Domain Ontology based Semantic Search for Efficient Information Retrieval through Automatic Query Expansion

Rashmi Chauhan[1], Rayan Goudar[2]
Dept. of Computer Science and Engineering
GEU Dehradun, India
{[1]rashmi06cs, [2]rhgoudar}@gmail.com

Robin Sharma[1], Atul Chauhan[2]
[1]Dept. of Informtion Technology
[2]Dept. of Computer Science and Engineering
GEU Dehradun, India

Abstract—to achieve semantic search, a search engine is needed which can interpret the meaning of a user's query and the relations among the concepts that a document contains with respect to a particular domain. We are presenting the skeleton of such a system based on ontology. In this system, a user enters a query from which the meaningful concepts are extracted; using these concepts and domain ontology, query expansion is performed. For all the terms (expanded and initial query terms), SPARQL query is built and then it is fired on the knowledge base that finds appropriate RDF triples in knowledge Base. Web documents relevant to the requested concepts and individuals specified in these triples are then retrieved. Finally, the retrieved documents are ranked according to their relevance to the user's query and then are sent to the user. If a user wants to find specific information; can search with another module of our system that works without query expansion. The approach of query expansion makes use of query concepts as well as synonyms of these concepts and the new terms relate with the original query terms within a threshold.

Keywords—Information Retrieval, Semantic Search, Automatic query expansion, semantic similarity, ontology, SPARQL, WordNet

I. INTRODUCTION

Information retrieval is the interaction between a user and an information retrieval system that consists of three parts a document representation, a user requirement and a matching function. Web documents are retrieved by the web crawlers and examined. The information collected from each web page is then added to the search engine index. When a query is entered to a search engine, it is checked against the search engine's index of all the web pages it has analyzed. The best urls are then returned to the user as hits, ranked in order with the best results at the top [7]. Though, the current information retrieval systems provide information for each domain but still it is difficult to provide the appropriate information to users what they seek. There are two fundamental issues with current information retrieval systems; information excess and information irrelevancy. The maximum results in the retrieved set are not relevant and do not meet user's intent. A number of the results may be needless. The searching systems can not be able to distinguish the meaning of the statement in terms of the user's requirement. For example, two different users may search for the term 'player' and may have dissimilar perception like one may be interested in 'player of any sport' and another one may be interested in a 'musical player'.

Most search engines do their text query and retrieval using keywords [7]. For example, consider the sentence "what is Query Expansion?" In this sentence, Query is a keyword. It'll be one of the keywords for a particular webpage in some search engine's index. Useful words and key phrases would be the keywords belonging to the index of web page [23].

The existing search engines provide a huge amount of information for any specific query but there are several problems with current searching systems as:

- Identifying hyponyms and synonyms for query keywords
- A lot of information retrieved i.e. information excess
- Poor precision and poor recall

Semantic search looks for improving search accuracy by understanding searcher intent and contextual meaning of terms as they appear in searchable data space, whether on the Web or within a system, to generate more significant results[2][3]. Query expansion allows to search on morphological difference of terms or word sense disambiguation [6].

Query expansion is a technique by using which any user query can be converted into several queries that are related to each other. It means that the basic query terms are replaced by a collection of some new terms and the previous original terms. For these terms the queries are generated that are used to obtain the more relevant results. The relevant terms can be found by using domain ontology [18].

Ontology is being increasingly used for building the applications for the specific domain. Ontology enables users to capture the semantic of the documents [18].

Several Techniques have been proposed by the researchers; but in terms of application in current IR systems they haven't integrated significantly with searching methodologies.

II. RELATED WORK

Various query expansion techniques and approaches designed to perform semantic search have been proposed by the researchers. We studied a number of techniques proposed and implemented for efficient information retrieval and automatic query expansion. The existing approaches and the systems are not suitable enough to get the relevant information. In [5], CROEQS, a semantically improved search engine is presented. It allows the user to query the annotated persons not

only on their name, but also on their roles at the time the multimedia item was broadcast. Authors have used ontology for query expansion as well as for maintaining user profile.

In [6], the meaning of context in relation to ontology based query expansion is examined and a review of query expansion approaches is performed. The various query expansion approaches include relevance feedback, corpus dependent knowledge models and corpus independent knowledge models. All techniques consist of their own advantages and pitfalls. Case studies specify query expansion using domain-specific and domain-independent ontologies are also presented.

In [8], an automatic query expansion method is proposed in which user requests are expressed in natural language. This method creates database queries with suitable and relevant expansion in the course of knowledge in ontology form.

In [11], the semantic retrieval system of sports information using SPARQL is developed. It recognizes the quick retrieval at semantic level according to the relations of "synonymy of", "kind of" and "part of" between sports concepts.

In [12], a hybrid approach of personalized Web Information Retrieval is proposed. In this, ontology for retrieval of user's context is used and a user profile is being maintained.

An automatic rule acquisition procedure using rule ontology RuleToOnto has been proposed in [13]. This system represents information about the rule components and their structures. The rule acquisition procedure consists of the rule component identification step and the rule composition step.

In [14] a method to compute semantic similarity is proposed based on Wordnet [4].

Various techniques for information search through web have been listed in [15]. The limitations in information retrieval schemes have been discussed.

Through this survey we have identified several limitations in existing approaches. To overcome the shortcomings of existing information retrieval, we have presented architecture for semantic information retrieval and from the aspects of query expansion we have proposed an algorithm for query expansion. To limit the expanded results we have used the model of semantic similarity from our previous paper [1].

III. PROPSED SYSTEM

We have proposed the skeleton of a semantic search engine that follows automatic query expansion. We have presented a new technique for ontology based query expansion that has been integrated with the help of specified tools, to our search scheme. The overall architecture is described below.

1. Overall System Architecture

We have presented the overall architecture for our system as Fig.1. The system consists of four basic modules: Domain Ontology Construction, User Interface, Query Handling and Semantic Search Module.

1) Domain Ontology Construction

Ontology has been created for a particular domain (sports here) and is used to model the knowledge for this domain in terms of Concepts (various terms of a specific domain) Relationships between concepts and relationships of concepts.

2) User Interface

The user enters the query representing what information he wants to get. After the processing of other modules the user will get the information accordingly.

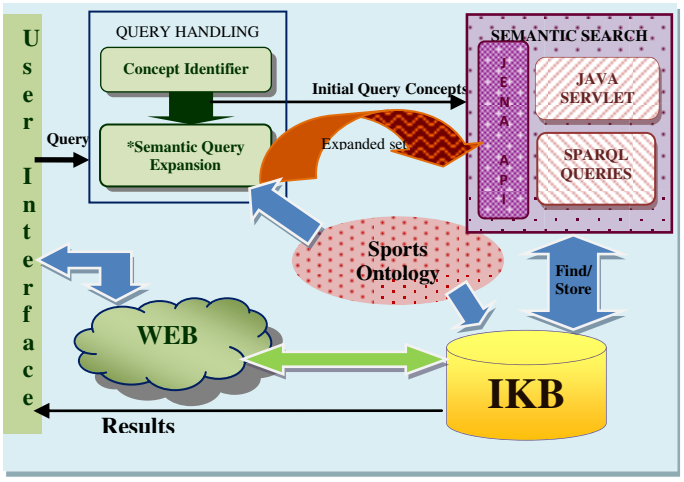

Fig. 1: Overall system Architecture

3) Query Handling

The query entered by the user through user interface is handled by query handling module. The meaningful concepts are extracted through concept identifier and expanded through semantic query expansion section III-B and section IV-C.

4) Semantic Search Engine

The results obtained for each concept (seed and expanded) through SPARQL queries are sent to user section IV-D.

4) Semantic Query Expansion

We are performing query expansion in two rounds:

1) In the first round the first expansion occurs. The synonyms of seed query concepts are obtained using a lexical database dictionary WordNet with the help of java enabled WordNet API.

2) In the second round, the second expansion occurs. For the second expansion, we have proposed a mathematical model to calculate semantic similarity between two concepts in our previous paper and in this paper we are proposing a new algorithm for query expansion which is based on ontology.

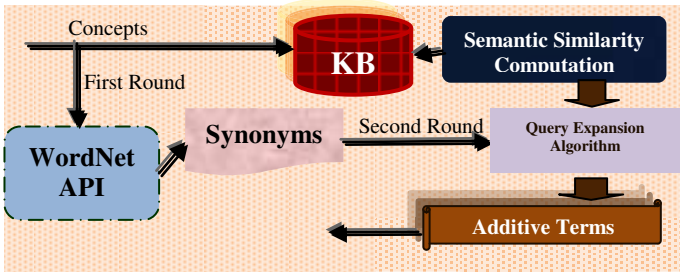

Fig.2: Semantic Query Expansion

Semantic Similarity Computation

Semantic similarity is the weight of relatedness between any two concepts; so that the relation between terms can be examined. Each concept is related to each other in ontology hierarchy; so semantic query expansion can be performed through the concepts that match with query keywords. To accomplish this intent, semantic similarity computation between concepts is needed. We proposed a mathematical model to compute semantic similarity in our previous paper [1]. It considers three main aspects: semantic distance between concepts, Layer factor, and degree of upper concepts [1].

Thus semantic similarity between concepts c1 and c2:

SSim (c1, c2) = 0 if dist (c1,c2) = ∞
SSim (c1, c2) = 1 if dist (c1,c2) = 0 or c1=c2
Otherwise, SSim (c1, c2) = SSm
Where,

$$\mathbf{SSm} = p * 1/(1 + a * dist(c1, c2)) + q * 1/(|L(c1) - L(c2)| + 1) + r * |comupc(c1,c2)| / |upc(c1, c2)|$$

and p + q + r = 1

2. Query Expansion Algorithm

We have proposed an algorithm for query expansion. This algorithm has been used for second round of query expansion. Let Qcon be the set that contains the combination of seed query concepts (Con) and the synonyms of these concepts (Syn). S_{class} and S_{ind} are the sets that contain the categorized concepts that are classes and individuals matched in our ontology respectively. The set CLASS contains the collection of expanded classes and INDV contains the corresponding individuals. Finally, QE will contain all the expanded terms. The algorithm is as follows:

ALGORITHM

Input: Qcon = Syn U Con
Output: Sets of CLASSES and INDIVIDUALS

1) Qcon = Syn U *Con i.e. {q1,q2,q3,…….qn}*
2) for each concept q_i in Qcon, categorize the words according to the RDF database

S_{class} = $\{cl_1, cl_2, \ldots\ldots cl\}$
S_{ind} = $\{Id_1, Id_2, \ldots\ldots Id\}$

3) for each concept cl_i in S_{class} do
  for each concept c_j in RDF do
    Find SSim(cl_i, c_j)
      if SSim(cl_i, c_j) ≥ *0.4 then*
        add c_j to S_{exp}
      otherwise, break;
    end for
  end for
4) CLASS = S_{class} U S_{exp}
5) for each class in set CLASS do
  Ind = getIndividual(CLASS)
  end for
6) INDV = S_{ind} U Ind
7) QE = CLASS U INDV

Finally, as an output of the algorithm we will have all the classes related to the entered query and corresponding individuals according to user's intent. These expanded terms and the initial query terms will be used to form the queries so that the documents will be selected on the basis of these new queries. These results will be ranked according to the semantic similarity values of the expanded classes so that the most relevant information would be displayed on top position. The results obtained thus will be more relevant.

IV. Implimentation Details And Experiments

The proposed system is implemented with appropriate tools and the proposed techniques. The details of entire work are discussed in brief in the following sections. We performed various experiments to our system and then computed precision and recall. We will discuss about all the modules individually in the following sections.

A. APPROACH

As soon as user queries the system, the entered query is firstly pre-processed and expanded by some relevant concepts as:

The request goes to the web server. On the web server the request is processed by the java Servlet. The Servlet first categorizes the words in query on the basis of its type according to the database. If the query consist the word which is a class in the database or an individual, it splits it into two different categories i.e. class and individual. Then the similarities of classes presented in the set of classes are calculated with the classes in the database with the help of proposed formulae. If the similarity value comes out to be more than 0.4, it is added to the set of expanded concepts.

When the final expanded set is prepared, then SPARQL queries are fired on the classes and individuals present in the sets and desirable output is sent back to the client application.

B. Tools And Technologies

We have used Protégé tool to create ontology for sports domain. We created user interface in html and Java Script. The user enters the query through user interface and will get the results accordingly. The query will be expanded using the mathematical formula given in the section III-B. For all of the purposes of query expansion and information retrieval, we have used Jena API with Java eclipse. We have used WordNet API to fetch the synonyms of query concepts. WordNet is a lexical database that stors of a lot of words with their possible forms in the form of ontology. The Java API of WordNet is available; we have integrated this existing API to our approach of Query Expansion.

C. Module 1: Sports Ontology Construction

The ontology for storing information about various sports is constructed that consist of various classes, individuals and properties with their attributes and relationships.

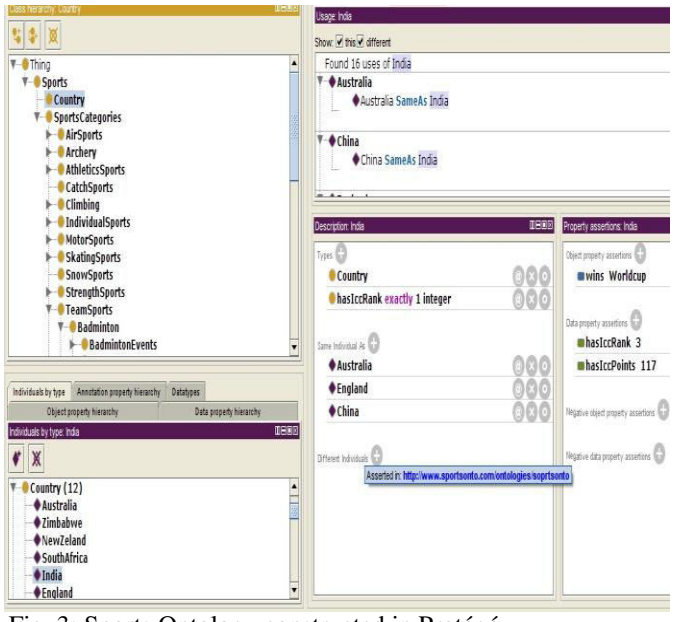

Fig. 3: Sports Ontology constructed in Protégé

The snapshot of the constructed ontology for sports in protégé is given in Fig.3. There are various categories of sports that contain their underlying sports/games and the corresponding events. A number of individuals have been created for various classes. We have instantiated our ontology with the information for various terms of the domain. We applied various restrictions on the data properties as well as the object properties for these individuals.

D. Module 2: User Interface

The user enters the query related to sports domain. He/she can

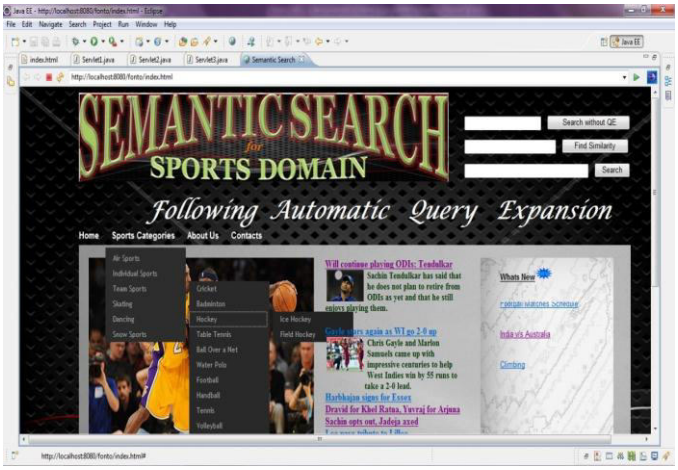

Fig. 4: A partial view of User Interface

select the available category to make the search more specific.

There is a common web page info.html that describes the sports categories in brief. From this the user can get an idea about the domain so that he can enter the appropriate query. There is another search box for semantic similarity. The user can see the various related classes existing in ontology with their similarity with reference to the entered query terms. If the user wants to search some specific information then he can perform search without query expansion. There is a separate search box for semantic search without query expansion.

E. Module 3: Query Handling

The query entered by the user is preprocessed and the concepts are extracted. Then the Synonyms of these concepts are found for **first QE.** We used WordNet API to get the synonyms.

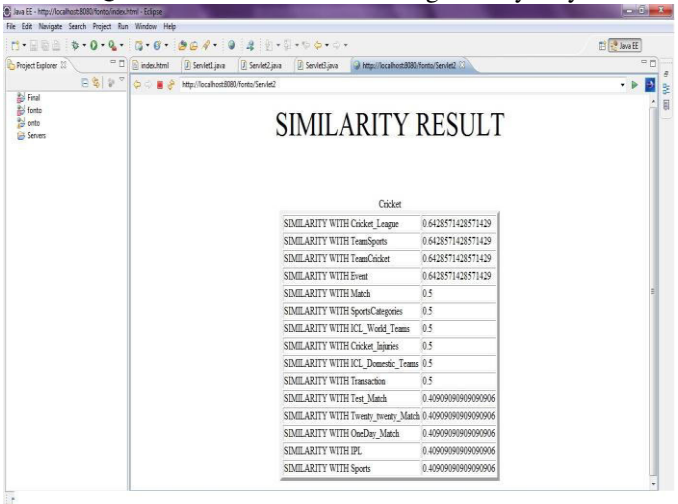

Fig. 5: "what are the various classes related to Cricket?"

The factors semantic distance factor, layer factor and the common upper concepts factors are implemented through Jena API in java. Fig.5 shows the results obtained for the query ***"what are the various classes related to Cricket?"*** In this Fig. the results show the classes related to Cricket class having the semantic similarity greater than 0.4.

In this query we want to see the concepts/classes related to the entered semantic concept Cricket. Using the mathematical model that we have proposed and implemented, all the related classes will be retrieved along with their SSim value. By this user can get an idea about the corresponding domain ontology; so that the next query can be entered in a proper form.

For the second round of QE, the proposed algorithm is implemented in java.

F. Module 4: Semantic Search Engine

We have used Jena API as the interface between user and the KB/RDF. Jena is a Java API that is used to manipulate RDF data. It uses SPARQL queries to perform any operation on RDF/Ontology. SPARQL (pronounced "sparkle", SPARQL Protocol and RDF Query Language) is an RDF (Resource description framework) query language. It is a query language for databases that is used to retrieve and manipulate data stored in RDF format. SPARQL query consists of triple patterns, conjunctions, disjunctions, and optional patterns. A SPARQL query is having the following form:

SELECT ?subject ?object WHERE { ?s rdfs:p ?o }

The semantic results are sent by the Java Servlet to the user.

If a user enters a query like "List of recent *IPL matches"* then firstly the meaningful words will be extracted by removing raw keywords like 'of' and the extracted concepts are: (list, recent, match). The synonyms for these concepts will be extracted as results of first round query expansion. These concepts will be matched to our ontology knowledge base and the second algorithm will be applied. For the output of second round of query expansion the SPARQL query will be fired through JENA and the following results will be retrieved. The retrieved results are shown in Fig.6.

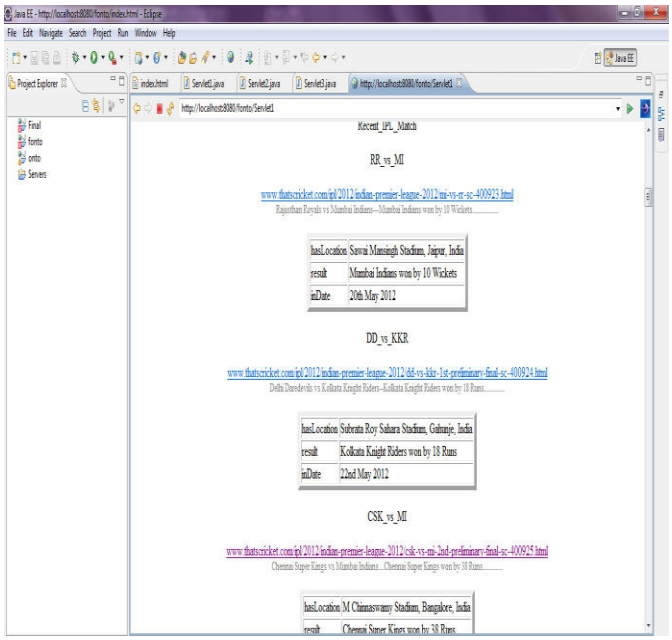

Fig. 6: the results for the query "List of recent IPL matches

G. Experiments and Performance analysis

As we discussed earlier, our system performs the domain specific search using automatic QE. We tested our system for a number of queries related to sports domain.

Queries Tested through the system to analyze the performance

1) What are the various events in Cricket?
2) What are recent IPL matches?
3) List the players of IPL 2012 Cricket Team
4) What is the ICC rank of England?
5) What are the details of The OVAL 2012?
6) Top 10 ICC ranked country name?
7) What type of cricket injuries can occur while playing cricket?
8) List the test match held in 2012.
9) Types of cricket League?
10) What are the types of skating sports?

1. *"What are the various events in Cricket?"*

If a user wants to know the events of cricket, then not only the events are his intention. He may be interested in getting more information about the cricket domain. If we consider the keyword-based search (e.g. Google) then there are a lot of results containing the well known term cricket and events. The required information exists only in some specific documents; so most of the results retrieved may be useless from user's perspective. So we are performing Semantic Query Expansion. We have stored the URI's of the web documents as data properties for a particular individual in ontology having about 1-4 documents for each individual of the classes related data to teamsports. We stored information about 580 web documents to create the sample knowledge base in the form of RDF/XML data. JENA is used to manipulate this knowledge base.

Precision and Recall are often used when evaluating a search engine's effectiveness. The precision is expressed as the ratio of relevant results with respect to the total retrieved results of a query entered by the user [23]. A simple binary relevance judgment is the most common approach to examine the precision of a search engine; as the result for a query may be either relevant or not. Let X_{rel} be the number of relevant results and X_{nonRel} is the number of non-relevant results. Then precision of our system P can be determined as follows:

$P = X_{rel} / X_{rel} + X_{non\text{-}relevant}$

Similarly, the recall is expressed as the ratio of relevant results with respect to the total relevant information existing in the data set for a query entered by the user. Consider Y_{rel} be the number of relevant results retrieved for a query and Y is the total number of relevant information presented in the data set. The recall of our system can be computed as: $R = Y_{rel} / Y$

For Query2, there are the concepts recent, IPL and matches. In ontology, recent_IPL_match is a class that consists of instantiated individuals regarding various match information. IPL and match are also classes in the data set. So, the results are displayed for all the three classes. We have stored 20 recent IPL matches information through the class recent_IPL_match.We have stored the record of 35 one day matches of 2012, 40 of 2011 and 25 for 2010 in history of OneDay_match class which is the subclass of match class.

In case of without query expansion the results will be displayed only for the information limited to three classes existing in seed query. Thus, P = (7 + 5 + 9) / (21 + 24) = 0.47

But, there is some more information stored in the data set that can be relevant and has not been retrieved directly. So the recall will be poor, as: And R = (7 + 5 + 9) / 48 = 0.43

Similarly, precision and recall for 10 sample queries are listed in tableI. The plotted the graph is also given for the same.

Table I: Precision and Recall for 10 sample queries

QUERY	PRECISION		RECALL	
	Without Query Expansion	With Query Expansion	Without Query Expansion	With Query Expansion
Query1	0.67	0.84	0.55	1
Query2	0.47	0.92	0.43	1
Query3	0.65	0.89	0.53	1
Query4	0.6	0.94	0.6	1
Query5	0.74	0.86	0.7	1
Query6	0.55	0.77	0.4	1
Query7	0.77	0.86	0.46	1
Query8	0.64	0.88	0.5	1
Query9	0.44	0.91	0.56	1
Query10	0.83	0.85	0.3	1
AVG	**0.622**	**0.87**	0.50	**1**

In case of Query expansion the results are retrieved for all the three classes of seed query along with the information stored through various classes in ontology that are semantically related to these three classes. Thus, all the relevant information will be retrieved and the recall will be 1 i.e. 100%. But the results can consist of some irrelevant information because the expanded results are also being displayed. So, the precision will be lesser than recall. Thus,

P = (132+173+122) / (430 + 35) = 0.92

R = (132 + 173 + 122) / 427 = 1

H. Experiments for sub-domains of our sports ontology

The efficiency of the system is tested for three sub-domains i.e. Cricket, Hockey and Football for the ideal case (minimum one query concept is presented in the query entered by the user). For example, 750 data elements were stored for Cricket. Query Expansion based information retrieval approach for Cricket and 776 documents retrieved for this particular search. In these results some of the documents were irrelevant also. We analyzed that 748 were relevant. So, the number of irrelevant results = 776 – 748 = 28

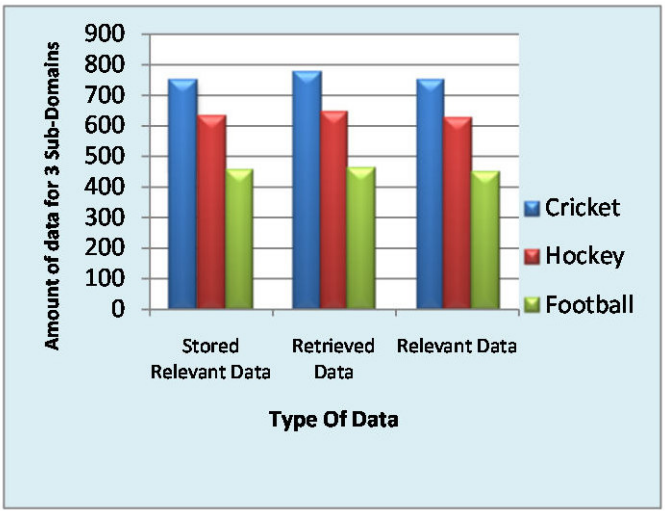

Fig. 7: Amount of data elements for three sub-domains

Precision for the Cricket = (748/776) * 100 = 96.39%

Recall for the Cricket = (748/750) * 100 = 99.73%

Similarly, we computed precision and recall for hockey and football in ideal case (see TableII). Finally, we have calculated the average Precision of three sub-domains as:

(748 + 625 +451) / (776 + 643 + 462) * 100 = 96.97%

Similarly, Average Recall for these three sub-domains as:

(748 + 625 +451) / 750 +632 + 453) * 100 = 99.40%

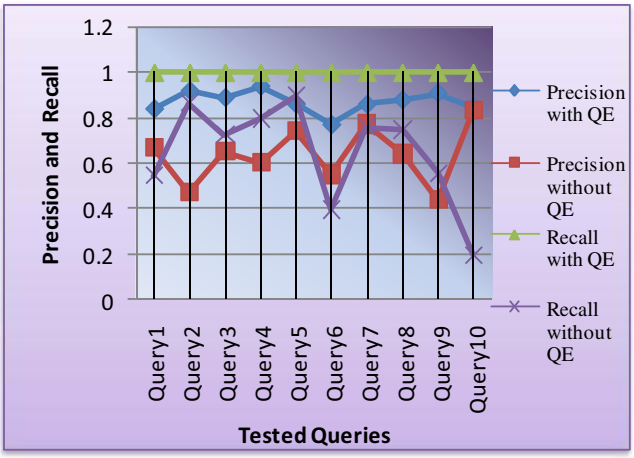

Fig. 8: Precision and Recall for 10 sample Queries

Table II: Precision and Recall for 3 sub-Domains

Term	Cricket	Hockey	Football
Stored Relevant Data	750	632	453
Retrieved Data	776	643	462
Relevant Data	748	625	451
Precision	96.39%	97.20%	97.62%
Recall	99.73%	98.89%	99.56%

Considering these statistics we constructed charts to represent it in the graphical form in (Fig.8, Fig.9) as follows:

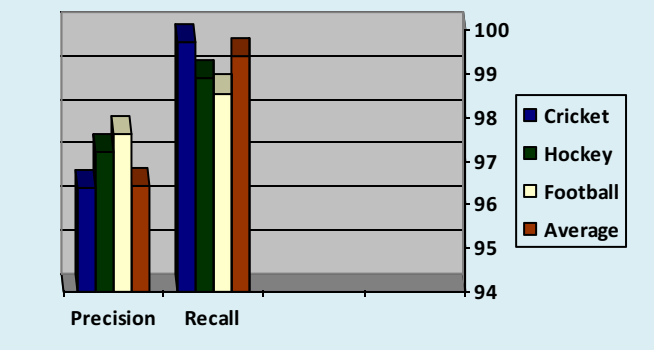

Fig.9: Amount of data elements for three sub-domains

The experiments show that the recall value for any particular sub-domain or for a specific query is higher than precision; as we are performing query expansion, so all the relevant information is extracted to the results. The user can also perform limited search without query expansion. The performance of system is good as it provides results with query expansion as well as specific search without query expansion.

Thus, our system provides the domain specific semantic search with higher percentage of relevant information.

V. Conclusions & Future Scope

A domain specific semantic information retrieval system has been implemented using appropriate tools and the proposed technique of ontology based automatic query expansion. The approach is different from others as:

1). It utilizes the query concepts as well as the synonyms of these concepts to perform Query Expansion.

2). the new terms are added only if consisting of a Similarity within a threshold.

3). only the relevant documents will have top rank.

Putting together the technique of Automatic Query Expansion and semantic search made the application more efficient. The system is domain specific as, we have targeted the single domain (sports domain) but in future the method can be applied for a multi-domain platform using various domains so as to make the effective semantic search for the entire web.

References

[1] Rashmi Chauhan, Rayan Goudar, Rohit Rathore, Priyamvada Singh, and Sreenivasa Rao *"Ontology Based Automatic Query Expansion for* Semantic Information Retrie*val in Sports Domain"* ICECCS 2012, CCIS 305, pp. 422–433, 2012. © Springer-Verlag Berlin Heidelberg.

[2] A. Sheth, C. Ramakrishnan. "Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis." In IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real, December 2003.

[3] Junaidah Mohamed Kassim, and Mahathir Rahmany, "Introduction to Semantic Search Engine," International Conference on Electrical Engineering and Informatics, pp. 380-386, Selangor, Malaysia, 2009.

[4] G. Miller, "WordNet: A Lexical Database for English" Communications of the ACM, vol. 38, No. 11, pp. 39-41, 1995

[5] Stijn Vandamme Johannes Deleu Tim Wauters Brecht Vermeulen Filip De Turck, "CROEQS: Contemporaneous Role Ontology-based Expanded Query Search", 2009 International Conference on Communication Software and Networks, IEEE.

[6] J. Bhogal, A. Macfarlane, P. Smith, "A review of ontology based query expansion" Information Processing and Management (2007), Elsevier.

[7] Latifur Khan, Dennis McLeod, Eduard Hovy, "Retrieval effectiveness of an ontology-based model for information selection", The VLDB Journal (2004) 13: 71–85 / Digital Object Identifier, Springer-Verlag 2003.

[8] Chiraz Latiri, Hatem Haddad, Tarek Hamrouni, "Towards an effective automatic query expansion process using an association rule mining approach", LLC, Springer 2011

[9] L. Ding, T. Finin, A. Joshi et al., "Swoogle: A Search and Metadata Engine for the Semantic Web", ACM International Conference on Information and Knowledge Management, 2004, pp. 652-659.

[10] Dou Hao, Wanli Zuo, Tao Peng, Fengling He, "An approach for calculating semantic similarity between words using WordNet", 2011 Second International Conference on Digital Manufacturing & Automation, IEEE 2011

[11] Jun Zhai, and Kaitao Zhou, "Semantic Retrieval for Sports Information Based on Ontology and SPARQL," 2010 International Conference of Information Science and Management Engineering, pp. 395-398, IEEE

[12] Namita Mittal, Richi Nayak, MC Govil, KC Jain, "A Hybrid Approach of Personalized Web Info. Retrieval",IEEE International Conference on Web Intelligence and Intelligent Agent Technology, 2010

[13] Sangun Park and Juyoung Kang, "Using Rule Ontology in Repeated Rule Acquisition from Similar Web Sites", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.24, NO.6 IEEE 2012

[14] Wang Wei, Payam Barnaghi and Andrzej Bargiela, "Probabilistic Topic Models for Learning Terminological Ontologies" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 7, IEEE, JULY 2010.

[15] Arasu, A.,Cho,J.,Garcia-Molina,H.,Paepcke,A., Raghavan, "S.Searching the Web", J. of ACM Transactions on Internet Technology 2001