Semantic based Personalized Framework for Information Retrieval

K. Saravanakumar
Asst. Professor
School of Information
Technology and Engineering,
VIT University,
Vellore-632014, India

Mahesh Moturi
Master of Technology in
Information Technology
School of Information
Technology and Engineering,
VIT University,
Vellore-632014, India

ABSTRACT
Traditional information retrieval systems are mostly keyword-based and retrieve documents or information by matching keywords. These systems lack a meaningful description for information, so it is difficult for users to find more relevant information. To provide what a user really needs, a framework of information retrieval based on semantics has been proposed in this paper. In this framework the semantics in the user query are identified and these are summarized according to the context. Then the results are classified into possible domains or groups and displayed to user according to his choice from domain the results are re-ranked. By this framework we provide users a convenient and more precise search service with personalization.

General Terms
Information Retrieval, Clustering, Ranking

Keywords
Semantic Identification, Personalization, Query processor

1. INTRODUCTION
The rapid growth of the World Wide Web has led to vast amount of data available on the web and also the available information increases rapidly. For such contents we need an information retrieval mechanism. Information retrieval is a field that deals with structure, storage, organization, searching and retrieval of information. IR has changed considerably in recent years with the expansion of the World Wide Web. Currently keyword-based information retrieval which performs keyword searching in documents by matching the keywords that users specify in their queries, these systems fail to represent the complete semantics in the query. [1] describes about the disadvantages of the keyword based systems and introduce a semantic based information retrieval mechanism where they extend the user query by introducing additional semantics related to query and then try to retrieve most appropriate results for the user. Many proposals have been made for semantic based information retrieval, a semantic enable information retrieval mechanism which features information retrieval based on semantics consisting of elements like subject, predicate and object has been developed using Word-net a lexical database for English[2]. Information retrieval based on user behavior is one approach where user profiles are used to know the behavior or the search pattern of the user for personalization of the results [3]. IR using user history based on previous search would help more in personalizing as well as improving the search [4]. The user profile is dynamic in nature due to changing of preferences of the user thereby trying to improve the search performance of each user through automatic creation, maintenance and personalization of user preference profiles that include search pattern for each user [5]. Many studies have been made in the field of information retrieval [6] describes about the various information retrieval mechanisms which are used including the keyword based also the concepts related to evaluation and operations on the query and different interface and languages that are used for information retrieval, Many models of information retrieval have come up due to increase of internet usage as well as increase in various types of information.

2. PROPOSED MODEL
The proposed information retrieval model as in Fig:1 consists of preprocessing and clustering modules, all the entities present in each module contribute equally to the model. The in detailed functions of the modules and entities present in the modules are explained in proceeding sections. Following are the modules and their entities present in our model along with their basic function.

2.1 Preprocessing Module
In this module the query is given by the user to our model then that query is subjected to preprocessing where the meaningless words in query like neuter pronouns, symbols are removed and semantics in the word and content are known and are summarized according to the context. Following are the entities in the module

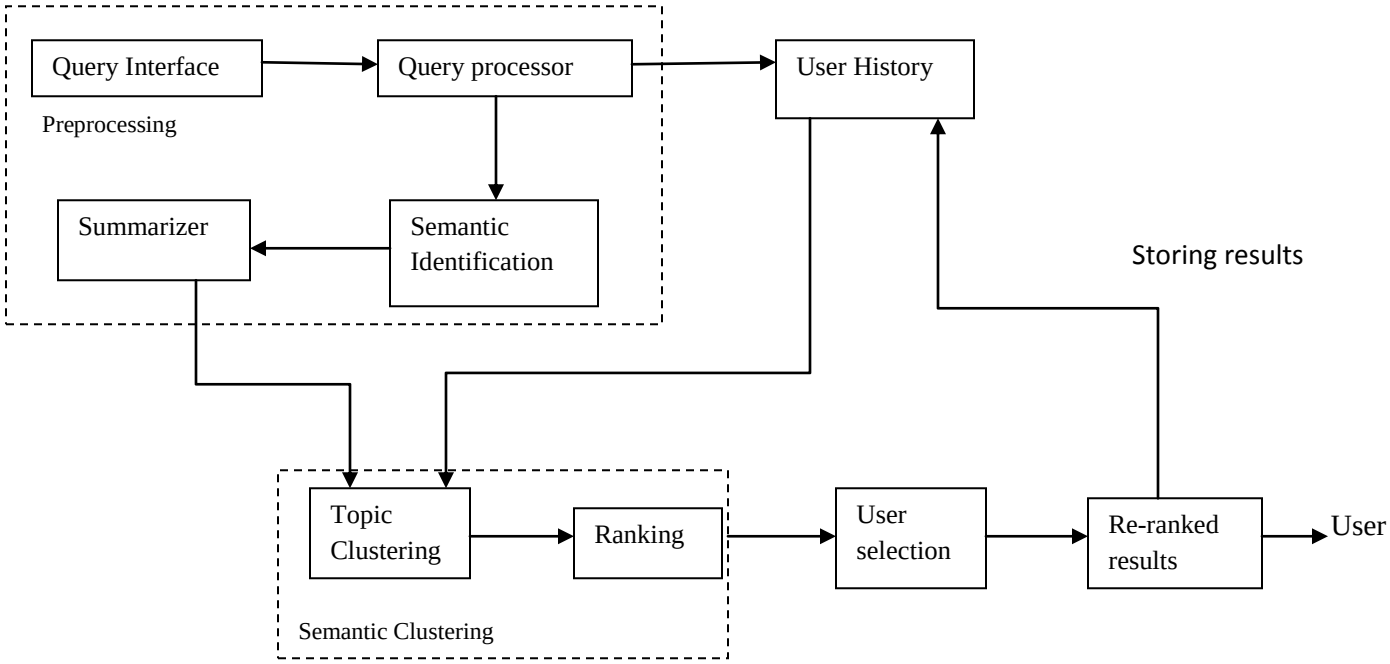

Fig 1: Architecture

2.1.1 Query Interface

The query interface is the GUI between the user and the system, in this the user can give query to the system which will be forwarded to the succeeding sections of the system. QI is the simple and user friendly interface.

2.1.2 Query Processor

The query processor helps in removing the unwanted and meaningless words like neuter pronouns, articles, and symbols from the query thereby reducing the amount of words that would need processing. It also reduces the word to its stem by stemming so that the searching and indexing of the word becomes easy.

2.1.3 Semantic Identification

In this section those words left after stemming and elimination of the stop words their meaning or semantics are known. These semantics of a word can be identified by English lexical database like Word-Net. This Word-net is freely available on the internet, many APIs are also available. Using word-net we can find the synonyms, hyponyms, domain terms of a particular given word.

2.1.4 Summarizer

Summarizer is used to summarize the semantics obtained in the semantic identification step according to the context. The semantics obtained are summarized accordingly to the context of the content. Summarizer scores each sentence in the content available and picks top 30 percent of high scored sentences and summarizes them with the semantics obtained.

2.2 User History

User history is useful in personalized information retrieval where a user profile is also maintained to keep track of the search pattern of the user. User history is a set of queries used in previous searches of the user. The terms used in queries can be classified into concept hierarchy by word-net. Fig2 shows the sample concept hierarchy used in word-net. A user profile is also maintained where user interests are stored it is a two layer hierarchal structure where top layer is a domain layer that includes web search results that were selected by user and bottom layer includes search results that was selected by the user.

2.3 Semantic clustering

In this the results are clustered into domains they belong to i.e. the summarized results are grouped into various possible domains. Then these domains are ranked and displayed to the user. The following entities are present in the Semantic clustering model.

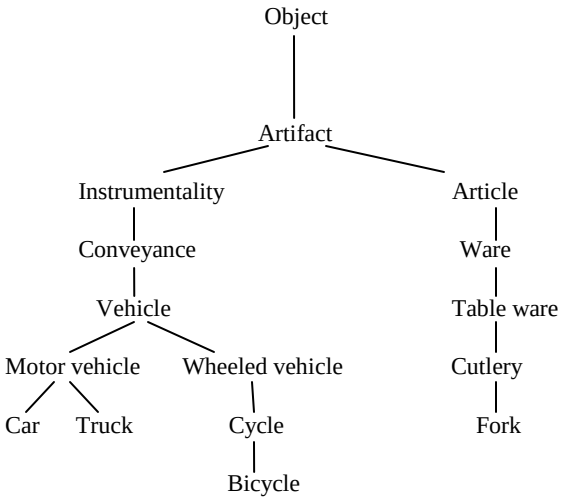

Fig2: Concept hierarchy

2.3.1 Topic Clustering

In this sub module the results obtained from the summarizer and the user history if any related to that search are grouped into clusters and are assigned a topic or a domain for e.g. if for given word "jaguar" , this word can be classified in both animal and cars domain, so results related to "jaguar" are grouped into these two domains accordingly.

2.3.2 Ranking

When the results are grouped into various domains or clusters the clusters are ranked, the ranking of clusters is based on many considerations like no of results and their weights etc in a particular cluster. Then these clusters are ranked and displayed to the user.

2.4 User selection and Re-ranking

The ranked clusters are displayed to the user and when the user makes his choice from the clusters the results are re-ranked according to the weight of each result and displayed to user there by achieving personalization

3. BASIC IDEA

The keyword based system in-spite of its merits regarding ease-of-use, fails to represent complete semantics of the content and leads to many problems like lack of content in query which led to retrieval failure, content identified through keywords did not meet always the user requirements[7],[8]. To overcome all these we have come with semantic based retrieval, Fig 1 represents the framework of proposed model. The methodology of the proposed model as follows :

When user gives query in the user interface the query is passed to the query processor where the query is processed. The query processor removes all the neuter pronouns, symbols etc by creating a stop word lists and the words remaining are subjected to suffix stripping algorithm if required this process is also known as stemming. Porter's algorithm which is a type of suffix stripping algorithm is used for stemming of word.Porter algorithm defines five successively applied steps of word transformation. Each step consists of set of rules in the form <condition> <suffix> → <new suffix>. For example, a rule EED→ EE means "if the word has at least one vowel and consonant also with EED ending, change that ending to EE". So "agreed" becomes "agree" while "feed" remains unchanged. The algorithm is very concise having around 60 rules and is very much readable for a programmer. In terms of computation complexity this is considered one of the efficient ones.

Then the semantics of the words left after processing are identified using Word-net 2.1, which is a English lexical database available on internet and also many of its API are also available. All the semantics identified are summarized with content. Summarizer is used to summarize the semantics according to the context of the content. In [9] the paradigms have proposed for extracting salient sentences from text using features like word and phrase frequency position in the text and key phrases to summarize the scientific documents. While summarizing the main concern is about what the summary content should be. Now after summarizing topic clustering takes place where results belonging to same content or same context are clustered and a topic name is assigned to them, also results from user history are also clustered if the search is previously made..

User history has a user profile which is a hierarchal multilayer structure and a user history which consists of the previous search of the user, this helps in determining the user interests and also helps in personalization. If a query is given for search by user then it is stored in user history, so next time same query given retrieves results from history also if new results found more weight will be given to results from history to that of new ones and are clustered together. Clustering assigns a set of observations to subsets, referred to as clusters, such that observations in the same cluster are similar according to pre specified criteria. Here we use hierarchal clustering to clusters the pages.

So the pages are clustered accordingly into various domains or topics and then these topics are ranked and displayed to user for his selection. The ranking is based on the Tf-Idf algorithm in which

$$\mathrm{TF}(k, t) = \log\left(1 + \frac{n(k, t)}{n(k)}\right)$$

Where n(k) is the number of terms in the document, n(k, t) is the number of occurrences of term t in document k and TF is the Term Frequency.

$$\mathrm{IDF}(t) = \frac{1}{n(t)}$$

The above equation is used to find the Inverse Document Frequency. Where n(t) is the number of documents that contain the term t. The relevance of document k to a set of terms Q can be calculated by

$$r(k, q) = \sum_{t \in q} \mathrm{TF}(K, t) * \mathrm{IDF}(t)$$

Based on this method ranking of clusters takes and when the results are displayed to the user he can choose from clusters upon his choice the results of the cluster chosen are re-ranked and displayed to the user. Then these results are also stored in user history so that when same query is given they can be tracked from user profile or history.

4. RESULTS

To show a comparative analysis between existing system and proposed system many queries have been taken. For the ease here we present the analysis of five queries, the results are shown in table.1.

Table 1. Comparative Analysis of the Systems

Query	Existing System (No of relevant results for given query)	Proposed system (No of relevant results for given query)
Query1	4	6
Query2	7	4
Query3	5	6
Query 4	3	3
Query5	5	7

These results are taken according to the result expected by the user i.e. when the user gives a query the relevancy of results expected by him. For e.g. when user gives query 1 in existing system he gets "10" results but only "4" results show the content needed by the user or we can say relevant to the user query, Whereas the proposed system shows "6" relevant results. The graph shows the comparison between the two systems. Here in fig 3 we can see that with some queries the existing system shows improved results than the existing system and with some queries it fares with existing systems.

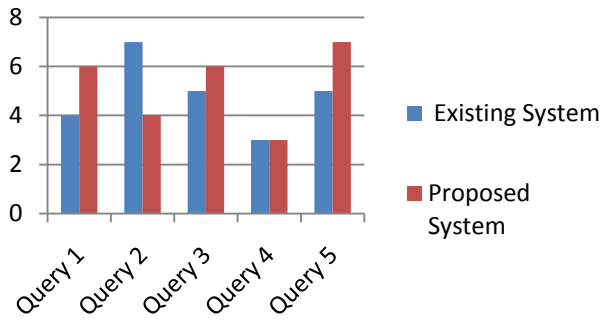

Fig 3: Comparison

5. CONCLUSION

The system model that we have proposed can be used for personalized information retrieval systems as well as intelligent information systems. Since it being a semantic based it will always have advantage over the keyword based systems as meaningful retrieval can be made. Our system should provide better efficiency in retrieving. The future work of our paper would be including lemmatization instead of stemming and also adaptability of dynamic changes to the user profile which changes with frequent changes in interests of user also suggestions for the sub queries related to main queries that can be generated.

6. REFERENCES

[1] Ming-Yen Chen, Hui-Chuan Chu ,Yuh-Min Chen " Developing a Semantic–enable Information Retrieval mechanism " Expert system with Applications 37(2010) 322-240

[2] P. Nithya, V. Vidya, Dr. L. Ganesan "Development of Semantic Based Information Retrieval using Word-Net Approach**"** Second International Conference on Computer and Network Technology.

[3] Hochul Jeon, Taehwan Kim, Joongmin Choi "Adaptive User Profiling for Personalized Information Retrieval**"** Third 2008 International Conference on Convergence and Hybrid Information Technology.

[4] Taehwan Kim, Hochul Jeon, Joongmin Choi**"**Personalized Information Retrieval using user history" 2008 International Conference on Multimedia and Ubiquitous Engineering

[5] Wang Hongsheng, Shu Xiaoming "Personalized Information Filtering Based on Semantic Similarity"

[6] Ricardo Baeza-Yates, Berthier Ribeiro-Neto "Modern Information Retrieval"

[7] Abdelali. A, Cowie. J, & Soliman. H. S, "Improving query precision using semantic expansion" 2007 Information processing and Management , 43, 705-716.

[8] Oh. H. J. Myaeng, & Jang. M. G "Semantic passage segmentation based on sentence topics for question answering " 2007 Information Sciences, 177, 3696-3717.

[9] B. Fung, K. Wang and M. Ester. "Hierarchical Document Clustering Using Frequent Itemsets".SDM' 03, pp. 59-70.