

<sup>1</sup>U.K.Sridevi,  
<sup>2</sup>N.Nagaveni,

## A Concept Relation Sub Graph in Semantic Web using Genetic Algorithm



**Abstract**— The development of search algorithms between web pages is motivated by the “related pages” queries of web search engines and web document classification. This paper proposes a method for preserving the relations between the concepts by using genetic algorithm based edge removal algorithm. The goal is to provide optimal concept relation graph based on keywords given by the user and comparing the solution from well known traditional keyword search method. The genetic algorithm based edge removal method removes the edges from the concept relation and generates the keyword relation pair. The retrieval model is based on the important factors of the structural elements, which are used to rerank the document retrieval by the standard weighting scheme. The structural field weights are combined with the annotation-weighting scheme to improve the relevance measures. The proposed method has been evaluated on USGS Science directory collection. Using Jena SPARQL query model and GATE Tool API, the documents can be retrieved from the corpora not only based on their textual content but also according to their annotation features and relations. The documents that are annotated with the concepts are retrieved with higher ranking. A preliminary experiment result shows that the proposed method may generate relevant document in the top rank.

**Keywords** – Semantic Web, Genetic Algorithm, Ontology, Information Retrieval, Semantic Annotation.

### I. INTRODUCTION

With the explosive growth of the World Wide Web, there is an increasing amount of information resources, of which most are represented as free text. A frequent approach to web page representation has been a bag-of-words representation of a document, in which all parts of a web page are considered to be of equal significance. However, unlike other text documents, web pages have certain characteristics, such as internal metadata, structural information hyperlinks and anchors, which could serve as potential indicators of subject content. For example, words from title could serve as potential indicators of subject content. The degree to which different web page elements are indicative of its content is referred to as significant indicator. The technique that determines the importance of different parts of a web page improves the retrieval performance. Today’s web search technologies rely on link analysis techniques that exploit the structure of the web to determine the important documents. Because of the limitations of traditional retrieval mechanisms, conventional direct keyword based information retrieval technology cannot meet the growing user retrieval need with semantic

knowledge. The keyword-based retrieval fails to integrate information spread over different resources.

In recent years the semantic web has been advocated as the extension of the current web. The semantic web aims to achieve better data automation, reuse and interoperability [4]. The main advantage of semantic web is to enhance search mechanisms with the use of ontology’s [1]. Ontology is a general description of all concepts as well as their relationship. The Resource Description Framework /Schema (RDF(S)) and Web Ontology Language (OWL) are W3C recommended data representation models which are used to represent the ontology’s [6]. The basic method for constructing the semantic web is to use the terms defined in ontology as a metadata to markup the web’s content.

The classical information models are Boolean model, Statistical model, Hyperlink based model [8]. The information retrieval model does not utilize semantics of the queries and document collection. There have been many works which employ the semantic web technologies for information and retrieval such as KIM [8]. A variety of aspects on improving search and ranking documents have been considered, such as concept based search of documents [7]. The research problem of improving relevance in search and ranking of documents requires techniques that consider the semantic annotation.

Over the last three years, a number of semantic web tools have been developed. Nevertheless, most of the currently available semantic web tools are limited to work only with few of the semantic web architecture just like RDF Editors or Ontology Editors [11]. The current research focuses on providing a semantic metadata to enhance the information retrieval and support e-business applications. Automatic

<sup>1</sup>U.K.Sridevi is with Sri Krishna College of Engineering and Technology,

Tamil nadu, India. She can be reached at [srideviunni@gmail.com](mailto:srideviunni@gmail.com).

<sup>2</sup>N.Nagaveni is with Coimbatore Institute of Technology, Tamilnadu, India.

She can be reached at [nagavenipalanisamy@yahoo.com](mailto:nagavenipalanisamy@yahoo.com)

creation of metadata for web pages resembles the task of semantic annotation in general [10]. A number of annotation tools for producing semantic markups exist such as SHOE, Protégé, OntoAnnotate, MnM [8]. There are several research projects about ontology-based information retrieval. The ontology definition of concepts can be used to describe the concepts and these concepts will be defined as a document class [3]. SEAL [6] was conceived for semantic search of knowledge on the web and also been used for sharing knowledge on the web.

KIM [8] introduces a holistic architecture of semantic annotation, indexing and retrieval for documents. It aims to achieve fully automatic annotation and to improve search and retrieval by integrating information extraction using GATE. Ontology based retrieval model complements KIM with a ranking algorithm specifically designed for ontology based retrieval model using semantic indexing scheme based on annotation weighting techniques. Genetic algorithms are generally quite effective for rapid global search to find solutions in nondeterministic problems. Genetic algorithm method enhances the efficiency and adaptability of a meta searching [9]. The distribution frequency of keyword improves the identification of an important document based on the query. The new similarity measures incorporate the tag structure weights in addition to standard weighting schemes.

In this paper, it is proposed that an ontology based retrieval model for the exploitation of domain ontology's and knowledge bases, to support semantic search in document repositories. The search system takes advantage of ontology based semantic annotation and it includes the weights of document structure in ranking. The collection of documents from USGS Science directory is used as test data set for evaluation. The approach combines the structural weights with annotation scheme to rank the annotated documents. The structural mining approach and probability based annotation scheme significantly improves the retrieval performance especially for the top ranked document. Experimental results indicate that combining the document structural relevance weight leaned using genetic algorithm in ontology based semantic annotation weights improves the retrieval performance.

## II. DEVELOPMENT OF ANNOTATION MODELS

Semantic annotation is about assigning to the entities in the text links to their semantic description. Annotation provides additional information about Web contents so that better decision on the content can be made. Annotation ontology tells what kind of property and value types should be used in describing a resource. The need of services for supporting the usage of domain ontology's is employed for the annotations. To improve the recognition of important indexing terms, it is possible to weight the concepts of a document in different ways. For example, in topic indexing, concepts that form semantically related terms, gain more weights. Although various annotation systems and methods have been developed, the question of how to easily and cost effectively produce quality metadata still remains largely

unanswered. Dublin core annotation mainly describes the properties of the document itself without providing too many details about its content. Ontology based annotations are instead developed to describe the content of the document and not its general properties. The manual annotation of document is a high cost and error prone task. To alleviate this task, an important effort is currently being made in automation of document annotation and the result is some degree of automation. However there is still some work to do to achieve a complete automation of the annotation. The classical information retrieval model is incapable of supporting logical inference.

In general, most of the work about semantic annotation requires some predefined ontology's to extract, define and relate the annotation. The models of automatic semantic annotation are ontology driven semantic tagging and semantic meta data generation. The automatic semantic meta data generates meta data that can semantically describe the content of annotating the page. The generated meta data includes ontology by system defining its own semantic categories or a system relies on some predefined ontology. The annotation process of a web page is based on annotating a web page with ontology and adding the relations between individuals [5]. The ontology-based information retrieval based on vector space model describes the semantic annotation scheme of KIM platform [2]. It has reused automatic concept to label mapping available from the KIMKB [8]. In fully automatic annotation systems like KIM architectures support instance identification in a restricted predefined ontology model. Ontology based semantic annotations are needed when building the semantic web. The ontology-based information retrieval recognizes the relations among terms by referring to the ontology. Creating ontology is not an easy task and obviously there is no unique correct ontology for any domain. The real quality of ontology can be assessed only for its use in real application.

### A. Document annotation

Document is composed of many terms and important words are spread out as documents. The importance of significance indicators assigned to the web elements like title, heading, bold, anchor improves the ranking of the web documents. Unlike text documents, web pages have certain characteristics such as structural information, hyperlinks and anchors which could serve as potential indicators of subject content. The relevance score of the document is assigned based on which term is matched and that part of the web page in which the match is found. The annotation process of a web page is done with concepts of the ontology. Then, relations between individuals are discovered and instances are added. The document is annotated based on the following steps using GATE tool.

1. Creation of ontology based document should be preprocessed to obtain semantic annotations. Indexing of the document should be done.
2. Examine the location of the annotated instance in the document.
3. The annotation weights are calculated by combining the frequency and structure weights.
4. Retrieval of documents related with query and ranking the document based on relevance of the query.

The effective information retrieval using Web structures is shown in Fig.1.

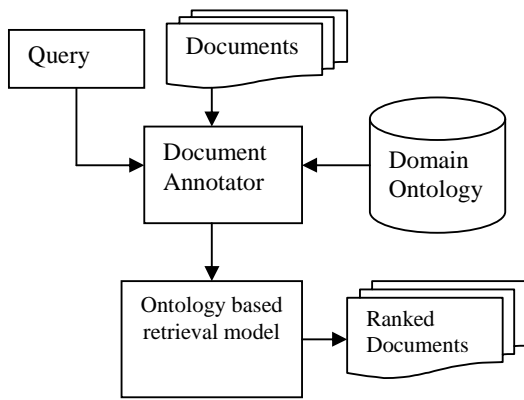


Fig.1. Ontology based retrieval model

Research in structural weights has suggested using document structures for document ranking. Genetic mining of HTML structures uses the HTML tag weights to improve the performance of document retrieval system [9]. The term that exists in title, bold and anchor tags add more weight to the document than the other terms. The document retrieval performance is improved depending on the structural importance of the document

**B. Semantic Annotation Weights**

Web search engines provide advanced features in that a user can specify how a query is matched with the title of the page, text of the page, URL and links to the page, anywhere in the page. Although it is worthwhile to investigate how different matching affects the accuracy of the similarity of query to the document of retrieval results using tag weights [9]. Suppose a document collection on the Semantic Web is  $D = \{d_1, d_2, \dots, d_n\}$ . The number of occurrences of an instance in a document is primarily defined as the number of times the label of the instance appears in the document, if the document is annotated with instance and zero otherwise. To extend vector space model to support structured ranking occurrences within each document, a structure must be included. The weight of a term in a document is basically computed by the classical term frequency and inverse document frequency. Term frequency (tf) is the number of times that a term  $t$  appears in a document. The inverse document frequency (idf) is the inverse of document frequency in the collection that contains term. The scheme is defined in Eq.(1)

$$wk = tfk \cdot idfk \cdot \sum_{i=1}^{i=m} loc[tfk] \tag{1}$$

$$Wdk = wk * wq \tag{2}$$

In Eq. (1)  $wk$  is the weight of  $k^{th}$  term in the document,  $tfk$  is the frequency of the  $k^{th}$  term in document,  $N$  is the total number of documents in the collection and  $idfk$  is the inverse document frequency annotated with  $k^{th}$  term. The structural

document field weight is  $loc[tfk]$  for every tag position in the document and  $m$  is the total number of tags in the document. If the term annotated with the concept in particular tags such as <HTML>, <TITLE>, <B> the weight of the tag is included along with the term weight.  $wq$  is the query weight and  $Wdk$  is the weight of an annotated document instance. In Eq.(2) the annotation weight is calculated.  $C_i$  is the concept weight and  $n$  is the total number of concepts in ontology. The probability of the term annotated with a concept using bayes rule is given by  $Wb$  in Eq. (3)

$$Wb = \sum_{i=1}^n P\left(\frac{wdk}{C_i}\right)P(C_i) \tag{3}$$

In Eq. (4), the score for the document is calculated using the ranking model that combines the concept frequency and the probability of encountering an instance.

$$Score(d, q) = Wdk + Wb \tag{4}$$

The keyword based analysis collects the set of keywords or terms that occur frequently together and then finds the correlation among them. The semantic information retrieval KB has been built and associated to the information document base by using domain ontology that describes the concepts. The query model can employ to find and manipulate the needful data from the annotated documents.

**III. DEVELOPMENT OF CONCEPT RELATION GRAPH USING GENETIC ALGORITHM**

In Semantic Web, RDF is represented in concept graph. The query keyword concept and the relations between them are considered when retrieving a web page. In keyword based searching only matching keywords in the document is considered and the relation between them is lost. But based on the keyword and relation pair the relevant pages can be retrieved.

A concept relation graph  $G(C, R)$ , where the vertex set  $C = \{c_1, c_2, \dots, c_n\}$  and  $R$  is the relation between the concept. The edges represent the number of relationships between the concepts. Suppose that the keywords submitted by a user are  $k_1, k_2, \dots, k_n$  and that corresponding concept set is  $C = \{c_1, c_2, \dots, c_n\}$ .  $G_1, G_2, \dots, G_p$  are the concept relation sub graph of  $G$ . Keyword pair is created based on the keywords submitted by user. The concept relation sub graph is generated based on the minimum edge removal method. In graph representation removing an edge from a graph means fetching a sub graph from a graph [11]. In the minimum edge of the graph  $G = (V, E)$ , partition  $V$  into disjoint subsets  $U$  and  $W$  such that  $e(U, W)$ , i.e., the number of edges  $\{(u, w) \in E \mid u \in U, w \in W\}$  is minimized.

A genetic algorithm is applied to find a sub graph of the graph based on concept and its relation. Each candidate is expressed as a chromosome represented as a bit string containing as many bits as the number of vertices in the sub graph. The  $j^{th}$  bit in a bit string identifies the group into which the  $j^{th}$  vertex of the graph is placed in the sub graph. Fitness refers to the actual rate that an individual ends up being sampled in contributing to the next generation. The

fitness value of candidate solution is the total number of edges that is connected to different group. The optimal fitness is the minimum number of edges connecting to the concepts that are not related to keywords.

Let  $N$  be the size of the population. The initial population of size 8 is generated randomly. The genetic algorithm uses crossover and mutation operators to generate the offspring of the existing population. Fitness is defined as total relations, where a total relation is the sum of relations that connects to the concept nodes that are in different concept group. The best fitness is the minimum value of  $R_p$ . The optimal fitness chromosome is chosen as the minimal sub graph. The Fig.2 shows the flowchart for genetic algorithm. Based on the query, retrieved page result set covers all the relations and arcs in the concept relation sub graph. The fitness value is calculated using Eq. (5)

$$Fitness = count(R_p) \quad (5)$$

Where  $R_p$  is the total number of relations connected to different concepts in the graph.

The GAEdgeremoval algorithm is as follows:

**Algorithm GAEdgeremoval ()**

**Input:** Incidence matrix of graph  $G (V, E)$  and the keywords concept.

Generate initial population  $P_s$  at random.

Evaluate the fitness of each chromosome  $C_i$ ;

$R_p$  is the total number of relations connected to different concepts in the graph.

Select the best individuals with fitness  $\geq$  average fitness;

**Repeat**

Generate offspring by applying crossover and mutation operator on individuals.

Evaluate the fitness according to the fitness function

Fitness= count ( $R_p$ )

Calculate the average and maximum fitness for the population

Until (optimal fitness is achieved)

For every edge in the  $R_p$

{Include the edge in the concept relation sub graph;

Each edge measures the correlation between the nodes;

Insert each  $R_{ij} \in R_p$  into Keyword relationship set;

}

**Output:** Return Web page that contains Keyword and its relation pairs generated from sub graph. The Property-keyword pair from concept relation sub graph is generated based on the GAEdgeremoval algorithm.

Let  $X$  and  $Y$  be two selected parent chromosomes, which are represented respectively as follows:

$$X = \{x_1, x_2, \dots, x_l\},$$

$$Y = \{y_1, y_2, \dots, y_l\}.$$

The selected parents produce offspring by crossover.

Let  $Z$  be the offspring generated:

$$Z = \{z_1, z_2, \dots, z_l\}.$$

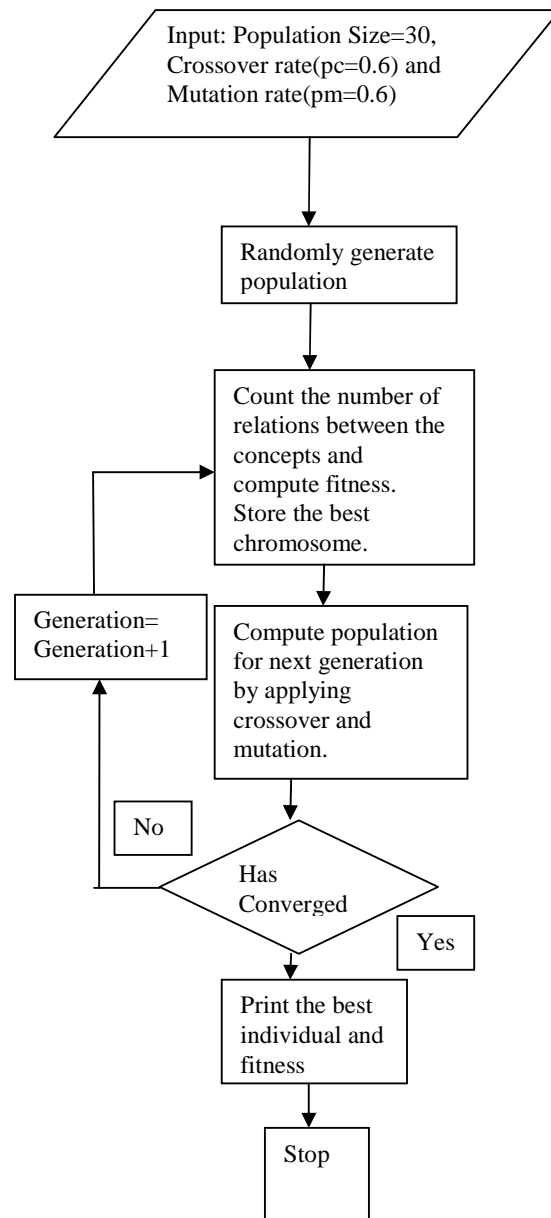


Fig.2. Flowchart for Genetic Algorithm

For the initialized offspring, the weight vectors of the offspring are defined in Eq. (6) and Eq. (7)

$$z_i = \begin{cases} z_i, & \delta > pc \\ \frac{(x_i + y_i)}{2}, & \delta \leq pc \end{cases} \quad (6)$$

Where  $\delta$  is the random number between 0 to 1 and  $pc$  is the crossover probability.

$$z_i = \begin{cases} z_i, & \alpha > pm \\ random(a, b), & \alpha \leq pm \end{cases} \quad (7)$$

where  $\alpha$  is the random number between 0 to 1 and  $pm$  is the mutation probability.

The crossovers are ranged from 0.5 to 1. The crossover used is the arithmetical crossover, which assigns the average

of two parents for each location of the offspring. The mutation operator preserves the diversity among the population which is very important for the search. For the experiment the mutation are ranged from 0.5 to 0.7.

#### IV. RESULTS AND DISCUSSIONS

The environmental science documents collected from USGS repository are annotated using Ontology and it is represented in RDF form using GATE Tool API. Using Jena SPARQL query model, the documents can be retrieved from the corpora not only based on their textual content but also according to their features or annotations. The concept relation graph generates the minimum edges covering the concept. The documents that are annotated with the concept are retrieved with higher ranking. The concepts will be defined as the document class and some attributes, which describe the document information. The predefined base ontology described based on USGS Scientific directory provides the basis for the semantic indexing of documents with no embedded annotations. Documents are annotated with concept instances from the KB by creating instances of the annotation class. The semantic information retrieval KB has been built and associated to the information document base by using domain ontology's that describe the concepts.

Documents are annotated with concept instances from the KB by creating instances of the annotation class. The query weights can be set by the user or automatically derived by concept frequency analysis [8]. In order to evaluate the model, precision and recall measures are applied. First the collection of HTML documents from USGS Science directory, a collection of 20 queries and a collection of relevant documents for each query are prepared for experiment. Table 1 shows the set of sample queries. The documents were annotated and stored. The GATE tool is used for the implementation. The weight for the query term is assigned and the query is run against the document taken forms USGS and returns the relevant information available. The similarity degree of each document is evaluated using the ranking model. It shows that weights learned using genetic algorithm with the document annotation has increased the precision of retrieved documents.

TABLE I  
SAMPLE QUERIES

minedrainage:Coal
contaminationpollution:Water Quality
contaminationpollution:Acid Rain
content:Water Quality
content:Acid Rain
environmental pollution
toxic

Once the experimental setting has been made, it is tested with the IR functionality in GATE. GATE comes with a full-featured Information Retrieval (IR) subsystem which uses the most popular open source full-text search engine – Lucene. In Gate IR the documents can be retrieved from the corpora not only based on their textual content but also according to their features or annotations. The ontology

annotation tool plug-in set available in GATE enables a user to manually annotate a text with respect to one or more ontology's. The required ontology must be selected from a pull-down list of available ontology's. The documents that are annotated with the concepts are retrieved with higher ranking.

The system takes the query and it is executed against the knowledge base and returns the matching documents. A query weight gives the importance of the concept in the information needed by the user. Several measures such as precision and recall are used to evaluate the performance of document retrieval. Precision is defined as the proportion of retrieved documents that are relevant. Recall is defined as the proportion of relevant documents that are retrieved

TABLE II  
TOP 5 DOCUMENT RANK COMPARISON

Document Name	Document rank without annotation weights	Ontolog based rank with GA concept relations
ten	0.0203	0.0246
nine	0.063	0.0644
second	0.184	0.1911
first	0.05	0.0581
five	0.084	0.0869

The observed results show the different levels of performance for different cases. The semantic information retrieval combined with the structural information improves the document ranking.

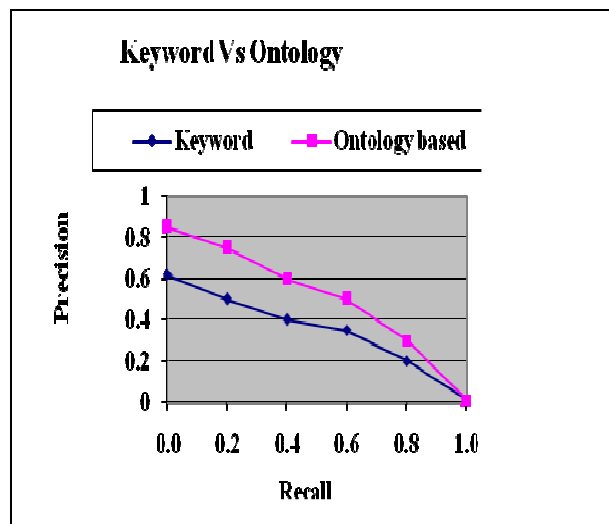


Fig.3. Performance Comparison of Keyword and Ontology based GA retrieval

Instead of simple keyword index lookup, the semantic search system processes a semantic query against the KB, which returns the relevant document. Better precision is achieved by using structured document annotation weight and genetic algorithm learning the concept relations. Table II shows the performance of retrieval based on keyword and with genetic algorithm for the query "minedrainage:Coal".

The precision and recall comparison of keyword and

ontology based retrieval system that uses genetic algorithm is depicted in Fig.3. Better precision is achieved by using structured document annotation weights learned by genetic algorithm. Table III shows the comparison of precision between keyword and genetic algorithm based document retrieval. The average precision of the top 10 documents for keyword retrieval is 0.376 and genetic algorithm based retrieval is 0.4119. The average precision is increased to 10% for the first top 10 documents.

TABLE III  
COMPARISONS OF AVERAGE PRECISION

Average Precision	Keyword retrieval	Genetic Algorithm based ontology annotation	Ratio
Top 10	0.3761	0.4119	1.09519
Top 20	0.1983	0.2074	1.04589

V. CONCLUSION

As an extension of the current web, semantic web provides a structured data and framework for knowledge representation of web information. It provides a technique to generate the metadata that semantically annotate a web page. The method combines ontology based annotation and uses genetic algorithm to form a concept relation graph based on the user keyword. The relations among the keywords are recorded and concept relation pair is generated. The web pages returned by this algorithm will be closer to the user’s intention since property keyword pair includes semantics of keywords and its relations. The approach can be seen as an evolution of the keyword based indices that is replaced by ontology based KB and a semiautomatic document annotation weighting procedure that improves the retrieval performance.

VI. ACKNOWLEDGMENT

The authors express their sincere thanks to the Management and Principal for their encouragement and support.

VII. REFERENCES

[1] Ahu Sleg, Bamshad Mobasher and Robin Burke, “Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search”, *IEEE Intelligent Informatics Bulletin*, 8, pp.7- 18, 2007.

[2] Dik L. Lee, Huei Chauang and Kent Seamons, ” Document Ranking and the Vector Space Model” , *IEEE Software*, pp.66-75, 1997.

[3] Guha, R, McCool,R and Miller,E ,” Semantic Search”, *International Conference on World Wide Web*, pp 700-709, 2003.

[4] Jose A.Alonso-Jimenez, Joaquin Borrego-Diaz, Antonia M. Chavez Gonzalez, Francisco and J. Martin-Mateos, “Foundational challenges in Automated Semantic Web Data and Ontology Cleaning”, *IEEE Intelligent Systems*, 42-52, 2006.

[5] Maedche, A, S.Staab, N.Stojanovic, R.Studer, Y.Sure, ,”Semantic portal: The SEAL Approach”, *Spinning the Semantic Web*, pp. 317-359, 2003.

[6] Mehrnoush Shamsfard, Azadeh Nematzadeh and Sarah Motiee., “ORank: An Ontology Based System for Ranking Documents”, *International Journal of Computer Science*, 1,225- 231, 2006.

[7] Pablo Castells, Mriam Fernandez, David Vallet,” An Adaption of

the Vector Space Model for Ontology based Information Retrieval”, *IEEE Transaction on Knowledge and Data Engineering*, 19(2), pp.261-22,2007.

[8] Sun Kim , Byoung-Tak Zhang, “ Genetic Mining of HTML Structures for Effective Web Document Retrieval”, *Applied Intelligence*, 18, pp.243-256, 2003.

[9] Wang Wei, Payam M.Barjaghi, Andrzej Bargiela, “Semantic enhanced information search and retrieval”, *Sixth International Conference on Advance Language and Web Information Technology*, pp218-223,2007.

[10] Yufei Li, Yuan Wang, and Xiaotao Huang, “A Relation- Based Search Engine in Semantic Web”, *IEEE Transaction on Knowledge and Data Engineering*, 19,273-282, 2007.

VIII. BIBLIOGRAPHIES

**U.K.Sridevi** received her M.C.A from Bharathiar University, India in 2001 and M.Phil degree from Manonmaniam Sundaranar University , India in 2003. Since 2008, she has been a senior lecturer in Department of Applied Sciences , Sri Krishna College of Engineering and Technology, India. Her research interests are in the areas of web mining, semantic web, data mining, and artificial intelligence. She is currently working towards her PhD degree at Anna University, India.

**N.Nagaveni** received the B.Sc., M.Sc and B.Ed degrees in Mathematics from Bharathiar University, India in 1985,1987 and 1988 respectively. M.Phil degree from Avinashilingam University, India in 1989 and M.Ed degree from Annamalai University, India in 1992. And Ph.D degree in the area of Topology in Mathematics from Bharathiar University, India in 2000. Since 1992, She has been with the Department of Mathematics, Coimbatore Institute of Technology, Coimbatore Tamil Nadu, India where she is currently as Assistant Professor. She is engaged as a research supervisor and her research interests includes topology, fuzzy sets and continuous function, data mining, distributed computing, web mining and privacy preservation in data mining. She is the member in Indian Science congress Association (ICSA). She has been presented many research papers in the annual conference of ICSA. She has been published many papers in the international and national Journals