# Ontology-Based Web Query Classification

Bharat Katariya

Computer Engineering

Guided By

Prof. Mitula Pandya [M.E, C.E]

Alpha college of Engineering & Technology,

Khatraj, Gandhinagar – 382721

**Abstract-** Now a day's use of internet becomes very popular. Everyone gets their need by searching on internet. User can access information easily by web searching tools like a search engine. (Google) According to one survey nearly 70% of the people use a Search Engine to access the information available on Web. If Query submitted by the user to the search engine is too short and ambiguous then it can create problem to getting right documents. Query classification is one technique of Data Mining in query should classify to the number of predefined categories. Query classification use ontology as a model to retrieve the document. Ontology is used in information retrieval system to retrieve more relevant information from a collection of unstructured information source. The main objective of this work is to use ontology concepts as categories. The concepts of ontology will be used as training set. Each document is ranking by the probability. The semantic relations will solve the problem of ambiguity. Web query classification is used to ranking the page(Search engine optimization).

**Key Words-** Classification, Ontology, Web query, categories of query, Irrelevant link, Duplicate link

## I. Introduction

Web query classification is a part of web mining. Web mining include data cleaning, data integration from multiple sources, warehousing the data, data cube construction, data selection, data mining, presentation of mining results, pattern and knowledge to be used to stored into knowledge base. Data mining can mined various types of data like relational database, data warehouse, transactional database, object-relational databases, heterogeneous databases, multimedia database, text databases, World-Wide Web etc. There are mainly five function of data mining. (1.Generalization, 2.Association and correlation analysis, 3.Classification, 4.Cluster analysis, 5.Outlier analysis) [1] Classification is one of the major function of data mining.

There is a large set of data and vast data available on the internet. The increasing of information on the internet (World Wide Web) has made the field of Information Retrieval (IR) more critical than ever. The main problem of information retrieval systems is to handle large amount of data and satisfy users by giving them with relevant information by their needs. The first step of good Information retrieval system is query understanding. Existing search engines mainly focus on basic term-based techniques for general search, and do not attempt query understanding (Keyword based search engine). Query understanding is done by semantic rules, extracting domain terms, and user's basic information like place, time, age, occupation etc understanding what are the requirements of user and providing that requirements is most important for developing successful Web search engine. Therefore, a search engine (Like Google) that can successfully map incoming search queries from the user to specific content can improve both the efficiency and the effectiveness of web searching. Search engines become one of the most popular tools for web users to find their desired information. If user searches information, he has an idea of what he wants but user usually cannot formalize the query. As a result, understanding the nature of information need behind the queries issued by Web users have become an important research problem.

Query classification is a two step process. First is learning step where a classification model is constructed. Second one is Classification step where the model is used to predict class label for given data. Classifying web queries into predefined target categories, also known as web query classification. There are several major difficulties which are needed to consider in query classification. Most of queries are short and query terms are noisy. A second difficulty of web query classification is that a user query often has multiple meanings. Web query classification aims to classify user input queries into a set of target categories. Query Classification has many applications including page ranking in Web search (Search Engine Optimization (SEO)) targeted advertisement in response to queries, and personalization. In web query classification first input query is extracted in domain terms. All are domain terms find relevant document from collected document on internet by their category. After that all are retrieve documents connected with appropriate documents and give to the user's screen.

The rest of the paper has been organized as follow. Section II presents the some related work in query classification. Section III describes the proposed system. Section IV describes Experiments and results. Conclusion is explained in Section V. References are listed in Section VI.

## II Related Work

Classification of web query to the user intendent query is major task for any information retrieval system. MyoMyo ThanNaing[9] proposed Query Classification Algorithm. To classify the web query inputted by the user into the user intended categories, MyoMyo ThanNaing use the domain ontology. Ontology is useful to matching of retrieve category to target category. User query are extracted in Domain terms are used as input to the query classification algorithm. Matched terms of each domain term are extracted in further sub category. Compute the probability for matched categories. Then all documents are ranked by their probability and displays to the user's desk.

Ernesto William De Luca and Andreas Nürnberger [2] proposed method of web query classification using sense folder. In this method the user query is separated in small terms. These small terms are matched with target categories using ontology. Ontology is set of rules. Word vectors (prototypes) are used to create semantic category. Then Search results are indexed by using sense folder. At last retrieved documents are displays to the user desk.

Suha S. Oleiwi, Azman Yasin [3] proposed method of web query classification using Ontology and classification. All are retrieve documents are indexed according to their probability. Probability depends on how often the documents are search on web by user.

Web query classification method to classifying user queries into a given target category. Lovelyn [4] proposed Web Query Classification method based on normalized web distance. In this system, intermediate categories are mapped to the required target categories by using direct mapping and Normalized Web Distance (NWD). The categories are then ranked based on three parameters of the intermediate categories namely, position, frequency and a combination of frequency and position. In the system Taxonomy-Bridging Algorithm is used to map target category. The Open Directory Project (ODP) is used to build an ODP-based classifier. This taxonomy is then mapped to the target categories using Taxonomy-Bridging Algorithm. Thus, the post-retrieval query document is first classified into the ODP taxonomy, and the classifications are then mapped into the target categories for web query.

Another study proposed an algorithm named Query-Query Semantic Based Similarity Algorithm (QQSSA). This algorithm works on a new approach it filters and breaks the long Query into small words and filters all possible preposition, conjunction, article, special characters and other sentence delimiters from the query. And then expand the query into logically similar word to form the collection of similar words. Construct the Hyponym Tree for query1 and query2 etc. And based upon some distance measure he classifies the query.

Another approach is Classification methodology by S. loelyn Rose, K R Chandran and M Nithya [5]. The classification methodology can be fragmented into the following phases. Feature Extraction, and Mapping intermediate categories to target categories The features extracted in the first phase are mapped onto various target categories in this second phase by Direct Mapping, Glossary Mapping, Wordnet Mapping, Semantic Similarity Measure.

## III. Proposed method

Ontology based web query classification can help to solve the problem of query classification. All retrieve results are display by ontology model using probability of each document. Ontology is a set of concepts, semantic rules and few most popular user queries. Web query input by the user is extracted in domains means that query is separated in small parts. Then all are domain terms search in collection of documents to retrieve appropriate results. All are results are combined into set of documents. After that all are documents are sorting by their probability and display to the user's desk.

Fig 1 shows implementation strategy of our work. We use two different algorithms. First one is irrelevant link removal algorithm and second one is duplicate link removal algorithm.

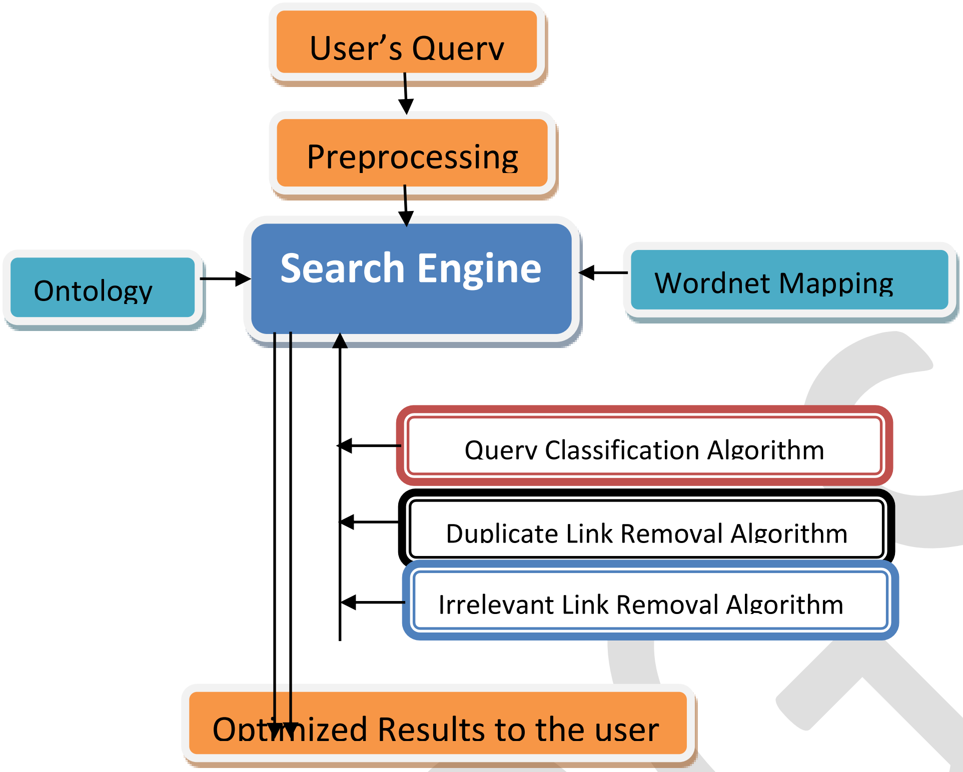

User's Query
Preprocessing
Ontology
Search Engine
Wordnet Mapping
Query Classification Algorithm
Duplicate Link Removal Algorithm
Irrelevant Link Removal Algorithm
Optimized Results to the user

Fig. 1 Implementation Strategy

Here Search engine work offline. Search engine is buildup using PHP and JAVASCRIPT technology. To use this search engine first we have to do Indexing of documents. Indexing can done using collect all offline webpages. First we have to select our related offline webpages and then we can index it. After that we can put the query into appropriate field as shown in snap 2.

**Irrelevant Links Removal Algorithm :**

```
Input query = IQ
Resultant query = RQ

        if (inputted query length > 2)
            then
            {
            perform matchingIQ  with  RQ;
        //* Calculating that how many words from input query are match with
    resultant query *//

                if (IQ = RQ)
                {
                Calculate link into the output result to the user
                }
            else
                if (IQ match with RQ, as two or more than two words)
                {
                Calculate link into the output result to the user
    //* Mined as user relevant result *//
                }
            else
                if (IQ match with RQ, as only one words)
```

```
                                        {
                                        Remove link from the output result to the user
                              //* Mined as user irrelevant result *//
                                        }
          else
          {Remove Link}
                              }
```

Irrelevant link removal algorithm help to remove unnecessary results from search result list. It can work efficiently if query length is more than two or three words. And Duplicate link removal algorithm remove same link (one of them) from search result list. Both algorithms are used to get optimized results.

**Duplicate Links Removal Algorithm :**

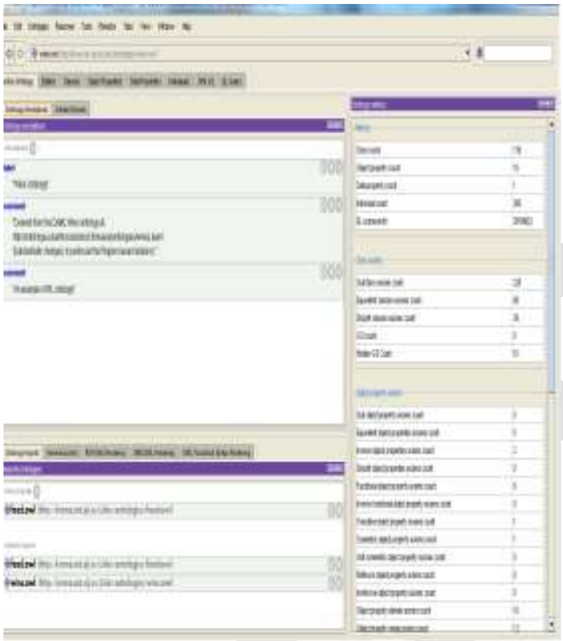
```
➔    In search results, first sort all URL in one sequence.

➔    Give ID to all URL as {U1, U2, U3,…….. }

➔    Compare all URL with each other using string matching function.

➔    function GetOptionFromUrlStr (str, optionName)

          {………URL matching

                    functionality………}

          string 1 = "U1"

          string 2 = "U2"

          strcmp(string1, string2, %)

                    //* If  there are 10 URLs then  computation performs like this

                         U1=U2, U1=U3…………. → 10 scan

                          U2=U3, U2=U4……….. → 9 scan

                           Total scan = n(n+1)/2 = 10(10+1)/2 = 5(11) = 55      *//

➔         If  string ni = string nj

                              //  If two URLs have same link remove any one of them from list

➔         Then remove any one link from search results .

// It can reduce ambiguity

➔Retrieve all remain results.

➔         Duplicate link example :  Salesmen job : URL 1→http://www.simplyhired.com/a/jobs/list

➔         Sales and purchase job : URL 2  →http://www.simplyhired.com/a/jobs/list
```

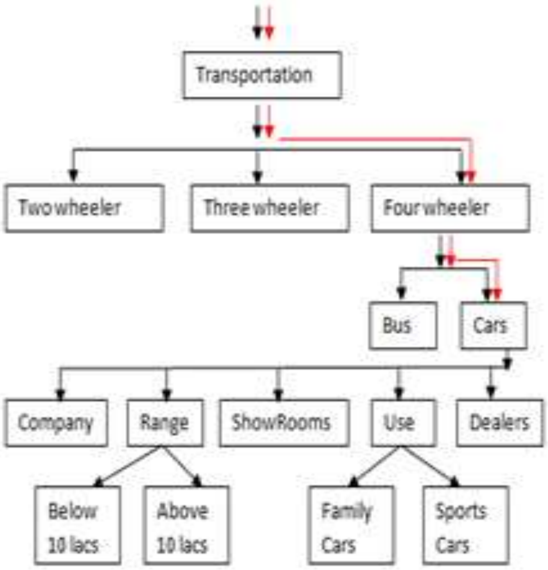
➔ Both above two links are shown in search result list for query “sales and purchase men's job” with two different tile tags.

## IV. **Experiments and Results**

We used two different tools to performing our experiments. First one is Protégé tool used for construction of ontology and second one is custom search engine build up using php and javascript. Custom search engine contain three basic algorithm first one is query classification algorithm proposed by MyoMyo Thannaing, second one is irrelevant link removal algorithm for optimize results and third one is duplicate link removal algorithm.

Protégé is used for developing ontology. It create OWL (web ontology language) file. It can help to getting relation between documents according to its semantic relation. Snapshot 1 shows the relationship between class, attributes and concepts. And snapshot 2 represent resultant query. (How input query should be mined in target category?)



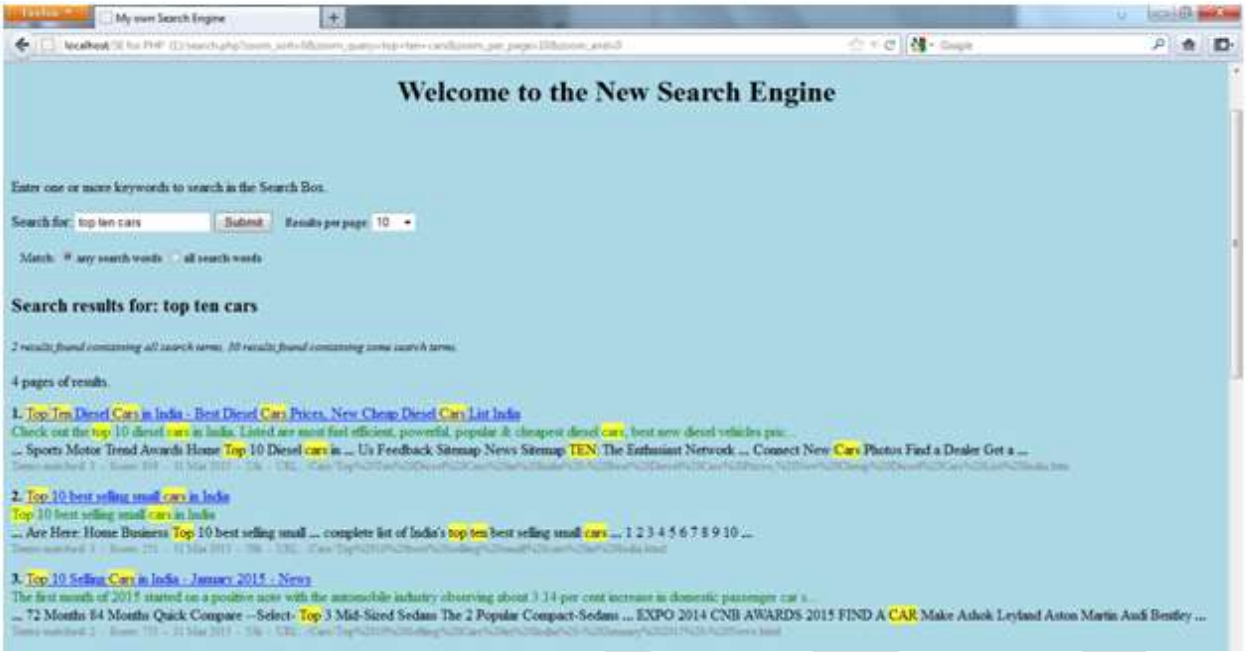


Snap 1 Class, Instance and attributes

Snap 2 Resultant query

In our method first we index all require documents. Web documents contain many forms of documents like .pdf, .doc, .ppt, .html, .xml etc . We can index many of them. Snapshot 3 shows the Search engine. In the search engine we can put our query and get appropriate results according to query. Search engine use query classification algorithm to classify the documents into predefined categories.

Snapshot 3 also represents the retrieve results from search engine for the user query “top ten cars”. All the words from user query are shown in bold and color text format.



Welcome to the New Search Engine

Enter one or more keywords to search in the Search Box.

Search for: top ten cars Submit Results per page: 10

Match: ● any search words ○ all search words

Search results for: top ten cars

*2 results found containing all search terms. 10 results found containing some search terms.*

4 pages of results.

1. Top Ten Diesel Cars in India - Best Diesel Cars Prices, New Cheap Diesel Cars List India
Check out the top 10 diesel cars in India. Listed are most fuel efficient, powerful, popular & cheapest diesel cars, best new diesel vehicles pric...
... Sports Motor Trend Awards Home Top 10 Diesel cars in ... Us Feedback Sitemap News Sitemap TEN The Enthusiast Network ... Connect New Cars Photos Find a Dealer Get a ...
[illegible]

2. Top 10 best selling small cars in India
Top 10 best selling small cars in India
... Are Here: Home Business Top 10 best selling small ... complete list of India's top ten best selling small cars ... 1 2 3 4 5 6 7 8 9 10 ...
[illegible]

3. Top 10 Selling Cars in India - January 2015 - News
The first month of 2015 started on a positive note with the automobile industry observing about 3.14 per cent increase in domestic passenger car s...
... 72 Months 84 Months Quick Compare --Select- Top 3 Mid-Sized Sedans The 2 Popular Compact-Sedans ... EXPO 2014 CNB AWARDS 2015 FIND A CAR Make Ashok Leyland Aston Martin Audi Bentley ...
[illegible]

Snapshot 3 Custom search engine with results

**Calculation of Precesion and Recall**

For query "Web query Classification"

Total result = 14 True Positive = 13 False Positive = 02 False Negative = 01

Precesion = True Positive/(True Positive+False Positive)
= 13/(13+02)
= 13/15
= 0.867

Recall = True Positive/(True Positive+False Negative)

= 13/(13+01)

= 13/14

= 0.982

We can see that there are good query results according to precesion and recall.

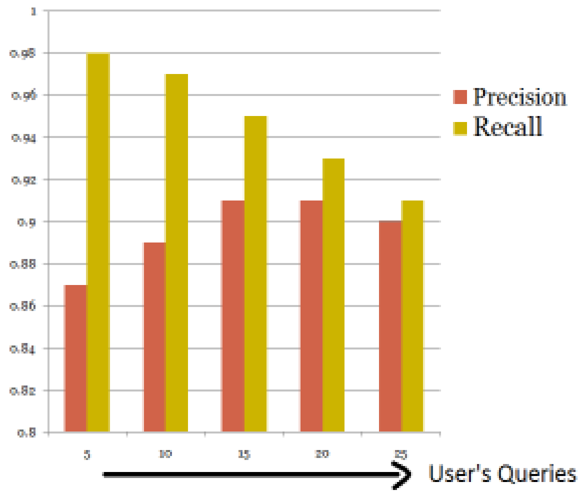


Fig. 2 Precision & Recall

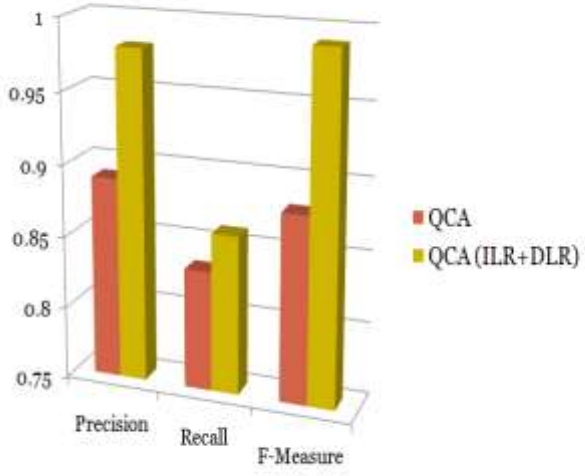
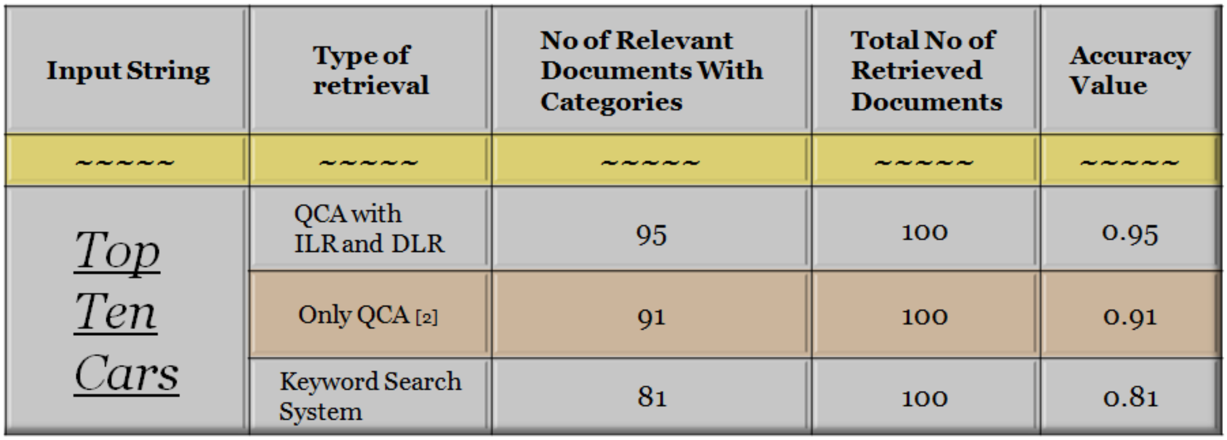


Fig. 3 Comparison with QCA



| Input String | Type of retrieval | No of Relevant Documents With Categories | Total No of Retrieved Documents | Accuracy Value |
|---|---|---|---|---|
| ~~~~~ | ~~~~~ | ~~~~~ | ~~~~~ | ~~~~~ |
| *Top Ten Cars* | QCA with ILR and DLR | 95 | 100 | 0.95 |
| | Only QCA [2] | 91 | 100 | 0.91 |
| | Keyword Search System | 81 | 100 | 0.81 |

Table 1 Comparison with QCA and Keyword based search Engine

## V Conclusion

Query classification is one technique in which query should classify to the number of predefined categories. Query classification use ontology as a model to classify the input search queries. Ontology is used in information retrieval system to retrieve more relevant information from a collection of unstructured information source. Here we used two different algorithms which can helps to getting optimize user relevant results.

In future work we want to develop Information retrieval system using large ontology which contains all possible semantic meanings and all query words. And also want to improve accuracy of information retrieval system**.** In future We also want to Develop Global Semantic search engine based on Ontology.

## REFERENCES:

[1] Data Mining Concepts and Techniques by Jiawei Han, Micheline Kamber, Jian Pei - Book

[2]Ontology-Based Web Query Classification for Research Paper Searching , by MyoMyo ThanNaing, International Journal of Innovations in Engineering and Technology (IJIET) , Vol. 2 Issue 1 February 2013,

[3] Ontology-Based Semantic Online Classification of Documents: Supporting Users in Searching the Web by Ernesto William De Luca and Andreas Nürnberger, IJCST, 2012

[4] Web Query Classification to Multi Categories Based on Ontology by Suha S. Oleiwi, Azman Yasin, International Journal of Digital Content Technology and its Applications(JDCTA) Volume7, Number13, Sep 2013

[5] An Efficient Approach to Web Query Classification using State Space Trees by S.Lovelyn Rose, K.R.Chandran, M.Nithya, ISSN : 2229-4333, International Journal of Computer Science and Technology (IJCST), June-2011

[6] A Simple Model for Classifying Web Queries by User Intent by D. Irazú Hernández, Parth Gupta, Paolo Rosso, and Martha Rocha, IJIET, 2012

[7] A Markovian Approach for Automatic Web Query Classification, by S.Lovelyn Rose, Dr. K. R. Chandran, S.Suriya , ,IJECT Vol. 2, Issue 3, Sept. 2011

[8]http://en.wikipedia.org/wiki/Category:Computer_science

[9]Classification-based Retrieval Methods to Enhance Information Discovery on the Web, by Yogendra Kumar Jain and Sandeep Wadekar, IJMIT Vol.3, No.1, February 2011

[10]Context-Aware Query Classification, by Huanhuan Cao, Derek Hao, Dou Shen and Daxi Jiang, SIGIR'09, July 19–23, 2009

[11]An Ontology-based Webpage Classification Approach for the Knowledge Grid Environment by Hai Dong, Farookh Hussain and Elizabeth Chang, 2009 Fifth International Conference on Semantics, Knowledge and Grid (IEEE-2009)

[12] Ontology Based Information Retrieval - An Analysis by Sakthi Murugan R and P. Shanthi Bala Dr. G. Aghila, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013

[13]http://www.cs.waikato.ac.nz/ml/weka/

[15]http://protege.stanford.edu/