Semantic Web Based Efficient Search Using Ontology and Mathematical Model

K.Palaniammal [1], Dr.S.Vijayalakshmi [2]

[1]Research Scholar, [2]Assistant Professor
Department of Computer Applications
Thiagarajar College of Engineering,
Madurai, TamilNadu, India-625015,
sudharoentgenphd@gmail.com[1]
svlcse@tce.edu[2]

***Abstract* —** The **semantic web is the forthcoming technology in the world of search engine. It becomes mainly focused towards the search which is more meaningful rather than the syntactic search prevailing now. This proposed work concerns about the semantic search with respect to the educational domain. In this paper, we propose semantic web based efficient search using ontology and mathematical model that takes into account the misleading, unmatched kind of service information, lack of relevant domain knowledge and the wrong service queries. To solve these issues in this framework is designed to make three major contributions, which are ontology knowledge base, Natural Language Processing (NLP) techniques and search model. Ontology knowledge base is to store domain specific service ontologies and service description entity (SDE) metadata. The search model is to retrieve SDE metadata as efficient for Education lenders, which include mathematical model. The Natural language processing techniques for spell-check and synonym based search. The results are retrieved and stored in an ontology, which in terms prevents the data redundancy. The results are more accurate to search, sensitive to spell check and synonymous context. This paper reduces the user's time and complexity in finding for the correct results of his/her search text and our model provides more accurate results. A series of experiments are conducted in order to respectively evaluate the mechanism and the employed mathematical model.**

Keywords- Ontology, Word Net, Semantic Web, Information Retrieval, Mathematical Model, Spell Check, Semantic Search, SPARQL, Natural Language Processing.

I. INTRODUCTION

Web search has been one of the motivations of the semantic web since it was envisioned. Semantic web was envisioned by Tim Berners-Lee as the extensions of the current web. It is necessary in the first place for mark-up data on the web semantically, so that they can be understood and processed by agents autonomously [7]. The semantic web vision is based on structuring the knowledge that is present in the current web so that it is understandable by machines without human intervention. Ontology is the core technology for the semantic web [10]. It removes the complexities to discover, present, gain, or maintain feasible electronic information on the web and it provides the method for a data representation to enable software products (agents) to provide intelligent access to heterogeneous and distributed information [16, 17].

The huge increase in the amount and complexity of reachable information in the World Wide Web caused an excessive demand for tools and techniques that can handle data semantically [5]. The existing web search systems are mostly keyword-based and identify relevant documents or information by comparing the keywords. Keyword-based search, in spite of its merits of apt query for data and ease-of-use, has flopped to illustrate the fulfil semantics held in the content and has led to the following problems [15]: (1) keywords could represent only fragmented meanings of the content, and the content determined through keywords did not always meet the queries requisites. The quarries had to screen retrieval results and correct keywords several times to obtain the required information. (2) Matched to a content, a query normally comprised fewer contents, which might lead to inaccurate retrieval results due to problems like insufficient information being used in the search method, incomplete query topics, and complexity in determining requisition features. (3) Forthcoming to synonym and polysemy in human language, information retrieval through keywords can only cover information enclosing the matching keyword, while other information with similar semantics but different keywords has been completely left out [12]. Based on the above analysis, in this paper, we develop semantic web based efficient search using ontology and mathematical model for education domain that handles the creation of ontology knowledge base, embedding spell-check, finding synonyms, proposing SCBR algorithm, querying the ontology using SPARQL. Our main concern is providing a scalable and efficient search with high retrieval performance.

This paper is structured as follows: In section II we give the details of related work to this research; In section III we present our approach to semantic web based efficient search using ontology and mathematical model; In section IV focuses on the working process details; In section V, we present on our system evaluation details; finally we conclude the paper and future work in section VI.

II. RELATED WORK

The traditional keyword-based search approaches are based on the vector space model tried by Salton et al. [21, 31]. In this model, documents and queries are simply represented as a vector of term weights, and the retrieval is done according to the cosine similarity between these vectors. Some of the important studies related to traditional search are [13, 18, 22, 23, 24, 25]. This approach does not require any extraction or annotation phases. Therefore, it is easy to implement, however, the precision values are relatively low. Wang et al. [1] tossed a semantic seek methodology to collect knowledge from normal tables, which has the three main steps: identifying semantic relationships between table cells; converting tables into data in the form of a database; and retrieving objective data by query languages. This work demonstrates how intelligent agents can extract the tabular information for answering queries with the assistance of ontological knowledge; the intelligent agents can distinguish concepts and instances in each table cell. Bhagwat and Polyzotis [3] proposed a file system search engine that integrates hypertext and web-based techniques by adopting a more structured view of file systems. It is based on a simple, yet powerful framework that automatically infers semantic relationships among files and thus transforms a conventional directory-based file system in a network of hyper-linked documents. Takama and Hattori [4] designed a method for mining association rules that reflect the behaviours of past users was proposed for an adaptive search engine. The logs of users' retrieving behaviours were described with the RDF model, from which association rules that reflect successful retrieving behaviours are extracted. A generalization law written with the RDF Schema was also proposed to absorb the variety of users' behaviours.

Semantic search is using an information extraction, there are many studies in this field main dissimilarities between these studies arise from the structure of sources, details of the extracted information and computational memory resources. NLP- based approaches are domain independent but use parse trees of sentences, pos taggers, chunk parsing, anaphora resolution, etc. in order to extract information. They need heavy computational processes [26, 27, 28, 29, 30]. There are some alternative information extraction methods such as pattern / rule-based information extractors against heavy computational costs. These methods are classified according to the creative forms of patterns and rules: automatic or manual. Automatic methods [9, 27, 32, 33, 34, 35, 36, 37] are superior compared to the manual ones considering the effort spent on the domain. On the other hand, they suffer from low precision-recall rates. Lee and Tsai [2] designed an interactive semantic search engine which collects feedback by means of selection in order to better capture users' personal concepts. It emphasized the importance of user feedback and concept modelling in web search. To succeed the concept-based semantic search, it presents an agent-based framework to overcome the difficulty of specifying appropriate queries to retrieve web contents semantically relevant to a user's need. Toma et al. [6] proposed a service ranking approach based on semantic descriptions of services non-functional properties was proposed. It introduces an approach for modelling and attaching non-functional property descriptions of services and goals. The proposed ranking mechanism makes use of logical rules describing non-functional properties of services and evaluates them using a reasoning engine.

The ontology based search systems are discussed in different aspects [5, 14, 38, 39, 40, 41, 42]. The survey of [5] deals with ontology based search and extraction system and its application in the soccer domain with three issues in semantic search, namely, scalability, and retrieval performance. In [41] proposes a knowledge based framework for integrating ontology based information personalized retrieval in support of reminiscence. The ambition is to benefit people in reminiscing, browsing and rediscovering events from their lives by considering their profiles and background knowledge and providing them with customized information retrieval. In [39] a domain ontology based information retrieval in the technique preparation process is proposed. From this survey, these ontology based retrieval systems are not used WordNet synonym sets (synsets). The drawback of this approach is the lack of semantics. In our system implemented NLP techniques for word semantics. The main idea is expanding indices and queries with semantics of the words to achieve better recall and precision. Our approach is shown to improve the retrieval performance. Dong [43] has suggested personalized information retrieval on the semantic web for hotel domain ontology. This paper should not be performed using large database and more queries. The interface for querying the ontological knowledge is one of the key points of information retrieval systems. This mechanism is storing and extracted data in RDF and OWL format and querying with RDF query languages such as RDQL or SPARQL. This approach offers the ultimate precision and recall performance. In our work functioned SPARQL query language for retrieving the information.

Our literature survey revealed that current studies on keyword-based searching are not mature enough: either they are not scalable to large knowledge bases or they cannot capture all the semantics in the queries. Our main contribution is to fill this gap by implementing a keyword- based semantic retrieval system using the semantic web techniques and mathematical model. In other words, we try to implement a system that performs at least as good as traditional approaches and improves the performance and usability of semantic querying. We tested our system in education domain to see the effectiveness of semantic searching over traditional approaches and observed a remarkable increase in recall and precision. In addition we indicated that our approaches can response difficult semantic queries, typing errors, imprecise data, which is not possible with other traditional methods.

III. OUR APPROACH TO SEMANTIC WEB BASED EFFICIENT SEARCH USING ONTOLOGY AND MATHEMATICAL MODEL

This paper is to develop a related reliable and an efficient search engine to retrieve the accurate results for the user's complex query. It even bears the human error in typing, and suggests the expected word to search for. It also aims at retrieving the same result for synonymous words which prevents the appearance of irrelevant search results. It provides the complete details for the query about the education domain with the correct URL and metadata in which to search for, which consumes more time in the syntactic search engine. The details are generated with the help of ontology and relations among classes, entities, individuals, data type properties, object properties, restrictions are also created. Hence now the user can query upon the information stored within the ontology. The querying of the ontology is supported towards the properties, classes, individuals and entities created in the ontology.

A. Overall System Architecture

The overall system architecture is shown in Fig 1. The system consists of ontology knowledge base, WordNet API and search model. Service knowledge base is to store domain specific service ontology's, and service description entity (SDE) metadata. WordNet API for meaningful search and search model is to retrieve SDE metadata for education lenders using query-concept matching model. The whole workflow of the system is as follows.

1) An education lender enters a set of key terms into the search engine interface.
2) Check spelling and find the synonym for each word from WordNet API.
3) Query and synonyms using SPARQL language is designed for that who have domain knowledge with regard to their service queries.
4) Match the query and concept with ontology knowledge base using (SCBR) algorithm is designed to who have not domain knowledge with regard to their service queries.
5) Choose concept is designed to human computer interaction.
6) Refers to Meta data.
7) Retrieve the metadata information from the ontology knowledge base and provide relevant information to education lenders.

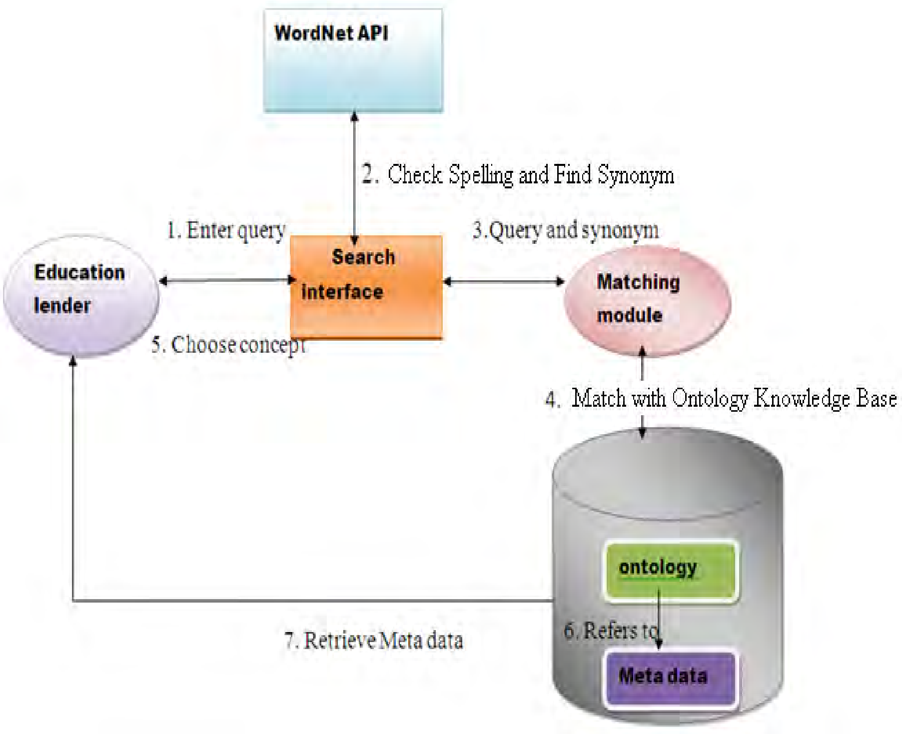

Fig.1 Overall System Architecture

The search interface sends each query term to the spell-check method for supposing the user enter words in incorrect or spelling mistake, the (spell check method) Java program providing suggestions and passes each of the words to Word Net API. If one query term can be retrieved from the API, the API returns its synonyms; otherwise, the query term is filtered. After the process has been completed, the search interface sends the query terms and their synonyms to the query concept matching model. The query-concept matching algorithm is run to compute the similarity values between the service ontology concepts stored in the ontology knowledge base and the query terms then provide relevant information to the user. Once the user selects a result, all its semantically relevant metadata will be retrieved from the ontology knowledge base.

IV. WORKING PROCESS

In this section, we introduce the mechanism of the semantic search model by describing the whole working process as follows:

A. Ontology Knowledge Base
B. Natural Language Processing Techniques
   - Embedding Spell-check
   - Finding Synonyms Using WordNet API
C. Search Model
   - Querying Ontology Using SPARQL
   - Query-Concept Matching Algorithm

A. Ontology Knowledge Base

The ontology knowledge base is created using Protégé OWL. It consists two main components are service ontology and metadata in which semantically related ontological concepts and metadata are linked by referencing their URL to one another and it has a two rules contained in the semantic relationship; the first one is a concept may semantically relate to arbitrary metadata and the second one is a metadata may also semantically relate to arbitrary concepts.

First we describe about the ontology is denoted as conceptualization of the service, in which identified by service name, service description and linked metadata. The service ontology is the incorporation of ontology name and a tuple where the elements of the tuple can be complex elements. The service name can be used to uniquely identify a service, and the service description refers to the definitional descriptions of a service. The normal form of a service description is a set of words like noun, adjective or adverb. A concept may have many service descriptions. The adventure of setting the property of service description is to compute the semantic similarity values between concepts and queries, linked metadata refers to the URL of semantically relates to metadata to a concepts. The ontology is the definition of service concept in the root of the service concept hierarchy. As leaf concepts all other concepts in this hierarchy automatically inherits its properties.

Second we respond metadata, the purpose of metadata is to bring about meaningful information with regarding the real environment. The metadata defined as linked concepts, service name, service address, contact details, and metadata descriptions. The linked concept refers to the URL of semantically related to concepts to the metadata. Service name refers to the name of the college or institution. Service address refers to the address where can be located. Service contact details refer to the information regarding phone number, fax number, website and so on. Service description refers to the detailed text description regarding the content of a service. This can be used for matching with concepts.

This paper made use of education ontology for querying upon the desired event, the required components to build up the ontology such as classes, instances and relationships are being created. The classes created are college, college type, school, university, district, etc., The subclasses created within the college are engineering, arts and science, law college, medical, polytechnic, institute etc., with regard to the metadata, currently no such ontology available, so we collect and develop resources from the (http://en.wikipedia.org/wiki, http://www.india.gov.in, http://www.studyguideindia.com/Colleges,http://education.tamilnadu.com, http://www.worldcolleges.info/) web sites [20]. The properties of the classes are created.

Properties are of two types:

- Data type Property: It is being used to set properties enhancing the existence of an individual.
- Object Property: It is used to create a relationship between two different class individuals.

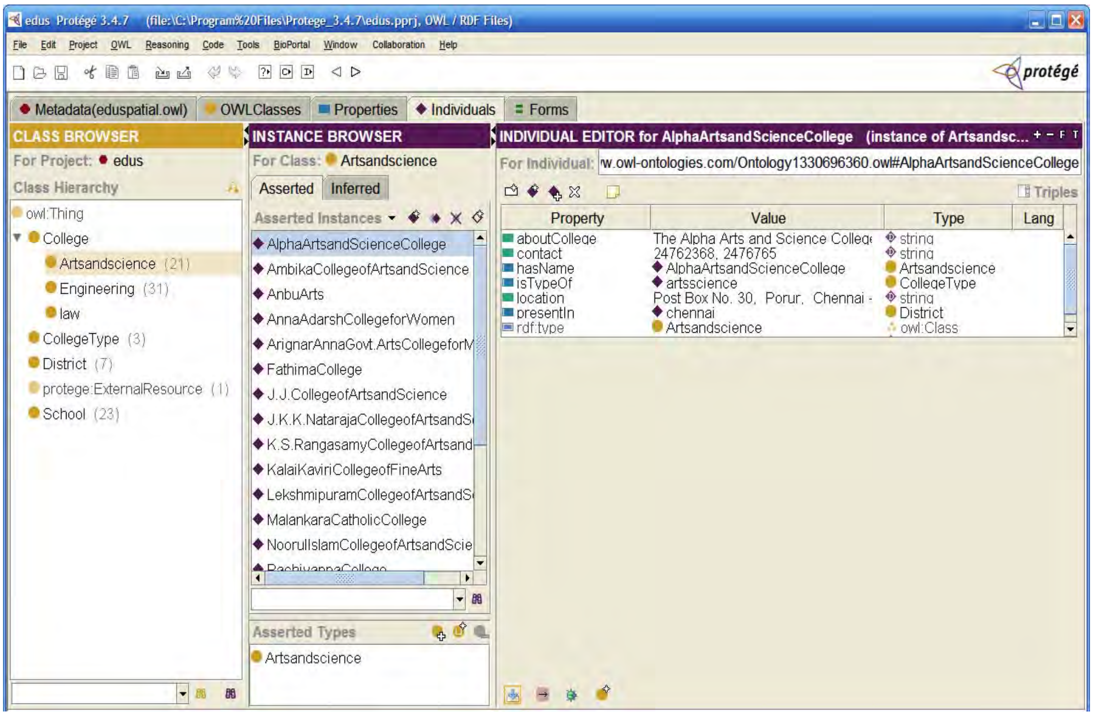

Fig. 2 Education Ontology

The data type properties created is about college, contact, location, URL. The object properties created are type of college, present in the location, the name of institution etc., The individuals are for each OWL classes are created for engineering college, arts and science college etc., In Fig .2 shows the sample ontology screen shot.

B. Natural Language Processing Techniques

The natural language processing techniques classified in to two sub-parts, which is embedding spell-check and finding synonym using WordNet API.

- Embedding Spell Check

```
Input
          A word
Output
          Retrieve the information from the ontology
Procedure:
Begin
          String word, getword syn
          // Get the word from the user
          Getword=spellcheck(word)
          syn=wordnet(getword)

          for i=0 to syn.length
                    Begin
                        if(syn=ontology)
                              Retrive the ontology using SPARQL
                              query with corresponding meta data
                         Else
                              Retrieve the ontology matched with getword
                        End if
          End for
End
```

Algorithm for spell-check

In this section we proposed spell check algorithm for proficient web search. Owing to the user enter words in incorrect or spelling mistake the method providing suggestions for unknown (misspell) words based on custom dictionary and system administrator can create a list of preferred words and assign higher weight to the list as a axiomatic mediation BasicSuggester can help as a spell checker. In this part spell-check is designed to it gets the query, and passes each of the words the spell-check method where the word is retrieved with the correct spelling and the correct word is being passed to get the synonym and the word with its synonyms is being passed to the similarity (Query-Concept) matching model to check for the presence of the word or its synonyms in the ontology and if present it is being retrieved along with the metadata.

- Finding Synonym Using WordNet API

The WordNet API employed for finding relevant meaningful search. In WordNet, the words and their relationships to each other are organized in a hierarchical manner similar to the taxonomies which may be found in the natural sciences. Words which are closely related to each other may be found in the same branch of the hierarchy's tree. Each word belongs to a set of synonyms, also known as a synset. These synsets are the foundation upon which the WordNet database is constructed. Formally, a synset is a set of one or more synonymous words that may be substituted for each other in context without changing the overall meaning of the sentence in which they are contained. Words which have multiple meanings or “word senses” appear in more than one synset. WordNet provides a polysemy count for each word which is used to track the number of synsets which contain the word.

Since different word types follow different grammatical rules, WordNet makes the distinction between four of the primary word types in the English language, which include nouns, verbs, adjectives, and adverbs. The noun category contains words which refer to entities, qualities, states, actions, or concepts, and can serve as the subject of a verb. Words classified as verbs may serve as the predicate of a sentence and describe an action, occurrence, or state of existence. Adjectives are words that may modify nouns. The final word classification stored in WordNet, the adverb, is similar to the adjective and contains words which modify word types other than nouns.

  - Importing WordNet

The JAWS runtime library, the API may be used by adding its Java Archive (.jar) file to the class path of the application which will be used WordNet.

  - Instantiating a WordNet Database

The WordNet Database class provides access to the information stored in the WordNet database and must be instantiated before use. A method, getFileInstance, returns an implementation of the class that works with the local WordNet database and may be used when creating a new instance of the WordNet Database class. Other than WordNet Database, another critical component of the JAWS API is the Synset interface. This interface represents WordNet's collections of related words, or synsets. These synsets are stored as an array of word forms. Several overloaded methods of the WordNet Database class known collectively as getSynsets can be used to retrieve synsets from the WordNet database by providing a starting word in the form of a string when the getSynsets is called. When instantiating a synset, the getSynsets method is used to populate the new instance of the Synset interface with WordNet information.

  - Retrieving Synonyms

The getWordForms method may be used to retrieve the individual groups of word forms for each synset stored as an element of this array, which may themselves be stored as arrays of strings containing all words similar to the original word.

C. Search Model

The search model classified into two sub-models which is querying ontology using SPARQL and query-concept matching algorithm using SCBR mathematical model. The SPARQL based matching for who have domain knowledge regarding their queries, the retrieval model allow to requesters to retrieve quickly and directly search metadata form the ontology knowledge base. The query-concept matching algorithm for the requesters who do not have domain knowledge regarding their queries, the retrieval model is used.

- Querying Ontology Using SPARQL

In this portion, SPARQL query is being used to retrieve relevant information from ontology [8]. SPARQL (Simple Protocol And RDF Query Language) is the same as in SQL and it is applicable to the users who already have some information about a service metadata attributes. Users can retrieve a metadata by querying any of it is attribute values, which is convenient and time saving search style.

- Query-Concept Matching Algorithm

This paper proposes SCBR, Semantic CBR (Cased Based Reasoning) Algorithm for query concept matching model, which is an enhanced version of Extended Case Based Reasoning (ECBR) algorithm [11]. It is expected that the SCBR algorithm is giving efficient search results than ECBR. The principle of the SCBR model is to seek the maximum similarity value between a query and their ontology knowledge base. If a query key term is contained in it, a value 1 will be awarded; if a meaning of a query key term is in it, a value 0.5 will be awarded; otherwise 0 will be awarded. Here we set optimal threshold value 0 to 1 with an increment of 0.1 at each time. A threshold value needs to be filter irrelevant data. We then obtain the performance concepts for each time of the variation of the threshold value. The query is then compared with metadata description property of each concept from the ontology. The maximum value between the query and any data description properties of a concept is considered as the similarity value between the query and the concepts. In addition, the paper adds spell-check method in this work in (section IV part-B) for improving the education search performance. The ECBR model not proposed spell-check method [11] [45]. The spell-check is designed such that it gets the query, and passes each of the words the spell-check method where the word is retrieved with the correct spelling and the correct word is being passed to get the synonym and the word with its synonyms is being passed to the query-concept matching algorithm to check for the presence of the word or its synonyms in the ontology and if present it is being retrieved along with the metadata. The SCBR model can be mathematically shown as:

$$sim(q,d) = \max_{md_i \ni d}\left(\sum_{k_{ih} \in md_i} \frac{f(q,k_{ih}) + f_m(m,k_{ih})}{\sum md_i} \right) \quad (1)$$

with,

$$f(q,k_{ih}) = \begin{cases} 1 \quad if \quad \exists q_{kt} \mid \left(\forall s_t, w_t(k_{ih}) = wt(q_{kt}) \wedge (q_{kt} \in q)\right) \\ \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (2)$$

$$f_m(m,k_{ih}) = \begin{cases} 0.5 \quad if \exists q_{kt} \mid \left(\forall s_t, w_t(k_{ih}) = wt(q_{kt}) \wedge (q_{kt} \in m)\right) \\ \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (3)$$

Where, q is a processed query, d is a result data, here d denoted as a concept c, md_i is a meta data descriptions property of data d, k_{ih} is a key involved in md_i, $\sum md_i$ is the sum of associated with md_i, q_{kt} is the query key term involved in md_i, s_t is the semantic term, w_t is a function that returns a weight associated with s_t, m is a meaning of query, sc_{ih} is will check the spelling of the queries provide by the user.

V. SYSTEM EVALUATIONS

The system evaluation is divided into two parts; A) System implementations and functional testing to validate the whole system. B) Evaluating the employed mathematical model by the means of simulation approach to test their performance with artificial data [44] [45].

A. *Prototype Implementation and Functional Testing*

In this section, we implement educational semantic search engine. We have run it with different input queries relevant to educational domain in Fig (3) displays sample screen shot with input as engineering retrieves the list of engineering college and when selected a specific college from the list gives its corresponding location, URL and contact followed by the mission of the institution.

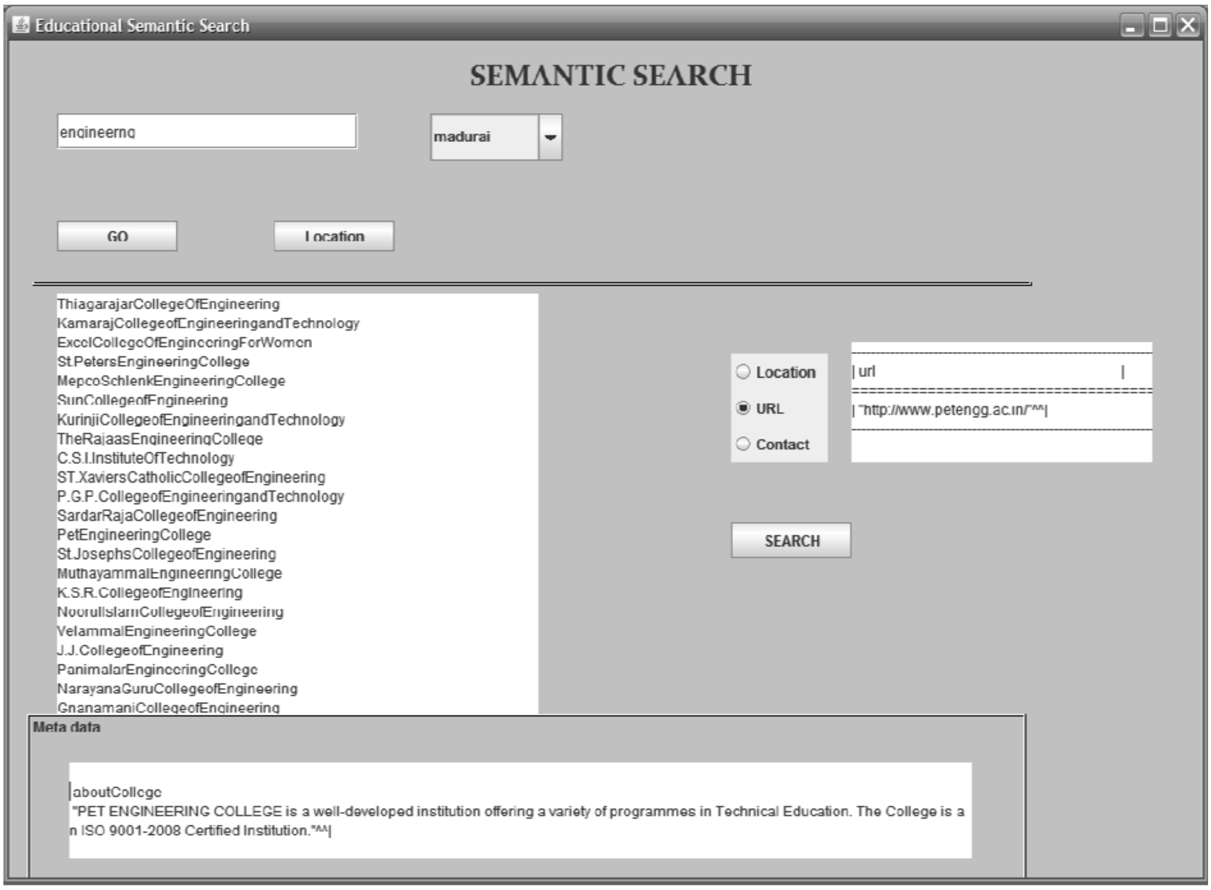

Fig.3 Sample Screenshot of the Educational system

When the required district to which the query term relevant to which the query term, it displays results corresponding to the district. Though the user enters the words with the wrong spelling, it retrieves the output with the correct word from ontology. The result is being retrieved the same for synonymous words.

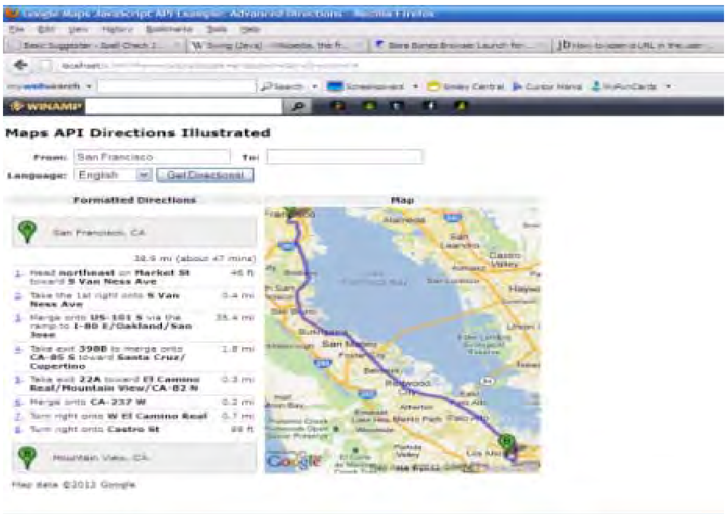

Fig.4 Output in Map

In Fig (4) the location button for when clicked with the required location of the college, outputs its route along with directions to reach it in a map and so on. The proposed work is being implemented with the available tools and environment as follows, Eclipse3.6.0: is an Integrated Development Environment for developing applications in Java, and it is used as the platform to develop educational search. Java Development Kit: JDK1.7 (Java software Development Kit) is used for the implementations of this framework. Protégé 3.4.7: is ontology editor, which is used as the platform for developing education domain ontology in the knowledge base. SPARQL (Simple Protocol And RDF Query Language): SPARQL is the same as in SQL and used to access more reliable and accurate results and querying the ontology. Protégé-OWL API: Protégé-OWL is an open source Java library for the web ontology language and RDF(S). Here the Protégé-OWL API is used to load OWL coded ontology. Jena: Jena is a Java framework for building semantic web applications and here is used to load RDF(S), SPARQL coded ontology. Java script, AJAX: is used for display the Google maps. Spell checker, BasicSuggester for spell checking the given input data and WordNet for finding relevant data.

B. Evaluating the Mathematical Model for the Semantic Search

As described previously, there are four models designed for semantic search, which are SCBR, ECBR, VSM, and LSI model in this work we concentrate on evaluating the performance of these models as follows.

- Extended Cased Based Reasoning Model (ECBR)

The ontology concept-metadata matching model is built upon an Extended Case-Based Reasoning (ECBR) model, which is an index-term-based set-theoretic model [11]. In the educational knowledge base, first of all, each educational ontology concept c is regarded as a body of plain texts comprised of concept description property(s).The ECBR model is used to calculate the similarity value of a concept c_j to a metadata m, which is represented as

$$sim\left(cd_j, sd\right) = \frac{\sum_{j=1}^{m} f\left(cd_{k,} sd\right)}{m} \quad (4)$$

$$f\left(cd_{k,} sd\right) = \begin{cases} 1 \quad if \quad \exists sd_k \mid \left(\forall k_i, g_i\left(cd_{kj}\right) = g_i\left(sd_k\right)\right) \\ \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (5)$$

where cd_j is the content of the definition of a concept, k_i is an index term; $cd_j = (cd_{k1}, cd_{k2} \ldots cd_{km})$, where cd_k is the index terms that occurs with cd_j, m is the number of index terms that occur with cd_j; sd is the content of the service description property regarding a service metadata; $sd = (sd_{k1}, sd_{k2} \ldots sd_{kn})$, where sd_k is the index terms that occur with sd, n is the number of index terms that occur with sd; g_i is a function that returns a weight associated with k_i.

- Vector Space Model (VSM)

The VSM is an algebraic model for representing text documents as vectors, and is a classical model for information retrieval [31]. The implementation details of the VSM in our query-concept matching process are introduced as follows:

In the educational knowledge base, first of all, each educational ontology concept c is regarded as a body of plain texts comprised of concept description property(s). Then, the term list is obtained from all the concepts in the educational ontology. Based on this list, each concept c is formed as a vector in which each element corresponds to term in the term list, and the weight of each element is computed by tf-idf. Similarly, a query q can also be seen as a concept corresponding to a vector. Thus, the relevance between a concept c_j and a query q can be calculated as the cosine of the angle between the two vectors.

$$sim\left(c_{j,} q\right) = \frac{\left|\overrightarrow{c_j} \cap \vec{q}\right|}{\left|\overrightarrow{c_j}\right| \times \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (6)$$

Where $\vec{c_j}$ and $\vec{q}$ are two vectors corresponding to c_j and$q$, respectively, $\lceil \vec{c_j} \rceil$and $\lceil \vec{q} \rceil$ are the norms of $\vec{c_j}$ and $\vec{q}$, t is the number of terms in the terms list of the education ontology, w_{ij} and w_{iq} are weights of each element of $\vec{c_j}$ and $\vec{q}$ corresponding to each term, since $w_{ij} \geq 0$, $w_{iq} \geq 0$, $sim\left(c_{j,} q\right) \rightarrow [0,1]$.

- Latent Semantic Indexing (LSI) model

The main idea of the LSI model is to map each concept and query vector into a lower dimensional space associated with concepts, which is used to information retrieval [13, 18] and the implementation details of the LSI in our query concept matching process as follows, In the Educational Service Knowledge Base, first of all, each educational service ontology concept c is regarded as a body of plain texts comprised of concept description property(s). Following that, an index term list is obtained from all the concepts in the educational service ontology. Based on the index term list, each concept c is formed as an array in which each element is obtained by tf-idf, and all the concepts in the ontology are formed as a term-concept matrix A. The term-concept matrix is then decomposed by the SVD approach, which can be mathematically represented by Equation (7)

$$A = U \sum V^{T} \qquad (7)$$

where U is the matrix derived from the term-to-term matrix given by AA^T, V^T is the matrix derived from the transpose of the concept to concept matrix given by A^TA, and $\sum$ is a r × r diagonal matrix of singular values where r = min(t, N) is the rank of A. Considering that now only k largest singular values of $\sum$ are kept along with their corresponding columns in U and V^T, the resultant A_k matrix is the matrix of rank k which is closest to the original matrix A in the least square sense. This matrix is given by Equation (8)

$$A_k = U_k \sum\nolimits_k V_k^T \qquad (8)$$

Where k (k < r) is the dimensionality of a reduced concept space.

Analogous to the concept, a query q can be formed as an index term-based array in which each element is the tf-idf weight between the query and a term from the index term list. The array can then be translated into the concept space by Eq. (9), and then compared with A_k by the cosine algorithm to calculate the similarity values of each concept, which can be represented by Eq. (10)

$$q' = \sum\nolimits_k^{-1} U_k^T q \qquad (9)$$

$$sim(c, q') = \frac{|A_k \cap q'|}{|A_k| \times |q'|} \qquad (10)$$

- Performance Indicators

In this paper define four performance indicators used in this experiment as follows,

$$Precision\ P = \frac{number\ of\ retrieved\ relevant\ data}{number\ of\ retrieved\ data} \qquad (11)$$

Precision is used to measure the preciseness of a search system [19]. In this experiment, Precision P is defined as the number of retrieved relevant data among the retrieved data.

$$Average\ precisions(Q) = \frac{sum\ (precisions\ @\ retrieved\ relevant\ data)}{number\ of\ retrieved\ relevant\ data} \quad (12)$$

Before we introduce the definition of mean average precision, the data of average precision should be defined. Average precision is the average of precision values at each retrieval relevant data for a query, given that these data are ranked according to their computed similarity values. This indicator is used to measure how quickly and precisely a search engine work [19].

$$Mean\ average\ precisions = \frac{\sum_{i=1}^{n} Average\ precision(Q_i)}{n} \quad (13)$$

Mean average precision refers to the average of average precision values for a set of queries and can represented as above.

$$Recall\ R = \frac{number\ of\ retrieved\ relevant\ data}{number\ of\ relevant\ data} \qquad (14)$$

Recall is used to measure the effectiveness of a search system [19]. In this experiment, Recall R is defined as the number of retrieved relevant data to total number of relevant data in the knowledge base.

$$F - Measure = \frac{2PR}{(P + R)} \qquad (15)$$

F-Measure combines precision and recall, in this paper is used as an aggregated performance scale for searchers and users can specify the preferred on recall or precision by configuring different weights. When the F-Measure value reaches the highest, it means the integrated value between precision and recall reaches to the highest at the same time [19].

- System Evaluation Results:

To evaluate, the performance of the SCBR model from the perspective of information retrieval with three models. The mechanism and algorithm concerning the model referred from [19]. Different queries are made to compare the performance of the system. All the parameter results are averaged by 100. These queries cover most of the general user requirements in the educational domain. A threshold values need to be configured to select the most similar concepts by filtering the concepts with the lower similarity values.

In addition, there are two major tasks involved in the experiment as follows: The first task is to find an optimal threshold for each IR model. The reason for this is that, in the search process, after the similarity values between a query and concepts are computed, a threshold needs to be determined for filtering the relatively dissimilar concepts to obtain the optimal performance for each model. Owing to the difference between each model, the optimal threshold could be different. To choose the optimal threshold, we utilize the F-Measure as the primary scale. The threshold scope is configured between 0 to 1 with an increment 0.1 at each time. The second task is to evaluate with four information retrieval algorithms and to choose the optimal thresholds with the overall performance of the search process, based on the same set of queries.

TABLE.I
Testing Results of Our Approach Semantic Case Based Reasoning (SCBR) Algorithm

Optimal Threshold Value	**Precision %**	**Mean average Precision %**	**Recall %**	**F-Measure %**
>0	32.98	82.98	80.50	46.79
>0.1	43.98	83.98	78.17	56.29
>0.2	45.21	85.21	77.17	57.01
>0.3	58.22	88.22	76.41	66.08
>0.4	61.18	91.18	75.34	67.52
>0.5	72.39	92.39	63.35	67.56
>0.6	84.34	95.35	52.29	64.55
>0.7	96.26	97.26	49.37	65.26
>0.8	97.37	98.37	48.53	64.77
>0.9	98.52	99.53	47.98	64.53

Table 1 presents testing results of SCBR model. It is observed that along with the increase of the threshold value the precision results a sharp rise. Mean average precision tests for the quickness and precisions of a search results. The mean average precision values from 82.98% to 99.53%, in contrast recall ranges from 80.50% to 47.98%. F-measure value ranges peek is 67.56% at the threshold value in 0.5. From the results the SCBR model is gain high level of mean average precision value. It is merits of this approach.

TABLE.II
Testing Results of Extended Case Based Reasoning (ECBR) Algorithm

Optimal Threshold Value	Precision %	Mean average Precision %	Recall %	F-Measure %
>0	12.39	70.98	75.26	21.28
>0.1	12.37	70.97	75.25	21.26
>0.2	17.21	71.22	74.60	27.98
>0.3	24.74	71.83	73.80	37.08
>0.4	28.02	78.18	65.81	39.30
>0.5	27.96	78.40	65.55	39.20
>0.6	66.46	90.65	41.44	51.05
>0.7	74.41	91.15	37.46	49.82
>0.8	79.45	90.66	36.64	50.14
>0.9	79.47	90.69	36.65	50.15

Table II presents testing results of ECBR model. It is observed that the precision, mean average precision, recall and f-measure values are less than SCBR algorithm.

TABLE.III
Testing Results of Vector Space Model (VSM) Algorithm

Optimal Threshold Value	Precision %	Mean average Precision %	Recall %	F-Measure %
>0	21.44	69.38	67.30	32.50
>0.1	23.27	70.66	64.75	34.23
>0.2	28.94	73.94	58.30	38.68
>0.3	40.20	76.77	50.01	42.72
>0.4	51.46	81.84	41.99	46.25
>0.5	68.55	87.00	33.05	44.59
>0.6	79.75	87.95	26.48	39.75
>0.7	76.95	79.24	13.39	22.82
>0.8	87.93	87.83	8.20	14.99
>0.9	87.93	87.83	8.20	14.99

Table III shows the testing results of VSM model. VSM precision and mean average precision basically experiences a consistent rise, and the only exceptions occurs when the threshold is 0.7 and the recall experiences fall almost linearly dropping from 67.30% to 8.20%. The highest f-measure are obtained at the threshold is 0.4.

TABLE.IV
Testing Results of Latent Semantic Indexing (LSI) Algorithm

Optimal Threshold Value	**Precision**	**Mean average Precision %**	**Recall %**	**F-Measure %**
>0	4.10	59.35	81.74	7.80
>0.1	20.23	65.94	70.10	31.40
>0.2	27.18	69.90	63.31	38.03
>0.3	32.13	74.05	56.61	41.00
>0.4	37.01	77.57	50.50	42.72
>0.5	42.39	82.10	44.01	43.19
>0.6	51.60	83.26	38.21	43.92
>0.7	59.98	85.11	31.55	41.54
>0.8	76.90	87.23	23.97	36.55
>0.9	76.89	87.22	23.95	36.54

Table IV gives the testing results of LSI model, the precision ranges from 4.10% to 76.89%, mean average precision ranges from 59.35% to 87.22%, recall ranges from 81.74% to 23.95% and F-Measure ranges from 7.80% to 36.54%. From the testing results it can be concluded that, the SCBR model have the highest scores on the performance. In addition the WordNet API and spell-check method enhances the performance of SCBR and our approach (SCBR) is providing more efficient search results compare than other three models.

VI. CONCLUSION

We have implemented a reliable and an efficient system, which suggests the user all the effective details to know about an educational domain. It is reliable because though it is being inputted with synonymous words and misspell, it retrieves the similar result and does not provide an irrelevant result. All the details can be retrieved in a single page, so it saves the user's time and inconvenience to move on to more pages to search for the right result. In addition, we designed a more efficient SCBR algorithm, an enhanced version of the ECBR algorithm, in structure to compute the similarity values between query and concepts. We compare the performance of the SCBR model with three information retrieval models. To address the defect of low recall rate that done in ECBR model [11]. We will modify ECBR algorithm to SCBR algorithm to obtain better performance. The system can be further refined with more words in the search interface which can yield more filtration of the query result. The system can be better used with more performance indicators which can better model user requirements.

REFERENCES

[1]. H. L. Wang, S. H. Wul, and K. K. Wang, "Semantic Search on Internet Tabular Information Extraction for Answering Queries", Proceedings of the ninth international conference on Information and knowledge management , ACM 2000, PP: 243 – 249.

[2]. Wei-Po Lee and Tsung-Che Tsai, "An Interactive Agent –Based System for Concept-Based Web Search" Expert Systems with Applications Volume 24, Issue 4, May 2003, PP: 365–373.

[3]. Deepavali Bhagwat and Neoklis Polyzotis,"Searching a File System using Inferred Semantic Links" Proceedings of the sixteenth ACM conference on Hypertext and hypermedia,2005, PP: 85-87.

[4]. Yasufumi Takama And Shunichi Hattori, "Mining Association Rules For Adaptive Search Engine Based On RDF Technology" IEEE Transactions on Industrial Electronics, VOL. 54, NO. 2, APRIL 2007,PP:790-796.

[5]. Soner kara, Ozgur Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Ciceki and Ferda N. Alpaslan, "An Ontology Based Retrieval System Using Semantic ndexing" Information systems Volume 37 Issue 4, june 2012, PP:-294-305.

[6]. Ioan Toma, Dumitru Roman, Dieter Fensel, Brahmanada Sapkota, and Juan Miguel Gomez "A Multi-criteria Service Ranking Approach Based on Non-Functional Properties Rules Evaluation", springer, Volume 4749,2007, PP: 435-441.

[7]. Shadbolt, N., Hall, W., Berners-Lee, T., "The Semantic Web Revisited", IEEE Intelligent Systems, Volume 21 Issue 3, May 2006 Page 96-101.

[8]. Dean Allemang, Jim Hendler "Querying the Semantic Web—SPARQL, Semantic Web for the Working Ontologist" (Second Edition), PP: 61-112, 2011.

[9]. Leva and F. Donida, "Multifunctional remote laboratory for education in automatic control: The CrAutoLab experience" IEEE Trans. Ind. Electron, Vol. 55, No. 6, 2008.

[10]. Alexandros Batzios, Pericles A. Mitkas., "WebOWL: A Semantic Web search engine development experiment" Expert Systems with Applications, Volume 39, Issue 5.PP: 5052–5060, 2012.

[11]. Dong, Hai and Hussain, Farookh Khadeer and Chang, Elizabeth," A Human-Centered Semantic Service Platform for the Digital Ecosystems Environment", World Wide Web. Volume 13 (1-2): pp. 75-103, 2010.

[12]. Ming-Yen Chen, Hui-Chuan Chu, Yuh-Min Chen,"Developing a semantic-enable information retrieval mechanism", Journal of Expert Systems with Applications, Volume 37,Issue 1, PP: 322–340, 2010.

[13]. Lim Bee Huang, Vimala Balakrishnan, Ram Gopal Raj," Improving the relevancy of document search using the multi-term adjacency keyword-order model",Malaysian Journal of Computer Science. Vol. 25(1), 2012,PP:1-10.

[14]. Q. Guo, M. Zhang, "Question answering based on pervasive agent ontology and Semantic Web", Knowledge-based Systems Volume 22 Issue (6) PP:443–448. 2009.

[15]. Ning Gao, Zhi-Hong Deng, Jia-Jian Jiang, Sheng-Long Lv, Hang Yu,"Combining Strategies for XML Retrieval", INEX'10 Proceedings of the 9th international conference on Initiative for the evaluation of XML retrieval: comparative evaluation of focused retrieval, Springer-Verlag Berlin, Heidelberg pp: 319–331, 2011.

[16]. G. Yuanbo, P. Zhengxiang, H. Jeff, LUBM: "A benchmark for OWL knowledge base systems", Journal of Web Semantics, Volume 3 (2–3) PP:158–182,2005.

[17]. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, Y. Katz, Pellet:, "A practical OWL-DL reasoned", Journal of Web Semantics, Volume 5 (2) PP: 51–53, 2007.

[18]. S.C. Deerwester, S.T. Dumais, G.W. Furnas, R.A. Harshman, T.K. Landauer, K.E. Lochbaum, L.A. Streeter, "Computer information retrieval using latent semantic structure", US Patent, Bell Communications Research, Inc., USA vol. 4839853, 1988.

[19]. R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", New York: Addison-Wesley, 1999.

[20]. K.Palaniammal, M.Indra devi, S.Vijayalakshmi,"An Unfangled approach to semantic search for e-tourism domain" ICRTIT-12 ,IEEE Explorer, PP: 130-135,2012.

[21]. G. Salton, A. Wong, C.S. Yang, "A vector space model for automatic indexing", Communications of the ACM Vol 18 (1975),PP: 613–620.

[22]. K.S. Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation vol 28 (1972) PP:11–21.

[23]. G. Salton, M.J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, 1983.

[24]. G. Salton, E.A. Fox, H. Wu, "Extended boolean information retrieval", Communications of the ACM Vol 26 (1983) PP: 1022–1036.

[25]. G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing and Management: An International Journal, Vol 24 (1988) PP:513–523.

[26]. M. Banko, O. Etzioni, "The tradeoffs between open and traditional relation extraction", in Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 28–36.

[27]. R.M. Casado, E. Alfonseca, P. Castells, "Automatic extraction of semantic relationships for WordNet by means of pattern learning from Wikipedia", In Tenth Int'Conference on Applications of Natural Language to Information Systems, NLDB 2005, Springer, 2005, pp.67–79.

[28]. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld, A. Yates, "Web-scale information extraction in knowitall: (preliminary results)", in WWW '04: Proceedings of the 13th International Conference on World Wide Web, ACM, New York, NY, USA, 2004, pp. 100–110.

[29]. M. Surdeanu, M. Ciaramita, "Robust information extraction with perceptrons", In proceedings of the NIST 2007 Automatic Content Extraction Workshop ACE2007.

[30]. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoffe," Semantic annotation, indexing, and retrieval", Journal of Web Semantics, Volume 2, Issue 1 2005, Pages 49–79.

[31]. Pablo Castells, Miriam Fernandez, and David Vallet " An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval " IEEE Transactions on Knowledge and Data Engineering, Volume 19, Issue 2, February 2007 PP 261-272.

[32]. R.J. Mooney, R. Bunescu, "Mining knowledge from text using information extraction", SIGKDD Explorations Newsletter, Vol 7 ,(2005) PP:3–10.

[33]. Schutz, P. Buitelaar, Relext: "A tool for relation extraction from text in ontology extension", in: The Semantic Web (ISWC 2005), 2005, PP. 593–606.

[34]. F. Ciravegna, S. Chapman, A. Dingli, Y. Wilks, "Learning to harvest information for the semantic web", in: European Semantic Web Symposium/Conference, vol. 3053, 2004, PP. 312–326.

[35]. Muslea, S. Minton, C. A. Knoblock, "Active learning with strong and weak views: a case study on wrapper induction", in: International Joint Conference on Artificial Intelligence, 2003, PP. 415–420.

[36]. Freitag, N. Kushmerick, "Boosted wrapper induction", in: Seventeenth National Conference on Artificial Intelligence, 2000, pp. 577–583.

[37]. P. Cimiano, S. Handschuh, S. Staab, "Towards the self-annotating web", in Thirteenth World Wide Web Conference Series, 2004, pp. 462–471.

[38]. Myungjin Lee et al., "Searching and ranking method of relevant resources by user intention on the semantic web" Expert Systems with applications, Vol 39 (2012) pp: 4111-4121.

[39]. Liu Xinhua,Zhang Xutang, Li zhongkai," A Domain Ontology-based Information Retrieval Approach for Technique Preparation" Physics Procedia, Vol 25, 2012, PP:1582 – 1588.

[40]. Karina Robles, Anabel Fraga, Jorge Morato, Juan Llorens, "Towards an ontology-based retrieval of UML Class Diagrams", Information and Software Technology 54 (2012) pp: 72–86.

[41]. Lei Shi, Rossitza Setchi," Ontology-based personalised retrieval in support of reminiscence", Knowledge-Based Systems, Volume 45, 2013, Pages 47–61, 2013.

[42]. Hong Zhou, Bing-wu Liu, Jun Liu, "Research on Mechanism of the Information Retrieval Based on Ontology Label", Procedia Engineering, Vol 29, (2012), pp: 4259-4266.

[43]. Donghee yoo, “ Hybrid query processing for personalized information retrieval on the semantic web”, Knowledge-Based systems, Vol 27(2012) PP: 211-218.
[44]. A.R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Raml, “Design science in information systems research”, MIS Quarterly Vol 28 (1) (2004)PP: 75–106.
[45]. Dong, Hai, and Farookh Khadeer Hussain. "Semantic service matchmaking for digital health ecosystems." *Knowledge-Based Systems* 24, no. 6 (2011): 761-774.