DEVELOPMENT OF SEMANTIC BASED INFORMATION RETRIEVAL USING WORD-NET APPROACH

Ms. P. Nithiya*, Ms. V. Vidhya, Dr. L. Ganesan

Address for correspondence
*P. Nithiya, PG Scholar Dept of CSE, Sri Venkateswara college of Engineering
Email nithiya@svce.ac.in
Ms. V. Vidhya, Asst. Professor
Dr. L. Ganesan, Prof. and HOD, Dept. of CSE, Alagappa Chettair College of Engg. & Tech, Karaikud

ABSTRACT
Semantic-based information retrieval mechanism that handles the processing, recognition, extraction, extensions and matching of content semantics to achieve the following objectives: i) to analyze and determine the semantic feature of the content and to develop a semantic pattern that represent the semantic features of the content. ii) to analyze user’s query and extend its implied semantics through semantic extension to identify more semantic features for matching. iii) to generate contents with approximate semantics by matching against the extended query to provide correct contents to the querist.

KEYWORDS: Information Retrieval, Keyword Extraction, Semantic Extraction

INTRODUCTION
A semantic enable information retrieval mechanism that features information retrieval based on semantics consisting of elements like subject, predicate and object, has been developed to effectively offset existing defects and constraints of the traditional keyword-based search, and consequently to enable the user to obtain required information with semantics based query. Information retrieval is a field that concerns the structure, analysis, organization, storage, searching and retrieval of information (Salton, 1989). The process of information retrieval can be made intelligent to help to the users in their search for information.

The existing information retrieval systems are mostly keyword-based and identify relevant documents or information by matching keywords. It failed to represent the complete semantics contained in the content (Oh, Myaeng & Jang, 2007). It retrieves more number of irrelevant documents too.

SEMANTIC BASED INFORMATION RETRIEVAL
Semantic information retrieval model handles the processing, recognition, extraction, extensions and matching of content semantics. It utilizes semantic pattern as the basis to transform query of the querist and the semantic features
in the content repository into semantic pattern to identify the corresponding contents by matching against the semantic pattern. The proposed system implements semantic based information retrieval using WordNet, a freely available lexical database for the English language, acts as a combination of a dictionary and thesaurus. The intention of WordNet is to map the relationships between words in a manner similar to the way the human mind stores and uses language. As the search queries used in the information retrieval process will often be entered as the query exists in the user's consciousness, WordNet is an ideal candidate for integration into a search engine, potentially adding features beyond the scope of typical engines.

In the proposed work, the user’s query will be analyzed in the semantic extraction and determination module to extract its semantic features for the purpose of determining contents of the query and representing them in a structured and materialized semantic pattern. In this module the semantic elements are identified and predicate in the content semantics and analyze their semantic relations, to be followed by the integration and simplification of semantic relations with Word Net. Now the semantic extension module will identify other potentially relevant semantic features based on semantic features of the query and include them into the query patterns. This will increase the number of semantic features in the query as the basis for matching. The next step is to extract various query patterns through semantic pattern extraction module. Semantic pattern is developed for each content in the content repository, to be followed by indexing based on semantic patterns. Indexing is done for fast retrieval of information. Finally pattern-matching module is utilized to identify the most approximate content among the content repository and submit it to the querist.

Fig. 1 Block diagram of the proposed system

SEMANTIC IR APPROACHES
In the proposed system, there are four modules, they are as follows:

1. Semantic Determination and Extraction.
2. Semantic Pattern Extension.

**Semantic Determination and Extraction**

Here the user’s query will be analyzed to extract its semantic features for the purpose of determining contents of the query and representing them in a structured, materialized semantic pattern. This module contains three parts.

**Content Preprocessing**

Preprocessing is a basic process in semantic retrieval that involved analyzing and retrieving contents from content repository and converting them into tokens. It is then processed with stop-word list to remove irrelevant terms like pronouns, determinants, articles and symbols and to convert variants of verbal nouns and participles into their original form for the purpose of reducing the volume of terms being processed. Retrieve nouns and verbs from these processed terms with part of speech analysis.

**Content Summarization**

The purpose of content summarization is to retrieve thinking the author wishes to express by collecting and extracting significant parts in the content. The purpose of content determining is to identify significant concepts and their distributions in the content based on content terms. i.e, to retrieve those parts with significant meanings in the content and minimizes repetitiveness in those parts and hence the keywords are extracted with parts of speech analysis by using tree tagger.

**Semantic Identification and Representation**

The process of semantic identification and representation started with analyzing the semantic relations between the elements, to be followed by the integration and simplification of semantic relations with WordNet 2.1. Initially parts of speech (pos) for each of the keyword are extracted. WordNet 2.1 will identify only noun, verb, adverb and adjective pos. Then the synset pattern is developed by considering all the pos that are generated. The semantic relations like kinds-of, part-of, higher and lower level of the keyword are identified and it is then represented as pattern associations like hypernyms, hyponyms, holonyms and meronyms. Noun synsets may be connected as hypernyms, hyponyms, coordinate
terms, holonyms, and meronyms. Verb synset relationships include hypernyms, troponyms, entailments, and coordinate terms.

![Fig. 2 Block diagram for Semantic determination and extraction](image)

### Semantic Pattern Extension

Semantic based IR performs search based on a query entered by querists. A query is usually composed of insufficient and fragmented descriptions, and this may lead to insufficient information for matching, difficulty in determining query topics and the consequent mismatch between the retrieval result and the querist’s requirements. To improve the query performance and to generate more content semantic features for comparison, semantic extension is to mine latent semantics of the query.

The semantic pattern extension module is to address the issue of insufficient amount of information from query content, latent semantic analysis (Landauer, Foltz, & Laham, 1998; Yeh, Ke, Yang, & Meng, 2005) was utilized to analyze latent semantics related to query content by comparing query content against content repository, thereby creating more semantic features for matching. This component solves the problem of query failure due to lack of required keywords in the query.

![Fig. 3 Block diagram of Semantic pattern extension](image)

### Semantic Pattern Extraction

Semantic representation is to determine the semantics of the content and constructing a semantic space that represents the content semantics. The semantics of the text could be marked in the three-dimension semantic space by means of topics and association type, after which the semantics could be represented by one point in the space. With this model, all semantics in the content could be plotted in the semantic space, thus recreating a
semantic pattern that represents the content semantic features. Here the figure 4 is represented to denote that all the keywords must be related to some topics and hence the semantic space is constructed. The keywords are related to one another by means of the following relations: is-a, part-of, extends, kind-of, higher level denoting the word and lower level of the word. These relations form the association type related to some topic. The semantic feature plays an important role here in this proposed work, i.e., noun and adjective were extracted and the patterns or the corresponding associations were identified. The pattern in turn passed on to the retrieval module. The following figure shows the three-dimensional view of topic association.

**Semantic Pattern Indexing and Matching**

Semantic pattern is developed for each content in the content repository, to be followed by indexing based on semantic patterns. In performing information retrieval, the document closest to the query was identified through semantic pattern, and then the most approximate content was identified by comparing with content semantic patterns in the corpus. The query pattern is classified to identify the contents most approximate to the query.

![Fig. 4 Semantic Space](image)

Finally the pattern matching component identifies the most approximate content in the content corpus and submits it to the querist. This module involves the following parts:

**Pattern Preprocessing**

It is the basic step in semantic pattern matching in which it filters out all the unwanted patterns in the extended query semantic pattern and process the patterns of type hypernym, hyponym, holonyms and meronyms.

**Pattern Classification**

Based on the semantic features, the patterns for the corpus content were generated. This is to be followed by matching contents in that corpus to reduce the need for matching large volumes of data.

**Pattern Indexing**
In this step, all the pattern documents are indexed for fast retrieval and to avoid duplication in matching. Matching is performed between the extended query pattern and the index which is developed to identify the document to which the query pattern belongs.

**Pattern Matching**
In pattern matching, the similarity between the query and each content in the corpus is computed and finally, such contents were sorted by the order of similarity to identify and submit the most approximate content to the querist.

**IMPLEMENTATION AND EXPERIMENTAL ANALYSIS**
To evaluate the proposed system in this study, a semantic based information retrieval system was implemented, using natural language query as the query input, and the results were then compared with those of a traditional keyword-based IR system to analyze the differences and relative strengths and weaknesses.

The proposed semantic based information retrieval system was implemented in Net Beans 6.0 and executed on a windows environment with a 3.2 GHz Pentium 4 CPU and 2 G Bytes of memory. User defined corpus were used as the source of experiment data.

Processes of this experiment began with the semantic extraction module to extract important semantics of the user query based on parts of speech using tree tagger and transformed them into semantic patterns.

Before generating the patterns, extension is done at the same time for each significant word.

On the other side, patterns were extracted for all the documents in the corpus.

**Fig. 5 Block diagram for Semantic Pattern indexing and matching**
Next, a total of 15 queries of one word, two word and three word queries were developed to be used by the system. Finally, the queries were fed into the semantic-based IR system to perform for matching and check if the search results matched the documents present.
in the corpus, and compare the results with those of a keyword-based IR system to analyze any differences between them.

The effectiveness of information retrieval system is usually measured by precision rate, which refers to the percentage of correct search results in an IR system.

**SAMPLE SEARCH QUERIES AND EXPECTED HITS**

The values displayed in the below tables shows the results of the evaluation and display the precision and recall rates for queries consisting of a single word using both semantic and keyword based IR. The results are to be read as follows. The first column, exp. hits, shows the documents of the corpus that are expected to be found by the search. The second column, hits, denotes the documents that have been presented as a result set for the query term. Finally, two columns display the precision and recall rates. F-measure can then be calculated. The sample queries were set as an input in the order of increasing length. The same is repeated for two word and three word queries as well.

**Table 1: Results of query matching for one word queries using semantic based IR**

<table>
<thead>
<tr>
<th>One word Queries</th>
<th>#Exp hits</th>
<th>#Hits(Total)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compiler</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>0.85</td>
<td>0.9</td>
</tr>
<tr>
<td>Design</td>
<td>6</td>
<td>9</td>
<td>1</td>
<td>0.9</td>
<td>0.95</td>
</tr>
<tr>
<td>Parser</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Grammar</td>
<td>9</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Code</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 2: Results of query matching for one word queries using keyword based IR**

<table>
<thead>
<tr>
<th>One word Queries</th>
<th>Exp hits</th>
<th>Hits(Total)</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compiler</td>
<td>7</td>
<td>9</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>Design</td>
<td>6</td>
<td>8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>Parser</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammar</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Code</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>0.6</td>
<td>0.8</td>
</tr>
</tbody>
</table>
The graph is plotted between the sample preset queries and the precision rate for both semantic based retrieval and keyword retrieval. Similarly comparison is done in terms of recall rate. As indicated in the below figure the semantic based IR system that performs matching based on semantic features. It demonstrated that, irrespective of the length of the user query, it gives response and as the length of the query increases, precision rate is higher.

**CONCLUSION**

In this study, the proposed approaches can efficiently and precisely perform semantic based information retrieval. It has been developed to offset the existing defects of the traditional keyword-based search, and, consequently, to enable the user to obtain required information with semantics-based query. Experiments in
this study also demonstrated that the proposed approach can efficiently and precisely perform semantic based information retrieval. In conclusion, the present work shows that adding search functions based on the implicit semantics. The work has been done with limited corpus and indexing, it can be extended by improving the corpus with a combination of clustering and indexing. In future, the work can be extended by integrating the LSA-based results into the metadata-based advanced searching would combine the advantages of latent semantics and metadata. Ontology based named entity retrieval may be new strategy to further improve the search facility.

REFERENCES