Information Retrieval on Text using Concept Similarity

Dr.Reshmy Krishnan
Muscat College
Muscat,Sultanate Of Oman
reshmy_krishnan@yahoo.co.in

***Abstract*—** Retrieving proper information from internet is a huge task due to the high amount of information available there. Identifying the individual concepts according to the queries is time consuming. To retrieve documents, keyword based retrieval method was used before. Using this type searching, the relationship between associated keywords can't be identified. If the same concept is described by different keywords, inaccurate and improper results will be retrieved. Concept based retrieval methods are the solution for this scenario. This gives the benefit of getting semantic relationships among concepts in finding relevant documents. Irrelevant documents can be eliminated by detecting conceptual mismatches, which is another benefit obtained from this. The main challenges identified are the ambiguity occurring due to multiple nature of words for the same concepts. Semantic analysis can reveal the conceptual relationships among words in a given document. In this paper the potential of concept-based information access via semantic analysis is explored with the help of a lexical database called WordNet. The mechanism is applied in the selected text documents and extracting the Synonym, Hyponym, Hypernym of each word from WordNet. The ranking will be calculated after checking the frequency rate of each word in the input documents and a hierarchy model will be generated according to the ranking.

Keywords- *Ontology, WordNet, Synonym, Hyponym, Hypernym, semantic analysis, Keyword based retrieval, concept based retrieval*

*****

I. INTRODUCTION

Although volume of Information available in www has been increasing continuously,most of the information is still unavailable to normal people due to the lack of proper techniques for Information retrieval.85% of the internet users are using Internet for Information retrieval.The Unstructured nature and huge volume of information in www has made it difficult for getting proper result while searching[1].The main issue related to the Information retrieval is poor quality of retrieved results.

The techniques used for Information retrieval was keyword based. This technique use keyword list for searching the contents of information. The main concern regarding this approach is the poor quality of the result .One of the reason for this concern is the vocabulary problem facing by the non-expert users. The keywords chosen by the users were often different from those used by the authors of the relevant documents. These problems are referred as synonymy and polysemy.

The information needs of people are in concept space. Keyword based access to information is sometimes unsatisfactory since it works in word space. Words represent concepts in human language but the mapping from words to concepts is many-to-many. That means one concept may be represented with many different words (synonym) and one word may represent many different concepts (polysemy). This mapping problem is known as Word Sense Disambiguation. Secondly, since concepts are abstract entities, representing them is another problem.

Concept-based information retrieval is an alternative IR approach that aims to tackle these problems differently. Concept-based IR represents both documents and queries using semantic concepts, instead of keywords, and performs retrieval in that concept space. This approach holds the promise that representing documents and queries using high-level concepts will result in a retrieval model that is less dependent on the specific terms used [2]. Such a model could yield matches even when the same notion is described by different terms in the query and target documents,thus alleviating the synonymy problem and increasing recall. Similarly, if the correct concepts are chosen for ambiguous words appearing in the query and in the documents, non-relevant documents that were retrieved could be eliminated from the results.

To tackle polysemy, the main proposed method was to apply automatic wordsense disambiguation algorithms to documents and query. Disambiguation methods use resources such as the Wordnet thesaurus [3] to find the possible senses of a word and map word occurrences to the correct sense. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

Section 2 reviews the state of the art in concepts based extraction from documents and section 3 sketches out our methodology for the generation of ontology from a text document by extracting semantic web concepts with the help of Word.Net in terms of Design, Implementation and results. Section 4 shows the conclusion and future works.

II. INFORMATION RETRIEVAL

The term Information retrieval(IR) refers to the access to Information and its representation. The key role of information retrieval process is to retrieve relevant information to a given request. The efficiency of the process is the retrieval of all relevant information available and rejection of all non relevant ones. Even though in reality, the results will contain both relevant and non relevant information, the aim is to achieve the ideal criteria. One information retrieval system could handle different information simultaneously. Majority of the information retrieval based on the text documents and hence can be named as text retrieval or data retrieval. The text retrieval incorporates all types of texts including complete articles ,books, web pages and minor fragments of texts such as sections, paragraphs ,sentences etc., Instead of retrieving information directly, the documents will be retrieved in IR process, from which information can be obtained. The basic model of Information retrieval can be shown as follows.

Fig 1.Information Retrieval

Queries are the requests to the system to get the results. Any one of the searching strategy can be used for searching from the internet. According to the queries of the user, documents will be retrieved from the storage using any appropriate search techniques. Storage comprises an abstract description of the input document .This description will be unstructured except for the syntax .The similarity between the given query and the stored documents will be checked in the matching process.

2.1 Concept-based Information Retrieval Model

In the cognitive view of the world, there exists the presumption that the meaning of a text (word) depends on conceptual relationships to objects in the world rather than to linguistic or contextual relations found in texts or dictionaries. A new generation information retrieval model is drawn from this view. We call it concept-basedinformation retrieval model. Sets of words, names, noun-phrases, terms, etc. will be mapped to the concepts they encode.

Generally, a content of an information object is described by a set of concepts in this model. Concepts can be extracted from the text by categorisation. Crucial in this model is existence of a conceptual structure for mapping descriptions of information objects to concepts used in a query. If keywords or noun-phrases are used, then they should be mapped to concepts in a conceptual structure. Conceptual structures can be general or domain specific, they can be created manually or automatically, they can differ in the forms of representation and ways of constructing relationships between the concepts. Naturally, the tools considered in this paper differ in this respect.

For establishing definitions of concepts it is necessary first to identify concepts inside the text and then classify found concepts according to the given conceptual structure. There are several ways of identification of concepts present in the text. This process is called categorization.Concepts can be identified also by using fuzzy reasoning about the cues (terms) found in the text for calculating likelihood of a concept present in the text.

After the concept is categorised, it can be given the definition by a classification process. Classification is determining where in the conceptual structure a new concept belongs. For this purpose, either an existing conceptual structure (like dictionary,thesaurus or ontology) or automatically generated one can be used. It is reported in many papers, that pre-existing dictionaries often do not meet the user's needs for interesting concepts, or ontology like WordNet does not include proper nouns.

III. STATE OF THE ART OF SEMANTIC EXTRACTION OF DOCUMENTS.

There is a strong requirement in the Information retrieval research area in recent years due to the enormous growth in the number of text databases available on-line and need for better techniques to access this databases[4][13]. Since the future web –semantic web-consists of pages containing texts and semantic mark up, the current IR techniques are unable to exploit the semantic knowledge within the documents and hence cannot give precise answers to precise queries [5].Information retrieval models can be distinguished such as Keyword-based Information Retrieval Model and content based IR model. In the first one, Information retrieval model is based on keyword indexing systems, frequency of occurrence of a keyword is taken into account[6][14]. Using the first one we can do data retrieval and latter gives us Information retrieval. As the name implies, the main task in information retrieval is to find information rather than data .Keyword based access can do the data retrieval which aims to provide data sets which fit the keywords of a query.

During the semantic web period, the meaning of a text or a word is depending on the conceptual relationships to objects in the world rather than to the contextual relations found in texts or dictionaries.The concepts of the words,names and nouns in the documents will be mapped to the concepts in wordNet.

A content of an information object is described by a set of concepts in Content based IR model. Concepts can be extracted from the text by categorization. The main problem facing is the non existence of a conceptual structure for mapping objects to concepts used in the user query.The nouns or names in the input documents should be mapped to concepts in wordNet in a conceptual structure. Since wordNet groups words together based on their meanings(synsets)[10],the groups can be interlinked using the relationships such as is-a and part-of/member-of. Since concepts are abstract entities, representing them is a big problem. Words represent concepts in human language but the mapping from words to concepts is many-to-many.That means one concept may be represented with many different words (synonym) and one word may represent many different concepts (polysemy)[7]. This mapping problem is known as Word Sense Disambiguation[8].

IV. METHODOLOGY

We are presenting a method for semantic concept extraction from the text document with the help of WordNet by reducing the above existing problems in this area. WordNet is such an existing general ontology from which a sub ontology can be generated[10]. Synsets are interlinked by means of conceptual-

semantic and lexical relations.WordNet can be queried according to the input text document and create classes of concepts based on the results of the query. Extraction of semantic concepts from the keywords is the initial phase of actual construction of the ontology which will be covered during the next phase of this project. To extract semantic concepts, a word in the text document is taken as input which one wants to improve the knowledge, WordNet is searched about this word and different meanings of words are taken from which initial documents are collected. Terms frequencies are then calculated and compared with each group and concept with highest frequencies will be displayed first. The second phase of the construction of ontology will be done using the result of the first phase.

In part of this study, we used WordNet 2.1 as our knowledge base. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept .

4.1 High Level Design

4.1.1. Extraction of semantic concepts from documents

To identify suitable concepts from WordNet by analyzing the text document is the main challenge.When retrieving/identifying concepts.it is important to make sure that irrelevant concepts should not be extracted and relevant concepts should not be discarded.Words can represent multiple concepts and different words can represent the same or very similar concepts.The input text documents should be analysed and process to extract relevant information.To retrieve semantic concepts form the document, a four-stage extraction process is invoked[3]. This includes: (1) concept selection, (2) relationship retrieval, (3) constraint discovery.

4.1.2 Term weighting

One of the simple representation of documents in information retrieval is a collection of terms corresponding to all the words contained in the documents.The classical approach for doing this is term weighting.weights indicate the frequency of words appearing in the document.The frequency(number of occurrences) of each word can be calculated by constant rank frequency law Zipf

$$\text{Frequency.rank} \approx \text{constant} \qquad (4.1)$$

Where rank is obtained by sorting words by frequency in decreasing order. Hence the frequency of a given word multiply by the rank of the word equal to the frequency of another word multiply by its rank. A method to find term weighting is term frequency $\boldsymbol{tf_{i,j}}$ where each word $\boldsymbol{t_i}$ is calculated as per the number of occurrences of the word associated with the term in document d_j. One popular global weight is inverse document frequency which assigns the level of discrimination to each word in collection of terms in a document. A word appearing in most items should have lower global weight than words appearing in few items.

$$idf_i = \log N/n_i \qquad (4.2)$$

here $\boldsymbol{n_i}$ are the No of item in which term $\boldsymbol{t_i}$ appear and N is the total number of documents in collection.The approach which states that a weight to each word in adocument depending not only on the local frequency of the word in the item,but also the resolving power of that word in the collection of document is known as ***tf.idf*** (term frequency-inverted document frequency).

V. IMPLEMENTATION AND RESULT

The retrieval of semantic concepts for the given text document have been implemented successfully using Java, the most powerful platform independent language .The retrieved semantic concepts will be used to generate the taxonomy for the ontology generation. JDK and Net Beans IDE 6.7.2 are used to develop the application. WordNet 2.1 is used as the knowledge source. The extraction of the required concepts has been done by using the following steps.

1. Text documents which are to be extracted are stored in a folder called input. Any number of text documents can be stored in the above folder. Fig 2.

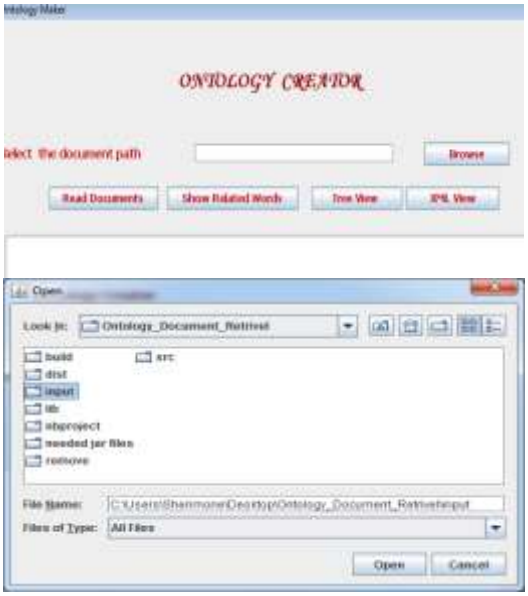

Fig 2. Selecting documents from input folder

All the text documents are read from the input folder and adding to the array list. The stream of texts is broken into words, phrases, symbols, articles, pronouns and prepositions. (Tokenization).Unwanted terms like articles, pronouns and prepositions etc. are removed from the array list. Stemming is used to generate a group of words of nouns from the present set of words. At the end of stemming process we get a group of nouns from all input documents. Frequency of the each word in the group is checked in each document and the whole documents using the formulae Ttf_idf= Math.log10 (tdf+1)*Math.log10 (N/NT). (fig.4). The word which gets highest frequency weight will come at the root of the taxonomy. Synonyms, hypernyms and hyponyms are extracted for each word with the help of WordNet by the usage of appropriate functions.

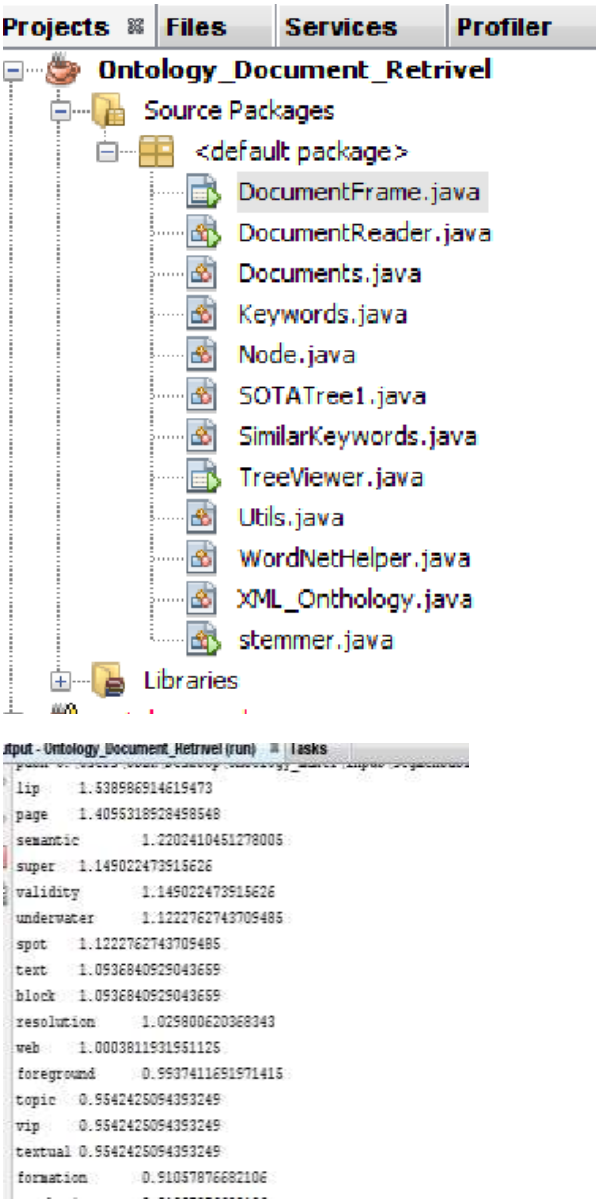

Fig. 3- calculation of frequency weightage

The sample input document is shown below (fig.4) from which the stemming done. The frequency weightage calculated is shown in the ontology creator. Synonym, Hypernym and hyponym are extracted for the word 'interval' is shown below. For the word 'interval', frequency calculated is 1.30102999 and synonyms extracted are time interval,separation,interval etc.. Hypernymas are credibility, credibleness, believability. Hyponym are effect and force.

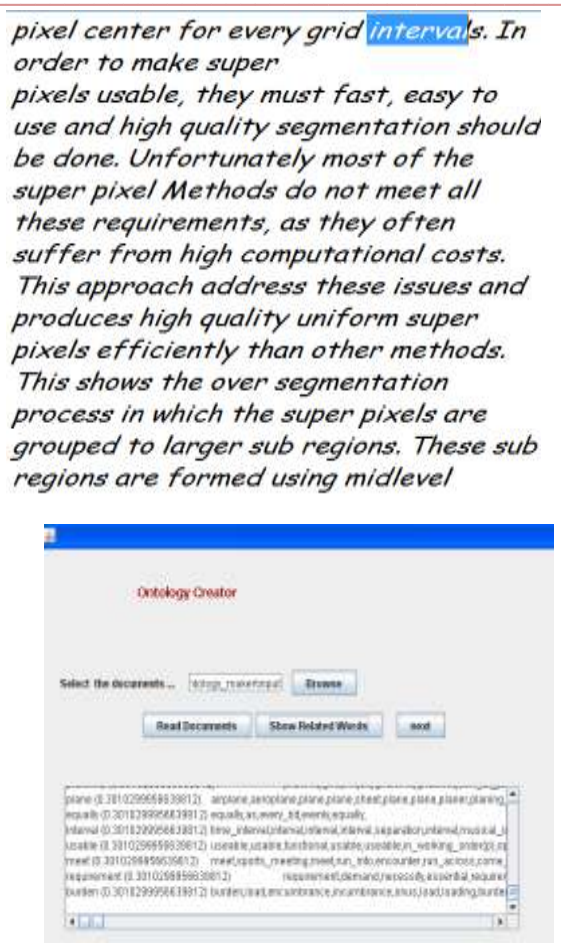

Fig.4.sample input document and extracted words

Synonym represents different words with almost with similar meaning. Hypernyms and hyponyms represent a general category and a specific instance of that category. A hyponym shares a type-of relationship with its hypernym. For example Toyota, Ford, Nissan are all hyponyms of Car (their hypernym) which in turn a hyponym of Vehicle. *Is-a* relationship is generally used to represent the hyponym and hypernym relationships. For example *Car is-a Vehicle* can be used to describe the hyponymic relationship between car and vehicle. WordNet 2.1 browser is used to find the synonyms, hyponym and hypernym of the input document. In the tree view,the frequency weighteage,synonyms,hypernyms and hyponyms are shown in hierarchical way.


```
<?xml version="1.0" encoding="UTF-8"?>
<Ontology>
  - <keyword>
      lip
      <similar>lip,sass,sassing,backtalk,back_talk,m
      <value>1.538986914619473</value>
      <hypernym>external_body_part,,</hypernym>
      <Hyponym>overlip,,</Hyponym>
    </keyword>
  - <keyword>
      page
      <similar>page,Page,Sri_Frederick_Handley_Pa
      <value>1.4095318928498548</value>
      <hypernym>leaf,folio,,summon,,</hypernym>
      <Hyponym>full_page,,</Hyponym>
    </keyword>
```

Fig.5 Tree view and Xml file

VI. CONCLUSION AND FUTURE WORK.

Keyword based Retrieval leads to inaccurate and incomplete results when different keywords are used to describe the same concept in the documents and in the queries. Concept based retrieval methods are the solution for this scenario. This gives the benefit of getting semantic relationships among concepts in finding relevant documents. Also elimination of irrelevant documents by identifying conceptual mismatches is another benefit obtained from this. The Initial step is the concept based extraction from wordNet. Words and phrases are the linguistic representatives of concepts. The extraction of the concepts is achieved by breaking into words, phrases, symbols, articles, pronouns and prepositions. (Tokenization).

Unwanted terms like articles, pronouns and prepositions etc. are removed from the array list. Stemming is used to generate a group of words of nouns from the present set of words. At the end of stemming process we get a group of synonym, hyponym and hypernyms of each word. Frequency of the each word in the group is checked. At the end of this phase, semantically related words and their relationship will be extracted from the input document with the help of knowledge base, WordNet. These concepts and their relationships are the source for automatic construction of ontology. The construction of ontology from the extracted words is identified as the future work of this paper.

REFERENCES

[1] Hele-Mai Haav, An Application of Inductive Concept Analysis to Construction of Domain-specific Ontologies. Akadeemia tee 21, 12618 Tallinn, Estonia ...

[2] W.Bruce Croft, What Do People Want from Information Retrieval? www.dlib.org dlib /november95 /11croft.html.

[3] Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, James Mayfield, Information Retrieval on the Semantic web, http://www.csee.umbc.edu/~finin//papers/cikm02/cikm02.pdf.

[4] Rifat Ozcan, Y. Alp Aslandogan, Concept-based Information Retrieval Using Ontologies and Latent Semantic Analysis, *www.cse.uta.edu/research/pblications/Downloads/CSE-2004-8.pdf*

[5] Hele-Mai Haav,,Tanel-Lauri Lubi, A Survey of Concept-based Information Retrieval Tools on the Web, 5th East-European Conference, ADBIS 2001 Vilnius, Lithuania: (2001) .

[6] Ide, N., J.Véronis. Word Sense Disambiguation: The State of the Art. Special issue of Computational linguistics on Word Sense Disambiguation, 24:1, Pages 1-40, 1998.

[7] Christian Safran, A Concept-Based Information Retrieval Approach for User-oriented Knowledge Transfer, Master's Thesis, 10th December 2005.

[8] http://wordNet.princeton.edu/

[9] Fensel, D]2001], Ontologies: Silver bullet for knowledge management and electronic commerce. Springer-Verlag, Berlin.

[10] Asuncion Gomez Perez and V. Richard Benjamins [1999], Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods. IJCAI- Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends.

[11] R. Bodner and F. Song[1996], "Knowledge-based Approaches to Query Expansion in Information Retrieval," in Proc. of Advances in Artificial Intelligence, pp. 146-158, New York,Springer.

[12] Lopez, M.F.[1999], "Overview of the methodologies for building ontologies". Proceedings of the IJCAI- 99 Workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, Sweden, August.

[13] G.Madhu and Dr.A.Govardhan Dr.T.V.Rajinikanth[2011] , Intelligent Semantic Web Search Engines: A Brief Survey.

[14] Henrick Bulskov Styltsvig.Ontology based Information Retrieval , http://coitweb.uncc.edu/~ras/RS/Onto-Retrieval.pdf.