# An Approach for Ontology Integration for Personalization with the Support of XML

S.Vigneshwari [1] and Dr. M. Aramudhan [2]

[1]Research Scholar, Department of Computer Science and Engineering, Sathyabama University, Chennai, Tamilnadu, India

jayam3@rediffmail.com

[2]Associate Professor and Head, Department of Information Technology, Perunthalaivar Kamarajar Institute of Engineering and Technology, Karaikal, Tamilnadu, India

aramudhan1973@yahoo.com

***Abstract* — Ontological way of knowledge representation is very much useful to the semantic web. In the modernized computer era, there is a need of a special technique for personalization. XML plays an important role in information retrieval systems and XML being a common format for information interpretation, it will be easy to understand as well as easy to construct. In this paper, a framework has been proposed for personalizing the web using XML based ontologies. This framework needs integration between global ontology and locally generated ontology based on user profiles. The relevant concepts between both the ontologies are identified, grouped together and ranked. Finally, the generated ontologies are evaluated using standard datasets, based on their semantic structures. The clustered concepts and query pairs are being analyzed with varying threshold limits. In addition, the performance metrics show that the ontology based techniques show a good precision, recall values for the user given data, when compared to text-based approaches.**

***Keywords*-** ontology, XOL, preprocessing, URL, W3C, ontology integration, information retrieval, relational databases, XML Table, UPO, OIS.

## I. INTRODUCTION

In the semantic world, concepts are given much more importance than usual lexicons. A semantic model is required for the identification of concepts, which are the user, preferred ones. Building of such a model should serve the semantic world with its enduring performances such as scalability and adaptability in a distributed environment. Ontologies are chosen as a best way for the construction of the semantic web. Conversion of ontological database into XML or XML based OWL is very easy and it can be done automatically. So XML based ontologies are preferred for web page personalization. Such an approach of ontology mining can be very much useful to the semantic society.

In this paper, the basic techniques and definitions related to ontologies like ontology alignment, ontology languages and OWL specifications are discussed in section III.

Section IV explains about XML data model. Here, the basics for the generation of XML graph, and accessing the data in XML database with the help of XML-QL, are being explained with example.

Section V A explains about the need for ontology integration. The structure of ontology integration framework is being explained in V B. Section V (C,D,E) plays an important role in understanding the relevance of concepts and grouping of concepts with the help of grouping algorithm, and the similarity of concepts between two different ontologies, with the help of Mutual information technique. Finally, in Section V F, the advantages of using similarity measures are being explained.

Section VI explains the experimental setup by the construction of a framework for XML based personalized ontologies. Section VI (B, C, and D) explains about how to preprocess the user input. The steps for preprocessing are explained using the probabilistic function of the weighted terms. In Section VI D, how the classification of text is done, is being explained. Section IV E shows, how the web search be personalized. In

Section VII results are displayed and analyzed in various aspects. Mainly Section VII A shows the analysis of the semantic structure of the constructed ontologies. Section VII C shows the analysis of similarity between the concepts with and without using ontologies. In Section VIII, conclusion and future work are given.

## II. RELATED WORK

Xiaohui Tao et al [3] proposed a personalized ontology model, which combines the search results from both world knowledge base and local instance repositories. Based on the exhaustivity and specificity of topics, the documents are organized into positive documents, negative documents and neutral documents. Here multidimensional ontology mining approach is followed. The generated ontology model is compared against the

benchmark models and the result show good improvement of performance compared to other models. This model plays an important role in the motivation behind research.

Peter D. Karp et al [4] suggested XOL an xml-based ontology exchange language, which is best suited for sharing ontologies in a distributed environment. XOL can also be used for translating the SQL query of a relational database into XQuery of XML database.

Vigneshwari et al [16] devised a technique to extract the interesting measures using ontology mining. Here the balanced mutual information is used for finding the similarity between two concepts in the same ontology.

Shashank et al [20] created a model, which assumes that the words are independent of each other. Instead of treating a document as a merely bag of words, the similarity or distance measures are calculated. This model shows that the concepts should be taken in a semantic manner instead of mere text.

Jayabharathi et al[8] created a hierarchical clustering algorithm using semantic similarity. This approach will have better clustering results than that of the partial clustering algorithms which uses merely cosine similarity measures.

Songyun Duan et al [13], suggested a cluster based approach to align different ontologies. In his approach, the medical datasets are analyzed, where internal partitioning of the same domain takes place. The ontology alignment is done based on the mapping across the sub- domains of the medical ontology. A similar approach is followed in our work, where the queries are left up to the user, and based on the user's query log, the global ontology is constructed, thereby following the URL, and the local ontology is automatically generated from the user given query tags.

Bagheri Hariri et al [21] developed a supervised neural network model to generate compound metrics for cross- ontology mapping. In our proposed work, based on the user search history, the first ontology is generated automatically. This is internally partitioned as clusters. Based on the user's continuous queries the second ontology is automatically generated from the previous one. The main scope of this paper is to understand the concepts, which are used for cross ontology mapping.

Heasoo Hwang et al [14], proposed a robust approach to organize user queries into groups dynamically and automatically. Search behavior graphs like query reformation graph, query click graph, query fusion graph were generated, and it was experimentally proved that query automation is very much useful for a collaborative search in his work. Dynamic query grouping also plays a significant role in organizing the queries given by the user. This is also important for the construction of ontologies.

Yanhui et al [9] proposed a flexible mechanism to integrate ontologies in multi ontology based system. A framework for ontology integration, which combined both ontology similarity measures and ontology integration algorithms, had been suggested in that work. The integrated ontology is evaluated and checked for consistency.

Narayana et al[22] proposed an approach to discover and rank semantic associations on semantic web and finally the associations are experimented using SWETO data set. The concepts are ranked according to the association ranking methodology.

## III. TECHNIQUES AND DEFINITIONS

### *A.Ontology definition*

Ontology is a representation of knowledge in a particular domain as a set of concepts and their relationships .Formal definition of ontology is given as follows.

Let *cls* be the class, *rel* the relationship between the classes, *attr* the attributes, and *ind* be the individual instances.The ontology O is defined as

$$O=( cls ,rel, attr, ind) \quad (1)$$

Based on the ontology definition, the major components are identified as classes, their relationships, the attributes and individual instances.

### *B.Ontology alignment or Cross ontology mapping:*

The other name for ontology mapping is called ontology alignment. It is a process, which is used to find out the relationships between the concepts. From the definition of ontology, the definition of ontology alignment can be derived. Let the first ontology be X and its components be ($cls_x$ ,$rel_x$, $attr_x$, $ind_x$) and the second ontology be Y ontology and its components be ($cls_y$ ,$rel_y$, $attr_y$, $ind_y$) , then cross ontology mapping is said to be the matching of the components of ontology X onto the components of the ontology Y. Various means of ontology mapping techniques are available, depending on the types of ontologies. Two major types of ontologies are homogenous ontologies and heterogeneous ontologies. If all the concepts are in the same domain then the ontology is called homogenous ontology. If the concepts refer different domain, then the ontology is called heterogeneous ontology. The query components of the ontologies may also belong to atomic type or complex type. Each and

every node in the ontology of a particular domain is called a concept. There are two types of relationships between the concepts in ontology. The relationships may be is-a relation or part-of relation.

*C.Ontology languages*

Early ontology languages were based on either HTML or XML. Previously XML based ontology exchange language (XOL) was used. XOL, follows generic approach of defining ontologies. In XOL, a single set of XML tags can be used to describe the ontologies. Then it was upgraded into Ontology Integration Language (OIL), which lays the foundation for merging of ontologies or ontology integration. Then the family comprising of knowledge representations called Web Ontology Language, (OWL) family was emerged, which is purely based on XML schemas. There are two specifications of OWL. They are OWL 1.1(2007) and OWL 2 (2009).This was recommended by World Wide Web Consortium (W3C ).

*D. OWL specifications*

OWL1.1 has its syntax defined in XML schema language and OWL 2 is an ontology language for the semantic web with formally defined meaning. OWL2 has XML serialization, which mirrors the structural specifications. OWL2 found its way into the semantic editors like protégé, and semantic reasoners like Hermit.

The following Figure 1, is an XML schema for some sample ontology classes like authors and books, defined by OWL 1.1

```
<?xml version="1.0"?>
…
…
…
<rdf:RDF>
<!-- http://www.semanticweb.org/viki/ontologies/2013/9/
            untitled-ontology-16#author -->
              <owl:Class rdf:about="http://www.semanticweb.org/viki/ontologies/2013/9/
                    untitled-ontology-16#author">
              <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/viki/
                 ontologies/2013/9/untitled-ontology-16#novel"/>
           </owl:Class>
</rdf:RDF>
<!-- Generated by the OWL API (version 3.3.1957) http://owlapi.sourceforge.net -->
   <!--                  http://www.semanticweb.org/viki/ontologies/2013/9/untitled-ontology-
16#author_name -->
   <owl:Class          rdf:about="http://www.semanticweb.org/viki/ontologies/2013/9/untitled-
ontology-
                                16#author_name">
      <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/viki/ontologies/2013/9/
                                untitled-ontology-16#text_Book"/>
   </owl:Class>
```

Figure 1. A sample XML schema defined by OWL1.1

## IV. XML DATA MODEL

*A. XML syntax and XML graph*

XML elements are tagged using user defined tag names. The syntax of XML tag is *<tagname>.....</tagname>*. Let us take a graph $G_x$ for the particular ontology X. The attributes of $G_x$ are vertices with distinct element identifiers, edges, labels.

$$G_x = (v,e,l) \quad (2)$$

where v is the set of vertices, e is the edge and l is the label. The types of edges in $G_x$ may be content edges or attribute edges. These edges can also be labeled using the attribute, l.

*B. XML Data Example*

book.xml

```
<document>
        <book >
    <type>
      <text_book>
        <author> Tanenbaum</author>
        <book_name>
              Computer Networks
        </book_name>
        <price>$6</price>
        <author>C J  Date</author>
        <book_name>
              Database Systems
        </book_name>
        <price>$5</price>
      </text_book>
      <novel>
        <author>Dan </author>
        <book_name>
              abc
        </book_name>
        <price>$10</price>
      </novel>
    </type>
  </book>
</document>
```

Figure 2. An example XML data file

Figure 2 represents a sample XML data file 'book.xml', for which the portion of the XML tree representation is given in Figure 3. An XML tree has nodes connected to each other by data labels. Each node of XML tree is having an element identity and the labels are the tags in XML data. An element identity is a numbered one. The XML tree has a single root, intermediate level of element identity nodes and leaf nodes. At the leaf level, the values of the XML tags are given.

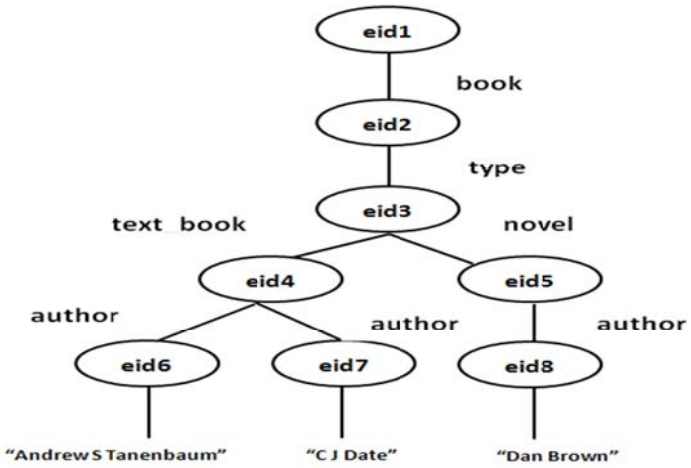


Figure 3. A portion of the sample XML tree

In Figure 3, eid'*n*' represents the element identity and eids are connected with each other using data labels.

*C. XML-QL Query Language*

The general syntax of XML QL is:

***where** (XML-pattern [Element-As $var]) *IN filename, (predicate)* **construct** XML-pattern/variable.*

For example let us take the sample query like "search for the author named Dan Brown who wrote the novel called 'Angels and Demons', then the syntax of XML QL will be

```
where  <type><novel><author>
            $N
       </type></novel></author>
in "book.xml", $N like *Dan* construct $E.
```

The results will be retrieved based on the query–to-data mappings.

*D. Creation of a simple XML table in native XML database*

The general syntax is:

*Create table name( doc XML)*

```
Select X.* from book, XML Table ('$t/type/text_book/author' passing doc as "t" COLUMNS author varchar (20) PATH 'type/text_book/author' AS X
```

On running the above query in DB2 the following COLUMN is created

TABLE I
A column in XML table

| **Name** |
|---|
| Tanenbaum |
| C J Date |

Table 1. shows a sample column in XML table. The XML table function has row-generating Xquery expression and the COLUMNS clause, which has one or more column generating expressions [5].

## V. ONTOLOGY INTEGRATION FRAMEWORK

*A. Need for ontology integration:*

Usually, the developers of ontology-based databases prefer to adopt existing ontologies than creating a newer one, since the user-developed ontologies are time consuming and tedious ones. Multidatabase query problem arises, when more number of ontologies exists for the same data type. It also leads to semantic mismatch problem. Various schemas are needed to be developed to follow semantics. This will lead to gather full knowledge about the database. Since distributed databases are shared ones, ontologies also needed to be developed in a shared manner.

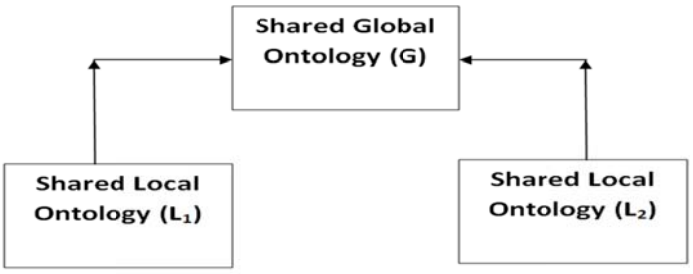


Figure 4. Global ontology shared by Local ontologies $L_1$ and L2

In Figure 4, G represents the shared Global ontology, and L1 and L2 represents the shared local ontologies. The above Figure illustrates the ontology integration, which is the mapping of the concepts between shared local and global ontologies.

To understand the above concept in a better way, let us consider two ontology classes like author class and book class as represented in Figure 5.

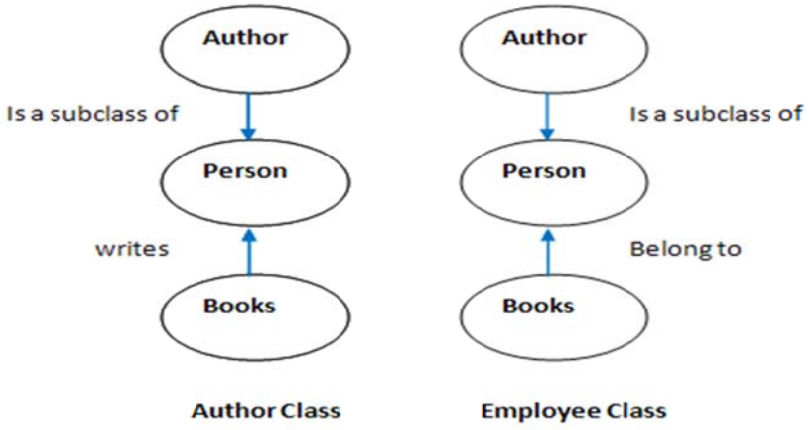


Figure 5. Representation of two distinct ontology classes 'author' and 'employee'

In Figure 5, two different ontology classes are introduced. One is the author class and the second one is the employee class. Both the classes have a common concept called Person. In this, we can merge the entity person, which leads to the integration of two distinct classes into a single one.

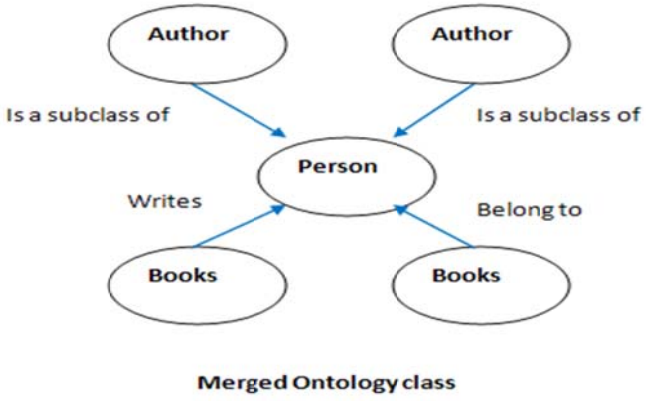


Figure 6. A merged ontology class

In Figure 6, the person concept, which is a common entity for the classes, author and employee, is being shared. From this, the idea of sharing of concepts between ontologies is clearly understood. This lead to the newer technique of ontology integration by finding the similarity measures between the concepts of the shared local ontologies, which are can be further shared by other ontologies.

*Structure of ontology integration framework*

The ontology integration framework consists of the following

- A global ontology, G
- A local ontology, L
- Mapping of L onto G, $M_{G,L}$
  - Assigning the semantics of Ontology Integration System(OIS) as :
    $Sem(G,L,M_{G,L})$ (3)
    Equation (3) represents the semantic function sem(..), which is the collection of G, L, $M_{G,L}$
  - Queries(Q) which are the collection of tuples of concepts extracted, on posing a query to the OIS
    $Q=\{c_1,c_2,\ldots c_n\}$ (4)
    Where $c_1,c_2,\ldots c_n$ are concepts which are the instances in the local ontology which also exist in the global ontology. But the depth of the concept may differ in both the ontologies [5].
  - Relationships between the concepts can be is-a, part-of, has-a etc.

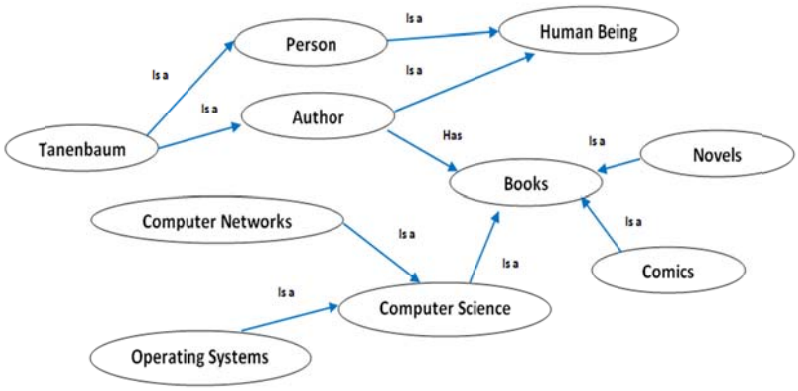


Figure 7. An example ontology describing author of a book

In Figure 7, sample ontology has been given which describes the author of book. In the above Figure, the relationships between the concepts such as is-a, has etc are described. Let us consider the example "Tanenbaum is an author (is a relationship) who has (has a relationship) written books in computer science (is a relationship)"

Multiple local ontologies can be mapped onto a global ontology.

$$Person(x) \leftarrow L_1(x)$$
$$Book(y) \leftarrow L_2(y)$$
$$Human\ Being(x) \leftarrow L_3(x)$$
$$Author(x) \leftarrow L_4(x)$$

In the above example, the concepts on the left hand side are global concepts and that on the right hand side like $L_3(x)$ are set of local concepts.

*B. Relevance of concepts*

Query relevance can be computed based on time-based relevance metric and by computing textual similarity.

$$Sim_{txt}(c_1,c_2)=\frac{|words(c_1)\cap words(c_2)|}{|words(c_1)\cup words(c_2)|} \quad (5)$$

Textual similarity is based on jacard similarity [7].The main issue in using textual similarity between the concepts, is lack of semantics. For example, bean and java bean can be identified as similar concepts and can be grouped under the same category. But if the bean is intended to be a vegetable, then both of the above concepts will come under different categories. So we need a semantic similarity measure to find the relevance between the concepts.

*C. Grouping of concepts*

Grouping up of concepts plays an important role in ontology integration. An automated, unsupervised, semantic approach is required which can be done dynamically with the support of ontologies.

***Algorithm for grouping the concepts***

***Inputs:*** $c_{gp}$ (Current concept group )for a single user in an existing ontology

$C_{cur}$ refers to the current concept

Set of existing concept groups $C=\{c_1,c_2,...c_n\}$

$Thr_{sim}\leq 0\leq Thr_{sim}\leq 1$, where $Thr_{sim}$ is the similarity threshold value

***Output:*** Concept group c ,that matches $c_{gp}$

(0) $c=\phi$
(1) $Thr_{max}=thr_{sim}$
(2) For i= 1 to m
  (2.1) if $sim(c_{gp},c_i)>Thr_{max}$
    (2.1.1) $c=c_i$
    (2.1.2)$Thr_{max}=sim(c_{gp},ci)$
  (2.2) End if
(3) if $c=\phi$

(3.1) $c=c \cup c_{gp}$
(3.2) $c=c_{gp}$
(4)End if
(5) return $c_{gp}$

*D. Calculation of concept similarity across different ontologies*

In a document vector model, mutual information is a non-negative and symmetric one. Let $I(c_g,c_l)$ be the mutual information between the concepts $c_g$ and $c_l$, where $c_g$ is a concept in the global ontology and $c_l$ is a concept in the local ontology. $I(c_g,c_l)$ is calculating using the following formula:

$$I(c_g,c_l)=\sum_{c_l \epsilon L} \sum_{c_g \epsilon G}^{P(c_g,c_l)} \left(\frac{P(c_g,c_l)}{P(c_g),P(c_l)}\right) \quad (6)$$

Here $P(c_{g,}c_l)$ is t.he probability of the co-occurrence of the concepts in both the global ontology and in the local ontology.

$P(c_g)$ is the probability of occurrence of the concepts in the global ontology alone and $P(c_l)$ is the probability of occurrence of the concepts in the local ontology alone .

*E. Advantages of using ontology based similarity measure*

When compared to the usual text based clusters, the main drawback is that, it is tedious to calculate the similarity or dissimilarity between the two concepts, as the requirement of results vary from user to user for the same query. The factor is that most traditional clustering algorithms treat the set of documents a mere bag of words. The main advantage of using ontology based similarity measure is that, since ontologies are created by domain experts, they will be more precise. When compared to other methods, ontology based methods are computationally more efficient. The integration of domain knowledge into data mining process is also made easier with the help of ontology mining.

## VI. Experimental Setup

*A. A framework for XML based personalized ontology*

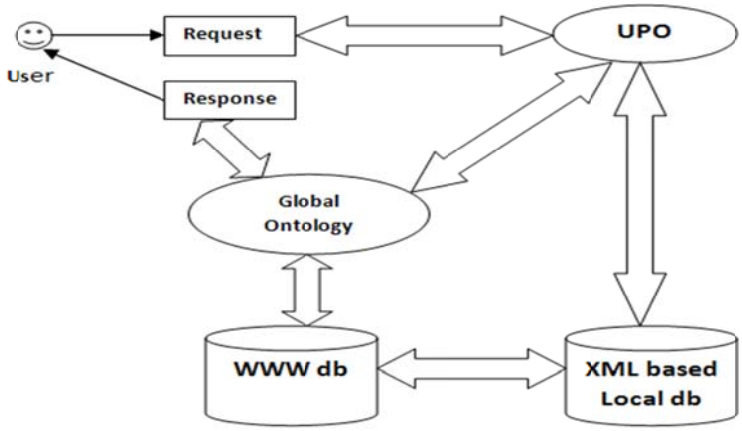


Figure 8. Proposed architecture for XML based ontologies

Figure 8 represents the framework for the construction of XML based personalized ontology. The request from the user is analyzed in the UPO (User Profiling Ontology). UPO has its backup in XML based Local database. In other words, UPO is nothing but the local ontology, being shared with the global ontology and the XML based local database. If the data is not present in the local XML database then the global information available from the World Wide Web(WWW) is taken and sent to the user, as well as the same is updated in the local database. Based on the continuous user queries, the web pages are ranked in the local database, thereby personalizing the web.

*B. Preprocessing the user input:*

The user logs are pre-processed into tokens. Using WordNet 2.1, the meaningful words are extracted and the global ontology is constructed.

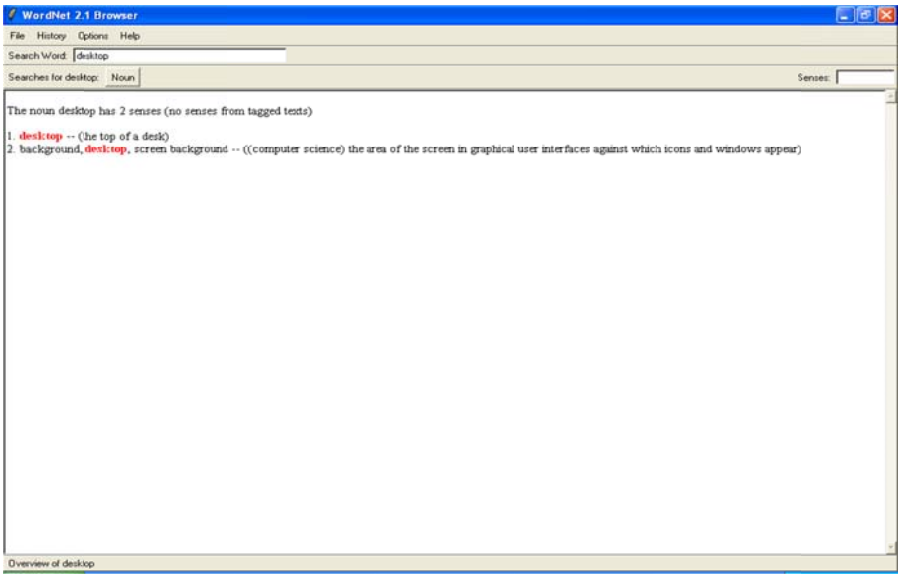


Figure 9. A sample WordNet browser

Figure. 9 shows a sample WordNet browser for browsing the word 'desktop'. The user queries are stored in tag-url format in the user log file, as given in Figure 8. Log files are extracted from the log base, which is otherwise the XML database, at different time periods by a single user.


Figure 10. A sample log file

Figure 10 shows the set of keywords and URLs browsed by a single user at different times. For experimental purpose, the pages from Wikipedia are considered in this file.

*C. Steps for processing and extracting the useful information*

- Documents fetched from the web sites are preprocessed. Initially the documents are classified into positive and negative documents [10] and the weights of the terms are calculated.
- Let D be the document , ti be the terms chosen, fri is the frequency of those terms. Then the document ,

$$D=\{(t_1,fr_1),(t_2,fr_2)...(t_n,fr_n)\} \quad (7)$$

where $fr_i$ is frequency of the term $t_i$ .The document frequency in a semantic space is calculated as

$$F(D)=\{(t_1,wt_1),(t_2,wt_2)....(t_n,wt_n)\} \quad (8)$$

where $wt_i(1..n)$ is the weight distribution of terms $t_i$.

$$wt_i= \frac{fr_i}{\sum_{j=1}^{n} fr_j} \quad (9)$$

- A probabilistic function p for the terms selected is derived as p(t) where

$p(t)=\sum_{D\in Doc^{+}(t,wt)\in F(D)} support(D) * wt$ (10)

and $Doc^{+}$ is a positive document. The weight of the individual terms is calculated.

- Based on the weight of the positive documents the local XML database is mapped onto the global ontology

*D. Classification of the Text*

*1) Training the document sets:*

The document features are based on the user input. The training sets are extracted and sent to machine learning algorithm for pattern recognition.

Before the document features are extracted and trained, the raw data has to be preprocessed. Preprocessing comprises of segmentation of terms in each sentence, followed by tokenization, stemming, removal of stop-words, parts of speech tagging. After preprocessing, the individual terms are extracted and the relationships between the terms are identified. The related terms are clustered and classified into a particular domain. Figure 11 shows the preprocessing of documents and tokenizing the terms.

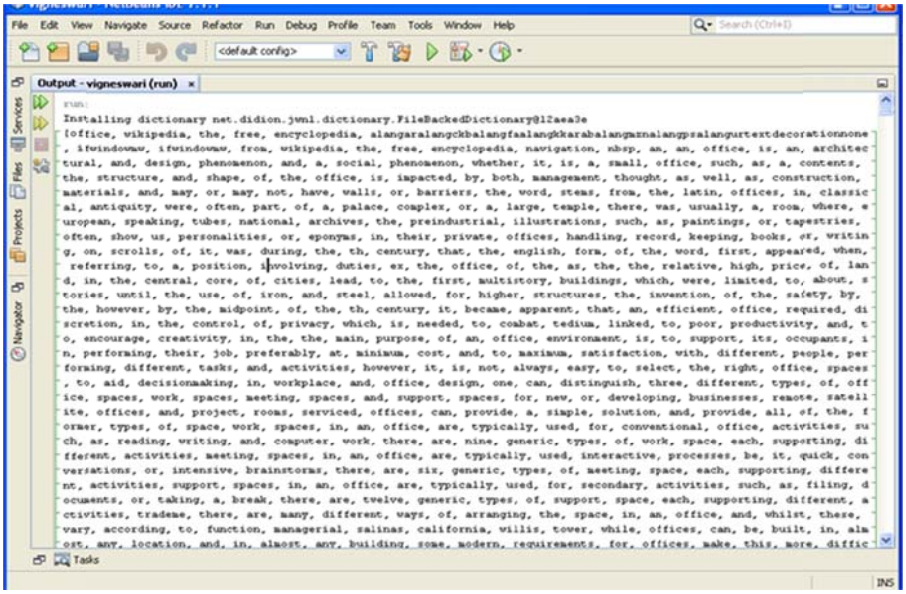


Figure 11. . Token extraction

*2) Relevant Pattern Prediction:*

From the results of the machine learning algorithms, a classifier model can be constructed and trained. This approach is a supervised learning approach. Grouping the terms is done, after stemming and stop word removal. The relevant concepts are predicted and clustered by k-means clustering algorithm.

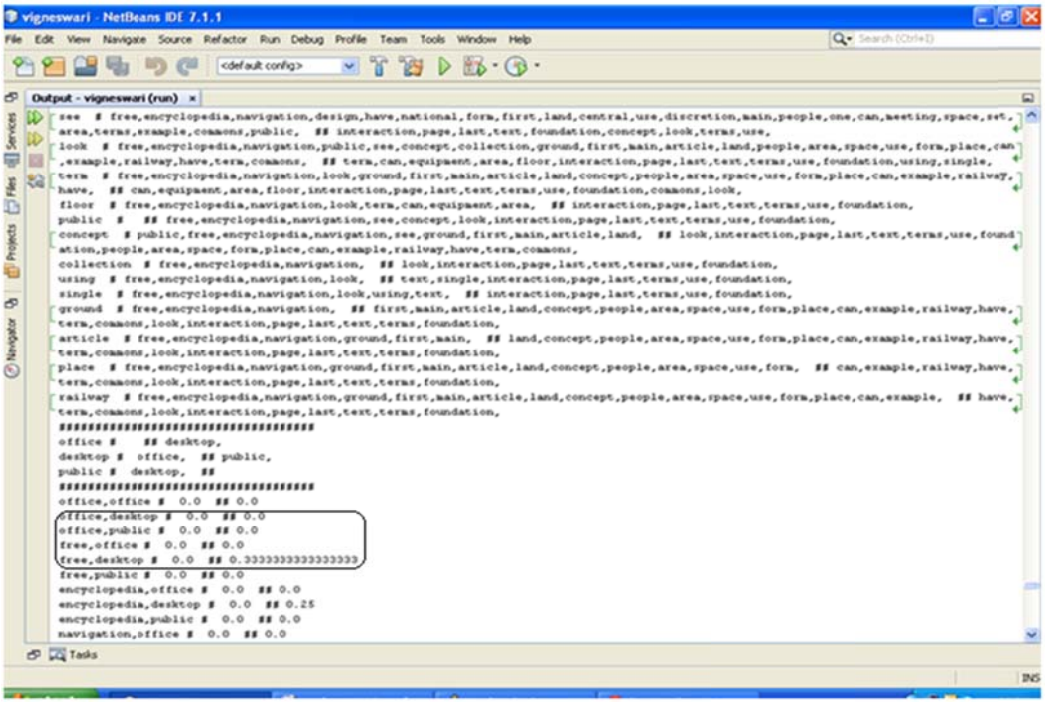


Figure 12. Ranking a grouping of concepts

Figure 12 shows the experimental results of the concept grouping. The concepts are grouped together using the grouping algorithm, as discussed in the above section(V-D). Before grouping up of the concepts, the concepts have to be ranked based on their weights. Weights are calculated based on the term frequency of the documents(eqn(9)), and then a probabilistic function is applied to the weights to find out their support in the document(eqn(10)) as discussed earlier, thereby making this in a semantic way. Then mapping up of the concepts between the local ontology and global ontology takes place,which is based on the mutual information between the concepts between the two ontologies(Eqn 6).

*E. Personalizing the web search*

The search history is updated in the users log base. The retrieved web pages are classified into relevant pages, non-relevant pages and ambiguous pages. Relevant pages, which are commonly visited by the user, are ranked top based on the search histories of the user. Usually in the Information Retrieval systems the stop words(*the, is, and,…*)are ignored. The stemmer is useful for stemming the words. A sample personalized web page can be created like the one shown in the following figure, Figure. 13.

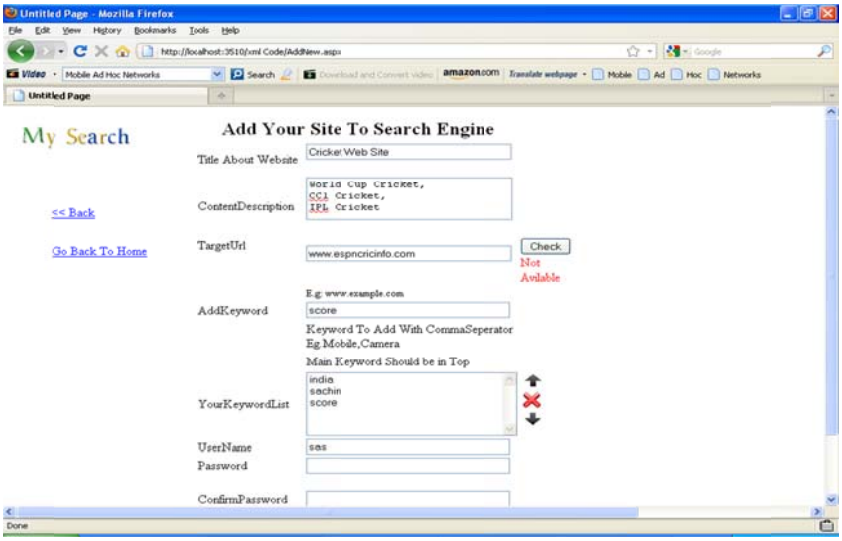


Figure 13. User Update on Personalization

## VII. Results and Discussions

### A. Analysis of the semantic structure of the constructed ontologies:

The tokens are classified using tokenizers, and the parse tree is generated. The meaningful terms and the relationship between the terms are identified. Finally, the precision, recall, F-Measure values of the relevant retrieved documents are calculated.

Precision is defined as the ratio of total number of relevant retrieved documents, to the total number of retrieved documents.

$$Precision=\frac{total\ number\ of\ relevant\ retrieved\ documents}{total\ number\ of\ retrieved\ documents} \quad (11)$$

Recall is defined as the ratio of total number of relevant retrieved document, to the total number of relevant documents in the xml database.

$$Recall=\frac{total\ number\ of\ relevant\ retrieved\ documents}{total\ number\ of\ relevant\ documents} \quad (12)$$

The weighted harmonic mean, *F- Measure* has been given in eqn (13)

$$F\text{-}Measure=\frac{2.precision.recall}{precision+recall} \quad (13)$$

### B. Datasets used

The datasets comprises of queries belonging to different semantic clusters. The datasets are taken from SWETO package. The corresponding web site is: "http://archive.knoesis.org/library/ontologies/sweto/".

The cluster results are obtained with the k-means clustering algorithm. The similarity of individual terms is taken and compared with that of the global ontology using balanced information concept.

TABLE II
A sample data set

| **Classes subsets** | **No. of Instances** |
|---|---|
| Cities, Countries, states | 2902 |
| Airports | 1515 |
| Companies and banks | 30948 |
| Persons and researchers | 307417 |
| Terrorist attacks and organizations | 1511 |
| Scientific Publications | 463270 |
| Journals conferences and books | 4256 |

Table 2. represent a sample dataset taken from SWETO as of January 2004 from LSDIS database. The datasets comprise of different documents, the corresponding classes, terms and size in Kilo bytes. After clustering, similarity is calculated. Then the Home Clusters and query pairs having similarity, is calculated for various threshold limits as shown in Table III.

TABLE III
Similarity table

| **Threshold** | **Total number of clusters with similarity** | **Total number of query pairs with similarity** |
|---|---|---|
| 0 | 60 | 200 |
| .25 | 30 | 100 |
| .5 | 15 | 20 |

The graphical representation of Table 3, shows the threshold variations in the clustering of queries

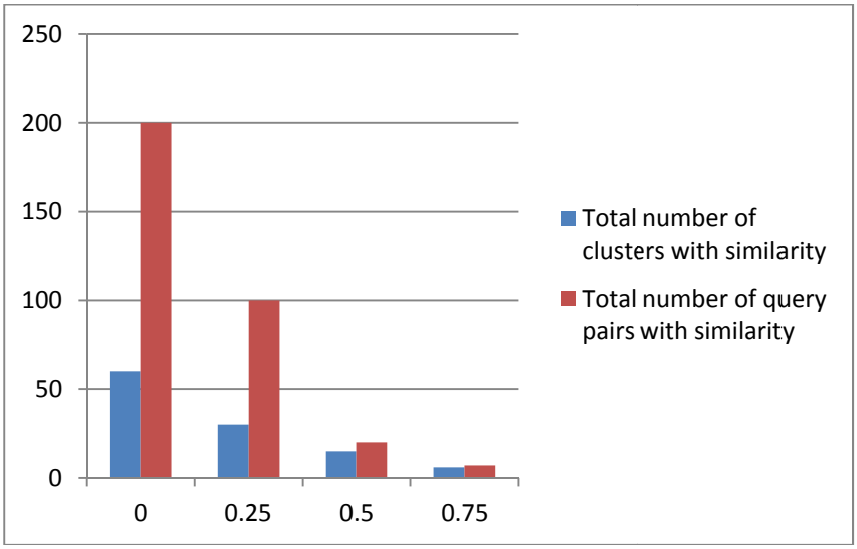


Figure 14. Threshold variations in query clusters

Figure 14 shows the threshold variations in using different query clusters. The results show that, on using lower threshold values, more similarity is found when compared to the higher thresholds. After clustering, the global ontology and the local ontologies are automatically generated. Figure. 15 shows a sample global ontology which has been generated automatically using protégé OWL software (OWL 1.1), since OWL supports XML.

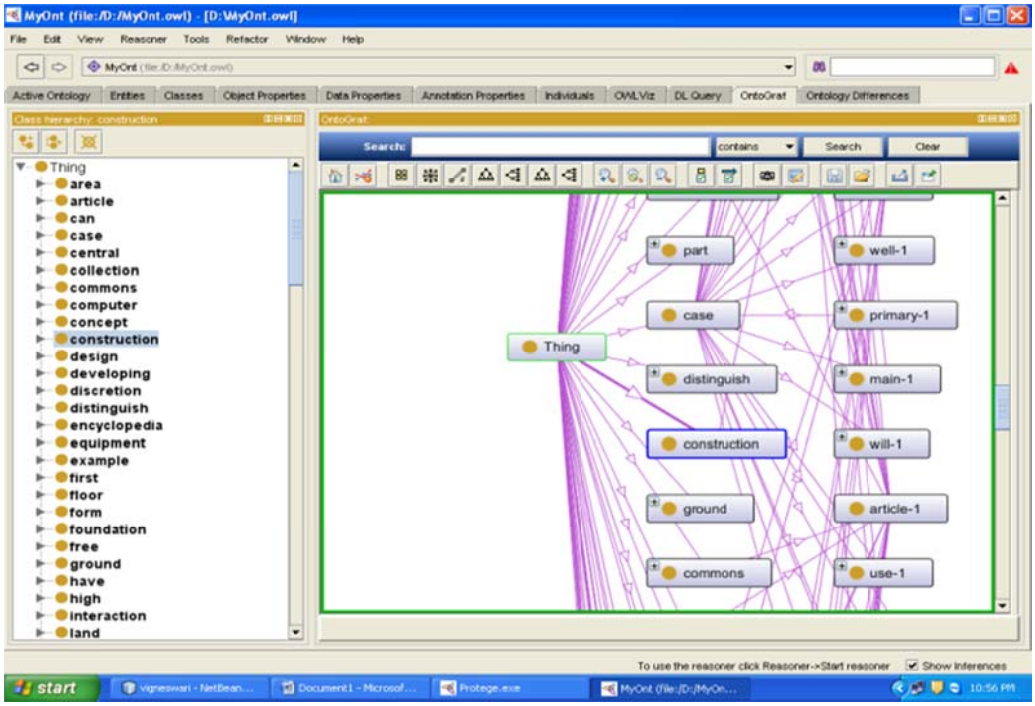


Figure 15. Global ontology

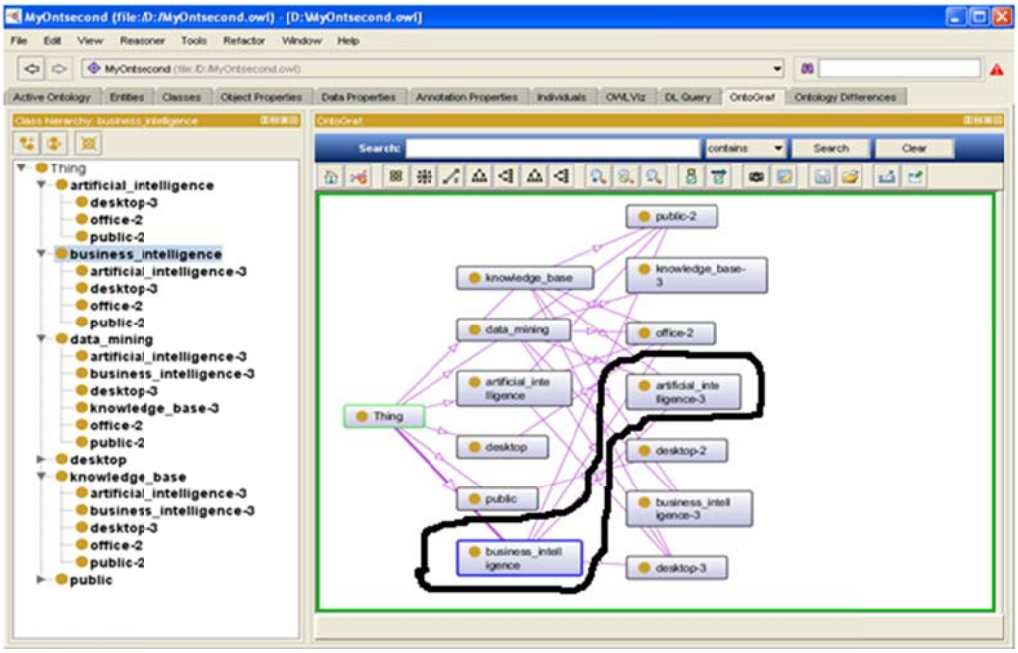


Figure 16. A sample integrated Local ontology

Figure 16 shows the generated local ontology. The sequence of the queries given by the, users are analyzed, mapped with the concepts in the global ontology and are ranked. In, the leaf nodes, the ranks are displayed. For example in the above Figure, the query-pair (business intelligence, artificial _intelligence) is ranked 3, based on the mapping with global ontology.

*C. Analysis of similarity between the concepts using ontologies and without using ontologies*

Some of the keywords searched by the user, and their precision, recall, F-measure using XML database using the ontologies and without using the ontologies is discussed below.

TABLE IV
Similarity measures using XML based ontology

| Keyword | University | Database | Novel | Text Book |
|---|---|---|---|---|
| **Precision** | .78 | .44 | .66 | .88 |
| **Recall** | .55 | .52 | .57 | .78 |
| **F-Measure** | 1.33 | .96 | 1.22 | .82 |

Graphical representation of the above data from **T**able 4 is given in Figure.17

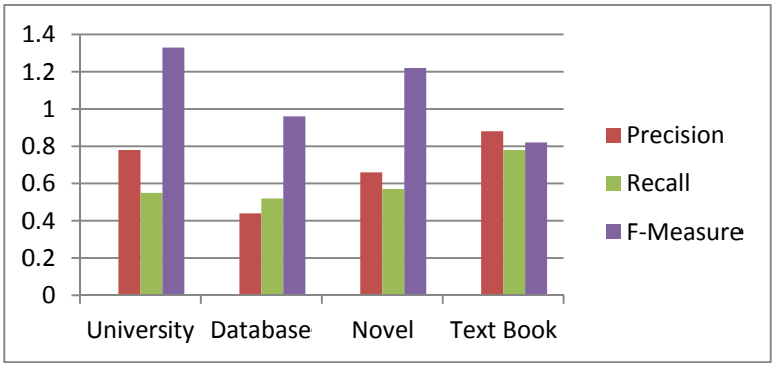


Figure 17. Precision, Recall, F-Measure values of retrieved documents from the database using ontologies

Similarly, the usage of k-means clustering algorithm or the mutual information to find the similarity of the same concepts can be found without using ontology, is discussed in Table 5.

TABLE V
Similarity measures using text based clustering without ontologies

| Keyword | University | Database | Novel | Text Book |
|---|---|---|---|---|
| **Precision** | .74 | .40 | .36 | .38 |
| **Recall** | .25 | .50 | .45 | .62 |
| **F-Measure** | 0.37 | 0.44 | 0.40 | 0.47 |

Graphical representation of the above data is given in Figure.18

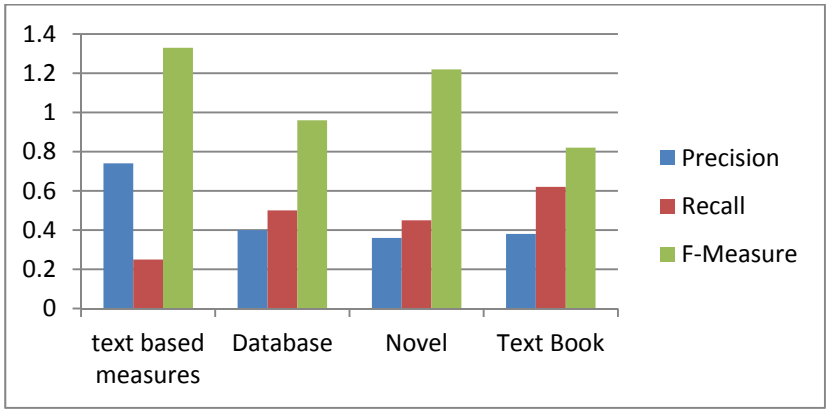


Figure 18. Precision, Recall, F-Measure values of retrieved documents without using ontologies

Average of precision, recall and f-measure values in using text-based approach and ontology mining approach as shown in the following Figure 19 indicates, that the ontology mining approach is good when compared to the text based ones.

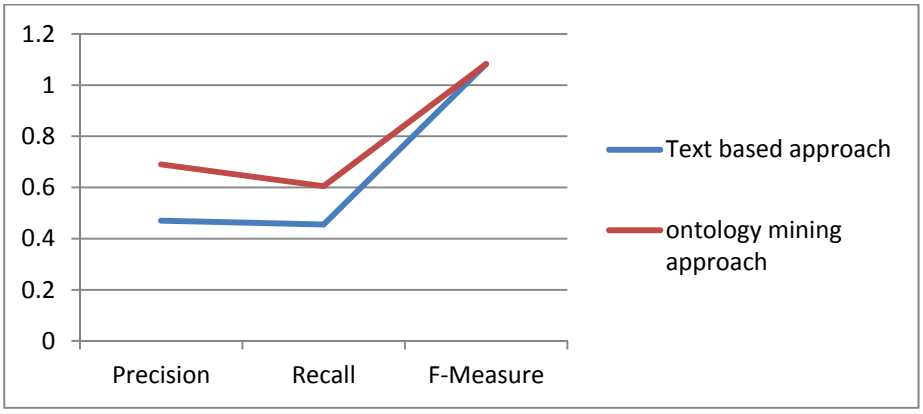


Figure 19. Comparison of similarity measures using ontologies and without using ontologies

## VIII. Conclusion and Future Work

XML based approach is globally accepted as a common one. XML based ontologies can be constructed to improve the overall performance of the system by calculating similarity measures. Results show that the clustering of the concepts using ontologies show a good precision and recall metrics when compared to the text based clustering measures. Usage of personalized ontologies can enhance the web based information retrieval in a more efficient way. More fast algorithms can be implemented and analysed as a future development of this

work. As well as, the relationships between the concepts have to be further strengthened in the locally generated ontology which is left as a future part of this work.

## ACKNOWLEDGEMENT

I am grateful to my research supervisor, Dr. M. Aramudhan for his collaboration and support during preliminary investigations of this work. I would like to thank the reviewers for the efficient preparation of this paper. I would also like to thank Sathyabama University, for providing facilities for the eminent preparation of this paper.

## REFERENCES

[1] Ming Mao, Yefei Peng, Michael , "An adaptive ontology mapping approach with neural network based constraint satisfaction", *Journal of web semantics* Vol 8, No 1 (2010)

[2] A.V. Aho, J.E. Hopcroft and J.D. Ullman, "On finding lowest common ancestors in trees," *Proceedings of 5th annual ACM symposium On Theory of Computing*, 1973.

[3] Xiaohui Tao, Yuefeng Li, and Ning Zhong. "A Personalized Ontology Model for Web Information Gathering", *IEEE transactions on knowledge and data engineering*, Vol. 23, No. 4, pp. 496-511, 2011.

[4] Peter D. Karp, Vinay K. Chaudhri, and Suzanne M. Paley. "A Collaborative Environment for Authoring Large Knowledge Bases," *Journal of Intelligent Information Systems*,Volume 13 Issue 3, Nov.-Dec. 1999 Pages 155-194

[5] S.Vigneshwari, Dr.M.Aramudhan." An ontological approach for effective knowledge engineering", *Proceedings of International Conference on Software Engineering and Mobile Application Modelling and Development, ICSEMA-2012*, Chennai , India, pages 331–341, 2012

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft."When is 'nearest neighbor' meaningful", Proceedings of International Conference on Database Theory ,ICDT-1999, Jerusalem, Israel,1999, pages 217–235, 1999

[7] Tversky, A., "Features of Similarity", *Psychological Review*, 1997 84(4):P. 327-352

[8] J.Jayabharathy, S. Kanmani and A.Ayeshaa Parveen,"A survey of Document Clustering algorithms", *Journal of computing*, Vol 3, Issue2,February 2011

[9] Yanhui Lv ,Chong Xie,"A Framework for Ontology Integration and Evaluation",*Proceedings of Intelligent Networks and Intelligent Systems (ICINIS)*, 2010 Nov. 2010 Page(s): 521 - 524

[10] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Kohler, Anand Kumar, Jane Lomax, Chris Mungall, Fabian Neuhaus, Alan Rector, Cornelius Rosse."Relations in Biomedical Ontologies", *Genome Biology*, 2005, 6 (5), R46

[11] S.Vigneshwari, Dr.M.Aramudhan," Enhancing the keyword search on XML documents thereby personalizing the web-Acomparative approach", *CIIT International Journal of Data Mining and Knowledge Engineering*, June 2012, 5(6):34-39 ISSN 0974 – 9683.

[12] M.Berry, M. Browne,"Lecture notes indata mining", *World scientific publishing company*, 2006

[13] Songyun Duan, Achille Fokoue, Kavitha Srinivas, Brian Byrne. "A Clustering-based Approach to Ontology Alignment", *Proceedings of the 10th international conference on The semantic web* ISWC'11- Volume Part I Pages 146-161 ISBN: 978-3-642-25072-9

[14] Heasoo Hwang, Hady W. Lauw.Lise Getoor, Alexandros Ntoulas, "Organizing User Search Histories",*IEEE Transactions on Knowledge and Data Engineering*, Vol. 24. No.5 May 2012, pp. 912-925

*[15]* S.Vigneshwari, Dr.M.Aramudhan."An approach to personalize the web using XML based ontologies", *Proceedings of World Congress on Information and Communication Technologies (WICT), 2012 , PP. 759 - 762*

[16] S.Vigneshwari, Dr.M.Aramudhan. "A Technique to Ontology Mining for Semantic Web Information Extraction",,*European Journal of scientific research*,Volume 94, Issue 1, January 2013, PP 49-60.

[17] L. Chen and Y. Papakonstantinou, "Supporting Top-K Keyword Search in XML Databases," *Proceedings of 26th International Conference on Data Engineering*, 2010.

[18] Raymond Y.K. Lau, Dawei Song, Yuefeng Li, Terence C.H. Cheung and Jin-Xing Hao, "Toward a Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning", *IEEE transactions on knowledge and data engineering*, Vol. 21, No. 6, June 2009.

[19] Paul Buitelaar, Philipp Cimiano and Bernardo Magnini, "Ontology Learning from Text: An Overview", *DFKI, Language Technology Lab AIFB, University of Karlsruhe*, Vol. 3, pp. 1-10, 2003

[20] Shashank Paliwal, Vikram Pudi,"Utilizing Term Proximity Based Features to Improve Document clustering"In the proceedings of The International Conference on Knowledge Discovery and Information Retrieval , KDIR, 2011

[21] Babak Bagheri Hariri, Hassan Abolhassani, Ali Khodaei,"A new Structural Similarity Measure for Ontology Alignment", *Proceedings of the 2006 International Conference on Semantic Web & Web Services* 2006: 36-42 ISBN 1-60132-016-7

[22] S. Narayana, A. Govardhan, G.P.S. Varma," Discovering and Ranking Semantic Associations on the Semantic Web*", International Journal of Computer Science and Management Research*, Vol 1 Issue 5 December 2012,ISSN 2278-733X.pages:1099-1102

[23] http://archive.knoesis.org/library/ontologies/sweto/