International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 3, March - 2013

Semantically Classified Web Portal For Engineers

V. M. Yazhmozhi
Kamaraj College of Engineering and Technology, Virudhunagar – 626001, India.

K. Ramya Soundarya Lakshmi
Kamaraj College of Engineering and Technology, Virudhunagar – 626001, India.

S. Sreessruthi
Kamaraj College of Engineering and Technology, Virudhunagar – 626001, India.

Abstract

Engineers want to stay connected with everything that is new in this scientific world. As web has become huge and dynamic, at times they get irrelevant data during search. In this paper we offer a semantically classified web portal for the engineers which provide easy access to knowledge in a semantic way. In our portal, we provide three sections namely news, books and Research articles so that the engineers can just register with our portal and update their domain knowledge in a semantic way. To achieve semantics, we propose a semclassifier algorithm, which preprocess each document in the input document set and create a word dictionary (.xml) which will be compared with the domain dictionary (.xml) which contains the keywords in each area under an engineering field. It ranks the documents based on relevancy factor. The experimental result assures that our algorithm provides 93.3% accuracy in classifying the document.

Keywords *– Word dictionary, Concept weight assignment, Ranking, Preprocessing, Classification, Domain dictionary.*

I. INTRODUCTION

The colossal growth of the World Wide Web containing structured, semi-structured and unstructured documents insists us to work on Semantic Web as we get irrelevant data at times. This may be wastage of time for the engineers. To update themselves, they need to stay connected with the technical world. Semantic web always provides seamless access to the most relevant information. We provide a web portal for the engineers which contain three sections (Books, News & Research articles). In our work, we achieved semantics in our portal by pre-processing each document and forming a word dictionary of the same. We then compare it with the domain dictionary and calculate the concept weights for each document and classify them according to the field and area. Once this classification is done by our semclassifier algorithm, the engineers who are registered with our portal by specifying their field and area of interest get more relevant information without any need to search for the data they need.

As our portal is concerned, we had concentrated on text mining in which data pre-processing is an important task to be done. We followed different approaches for pre-processing pdf, word and HTML documents. The word dictionary which contains the "word frequency" will be in a same format for word, pdf documents and a different format is followed for HTML documents.

Outline of the paper

Section II presents the related work. Section III presents the Architectural design of the proposed work. Section IV presents the semclassifier algorithm. Section V presents the experimental results. Section VI presents the Performance Evaluation and finally Section VII presents the Conclusion and future work.

II. RELATED WORK

World Wide Web is a collection of vast and heterogeneous data. So the user has to rely on any of the portals to keep them updated. In many of the existing portals, the results often given are more irrelevant than our desired need. The user who is not familiar with particular domain would find it very difficult to determine the result that is required for him/her. In [1] and [2], authors' has used an ontological approach to overcome the syntactical information retrieval. Here the information access will be difficult, during the information growth. In [1], in one of the phases, word stemming has been performed. Stemming sometimes could change the actual meaning of the word. Recently, several tools for Ontology construction have been developed. However, these tools do not provide any services for Knowledge

Engineers to share or work together and reuse their work. In [3], WordNet is used to sense the word and thus enhance the information that is been retrieved. But WordNet do not provide classification of word senses for technical terms. In [4], a semantic web portal is designed to simplify the ontological engineering process. For ontological crawling and classification RDF is used. Here during ranking, there might be a chance for cyclic reference to be occurred. Also, certain data cannot be easily represented in RDF. In [5], a web portal is designed to provide the end user and application an integrated access to information. The semantic data retrieval is achieved by extracting the metadata from various data sources. In [6], OWL is used to describe the semantics of knowledge in a machine accessible way. The mapping of OWL on logics provides formal semantics and reasoning support. In [7], a semantic web portal is designed which provides the functionality for semantic visualization and search functions of RDF triples. This portal can be easily deployed for medium sized domain. In [8], a website is created for a library portal. This website provides access to various information resources all in one place to serve the needs of specific community.

III. ARCHITECTURAL DESIGN

In our proposed work, we maintain a repository of structured and unstructured documents which will be uploaded by the admin. The input document set is pre-processed. In pre-processing of word, pdf and HTML documents extraction of text from those documents is the first and foremost step. For pdf, word, HTML documents we used PDFBox-0.7.3, Spire.doc, HTMLAgilityPack respectively to extract text. Stop words are those which are not meaningful. After text extraction we remove the stop words like {“and”, “or”, "a", "an", "the", etc…..}. After removal of stop words, tokenization is performed which involves getting individual words. Then with the words obtained from tokenization a word dictionary is formed which contains the “word frequency”. The word dictionary for pdf, word documents is shown in Fig.1 and the word dictionary for HTML documents is shown in Fig.2. In HTML documents the tags which are text based like <meta>, <title>, <p>, <b>, <i>, <td>, <a>, etc… are examined for text extraction.

```
<? Xml version="1.0" encoding="utf-8"?>
<!--Word dictionary of ELMASRI - DBMS.pdf-->
<words>
<word name="table" count="59" />
<word name="entity" count="583" />
<word name="database" count="2674" />
<word name="systems" count="688" />
<word name="relational" count="984" />
<word name="queries" count="409" />
<word name="fundamental" count="24" />
<word name="preface" count="2" />
<word name="contents" count="38" />
<word name="edition" count="51" />
<word name="guidelines" count="25" />
<word name="using" count="586" />
<word name="book" count="117" />
<word name="about" count="119" />
<word name="authors" count="5" />
<word name="part" count="276" />
<word name="basic" count="199" />
<word name="concepts" count="364" />
<word name="chapter" count="815" />
<word name="tuple" count="689" />
<word name="users" count="328" />
 …….
 …….
</words>
```

Fig. 1: Word dictionary format for pdf, word documents

```
<? Xml version="1.0" encoding="utf-8"?>
<!--Word dictionary of database.htm-->
<words>
  <title>
    <word name=”introduction” count=”1” />
    <word name=”databases” count=”1” />
  </title>
  <paragraph>
    <word name=”domain” count=”5” />
    <word name=”relational” count=”2” />
    …….
  </paragraph>

 Similarly for all text containing tags
</words>
```

Fig. 2: Word dictionary format for HTML Documents

The word dictionary thus obtained is then compared with the domain dictionary to obtain the concept weight based on total hits or matches. The sample of domain dictionary is shown in Fig.4

With the concept weight obtained, we will be able to classify the document. We can identify the field and subfield to which the document belongs to. The documents that are classified are stored into the database along with field and subfield id in the domain dictionary. Then the documents are ranked and they are

presented to the users in the appropriate category.

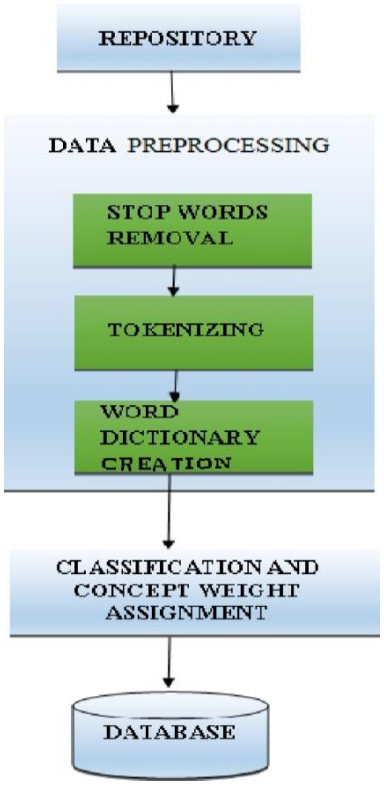

Fig. 3: Architectural design

```
<? Xml version="1.0" encoding="utf-8"?>
<fields>
<field name="cse" id="1">
<subField name="data mining" id="1">
      <term name="data"/>
      <term name="mining"/>
      <term name="pre-processing"/>
      ……….
</subField>
<subField name="databases" id="2" >
     <term name="dictionary" />
     <term name="alter"/>
     <term name="drop"/>
     <term name="grant"/>
     <term name="revoke"/>
      ……….
</field>
      ……….
<field name="mining" id="2">
<subField name="mining method" id="1">
      <term name="mineral"/>
      <term name="extraction"/>
      <term name="discovery"/>
      <term name="surface"/>
      <term name="blasting"/>
      ……..
</fields>
```

Fig. 4: Domain dictionary sample

IV. SEMCLASSIFIER ALGORITHM

ALGORITHM	: SemClassifier
INPUT	: Document D
OUTPUT	: Classified result (Documents category)

Step **1** : Initialize concept weight CW= 0.

Step **2** : Initialize field=0,subfield=0
Step **3** : The document D is pre-processed.

Step **4** : Extract the text from the document, based on the extension of the document.

Step **5** : Remove the stop words {SW1,..., SWn }.

Step **6** : Tokenize the remaining words and obtain the tokens {TK1,….,TKn}.

Step **7** : Construct the word dictionary, a xml file with the "word frequency" frequency" {FQ1,…FQn}.

Step **8** : Compare the word dictionary with the domain dictionary.

Step **8a** : for each field (fid) in the domain dictionary and for each subfield (sfid)

Step **8b** : Compare the {TK1,…TKn} with the terms in subfield

Step **8c** : If a match is found, then the matched token's FQ is added to concept weight CW.

Step **8d** : If CWtemp greater than CW, then

CW=CWtemp;
field=fid;
subfield=sfid;

Step **9** : The field and subfield is obtained and the document is classified.

When an engineer logs into the portal, the documents are ranked (i.e.) the concept weights are sorted and the semantically classified and ranked documents are provided to the engineers as per their field and subfield (i.e.) area of interest.

V. EXPERIMENTAL RESULTS

This section presents the experimental results for the proposed work. We used data mining, database concepts, etc., under the computer science domain for the testing purpose. To implement the module for the portal, we used C#. A web interface is also created. Only the administrator has the authority to upload the document to the repository. It is shown in Fig.5.

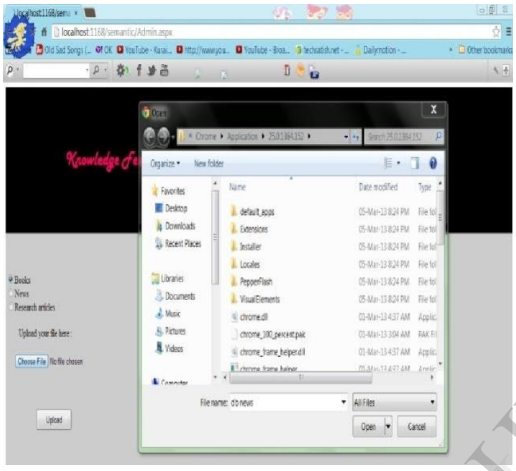

Fig. 5: Uploading document

When a user registers to this portal, he/she has to specify his/her area of interest, as shown in Fig.6. When the user logs in, he/she can view the books, news and research articles related to his/her interest. List of books retrieved that is related to user interest is shown in Fig.7. The resultant document is shown in Fig.8.

VI. PERFORMANCE EVALUATION

Performance evaluation of the proposed approach is done based on classification context scenario. Precision, Recall, Accuracy and F-measure are the major terms used for classification based performance.

Precision is the probability that a retrieved document is relevant. Precision is calculated based on the formula

Precision=tp/(tp + fp)

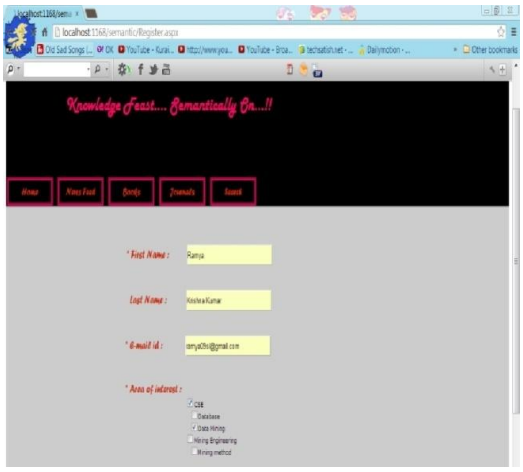

Fig. 6: User registration

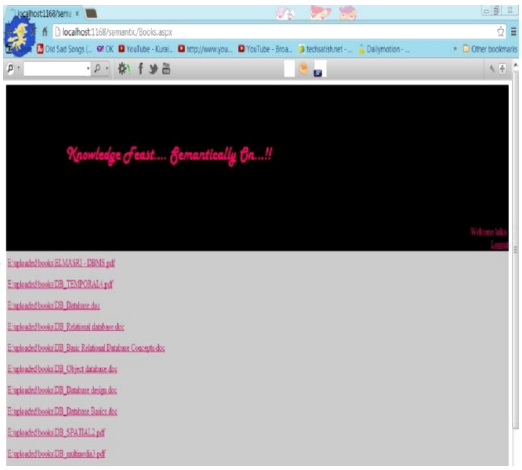

Fig. 7: List of books retrieved

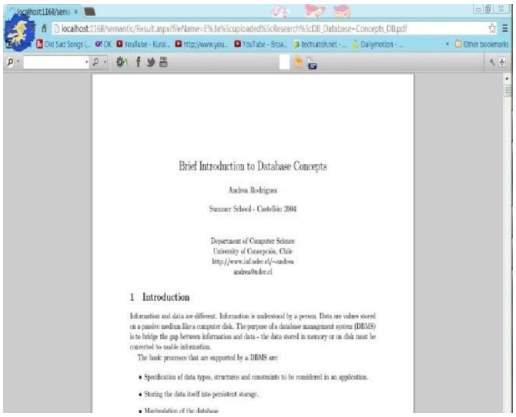

Fig. 8: Resultant document

Recall is the probability that a (randomly selected) relevant document is retrieved in a search. Recall is calculated based on the formula

Recall = tp/(tp + fn)

Accuracy is a weighted arithmetic mean of Precision and Inverse Precision as well as a weighted arithmetic mean of Recall and Inverse Recall. Accuracy is calculated based on the formula

Accuracy = (tp + tn)/(tp + tn + fp + fn)

Where,

tp – True Positive(Correctly retrieved)
tn – True Negative(Correctly rejected)
fp – False Positive(Incorrectly retrieved)
fn – False Negative(Incorrectly rejected)

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure. F-Measure is calculated based on the formula

F = (2*Precision*recall)/(Precision+recall)

The performance measure of proposed system is plotted in Fig.9.

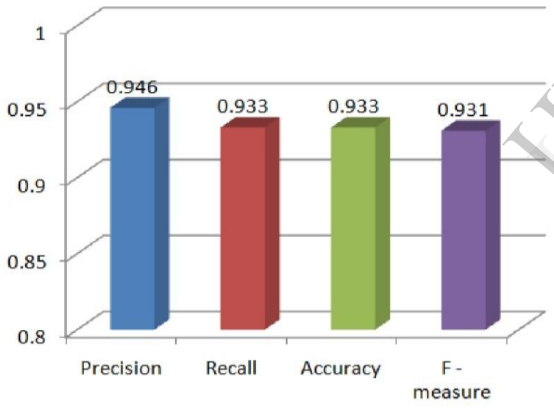

Fig. 9: Performance of Proposed System

From the performance measure, it is understood that the accuracy of the proposed system is 93.3%, which is high compared with existing approaches.

VI. CONCLUSION

In this paper, a web portal has been designed for engineers. Since all the documents are accurately classified according to their domain using the semclassifier algorithm, the engineers have seamless access to their domain knowledge. The experimental results have also proven that the system is more accurate. This system can also be extended to other domains like medicine, law, etc.

REFERENCES

[1] Dr.G.R.Karpagam and J. Uma Maheswari, "A Conceptual Framework for Ontology based Information Retrieval", International Journal of Engineering Science and Technology Vol. 2(10), 2010, pg no: 5679-5688.

[2] NeelMadhav Gantayat, Shridhar Iyer, "Automated building of domain ontologies from lecture notes in courseware", 2010 IEEE conference on Technology for Education pg no: 89-95.

[3] Peter M. Kruse, André Naujoks, Dietmar Rösner, Manuela Kunze, "Clever Search: A WordNet Based Wrapper for Internet Search Engines", Computing Research Repository – CORR, vol. Abs/cs/050, 2005.

[4] Chintan Patel, Kaustubh Supekar, Yugyung Lee, E. K. Park, "OntoKhoj: A Semantic Web Portal for Ontology Searching, Ranking and Classification", WIDM '03 Proceedings of the 5th ACM international workshop on Web information and data management, pg no: 58-61.

[5] Yuangui Lei, Vanessa Lopez, and Enrico Motta, "An Infrastructure for Building Semantic Web Portals", International Workshop on Web Information Systems Modeling (WISM 2006), pg no: 283-308.

[6] Grigoris Antoniou and Frank van Harmelen, "Web Ontology Language: OWL", Handbook on Ontologies in Information Systems, Springer-Verlag, pg no: 76-92.

[7] Ying Ding, Yuyin Sun, Bin Chen, Katy Borner, Li Ding, David Wild, Melanie Wu, Dominic DiFranzo, Alvaro Graves Fuenzalida, Daifeng Li, Stasa Milojevic, ShanShan Chen, Madhuvanthi Sankaranarayanan, Ioan Toma, "Semantic Web Portal : A Platform for Better Browsing and Visualizing Semantic Data", 2010 International Conference on Active Media Technology, Toronto, Canada (AMT 2010), pg no: 448-460.

[8] Tamar Sadesh and Fenny Walker, "Library portals: toward the semantic web", New Library World, Vol.104 Iss: 1/2, pg no: 11-19.

[9] Ina O'Murchu, Anna V. Zhdanova, John G. Breslin, "Semantic Community Portals", Encyclopaedia of Portal Technology and Applications (Ed.: Tatnall, A.), Idea Group Publishing (2006).

[10] Martin Hepp, Katharina Siorpaes, Daniel Bachlechner, "Towards The Semantic Web In E-Tourism: Can Annotation Do The Trick?", Proceedings of the 14th European Conference on Information System (ECIS 2006), June 12-14, 2006, Gothenburg, Sweden.

[11] Rah mat Hidayat, Yazrina Yahya, Shahrul Azman Mohd Noah, Mohd Zakree Ahmad, Abdul Razak Hamdan, "Semantic Web Portal in University Research Community Framework", International

Journal on Advanced Science Engineering Information Technology, Vol.2 (2012) No.6. pg no: 39-43.

[12] Holger Lausen, Michael Stollberg, Rubén LaraHernández, Ying Ding, Sung-Kook Han, Dieter Fensel, "Semantic Web Portals – State of the Art Survey", Journal of Knowledge Management, Vol. 9, No. 5. (May 2005), pg no: 40-49.