Metadata Generation for the Web Documents using WordNet Ontology

K.Saruladha[1], G.Aghila[2], P.Giridhar Reddy[3], M.Abinesh[4], N.Dhanigai Arasan[5] and S.Sujesh[6]

[1,2,3,4,5,6]Pondicherry Engineering College, Puducherry, India
Email:[1] charusanthaprasad@yahoo.com
Email: [2]aghilaa@yahoo.com, [3]merashgiri@gmail.com,
[4]abinesh.2@gmail.com,[5] ndhanigai@gmail.com,[6] sujesh88@yahoo.co.in

***Abstract*—The objective of this paper is to generate metadata for the web documents. Metadata plays a key role in the indexing of web documents in information retrieval systems. Metadata is generated by finding the dominant concepts of the documents, which are obtained by the using the semantic similarity measures(Resnik, Lin, and Jiang & Conrath). This metadata could aid better retrieval effectiveness as semantically identified dominant concepts could better contribute to relevance than keywords. In the proposed method corpus independent information content measure is used along with the WordNet as the underlying ontology and is implemented using JAWS. In calculating these corpus independent information content values we make use of the hyponyms and the measure proposed by Seco. As wordnet is a light weight terminology oriented ontology covering almost all the domains, it is a generalized method to generate metadata. Representativeness is calculated for every noun present in the document and the noun with high representativeness is included in the metadata.**

***Index Terms*—metadata, semantic similarity, wordnet ontology, hyponym, meronym, Information content.**

I. INTRODUCTION

In the semantic web information is given with well-defined meaning. Using the keywords as metadata is not an efficient approach. Most of the time using this approach leads to a high recall (ratio of number of relevant documents retrieved to the total number of relevant documents in the collection) and a weak precision (ratio of number of relevant documents retrieved to the total number of documents retrieved).

For effective retrieval of documents from web several semantic measures and natural language processing techniques are used. In [1] the author came up with a semi-automatic process to index the semantic documents using terminology oriented ontology of a particular domain. The main drawbacks of this technique are that the ontology is a domain specific one and is not applicable for general purpose and there are many better semantic similarity measures which can give better results than the semantic measure used in it.

In this paper a generalized automatic mechanism to find the metadata that represents the document more by using

WordNet [5] ontology is proposed. Better semantic similarity measures proposed by *resnik* [2], *jiang & conrath* [3] and *Lin* [4]. H*yponym* (specification of a given word) and *meronyms* (part of a larger whole) [6] are used in the similarity measures moving a step ahead of the existing synonyms. By assessing the results the proportions which give better results are evaluated. Relationships which contribute more to the semantic similarity are also evaluated.

Section II explains all the existing semantic similarity measures and tools used in the design for generation of metadata. Section A in part II presents the architecture of the proposed design. The proposed algorithms designed for each module is discussed in Section III. Section IV presents the snapshots of the outputs.

II. RELATED WORK

The main objective of this paper is to present the design of a mechanism that allows in finding the metadata of a document using Wordnet ontology, which in turn can be used as metadata for the effective retrieval of documents from the web.

Currently most of the documents are indexed and retrieved by using centralized databases and simple keywords. Since the semantic web is evolving it is very difficult to retrieve desired documents by using keywords alone. This lead to the importance of semantic similarity measures in the natural language processing.

Desmontils et. Al. [1] used a similarity measure similar to that of edge counting but taking the distance from the root to the word. This is applicable in the taxonomy with a single root alone. One more drawback with this is that all the relationships are measured as equal distance. From the results obtained from the research it is proven that IC based semantic similarity measures yield better results. Three IC based semantic similarity measures viz., Resnik measure, Lin measure, Jiang & Conrath measure are reported in the literature.

In 1998 Resnik [2] proposed a semantic similarity measure for IS-A taxonomy using the notion of Information Content (IC). In this method the similarity of two concepts is the extent to which they share information in common. In an IS-A taxonomy it is the super concept that subsumes both the concepts. ie., the common parent concept of the two concepts which has maximum information content is sed to represent both the concepts.

Formally it is defined as

$$Sim(c_1,c_2) = \max_{C \in S(C1,C2)} [- \log P(c)]$$

Where $S(c_1,c_2)$ is the set of concepts that subsumes both c_1 and c_2 and P(c) is the probability of finding an instance of concept c in the considered taxonomy.

In [3] J.J. Jiang and D.W. Conrath proposed a similarity measure comprising of both node based approach (ie., information content) and edge based approach(ie., distance between the concepts). In the edge based approach the distance between the two concepts in the taxonomy is calculated. The least the distance between them the most they are related. But care is to be taken while calculating distance as different weights are to be assigned for different types of relations. It also depends on network density of the taxonomy and the closeness of the child and parent when compared to the closeness of all other children of the parent. After considering all these they proposed a formula to calculate the similarity between two concepts.

$$Dist(w_1,w_2) = IC(c_1)+IC(c_2)-2\times IC(LSuper(c_1,c_2))$$

Where $c_1$and $c_2$are the set of concepts represented by words w_1 and w_2 respectively. $LSuper(c_1,c_2)$ is the lowest super-ordinate of c_1 and c_2.

Dekang Lin came up with a similarity measure in information theoretic terms to achieve universality. It is applicable as long as there is a probabilistic model. He defined similarity in terms of commonalities and differences between the concepts. The more the commonalities are, the more their similarity is. In the same way the more the differences are, the less their similarity is. No matter how much similar they are the maximum similarity is achieved only when both are identical. The similarity between any two concepts A and B is measured by using the formula

$$Sim(A,B)= \log P(common(A,B))/ \log P(description(A,B))$$

When it comes to semantic similarity in taxonomy the similarity is calculated as

$$Sim(x_1, x_2) = 2\times \log P(c_0)/(\log P(c_{x1})+\log P(c_{x2}))$$

Where c_{x1} and c_{x2} are the concepts of x_1 and $_{x2}$. c_0 is the most specific class that subsumes both c_{x1} and c_{x2}.

The semantic similarity measure used by Desmontils & Jacquin is not a traditionally proven method. Though literature has reported many semantic similarity approaches, the Information content based approaches report highest correlations against human judgements. Jiang & Conrath has the highest correlation of 0.859. These traditional methods compute IC based on brown corpus. Pirro has experimented IC based approaches by incorporating corpus independent IC considering hyponym relations. The study is aimed at identifying the best semantic similarity which could identify metadata computationally.

Semantic similarity methods and Information content calculation

The proposed method is underpinned under the following design considerations. We make use of the above three information content based semantic similarity measures instead of the measure used by Desmontils[1] and analyze how efficiently the dominant concepts are identified and used as metadata to aid better retrieval of documents from the web.

Information content computation used by Resnik, jiang and lin use Brown corpus statistics to quantify the informativeness of WordNet concepts. But the proposed method uses the corpus independent IC computation proposed by Seco. In calculating the information content values of the hierarchical semantic relations we make use of hyponyms and meronyms.

A. Architecture used for metadata generation

The HTML documents and the text documents are preprocessed in which all the html markers are removed and the plain text is tagged using a Parts-of-Speech(POS) tagger. From this tagged text the nouns are extracted and their weighted frequencies are calculated by using the weights assigned to the different tags. The weights used for various tags are tabulated in table 2. For these nouns IC values are calculated by using the IC computation method proposed by Seco [7] and similarity between every two nouns is calculated using the three semantic similarity measures. The cumulative similarity of a noun with respect to a particular document is calculated by taking the weighted mean of all the similarity values of that noun with every other noun in the document. The mean of this cumulative similarity and the noun frequency gives the representativeness of a noun for that document. The nouns with high representativeness are selected as metadata of that document. **Fig.2** represents the architecture of the proposed method and **Table.1** mentions the functionalities of the various modules.

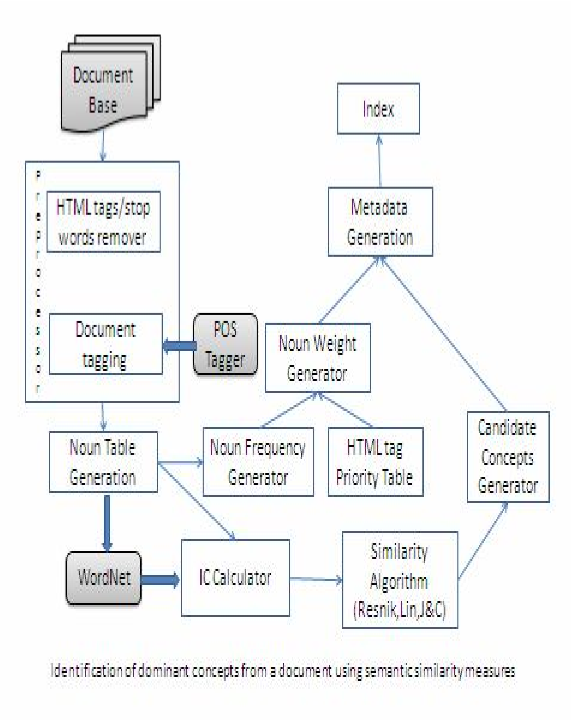


Figure 2

Functionalities of various module

Module	**Input**	**Output**	**Description**
HTML Parser	HTML documents	Parsed text	Document is parsed and the contents of the selected tags are stored along with the tags.
POS Tagger	Parsed text and text documents	Tagged text	The text passed to this module is tagged with its appropriate parts-of-speech.
Noun Extractor	Tagged text	Nouns	All the nouns are extracted from the above tagged text.
Noun frequency Generator	Nouns	Frequency	Frequencies of all the nouns present in different tags in the document is calculated.
Information Content (IC) calculator	Nouns	IC values	IC values of all the nouns are calculated using the WordNet ontology.
Intra-noun similarity measure	Nouns, IC values	Cumulative similarities	It takes nouns and its IC values as input and finds the similarity between every two nouns. The weighted average of these similarities gives the cumulative similarity of the word with respect to that particular document.
Noun weight Calculator	Noun frequencies and tag weights	Noun weights	It calculates noun weights based on the presence of nouns in different tags.
Metadata generation	Noun weights, cumulative similarities	Metadata	Representativeness is calculated for every noun and the nouns with high representativeness are selected as metadata.

Table.1

Tag weightage table

TAG	**WEIGHTAGE**
TITLE	10
A	8
IMG	2
I	2
B	2
H1	3
H2	3
H3	3
Font size 7	5
Font size +4	5
Font size 6	4
Font size +3	4
Font size 5	3
U	2

Table.2

Tools used for Implementation

POS Tagger: For the extraction of concepts from the document we need a tagger which can tag every word in the document with its appropriate parts of speech. So we use Stanford Parts-of-Speech(POS) tagger[8] for tagging the whole text. We make use of this tagger to identify the nouns in the HTML and the text documents and extract these nouns to form concepts.

WordNet: WordNet[9] is a light weight lexical ontology which contains all the available concepts which are connected together semantically. WordNet is organized in a way that all the concepts are arranged with “IS-A” semantic link. This semantic link not only gives the synsets(synonym sets) of a particular word but also providing various other semantic relations(hyponym, hypernym, meronym, holonym). We make use of WordNet to find hyponym and meronym of a given word which are used in calculating the Information content value.

JAWS: We make use of Java API for Searching in WordNet (JAWS)[10] and retrieve data from the WordNet database. The directory where the WordNet software is installed in the system is to be given as input to the API. The directory given also contains the WordNet database in an encrypted format. Using appropriate objects and methods which are provided in the API the required data can be retrieved from the WordNet database. Using WordNetDatabase.getFileInstance() method all the relations for a noun are identified and extracted.

III. ALGORITHM OF METADATA GENERATION (MDG)

The algorithm is designed to computationally identify metadata in the documents. The documents could be an HTML document or text document. The computationally identified dominant concepts will be stored along with the document as metadata which could improve the retrieval effectiveness of the search engines. To identify the dominant concepts the information content based semantic similarity methods proposed by Resnik, Jiang &Conrath and Lin are used.

```
  Algorithm MDG(documents) //Metadata
Generation //algorithm
  { //files[] is the array of documents for which
metadata
    // are to be determined.
    //maxcount[] is the array of dominant concepts.
    for i:= 0 to length(files) do
      { if(files[i] ends with “.htm” ) then
        nouns[]:=parser(files[i]); //for html documents
        else
        nouns[]:=postag(files[i],1); //for text documents
       }IC[]:=getIC(nouns[]);
   for i:= 0 to length(nouns) do
     {  for j :=0 to length(nouns) do
        { maxparent=parent(nouns[i],nouns[j])
    //get the parent with maximum IC value for each
   // noun-noun combination
   Resniksim[i][j]= - log(IC(maxparent));//resnik
similarity
   Cumualtivesimresnik[i]=sum[i]+resniksim[i][j];
   linsim[i][j]:=
2×log(IC(maxparent))/(log(IC(noun[i]))
+log(IC(noun[j])); // lin similarity
   cumualtivesimlin[i]=sum[i]+linsim[i][j];
   jandcsim[i][j]:=IC(noun[i]) + IC(noun[j])-
2×Ic(maxparent);  // J and C similarity
   cumualtivesimj&c[i]=sum[i]+linsim[i][j];
        } }representativeness();
   }
   Algorithm parser(file)
   { //preprocessing HTML documents.
    //taglist[] is the array of all tags present in the html
    //document.
    //ch gets the next character from the file which is
    //parsed.
     temp:= null; content:=null; //initialize
     while(ch ≠ EOF) do //get all characters one by one
                    //from file
      { if(ch:=’<’) then   //starting of a html  tag
        {  while(ch ≠ ‘>’ or ch ≠ ‘ ’) do  //get tag name
        {   temp+:= ch;
        }
      for i:=0 to length(tag) do
       {  if(temp:= taglist[i]) then //check for occurrence
              //of current tag in tag list
        {    while(ch ≠ ’<’) do
             content+=ch; //get the sentence within the tag
        nouns[]=postag(content,0);
        } } }return nouns;
   }

   Algorithm postag(content,f) //parts of speech tagger
   { //nouns[] is the array of all nouns obtained from the
     //tagged contents.
    if f := 1 then
      { //newcontents is the file which contains the
       //contents of each tag.
       content := newcontents;
      } tagged := gettagged(content);
    nouns[] = getnouns(tagged);
    Return nouns;
   }
   Algorithm getnouns(tagged)
   {//to extract nouns
     pos := 0; //initialize
     for i := 0 to length(tagged) do
```

```
      {    If (tagged[i] := '/' and tagged[i+1] :=
tagged[i+2] := 'N') then
             r :=i;
         while (tagged[r] ≠ ' ' and tagged[r] ≠ '$') do
        {     nouns[pos] +:= tagged[r]
                r--;
        }  nouns[pos] :=reverse(nouns[pos]);
      pos++;
     } return nouns;
   }
   Algorithm getIC(nouns[])
   {//MAXCON is the total number of synsets in
WordNet.
     //Hypo_no is the number of hyponyms available for
a
     //given word in WordNet.
   for i := 0 to length(nouns) do
   {Hypo_no := length(gethypo(noun[i]));
   IC[i]:= 1-(log(Hypo_no + 1)/log(MAXCON)) ;
   }return IC;
   }
   Algorithm parent(word1,word2) //finding the
parent of
   //two words in the taxonomy with maxIC
  //a and b contains all the hypernyms of word1 and
word2
      //respectively.
      k :=0, common[] //initialize
     for i := 0 to length(a) do
     {for i := 0 to length(b) do
      { if (a[i] := b[i] ) then
      common[k] := a[i];
      k++;
      break;
       }}if (length(common) > 0 ) then
       { maxparent :=getmax(common);
        } else if(com := null) then
        {     for i :=0 to length(a) do
              { for j :=0 to length(b) do
                  { maxparent=parent(a[i],b[j])
                   }}}}
   return maxparent;
   }
   Algorithm getmax(common[])
   {//max contains the position of the concept with
    //maximum IC value in common array.
    //temp[] contains the IC values of all concepts in
    //common array.
   temp[]=getIC(common);
     for i := 0 to lenth(temp) do
       { for j:= 0 to lenth(temp) do
          { If(temp[i]>temp[j])
                max=i;
             else
                max=j;
          } }
    return common[max];
    }
```

Metadata Generation:

181

the file, Dominant Concepts. The structure of the metadata is given below:

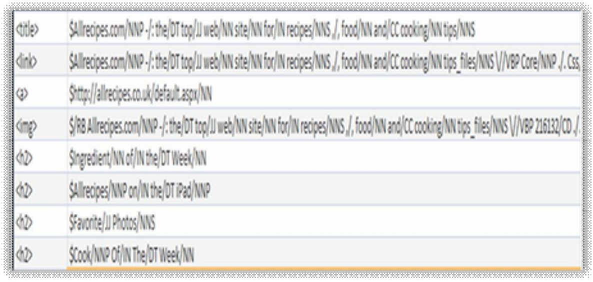
File ID	1
File Name	File 1
Author	Name
Metadata	Noun1, Noun 2, Noun 3

This metadata is used for indexing the documents and for the better retrieval of the web documents.

IV. SNAPSHOTS

SAMPLE OUTPUTS

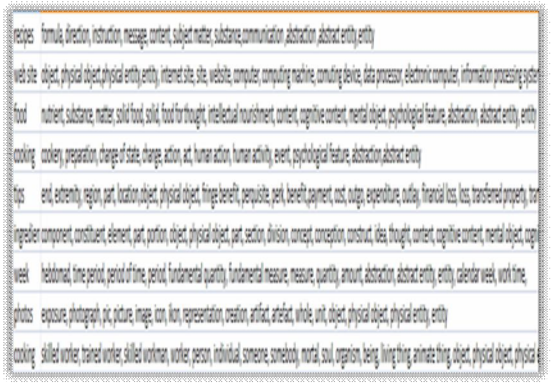
Tagged contents:



<title>	$Allrecipes.com/NNP -/: the/DT top/JJ web/NN site/NN for/IN recipes/NNS ,/, food/NN and/CC cooking/NN tips/NNS
<link>	$Allrecipes.com/NNP -/: the/DT top/JJ web/NN site/NN for/IN recipes/NNS ,/, food/NN and/CC cooking/NN tips_files/NNS \/VBP Core/NNP ./. Css
<a>	$http://allrecipes.co.uk/default.aspx/NN
<img>	$/RB Allrecipes.com/NNP -/: the/DT top/JJ web/NN site/NN for/IN recipes/NNS ,/, food/NN and/CC cooking/NN tips_files/NNS \/VBP 216132/CD ./.
<h2>	$Ingredient/NN of/IN the/DT Week/NN
<h2>	$Allrecipes/NNP on/IN the/DT iPad/NNP
<h2>	$Favorite/JJ Photos/NNS
<h2>	$Cook/NNP Of/IN The/DT Week/NN

Nouns:file1			
file	**Tag**	**Noun**	**Freq**
file1	Title	Allrecipes.com	3
file1	H2	Web	4
file1	H2	Site	4
file1	A	Recipes	6
file1	H1	Food	8
file1	B	Cooking	7
file1	Link	Tips	2

Hyponym for nouns:



recipes	formula, direction, instruction, message, content, subject matter, substance,communication ,abstraction ,abstract entity,entity
web site	object, physical object,physical entity,entity, internet site, site, website, computer, computing machine, comuting device, data processor, electronic computer, information processing syste
food	nutrient, substance, matter, solid food, solid, food for thought, intellectual nourishment, content, cognitive content, mental object, psychological feature, abstraction, abstract entity, entity
cooking	cookery, preparation, change of state, change, action, act, human action, human activity, event, psychological feature, abstraction,abstract entity
tips	end, extremity, region, part, location,object, physical object, fringe benefit, perquisite, perk, benefit,payment, cost, outgo, expenditure, outlay, financial loss, loss, transferred property, tra
ingredien	component, constituent, element, part, portion, object, physical object, part, section, division, concept, conception, construct, idea, thought, content, cognitive content, mental object, cogi
week	hebdomad, time period, period of time, period, fundamental quantity, fundamental measure, measure, quantity, amount, abstraction, abstract entity, entity, calendar week, work time,
photos	exposure, photograph, pic, picture, image, icon, ikon, representation, creation, artifact, artefact, whole, unit, object, physical object, physical entity, entity
cooking	skilled worker, trained worker, skilled workman, worker, person, individual, someone, somebody, mortal, soul, organism, being, living thing, animate thing, object, physical object, physical

V. CONCLUSION

Till now we worked out for few set of documents which we downloaded in the food domain and calculated the Information Content(IC) values for few concepts in the chosen documents. We are proceeding the same for large set of documents under the same domain. We are yet to calculate the similarity values using resnik, lin and J&C to see which similarity measure contributes to the better retrieval of the documents.

REFERENCES

[1] E. Desmontils & C. Jacquin, “Indexing a Web Site with a Terminology Oriented Ontology”. http://www.sciences.univ-nantes.fr/irin/indexGB.html,IRIN, Université de Nantes, 2004.

[2] P. Resnik, “Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language” *Journal of Artificial Intelligence Research*, 11:95-130, 1999.

[3] J.J. Jiang and D.W. Conrath, “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy” In *Proceedings of the International Conference on Research in Computational Linguistic*, Taiwan, 1998.

[4] Dekang Lin, “An Information-Theoretic Definition of Similarity”, *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin, July, 1998.

[5] G. A. Miller, “WordNet: an Online Lexical Database”, *International Journal of Lexicography*, 3(4), 1990, pp. 235-312.

[6]] JungAe Kwak and Hwan-Seung Yong, “Ontology matching based on hypernym, hyponym, holonym, and meronym sets in WordNet”, *International journal of Web & Semantic Technology (IJWesT)*, Vol.1, No.2, April 2010

[7] N. Seco, T. Veale, J. Hayes, “An intrinsic information content metric for semantic similarity in WordNet”, Proceedings of ECAI, 2004, pp. 1089–1090.

[8] POS Tagger, http://nlp.stanford.edu/software/tagger.shtml

[9] WordNet, http://wordnet.princeton.edu/wordnet/download/

[10] Java API for WordNet Searching (JAWS), lyle.smu.edu/~tspell/jaws/index.html

gopalax Publications