms in queries and the text collection can be ambiguous. Hence, WSD is needed to disambiguate these ambiguous terms. In most cases, documents in a text collection are full articles. Therefore, a WSD system has sufficient context to disambiguate the words in the document. In contrast, queries are usually short, often with only two or three terms in a query. Short queries pose a challenge to WSD systems since there is insufficient context to disambiguate a term in a short query.

One possible solution to this problem is to find some text fragments that contain a query term.

JIGSAW is a WSD algorithm based on the idea of combining three different strategies to disambiguate nouns, verbs, adjectives and adverbs. The main motivation behind our approach is that the effectiveness of a WSD algorithm is strongly influenced by the POS tag of the target word. An adaptation of Lesk dictionary-based WSD algorithm has been used to disambiguate adjectives and adverbs (Banerjee and Pedersen, 2002), an adaptation of the Resnik algorithm has been used to disambiguate nouns (Resnik, 1995), while the algorithm we developed for disambiguating verbs exploits the nouns in the *context* of the verb as well as the nouns both in the glosses and in the phrases that WordNet utilizes to describe the usage of a verb. JIGSAW takes as input a query term $d = \{w1, w2, \ldots, wh\}$ and returns a list of WordNet synsets $X = \{s1, s2, \ldots, sk\}$ in which each element si is obtained by disambiguating the *target word wi* based on the information obtained from WordNet.

E. Ranking Documents

IR systems must be designed to aid users in determining which documents of those retrieved are most likely to be relevant to given queries. Therefore document ranking is very important part. Most commercial text retrieval systems employ inverted files to improve retrieval speed. The inverted file specifies a document identification number for each document in which the word occurs. In order to improve retrieval effectiveness, vector processing systems employing similarity measures have been suggested and studied extensively. In a vector processing system, an expanded query (EQ) can be represented as vector <q1 , q2, qv> with original query terms and one depth descendants in KN. The similarity between EQ and documents can be computed in order to rank the retrieved documents in decreasing order of the query document similarity. Dij represents the weights of term *j* in document i. Qj represents the weights of term *j* in query q. tfDi(tj) represents the term frequency of term tj in document Di. idf(tj) is called the inverse document frequency of term tj and is set to log2(N/df(tj)). N is the number of documents in a collection df(tj) is the document frequency of term tj.

$$\text{Similarity (EQ, Di)} = \text{Dij} * \text{Qj}; \qquad (1)$$

$$\text{Dij} = \text{tf}_{\text{Di}}(\text{tj}) * \text{idf(tj)} = \text{tf}_{\text{Di}}(\text{tj}) * \text{log2(N/dfj)} \qquad (2)$$

Each document vector uses *tf * idf* weighting strategies.

5. EVALUATION

To evaluate the retrieval performance of the proposed system, three queries have been chosen which are given below.

I. Classification algorithms
II. Classify the clustering algorithms
III. Outlier analysis

Proposed system’s retrieval performance is compared with the information retrieval system which does not use Ontology.

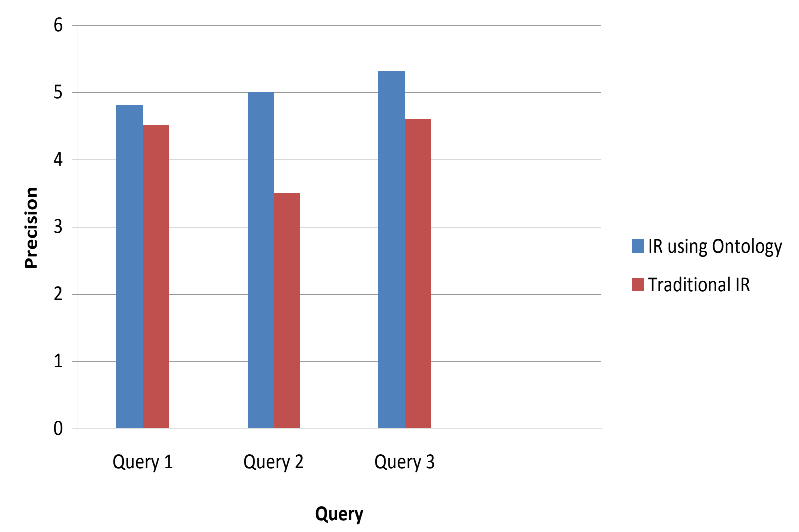

Fig.5 Performance Evaluation Chart

6. CONCLUSION

A semantic retrieval framework and its application in the data mining domain is created, which includes ontology development, semantic indexing, query expansion, word sense disambiguation and ranking documents. Ontology is created using Protégé tool and Apache Lucene is used to construct a semantic index, which is a scalable and high performance indexer. IJSAW algorithm is used to solve ambiguation problem.

These technologies are combined with the keyword-based search interface and obtained a user-friendly, high performance and scalable semantic retrieval system. Also the proposed system can answer complex semantic queries without requiring formal queries such as SPARQL.

REFERENCES

[1] V. Jain and S. K. Malik, "Using Ontologies in Web Mining for Information Extraction in Semantic Web : A Summary," 2010, pp.3–6.

[2] J. Domingue, D. Fensel, and J. A.Hendler, "Introduction to the Semantic Web Technologies," in *Handbook of Semantic Web Technologies*, Springer-Verlag Berlin Heidelberg 2, 2011.

[3] Y. Ding, C. J. Van Rijsbergen, I. Ounis, J. Jose, and S. W. Road, "ACM SIGIR Workshop on „ Semantic Web ‟, 2003.

[4] G. Madhu, a Govardhan, and T. K. V. Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey," *International journal of Web & Semantic Technology*, vol. 2, no. 1, pp. 34–42, Jan. 2011.

[5] R. Sujatha, "Semantic search engine : A survey," vol. 2, no. 6, pp. 1806–1811, 2011.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval : the concept and technology behind search*, Second. Addison Wesley, 2011.

[7] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," *Scientific American*, pp. 29–37, 2001.

[8] G. Stumme, a Hotho, and B. Berendt, "Semantic Web MiningState of the art and future directions," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 2, pp. 124–143, Jun. 2006.

[9] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press. Addison Wesley, 1999.

[10] R. Guha, R. McCool, and E. Miller, "Semantic search," *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, p. 700, 2003.

[11] G. Golovchinsky, A. Diriye, and T. Dunnigan, "The future is in the past : Designing for exploratory search," pp. 52–61.

[12] S. S. Al-rawi, A. T. Sadiq, and S. A. Hamad, "Design and Evaluation of Semantic Guided Search Engine." International Journal of Web Engineerung, 1(3), pp. 15–23, 2012.

[13] M. Unni and K. Baskaran, "Overview of approaches to semantic web search," vol. 2, no. 2, pp. 345–349, 2011.

[14] G. Sudeepthi, G. Anuradha, P. M. Surendra, and P. Babu, "A Survey on Semantic Web Search Engine," vol. 9, no. 2, pp. 241–245, 2012.

[15] T. R. Gruber, "Translation Approach to Portable Ontology Specifications by A Translation Approach to Portable Ontology Specifications," in *Technical Report KSL 92-71 Revised April 1993.*

[16] A. Farquhar, R. Fikes, and J. Rice, "The Ontolingua Server: a tool for collaborative ontology construction," *International Journal of Human-Computer Studies*, vol. 46, no. 6, pp. 707–727, Jun. 1997.

[17] M. R. Khondoker and P. Mueller, "Comparing Ontology Development Tools Based on an Online Survey," vol. 0958, 2010.

[18] X. H. Wang, T. Gu, D. Q. Zhang, and H. K. Pung, "Ontology based context modeling and reasoning using OWL," *IEEE Annual Conference on Pervasive Computing and Communications Workshops. Proceedings of the Second*, pp. 18–22, 2004.

[19] R. García-Castro and A. G.-P. and O. Muñoz-García, "The Semantic Web Framework: A Component-Based Framework for the Development of SemanticWeb Applications," *2008 19th International Conference on Database and Expert Systems Applications*, pp. 185–189, Sep. 2008.

[20] D. Vallet, M. Fernández, and P. Castells, "An Ontology-Based Information Retrieval Model," in *Proceedings of ESWC Conference*, 2005.

[21] A. Karthikeyan, "An Novel Approach Using Semantic Information Retrieval For Tamil Documents," *International Journal of Engineering Science and Technology*, vol. 2, no. 9, pp. 4424–4433, 2010.

[22] A. Bouramoul, M.-K. Kholladi, and B.-L. Doan, "PRESY : A Context Based Query Reformulation Tool for Information Retrieval on the Web," vol. 6, no. 4, pp. 470–477, 2006.

[23] S. M. Patil and D. M. Jadhav, "Semantic Information Retrieval Using Ontology and SPARQL for Cricket," vol. 4, no. 2, pp. 354– 363, 2012.

[24] M. Supiah, "Ontology –Based Semantic Search For Documents Related To Chilli Crop," National University of Malaysia, 2012.