Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING WEB QUERY PROCESS USING CONTEXT CYCLIC METHOD

[1]M.MANIKANTAN, [2]Dr.S.DURAISAMY

[1] Assistant Professor, Department of Computer Applications,
Kumaraguru College of Technology, Coimbatore, India.
[2] Professor & Head, Department of Computer Applications,
Sri Krishna College of Engineering and Technology, Coimbatore, India,
E-mail: [1]mani_dpm@yahoo.com , [2]sdsamy.s@gmail.com

ABSTRACT

The advent of internet besides the sophisticated and advanced web technologies have edified themselves to be a panacea for solving the complications in the process of information retrieval from various data sources like local, regional and other organizational information. Even though the utility of search engines is prevalently observed, the web users still encounter problems in fixing or retrieving the requisite information for their web queries. Accuracy and appropriateness of data-retrieval through web queries witness a great demand in the current scenario. This paper is an effort to elucidate the enhancement of the accuracy and appropriateness of the data-retrieval for the web queries posted by the users. A context sensitive model is proposed in this article to enhance the web query processing through intelligent tag assignment graph approach. The aspiration of this research is to provide a successful demonstration of the use of existing knowledge sources to enhance the content of web queries.

Keywords: *Web Query Processing, Information Retrieval, Tag Assignment, Search Engine, Context Mapping Model.*

1. INTRODUCTION

Information is the fulcrum of decision making in discrepant capacities conjugated with data warehousing processes. The prioritized demand of the business world in the current scenario is enhanced efficacy of retrieving information from heterogeneous and distributed data sources. Perhaps, the credibility of information retrieved from these sources may be integrated into a single system to help the user retrieve the aspired or anticipated information through a single query. However, Query expansion edified itself to be a feasible task even before the inception and existence of the World Wide Web. In fact, down the past a variety of noticeable techniques have been executed. Obviously, the significant connection underlying all of them is to extend the query with words related to the query terms. The utility and availability of thesaurus is one exemplification that helps in WordNet through expanding the query by adding synonyms [11]. Taking into account statistical information such as co-occurrences of words in documents or in fragments of documents from the searchable documents, the related terms can also be automatically explored.

The analysis of prior research opened an avenue for the application of different types of knowledge to elevate the processing of web querying with various levels of success. The efficacy of the semantic ingenuity, while processing queries to adequately supplement the essential information intended by the user has to surpass the previous searches [2]. Application of linguistic proficiency in query extension approaches [3]. The noteworthy facts are linguistic repositories lack semantic knowledge, so query expansion cannot deal with several issues: i) knowledge related to the domain of the query, ii) common sense inferences, and iii) the semantic relationships in which the concepts of the query can participate.

As the anticipated keywords of the user are not exactly available in data server, it may be stored in another related keyword in repositories. This empirical research, considers the query keyword as a mapping domain to formulate the possible keyword which is closely related to the need of the user besides, collecting the result in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

various keyword formats with all available repositories expected by the user.

1.1 Web Queries

The objective of a web query is to explore the best information the users look for. The best exploration is possible only through the application of comprehensive semantic discretion. Actually, semantics is defined as the meaning or connotation of words. Pursuing an effective research in the application of semantics depends on the utility and manipulation of keywords in both meaningful and useful ways [7].

On the other facet, the potential and noteworthy menaces of semantic search are the ambiguity of the language used in the queries, the hazardous partial knowledge of the user, and the predicament in determining the intension of the users. These reasons consequently may lead to the subsequent problems that hinder the processing of web queries.

1.1.1 Detecting intelligible queries

The paucity of guidelines to support the user in detecting the best jargon for a query edifies the inadequacy of systematic strategies of web search.

1.1.2 Resolving ambiguous terminology

Documents that deal with the same domain can use discrepant terminology to enunciate the same concepts.

1.1.3 Categorizing the relevance of results

Detecting the validity of the results for a query resumes a challenge because the validity of the result could be judged exclusively by the percentage expectation it meets out of the user.

Researchers have been pursued to curtail these problems. Interactive approaches, such as query refinement, help the user identify better terms for a query [2, 8].

1.2 Query Classification

The requirement of the users bifurcates, split into specific and general.

i) Specific query: This type of query procures deliberate information about expected query. Here the user restricts the domain or region for the heterogeneous sources to extract result.

Example 1:

"best cbse schools in Chennai"

This query aspires extracting the apposite results of all the cbse schools in Chennai based on the ranking for best school. Usually knowledge repositories use only intentional information to define their knowledge. In this type, more specific results are compatible to the user's query.

ii) General query: Inadequacy of clarity of the user over the input keyword leads to general web query. In such cases the results are arbitrarily mustered from various domains from which the user chooses the appropriate one.

Example 2:

"best schools in Chennai"

This query links itself to various domains to extract the results at random on the subsequent lines.

1. best matriculation schools in Chennai.
2. best cbse schools in Chennai.
3. best international schools in Chennai.

The user may intend any one of the domain results or a combination of results all the three domains. For queries that use intensive and extensive knowledge, the knowledge repositories that represent both of these types of knowledge may be deployed.

2. RELATED WORK

The previous web query processing techniques are oblivious about domain specific knowledge besides, they fail to infer the user's query and the intrinsic relationships between its terminologies [1]. The strategy of integrating linguistic and semantic information into a holistic repository could be useful to increase the contexts where knowledge in these repositories can be used successfully [2]. Most of the query extension approaches use only linguistic knowledge. Linguistic knowledge can improve information retrieval [3], however, linguistic repositories lack semantic knowledge, so query expansion may not embark on deal some issues like: 1.Knowledge

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

related to the domain of the query. 2. Semantic relationships in which the concepts of the query can participate. 3. Commonsense inferences.

In retrieval process heuristic approach was used in query processing. The devising of a heuristic-based methodology for an ingenious agent to process queries on the semantic web [4] is to take the processing into account the semantics of the user's request. Besides aiding proper interpretations relating to specific domain, ontology may also exhibit domain knowledge for information retrieval. Domain ontology may supply vocabulary for representing and communicating knowledge in that domain and a set of relationships that hold among the terms in that vocabulary [1]. In that work, they adopt OWL DL to represent domain ontologies. OWL DL is a sublanguage of OWL which can be directly mapped to Description Logics (DL).

An additional strategy to query expansion in answer retrieval is the utility of Statistical Machine Translation (SMT) techniques in bridging the lexical gap between questions and answers [5]. SMT-based query expansion is pursued by i) using a full-sentence paraphrases to introduce synonyms in context of the entire query[6], and ii) by translating query terms into answer terms using a full-sentence SMT model trained on question-answer pairs.

According to Stevenson (2003), the computer systems such as search engines overcome these limitations in constrained domains. The noteworthy situations here are the two approaches to addressing this challenge. The first requires designing a "semantic" web by enhancing web content with meta information such as tags as stated by Berners-Lee et al.(2001). The second approach suggests improvised exploration by integrating semantics into queries as on World Wide Web. Of course, the search engines attempt to terminate these predicaments through discrepant means. Recently, the web search is confided on for very many functions. So, the need for better search services is becoming increasingly important [9]. Due to the diversity of content and structure of the web, innovative techniques are needed to create more focused queries. Some query expansion and refinement techniques use conceptual fuzzy sets [10]. Although several approaches to improve web queries through query enhancement are reported in the literature survey, we discuss only the ones closet to our approach. [2].

3. PROPOSED ARCHITECTURE FOR CONTEXT MAPPING MODEL (CMM)

The roots of the a web query processing system evolve from the domain specific context sensitive mapping technique besides deploying a context cyclic graph algorithm to improvise the retrieval process either from indigenous or from other organizational information domains.

An exclusive utility of the user keyword to form the additional keywords so as to form the added query to filter appropriate results is the specialty of this model. Keyword generation and query formation seem to be the major aspects of this strategy. The system architecture design of the strategy proposed comprises three layers namely Application Layer, Query Processor and Database Layer.

The Application Layer, the Primary Layer, assumes the function of the User Interface in the execution of converting the user queries into added queries Whereas, the Query Processing Layer, the Immediate Layer, processes the user queries besides, the added queries and thus functions as a link layer to the application layer and database layer. The tertiary layer, titled database layer extracts relevant information from various Data Sources (DS) as exhibited in figure1, CMM System Architecture.

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

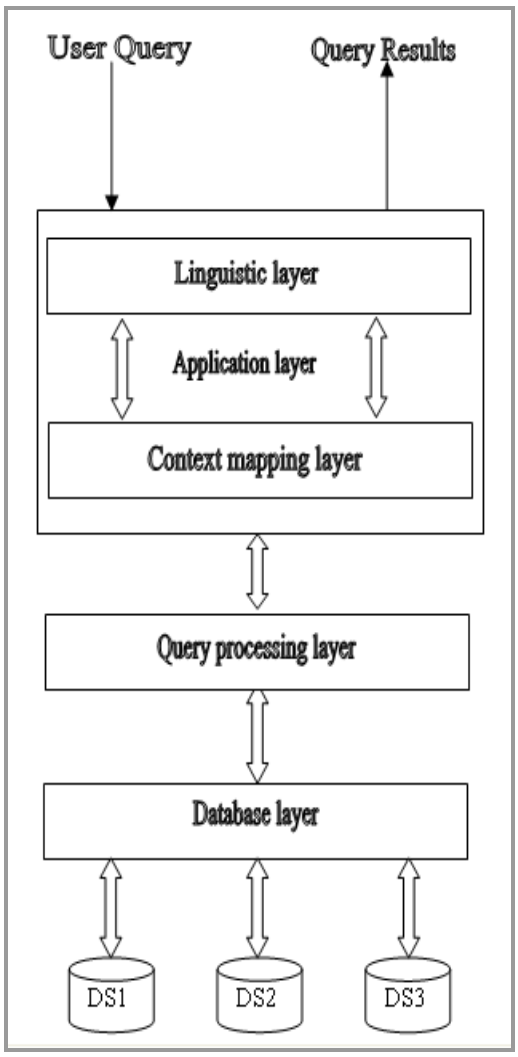

Figure 1. CMM Architecture.

The combination of linguistic module and context mapping module improve the efficacy of application layer. Here, Exploit algorithm plays a pivotal role in dividing the key words to analyze the user query. The processing of each individual keyword takes place in the linguistic module that sustains the requisite semantic support in the search. Consequently, the context mapping module generates and poses the possible alternate queries by using context cyclic graph mapping. This research exclusively concentrates on the deployment of specific query search of the users.

Example 3:

"best datamining papers"

In the above stated user query, the phrase "best datamining papers" perhaps, may be divided into three individual keywords like best, datamining and papers by using exploit algorithm. The subsequent illustration in figure2 edifies the deployment exploitation algorithm.

```
Start
assign x and i
x = "best datamining papers"
i = exploit[x, " "]
for (j=0; k=0; j<count (i); j++; k++)
{
        m[k] = i[j]
 }
End
```

Figure 2. Exploit initial query algorithm

Based on the algorithm, consider the initial query

X→ "best datamining papers"

The outcome of this user query on deployment of this algorithm was,

m[0] → "best"

m[1] → "datamining"

m[2] → "papers"

However, this outcome may be noted as

X ∈ {best, datamining, papers}

Consequent to the exploitation process, these individual keywords are collected and sent to the linguistic module. The keyword analyzer of this module generates the related keywords for the three divided keywords shown above, culled from the semantic layer that exists in this module. These three sets of keywords

are then sent to the keyword analyzer exclusively for fixing the appropriateness two different keywords per each key word that are identified. As a result, three different word pools are generated with the support of the linguistic module.

best → (keyword analyzer) → {better, good}

datamining→(keyword analyzer) →{miningdata, exploringdata}

paper → (keyword analyzer) → {article, journal}

pool1 {better, best, good}

pool2 {miningdata, datamining, exploringdata}

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

pool3 {article, paper, journal}

Perhaps, the word pool generation facilitates context mapping module which in turn facilitates the generation of synonymous words for various possible queries through matching the keywords in a cyclic way. The subsequent diagram, figure 3 is depiction of it.

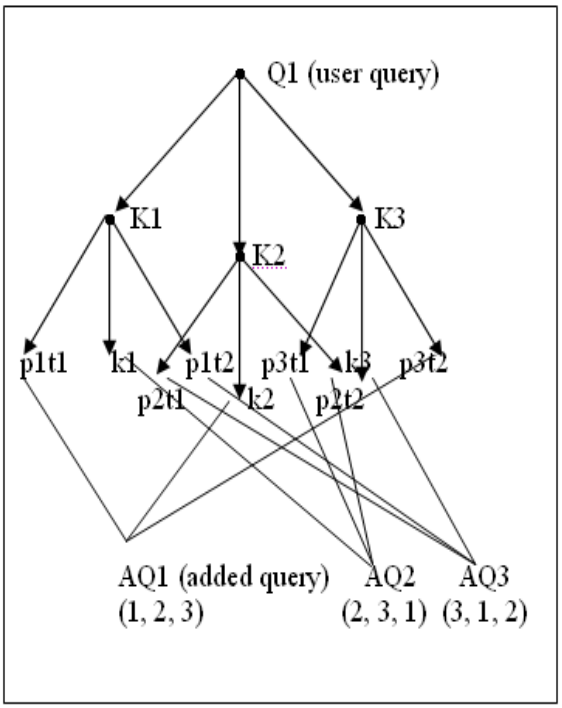

Figure 3. Context cyclic graph

The three different word pools are represented in three different sets as,

Poo11	Pool2	Pool3
better	miningdata	article
best	datamining	paper
good	exploringdata	journal

In each pool, the user word is named as 'k' and added word is named as't'. After converting the words into tag name, it is matched with another tag into other two pools.

Pool1(p1)	Pool2(p2)	Pool3(p3)
p1t1	p2t1	p3t1
k1	k2	k3
p1t2	p2t2	p3t2

The formation of the query was followed by the context cyclic graph (CCG) method. The query formation will be used in the cyclic method like, (1, 2, 3), (2, 3, 1) and (3, 1, 2).

(1, 2, 3) – p1t1 & k2 & p3t2.

(1, 2, 3) – better datamining journal.

The strategies used for the other two keyword formation are the same. Here, the user query is converted into other format of query. In each word pool, one keyword is used to form the added query.

User query: "best datamining papers"

Added query1: "better datamining journal"

The primary keyword of the pool helps in forming this, while the secondary keyword of pool2 and tertiary keyword of pool3 by using (1+2+3) formation.

(p1t1+k2+p3t2) - better datamining journal

Added query2: "best exploringdata article"

The secondary keyword of pool1 and tertiary keyword of pool2 and primary keyword of pool3 leads to (2+3+1) formation.

(k1+p2t2+p3t1) - best exploringdata article

Added query3: "good miningdata paper"

The tertiary keyword of pool1 and primary keyword of pool2 and secondary keyword of pool3 facilitates (3+1+2) formation.

(p1t2+p2t1+k3) - good miningdata paper

The subsequent diagram, figure 4 exhibits query processing.

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

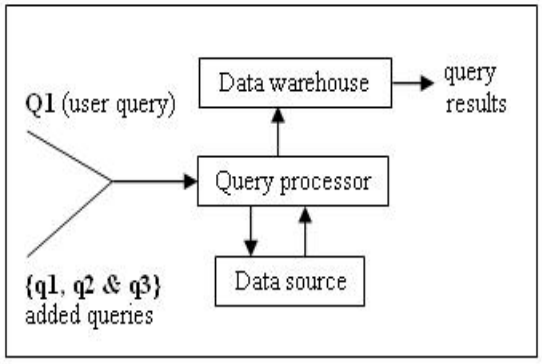

Figure 4. Processing Of User Query And Added Queries

The significance of prioritizing the prominence or recurrence of any one particular key word of the user's query is to restrict the search within user relevant domain. Ultimately, the user query and added query is sent to the query processing and the results are collected to the data warehouse for ranking the result to the user need. Then the most related information will get displayed to the user's view.

4. IMPLEMENTATION

Based on the framework introduced above, a prototype system is implemented for semantic search using java and xml as a middleware with MySQL and Apache servers at the backend. Compared to other search methodologies, this system can significantly improve the precision and depth of the content related results cater to the user need.

The phenomenal implementation of the techniques amalgamated with the efficacy in catering to the user's need to examine the correctness of the results besides, the fuzzy rough set approach enhances the scalability of the proposed web query processing.

Once the relationship between terms is known, either through a lexical aid such as WordNet, or automatically generated from linguistic module, the original query can be expanded in various ways. The straightforward way is to extend the query with all the words that are related to at least one of the query terms. As mentioned in the introduction, this corresponds to talking the upper approximation of the query. This link between query expansion and rough set theory has been established in [12, 13], even involving fuzzy logical representations of the term-term relations and the queries.

4.1 Fuzzy Rough Set Approach

Rough set theory [14] is an interesting candidate framework to aid in query refinement. Indeed, a thesaurus characterizes an approximation space in which the query, which is a set of terms, can be approximated from the upper and the lower side. By definition, the upper approximation will add a term to the query as soon as it is related to one of the words already in the query, while the lower approximation will only retain a term in the query if all the words it is related too are also in the query.

Let X denotes the universe of terms. A fuzzy set A in X is characterized by a X → [0, 1] mapping, called the membership function of A. For all x in X, A(x) denotes the degree to which x belongs to A. Furthermore, a fuzzy relation R in X is a fuzzy set in X x X. For all y in X, the R-forest of y is the fuzzy set Ry defined by

$$Ry(x) = R(x,y) \quad (1)$$

for all x in X. A fuzzy relation is called reflexive if and only if

$$R(x,x) = 1 \quad (2)$$

for all x in X. Moreover, R is called symmetrical if and only if

$$R(x,y) = R(y,x) \quad (3)$$

for all x and y in X. For A and B fuzzy sets in X, inclusion can be defined as

$$A \subseteq B \text{ if } (\forall\, x \in X)\ (A(x) \leq B(x)) \quad (4)$$

Table 1: Number Of Web Pages Found By Google

Keyword	Initial	I related	II related
School	1,120,000,000		
Educational institution		35,200,000	
Teaching learning institute			156,000,000
Datamining	43,900,000		
Miningdata		12,300,000	
Exploringdata			65,200,000

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Based on the above rough set equations, computing the upper approximations, will keep on using the t-norm besides its residual implicator as observed in this example shown in Table 1. It is represented in three divisions. The initial one is direct user query representation and the value of upper approximation of the example query. The first and second are the converted term value of upper approximation

In this example suppose we constructed a graded thesaurus taking the 0.5-level set of R, defined as

$$(x,y) \in R_{.5} \text{ if } R(x,y) \geq 0.5 \qquad (5)$$

The tight upper approximation of a fuzzy set A in the approximation space (X, R) is the fuzzy set R↓↑A defined by

$$R{\downarrow}{\uparrow}A(y) = \inf_{Z \in X} I_T\left(R_Z(y), \sup_{x \in X} T\left(R_Z(x), A(x)\right)\right) \qquad (6)$$

for all y in X.

This may be easily verified through

$$R{\downarrow}{\uparrow}A = R{\downarrow}(R{\uparrow}A) \qquad (7)$$

It is important to point out

$$A \subseteq R{\downarrow}{\uparrow}A \subseteq R{\downarrow}A \qquad (8)$$

always holds, guaranteeing that the tight upper approximation indeed leads to an expansion of the query[13], none of the original terms are lost.

Table 2: Graded Thesaurus

Keyword	Initial	I related	II related
School	1.0		
educational institution		0.64	
teaching learning institute			0.89
Datamining	1.0		
Miningdata		0.56	
Exploringdata			1.0

From the equations (1) to (8), the fuzzy roughest model clearly indicates that the query expansion based the upper approximation leads the lossless information of original data. In Table 2, the upper approximation values are converted into graded thesaurus and all the values are less than are equal to one.

5. RESULTS AND DISCUSSION

A good search technique will always focus on the accuracy of the results and the time taken to execute the query. Based on this, the retrieved query results are analyzed by the four categories namely i) Query processing domain ii) Query processing approach iii) Accuracy of the results percentage iv) Query processing time.

The proposed approach of CMM executes the user query in multiple domains simultaneously, whereas the direct method searches only a single domain for executing the query. In CMM, the processing domain will be split into multiple categories by an added query. This will increase the number of relevant results retrieved by the query.

Our approach uses the CMM, which generates the relevant added queries along with user query. Thus the result generated will be a combination of the results generated by the user query as well as the results of the added query. This will increase the search data because of the parallel query approach followed by CMM. The semantic knowledge used in CMM will reduce the unwanted results thereby increasing the accuracy percentage of generated result. The result percentage for the first query using direct method is 76%, whereas in CMM, it is 89% as shown in Table3. Collectively the result percentage is nearly 15% increased from direct method to CMM is shown in Table 3.

Table3: Improvised Specific Results Of CMM method

S No	Base query (direct method)	Relevance result percentage (Google)	Added query (cmm method)	Relevance result percentage (Google)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

1	best schools in Chennai	76%	i)top schools in madras ii)best educational institution in Chennai iii) good teaching learning institute in Chennai	89%
2	best data mining papers	74%	i)better datamining journal ii)best exploring data article iii)good miningdata paper	88%
3	best backend server	72%	i)good backbone node ii)better datastorage system iii)best database server	86%

The graph for Table 3 is given in figure 5.

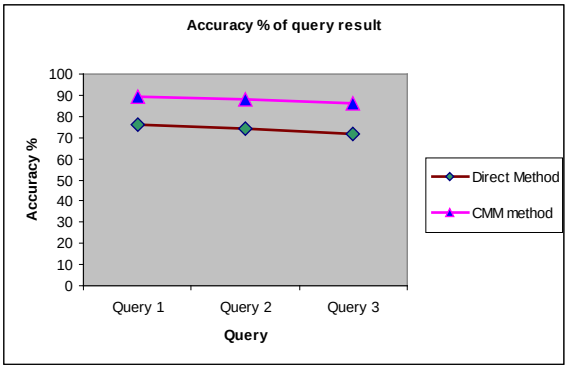

Figure 5. Comparison Of Accuracy Percentage For Query Result

The processing time is marginally increased in CMM method because of added queries. But this time factor is significantly less and it is in acceptable range only. The time duration for the two methods are shown in Table 4.It can be seen that the processing time for "best schools in Chennai" takes 9 seconds using direct method and 17 seconds using CMM. Collectively the results in Table 4 show that it is within the acceptable range only.

Table 4: Query processing time of CMM & direct method

SNo	Base query (direct method)	Processing Time	Added query (CMM method)	Processing Time
1	best schools in Chennai	9 seconds	i)top schools in madras ii)best educational institution in chennai iii) good teaching learning institute in Chennai	17 seconds
2	best data mining papers	11 seconds	i)better datamining journal ii)best exploring data article iii)good miningdata paper	18 seconds
3	best backend server	9 seconds	i)good backbone node ii)better datastorage system iii)best database server	16 seconds

The graph for Table 4 is given in figure 6.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

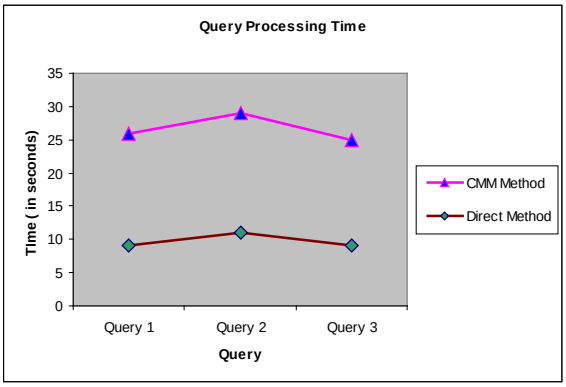

Figure 6. Comparison of query processing time

The comparison between both approaches based on the four categories is summarized in Table 5.

Table 5: Result comparison.

S.No.	Category	Direct approach	CMM approach
1	Query processing domain	Single domain	Multiple domain
2	Query processing approach	Direct method	Semantic method
3	Accuracy of the result percentage	Less than 80 %	Greater than 80%
4	Query processing time	Normal	Slightly increased

The above results indicate that the retrieving information in web, relevant to the user's query is more accurate when CMM method is used.

6. CONCLUSION AND FUTURE WORK

The methodology of this research has its roots in those prior researches on semantic, linguistic, knowledge repositories. This research thoroughly focuses to perform justice in the search of specific queries. This methodology has been implemented in a prototype and applied to web queries. The preliminary results from the prototype are quite improving.

This research contributes to the enhancement of web queries in several ways. First it includes the additional synonymous query of the user query by using the context mapping model. This may help to extract the similar information from various sources, which are stored in the different names. Above all, this research illustrates mapping model and keyword related mapping to represent the user's intentional query.

In this work, the academic domain is exclusively taken for implementation and result analysis. Future research work may be extended to the different types of domains and queries. Also the time taken for query processing could be more optimized.

REFERENCES

[1] W. Fang, L.Zhang, Y.Wang, S.Dong, "Toward a semantic search engine based on ontologies", *Proceedings of the fourth international conference on machine learning and cybernetics*, August 2005, Vol. 3:pp. 1913-1918.

[2] J. Conesa, V.C.Storey, V.Sugumaran, "Improving web-query processing through semantic knowledge", *Data and Knowledge Engineering*, 2008.

[3] William A. Woods, Lawrence A. Bookman, Ann Houston, Robert J. Kuhns, Paul Martin, Stephen Green, "Linguistic knowledge can improve information retrieval", *conference on Applied natural language processing* (ANLC '00), 2000, pp.262-267.

[4] A. Burton-Jones, V.C.Storey, V.Sugumaran and Sandeep Purao, "A Heuristic-based methodology for semantic augmentation of user queries on the web", *Springer-Verlag Berlin Heidelberg* 2003, Volume 2813, pp 476-489.

[5] S. Riezler, A. Vasserman, I. Tsochantarids, V.Mittal & Yi Liu, "Statistical machine translation for query expansion in answer retrieval", *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007.

[6] V.C.Storey, A. Burton-Jones, V.Sugumaran and S. Purao, "CONQUER: A methodology for context-aware query processing on the World Wide Web", *Information System Research*, 2008.

[7] Mohd K.Yusof, Ahmad F.Amri Abidin Sufian M.Deris and Surayati Usop, "Implementing of XML and Intelligent Algorithm for Improving Web Query Processing in Heterogeneous Database Access", *International Journal of*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Database Theory and Application, vol. 4, 2011.

[8] M.S. Khan & S.Khor, "Enhanced web document retrieval using automatic query expansion", *Journal of the American Society for Information Science and Technology*, 2004, pp 29-40.

[9] Steve Lawrence, "Context in web search", *The IEEE Data Engineering Bulletin*, 2000.

[10] T. Takagi, M. Tajima, "Query expansion using conceptual fuzzy sets for search engine", *Proceedings of International Conference on Fuzzy Systems*, 2001.

[11] E.M Voorhees, "Query expansion using lexical semantic relations", *ACM SIGIR conference on Research and Development in Information Retrieval*, 1994, pp 61-69.

[12] P.Srinivasan, M.E. Ruiz, D.H. Kraft, J.Chen, "Vocabulary Mining for Information Retrieval: Rough Sets and Fuzzy sets", *Information processing and management*, Elsevier, 2001.

[13] M De Cock, Chris Cornelis, "Fuzzy Rough Set Based Query Expansion", *Proceedings of Rough Sets and Soft Computing in Intelligent Agent and Web Technologies*, 2005.

[14] Z. Pawlak, "Rough Set theory and its applications to data analysis", *Cybernetics and Systems*, Taylor & Francis, 1998.