Automatic Spell Correction of User query with Semantic Information Retrieval and Ranking of Search Results using WordNet Approach

Kirthi J[1] , Neeju N.J[2] and P.Nithiya [3]

**[1]U.G student, Dept of Information Technology, Sri Venkateswara College of Engineering
Sriperumbudur, Tamil Nadu 602 105 , India**

**[2]U.G student, Dept of Information Technology, Sri Venkateswara College of Engineering
Sriperumbudur, Tamil Nadu 602 105 , India**

**[3]Assistant Professor, Dept of Information Technology, Sri Venkateswara College of Engineering
Sriperumbudur, Tamil Nadu 602 105 , India**

Abstract

The proposed semantic information retrieval system handles the following : i) automatic spell correction of user query. ii) analysis and determination of the semantic feature of the content and development of a semantic pattern that represents the semantic features of the content. iii) analysis of user's query and extension of implied semantics through semantic extension to identify more semantic features for matching. vi) generation of contents with approximate semantics by clustering the documents and matching against the extended query to provide correct contents to the querist. v) a ranking method which computes relevance of documents for actual queries by computing quantitative document–query semantic distance.

Keywords: *Information retrieval, Semantic retrieval, Semantic ranking, Text retrieval, Semantic extraction, Retrieval mechanism.*

1. Introduction

As large amounts of digital information become more and more accessible, the ability to effectively find relevant information is increasingly important. Search engines have historically performed well at finding relevant information by relying primarily on keyword based techniques.

Despite the ease of use,it fails to represent the complete semantics contained in the content. The technique also retrieves duplicate contents. Thus, refinement of user query becomes necessary to retrieve required contents. Moreover, Polysemy and synonymy are not considered.
So user query in such technique represent fragmented meaning of the content. Hence the relationship between queries and document limited to lexical analysis. So the above technique takes place without respect to semantics, or word meanings. This is perhaps due to the fact that the idea of meaningful similarity is naturally qualitative, and thus difficult to incorporate into quantitative processes.

Another existing technique is the semantic web which is mostly called as web crawlers. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site.

The Semantic information retrieval system is the emerging information retrieval technique that overcomes the above drawbacks by handling the processing, recognition, extraction, extensions and matching of content semantics to analyze user's query and extend its implied semantics through semantic extension to identify more semantic features for matching. (Ming-Yen Chen, Hui-Chuan Chu and Yuh-Min Chen, 2009)

The search results generated by the retrieval mechanism are not necessarily in the 'relevance order'. Hence, a ranking method is used which computes relevance of documents for actual queries by computing quantitative document –query semantic distance. Automatic spell correction of user query is included in this semantic system because spelling errors are common in user queries. An algorithm is used for comparing the preprocessed query words against a known list of correctly spelt words.

2. Overall Architecture of Semantic Information Retrieval System

The information retrieval system must analyze the word based digital contents stored in various formats like *.txt, *.html or *.xml and perform collection and conversion of digital contents in different formats before searching for the corresponding contents based on query entered by the querist so as to provide the querist the required information in a factual and accurate manner (O'Leary, 1999; Zantout and Farhi, 1999). To improve the efficiency of information query and to enable the effective use of information query mechanism by the user in of information query mechanism by the user in searching for information for different applications, a semantic-based approach must be adopted to enable understanding of human languages by computers, thus reducing the difficulty in discerning and extracting content semantics (Feng, Millard, Woukeu, & Davis, 2005). As shown in Fig 1, the three major tasks in the proposed semantic information retrieval system are automatic spell correction of user query, semantic information retrieval and ranking of search results.

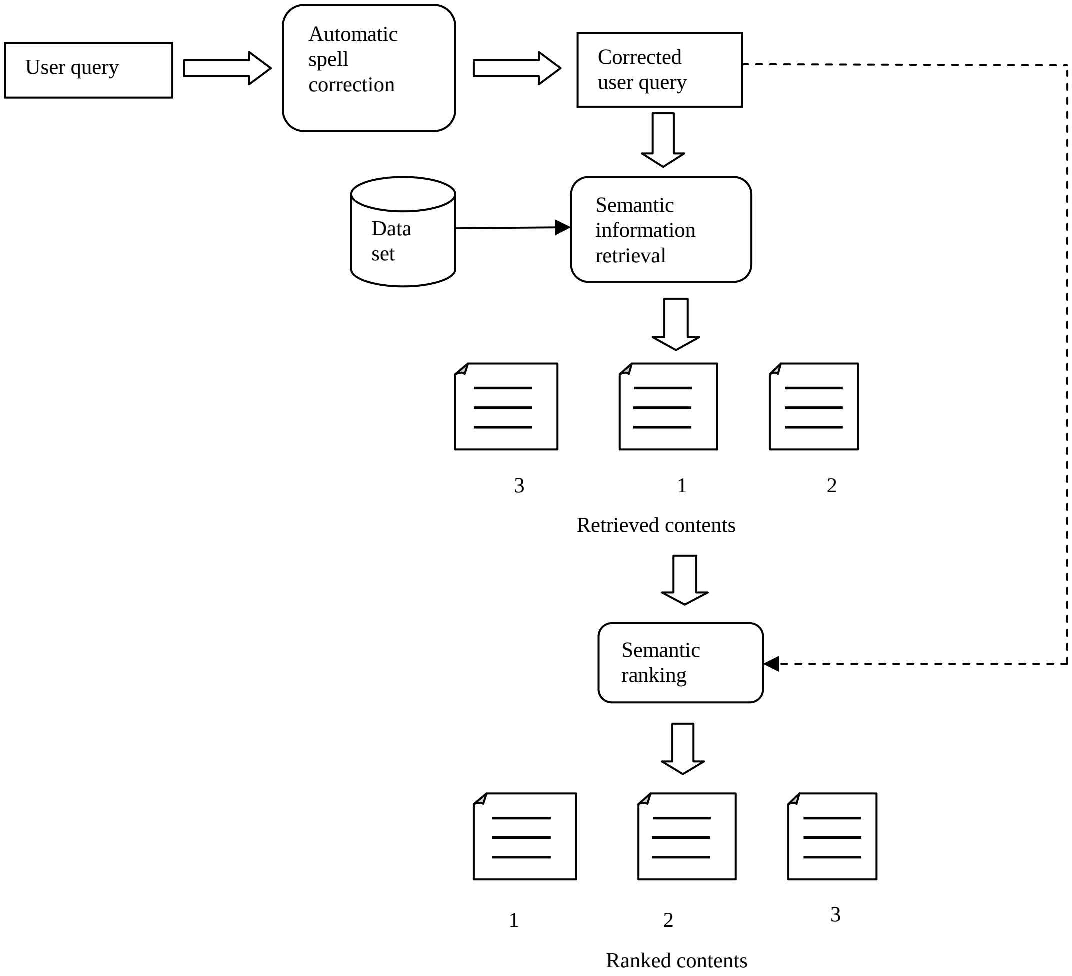

Fig 1. Overall Semantic Information Retrieval System

2.1 Automatic Spell Correction

User query is first scanned for errors. The misspelled words are identified and checked against the dictionary. Each misspelled word is replaced with the best suggestion from the dictionary. The corrected query is given as input to the Retrieval module.

2.2 Semantic Information Retrieval

In the semantic information retrieval mechanism, 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 component 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 WordNet. 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.

There are four modules, they are as follows:
1. Semantic Determination and Extraction.
2. Semantic Pattern Extension.
3. Semantic Pattern Extraction.
4. Semantic Pattern Indexing and Matching

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.

2.3 Ranking of Search Results

The search results generated by the retrieval mechanism are not necessarily in the ‘relevance order’. Hence, a ranking method is used which computes relevance of documents for actual queries by computing quantitative document–query semantic distance. A tool is used that identifies top 5 words that are semantically close to the document title .A matrix is constructed using the top 5 words across column and query words across row for each document. The matrix columns are filled with the semantic distance between each word and query word for the document. Average score ranging between 0 and 1 is computed for each document. Documents are sorted in ascending order, hence ranked**.**

3. Automatic Spell Correction of User query

Automatic spell correction task performs correction of misspelled words in the user query by comparing the preprocessed words against a dictionary of correctly spelt words.Fig.2 shows the detailed steps involved in performing spell correction of user query.

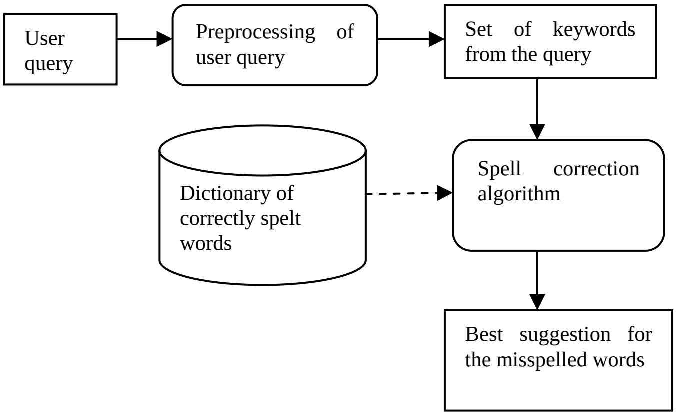
User query
Preprocessing of user query
Set of keywords from the query
Dictionary of correctly spelt words
Spell correction algorithm
Best suggestion for the misspelled words

Fig.2. Automatic spell correction of user query

A set of routines is used for scanning the entered user query and then preprocessing of user query is performed. In preprocessing, the keywords of the query are extracted i.e. the prepositions, articles, pronouns and punctuations are neglected.

The dictionary is a set of correctly spelt words with its frequency of occurrence in the dataset. The spell correction algorithm stores the dictionary words in ternary search tree format as it combines the compact size of a binary search tree with the speed of a digital search tree, and is therefore ideal for practical use in sorting and searching data. This data structure is faster than hashing for many typical search problems, and supports a broader range of useful problems and operations. Ternary searches are faster than hashing and more powerful, too. The theory of ternary search trees was described at a symposium in 1997 ("Fast Algorithms for Sorting and Searching Strings," by J.L. Bentley and R. Sedgewick). The algorithm described by Bruno Martins takes the misspelt word and frequency of its occurrence and returns the most similar word from the dictionary using the Levenshtein distance to measure the similarity. Levenshtein distance (LD) is a measure of the similarity between which we will refer to as the source string (s) and the target string (t). The distance is the number of deletions, insertions, or substitutions required to transform s into t. For example, If s is "test" and t is "test", then LD(s,t) = 0, because no transformations are needed and the strings are already identical. If s is "test" and t is "tent", then LD(s,t) = 1, because one substitution (change "s" to "n") is sufficient to transform s into t. The greater the Levenshtein distance, the more different the strings are. The measure is named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965.

4. Semantic Information Retrieval Mechanism

In the semantic information retrieval mechanism, there are four modules as shown in fig.2, they are as follows:

1. Semantic Determination and Extraction.
2. Semantic Pattern Extension.
3. Semantic Pattern Extraction.
4. Semantic Pattern Indexing and Matching

4.1. 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.

4.1.1 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. Since topics in a specialized field are usually expressed in nouns while their associations are expressed in verbs, it is therefore necessary to retrieve nouns and verbs from these processed terms with part of speech analysis before proceeding with the subsequent content summarization phase.

4.1.2 Content Summarization

The purpose of content summarization is to retrieve thinking the author wishes to express by collecting and extracting those parts with significant implications in the content based on significance of the content and the author semantics. The human perception and understanding of things is based on concepts, and human thoughts can be expressed in structured, written contents in a logical organization of words. 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.

4.1.3 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. Thus, the possible relationships shared between WordNet synsets vary depending on the type of word. Noun synsets may be connected as hypernyms, hyponyms, coordinate terms, holonyms, and meronyms. Verb synset relationships include hypernyms, troponyms, entailments, and coordinate terms. Possible adjective relationships are related nouns, similar to, and participle of verb. Adjective synsets may be related as root adjectives. Words may also be connected through lexical relations such as antonyms too.

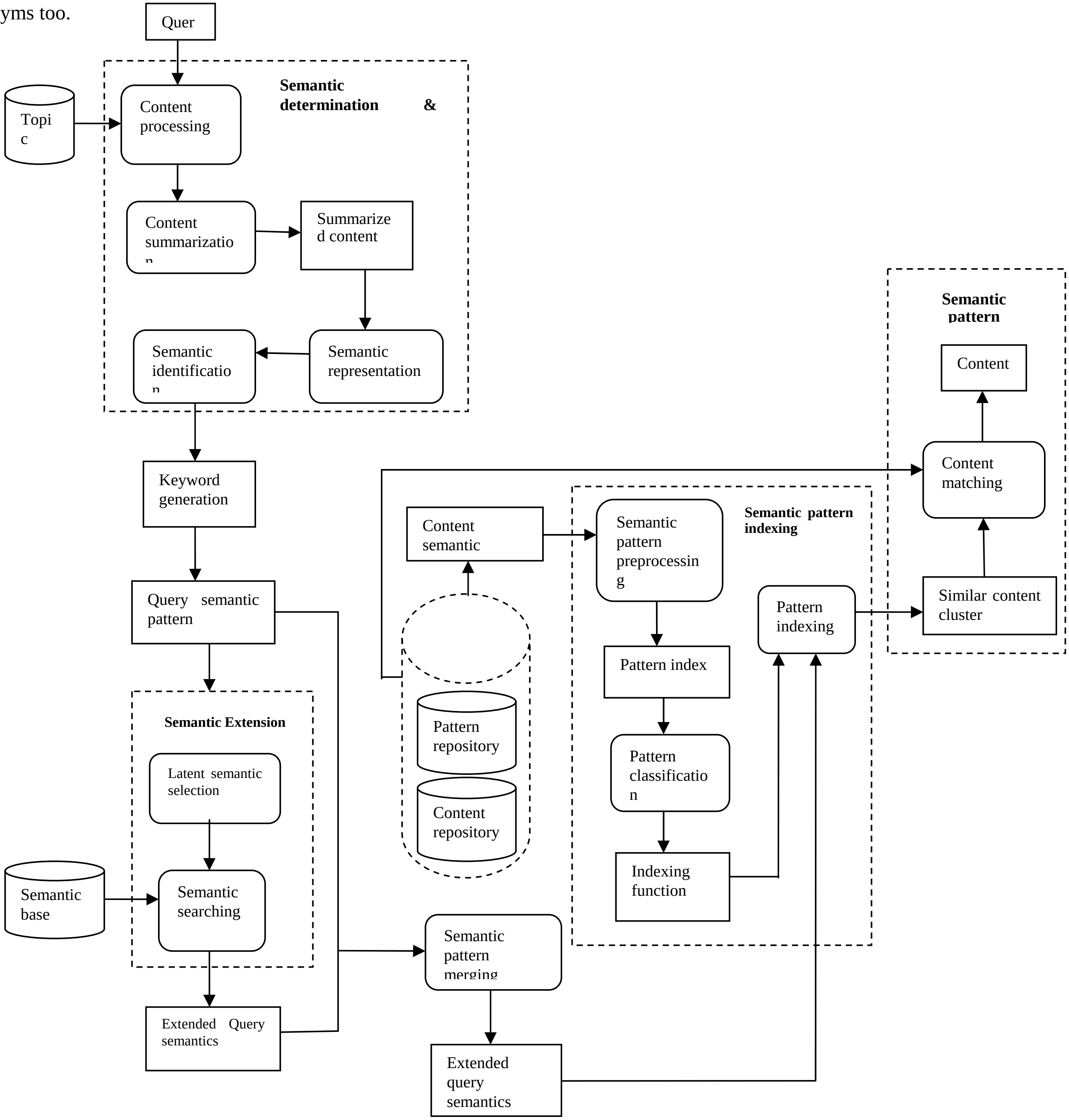

Fig.3 Semantic Information Retrieval mechanism

4.2 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 query performance through generating more content semantic features for comparison, semantic extension is to mine latent semantics of the query, and latent semantics in the query semantic pattern were extended to serve as the reference for comparison. The statistical latent semantic analysis is to identify latent semantic relations of the content.

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

4.3 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 records and represents the content semantic features. This semantic pattern could be utilized for operations and applications of the content and it could be converted and expressed in the forms of tag, image or database to meet content management requirements.

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 i.e., it should extract noun, verb, adjective, adverb and determinants. Then in this proposed work, mainly 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.

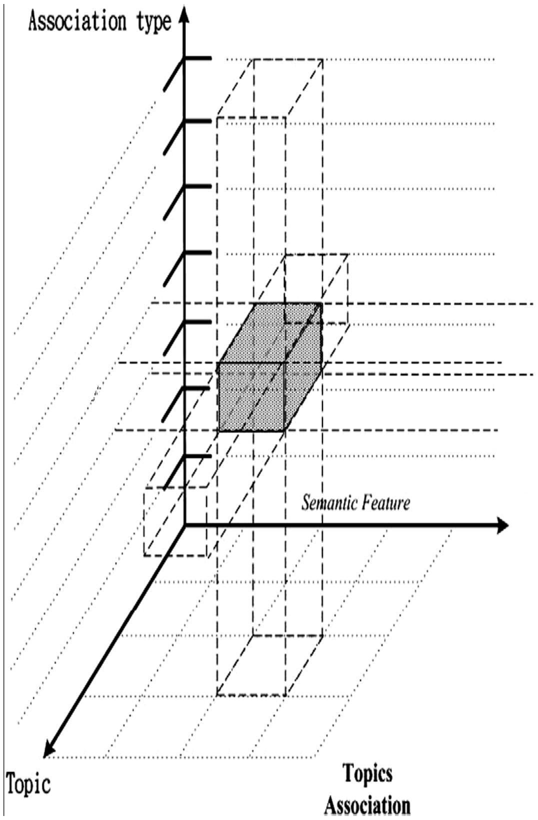

Fig. 4 Semantic Space

4.4 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. 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:

4.4.1 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.

4.4.2 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.

4.4.3 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.

4.4.4 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.

5. Ranking of Search Results

The semantic information retrieval mechanism generates the relevant contents for the given query. But the search results generated by the retrieval mechanism are not necessarily in the ‘relevance order’. Hence, a ranking method is used which computes relevance of documents for actual queries by computing quantitative document–query semantic distance. The ranking mechanism is shown in fig. 4.

The overall strategy to capture semantic similarity between query and document can be described as follows:

1. Document is represented by top N high frequency words, which are obtained by measuring the semantic closeness with the document contents.

2. Building a semantic similarity matrix R[m,N] of the preprocessed query key words(m) and the top N words that describe the document.

3. To compute the overall score, the matching average ranging between 0 and 1 is calculated for the matrix constructed.

There are several ways to determine the conceptual similarity of two words for constructing the R[i,j] matrix elements. . Topographically, this can be categorized as node based and edge based approaches, which correspond to the information content approach and the conceptual distance approach, respectively.

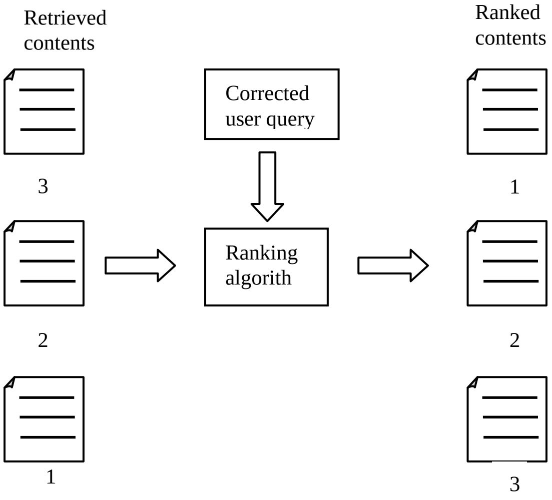

Fig.5 Ranking mechanism

The edge based approach is a more natural and direct way of evaluating semantic similarity in a taxonomy. It estimates the distance (e.g. edge length) between nodes which correspond to the concepts/classes being compared. Given the multidimensional concept space, the conceptual distance can conveniently be measured by the geometric distance between the nodes representing the concepts. Obviously, the shorter the path from one node to the other, the more similar they are.
For a hierarchical taxonomy, Rada et al. (1989) pointed out that the distance should satisfy the properties of a metric, namely: zero property, symmetric property, positive property, and triangular inequality. Furthermore, in an IS-A semantic network, the simplest form of determining the distance between two elemental concept nodes, A and B, is the shortest path that links A and B, *i.e.* the minimum number of edges that separate A and B (Rada et al. 1989).

With regard to network density, it can be observed that the densities in different parts of the hierarchy are higher than others. For example, in the plant/flora section of WordNet the hierarchy is very dense. One parent node can have up to several hundred child nodes. Since the overall semantic mass is of a certain amount for a given node (and its subordinates), the local density effect (Richardson and Smeaton 1995) would suggest that the greater the density, the closer the distance between the nodes (*i.e.* parent child nodes or sibling nodes).

6. Conclusions

In this study, we propose a novel and theoretically sound semantic information retrieval system. The proposed system automatically corrects the misspelled preprocessed words of user query, which is then used as an input to the semantic retrieval mechanism. In addition to semantic-based information retrieval, the proposed system has two significant parts: a semantic extension model to generate more semantics for matching, thereby solving the problem of insufficient information for query; and a semantic clustering model which uses bisecting k-means clustering algorithm and then performs content matching in that category, thereby improving matching accuracy. The retrieved contents are not necessarily in the 'order of relevance', hence semantic ranking mechanism is used to rank the retrieved contents. Semantic based information retrieval can be used in knowledge management, document management and other applications that require searching large quantities of information.

References

[1] Ming-Yen Chen, Hui-Chuan Chu and Yuh-Min Chen " Developing a semantic-enable information retrieval mechanism", Elsevier Journal on Expert Systems with Applications, Vol. 37 ,Issue 1, May 2009.

[2] P.Nithiya, V.Vidhya and Ganesan,L. " Development of semantic based information retrieval using WordNet approach", in Computer and Network Technology (ICCNT), 2010 Second International Conference , 2010.

[3] Nathan S Davis, " An analysis of document retrieval and clustering using an effective semantic distance measure", M.S. thesis, Department of Computer Science,Brigham Young University , Provo, Utah, United States,December 2008

[4] G.Salton,"Automatic Text Processing", Addison-Wesley, 1989

[5] Ahmed Abdelali, Jim Cowie and Hamdy S. Soliman, "Improving query precision using semantic expansion", Elsevier Journal on Information Processing and Management, 43, 2007,705-716.

[6] Ray Richardson and Alan F. Smeaton, "Using WordNet in Knowledge based approach to information retrieval", Working Paper, CA-0395, Dublin City university, Ireland,1995.

[7] Jay J. Jiang David .Conrath."Semantic Similarity on Corpus Statistics and Lexical Taxonomy", In Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997,

[8] Rada, R., H. Mili, E. Bicknell, and M. Bletner, 1989, "Development and Application of a Metric on Semantic Nets", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, 17-30.

[9] Feng, T., Millard, D., Woukeu, A., & Davis, H. (2005). Managing the semantic aspects of learning using the knowledge life cycle. In Proceeding of fifth IEEE international conference on advanced learning technologies (pp. 575–579). Kaohsiung.

[10] O'Leary, Daniel E. (1999). Internet-based information and retrieval systems.

` Decision Support Systems, 27(3), 319–327.

[11] H.Zantout, and M. Farhi"Document management systems from current capabilities towards intelligent information retrieval: An overview ", International Journal of Information Management, 19(6), 471–484.

[12] Michael Steinbach, George Karypis, and Vipin Kumar, "A Comparison of Document Clustering Techniques", Text Mining Workshop, in Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2000), Boston, MA, August 20-23, 2000

[13] Landauer,T. K., Foltz,P. W.,& . Laham,D.(1998), " Introduction to latent semantic Analysis, " Journal on Discourse Processes, , 25 , PP 259 -284.

Kirthi.J: She is pursuing her B.Tech degree in Information Technology (Sri Venkateswara College of Engineering, Tamil Nadu).

Neeju.N.J: She is pursuing her B.Tech degree in Information Technology (Sri Venkateswara College of Engineering, Tamil Nadu).

P.Nithiya : She has obtained her M.E degree in Computer Science Engineering and B.E degree in Computer Science .She is presently working in Sri Venkateshwara College of Engineering as Assistant Professor (Department of Information Technology).She has presented a paper titled "Development of semantic based information retrieval using WordNet approach", in Computer and Network Technology (ICCNT), 2010 Second International Conference .Her area of interest and research is in the field of Natural Language Processing.