IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-11 E-ISSN: 2347-2693

Information Extraction Using Text Mining by Keyword Ranking and Scoring

Priyanka Gonnade[1*], Sarika Bongade[2] and Tushar Mendhe[3]

[1*,2,3]*Depaprtment of Computer Science and engineering, RGCER, Nagpur, India*

Received: Oct/27/2014 Revised: Nov/10/2014 Accepted: Nov/22/2014 Published: Nov/30/2014

***Abstract*—** As the number of data is stored in a database, searching of a relevant data is the important issue in text mining. Though the today's searching method provides us the relevant data but the numbers of results are too big to find the useful data. The needs of the user vary from time to time and they require different information at every instant of time. Keywords are useful for scanning large documents in a short time. Extracting keywords manually are very difficult and time consuming process. In this paper, we present the technique that are most likely able to satisfy the user's needs and bring useful data in the top positions by extracting keywords from the data present in the database, scoring those keywords based on their occurrences and ranking the data based on keyword scores.

***Keywords*—**Extraction,Scores,Text Mining,Page Rank,Clustering,Open Calais

I. INTRODUCTION

Representing and sharing knowledge has always been a topic of interest to mankind. Indeed, it is one of the main contributing factors for the rapid development of humanity. Sharing and passing knowledge avoids reinventing the wheel, advances scientific insight, inventions, and improvements, and hence humanity as a whole [1].Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT)[2], refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system

Corresponding Author: *Priyanka Gonnade*

II. RELATED WORK

1. Knowledge Discovery from text

The problem of Knowledge Discovery from Text (KDT)[2] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization and question answering.

2. Text Mining For Information Retrieval

Propose an algorithm based on inverted index file. By using the range partition[3] feature of oracle, the space requirement of memory is reduced considerably as the inverted index file is stored on secondary storage and only the required portion of the inverted index file is maintained in the main memory. Fuzzy logic is applied to retrieve the selected documents and then suffix tree clustering is used to group the similar documents.

A method is proposed for learning web structure to classify web documents and demonstrates the usefulness of considering the text content information of backward links and forward links for computing the page rank score.
An approach to text document clustering[4] that overcomes the drawback of K-means and Global k-means as discussed is proposed which gives global optimal solution with time complexity of O (lk) to obtain k clusters from an initial set of l starting clusters.

Index term(s) of the document is/are identified based on key terms of each sentence and paragraph within the document. Rank of the sentence is computed based on the numberof matching terms between the document and sentence index terms.

3. Keywords Extraction Techniques

A lexical chain is a sequence of related words in writing, spanning short (adjacent words or sentences) or long distances (i.e., entire text). A chain is independent of the grammatical structure of the text and in effect it is a list of words that captures a portion of the cohesive structure of the text.

Word co-occurrence [3,5] is extensively used in various forms of research such as content analysis, text mining, construction of thesauri, ontology's, etc. Its aim is to find similarity between words or similarities of meaning among word patterns. The sentences in the document are considered as a set of words; it includes title of a document, section title and a caption.

A graph has been built after doing the basic text preprocessing operations such as stemming and stop words removal[4,6]. Only a single vertex for each distinct word is created even if it appears more than once in the text. Thus each vertex label in the graph is unique. There is a directed edge from the vertex corresponding to the term *x* to the vertex corresponding to term *y,* if a word *x* immediately precedes a word *y* in the same sentence somewhere in the document.

III. PROPOSED SYSTEM

Searching any data from the database gives bulk of result. These results may contain data which are not related with a given search and in some search it may require lot of time for data retrieval as well as for comparison. So we are developing an efficient data retrieval module which would overcome the problems of comparison and retrieval time.

Issues in Information Retrieval:

- Traditional Information Retrieval techniques become inadequate to handle large text databases containing high volume of text documents.
- Presently, while doing query based searching, search engines return a set of web pages containing both relevant and non-relevant pages, sometimes showing non relevant pages assigned higher rank score.
- A common problem of Information Retrieval is that users have to browse large number of documents containing both relevant and non-relevant documents before finding relevant documents.
- To fasten the process of document retrieval, text summarization technique is used, Ranking of documents is made based on the summary or the abstract provided by the authors of the document. But it is not always possible as not all documents come with an abstract or summary.

In the given database we have number of text based articles. The articles are in the form of questions, answers and related tags .The database considered in our project is relational database. The relational database consists of four tables: content table, keyword table, score table, rank table. From the content table we will find relative keywords and compare those keywords with other articles and form relative keyword index. This index will be sorted and optimized to find the degree of relativity among the articles. The degree of relativity will define the base line for keyword matching and search. Fig 1 shows the proposed architecture

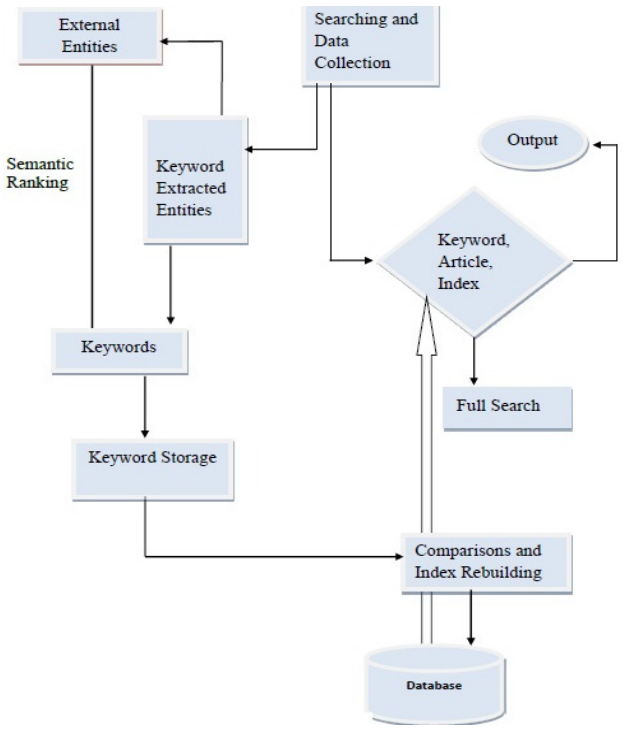

Fig 1. Proposed Architecture

In the proposed work we aim at creating four tables .These four tables are as follows:

- **Content table**

Content table consist of the articles in the form of questions, answers and related tags.

- **Keyword table**

Keyword table consist of the keyword extracted from the contents. Keyword is the smallest unit, which expresses meaning of entire document that also used for extracting exact information as per user requirements. Everyday thousands of books, papers, articles and documents are created and published. It is very difficult to go through all the text materials, so that there is a need of good information extraction or summarization method that provides the real contents of a given document. Various applications can take advantage of it such as information retrieval, automatic indexing, text summarization, classification, clustering, topic detection and tracking, web searches, report generation, filtering, cataloging, etc. Keyword extraction, also known as

key phrase extraction is an area of text mining that intends to identify the most useful and important words, phrases that are also called terms.

- **Score table**

Score table consist of the occurrence of the keywords in each article. This table will help us to rank the article as per the occurrence of the keyword in that article.

- **Rank table**

Rank table will consist of the articles as per the rank decided on the keyword score. The volume of information is increasing day by day so there is a challenge to provide proper and relevant information to the user. An efficient ranking of query words has a major role in efficient searching for query words. There are various challenges associated with the ranking of pages such that some pages are made only for navigation purpose and some pages of the do not possess the quality of self-descriptiveness. Ranking retrieval systems are particularly appropriate for end-users.

IV. RESEARCH METHODOLOGIES

Keywords are a set of major words in a document that give high-level description of the content for readers. Keywords are useful for scanning large documents in a short time. Extracting keywords manually are very difficult and time consuming process. Therefore, there is in need for process to extract keywords from documents automatically. Keyword extraction is a process in which a set of words are selected that gives the meaning of the whole document.

a. Open Calais

The Open Calais Web Service automatically creates rich semantic metadata for the content submitted – in well under a second. Using natural language processing (NLP), machine learning and other methods, Calais analyzes document and finds the entities within it. But, Calais goes well beyond classic entity identification and returns the facts and events hidden within text as well.

b. Yahoo Apies

The Term Extraction Web Service provides a list of significant words or phrases extracted from a larger content.

c. Custom Built

A lexical chain is a sequence of related words in writing, spanning short (adjacent words or sentences) or long distances (i.e., entire text). A chain is independent of the grammatical structure of the text and in effect it is a list of words that captures a portion of the cohesive structure of the text.

Lexical chains are used in different NLP problems such as word sense disambiguation, text segmentation, text summarization and topic tracing. WordNet dictionary is used to build the lexical chain that provides the word senses and semantic relations between words. Lexical chain builder uses Word relations, which are Synonym, Hyponym and Meronym to build a lexical chain.

Every node in a lexical chain denotes a meaning of a word, and each link can be synonym, hyponym or hyponym relation between two word senses. In this approach keywords are extracted using the following features that are,

- First occurrence position
- Word frequency
- Last occurrence position
- Lexical chain score of a word
- Direct lexical chain score of a word
- Lexical span score of a word
- Direct lexical span score of a word

Lexical Chain Score of a Word

A word can be a member of more than one lexical chain. The score can be assigned for these words. Then the word that as the maximum score is chose as the lexical chain score of the word. The score depends on the relations appearing in the lexical chain.

Direct Lexical Chain Score of a Word

This can be calculated by scoring only the relations that belong to the word.

Lexical Span Score of a Word

The span score of a lexical chain depends on the portion of the text that is covered by the lexical chain. This covered portion of the text is considered to be the distance between the first occurrence position of a lexical chain member (word) and the last occurrence position of a lexical chain member (word).The span score is computed by finding the difference between these two positions.

Direct Lexical Span Score of a Word

The score of the lexical chain with maximum score can be considered as the direct lexical chain span score of the word. This score can be computed as same as the lexical chain span score except that the words that are directly related with the word in the lexical chain. This technique uses statistical classifier to build decision trees that is to identify whether the word is likely one or not. The decision tree uses bagging; it is a process of classifying the objects with multiple classifiers. In bagging technique the average classification probability is used to classify the objects. Gonenc Ercan, IlyasCicekli has proposed this approach with a corpus for extracting keywords. Precision values are calculated for this system with all seven features gave the better results.

Stages:

- **Content Analysis**

To develop a system which can consume content in the form of text and generate a relative rank and index based on keyword analysis. We will use this system as a base for a search engine. In Content analysis we will extract keywords from the article.

- **Keyword Ranking**

a. String Equality and extensions:
In String Equality, while searching if that content is already has been searched, then it will show the output directly without comparing it with the other articles stored in the database. And in extensions we check all the possible suffixes that may occur in the keywords. For example, if the extracted keyword is eat then the entire suffix's of eat i.e., eaten, eating, ate will be checked.

b. Word co-occurrence: Word co-occurrence is extensively used in various forms of research such as content analysis, text mining. Its aim is to find similarity between words or similarities of meaning among word patterns. The sentences in the document are considered as a set of words; it includes title of a document, section title and a caption. The term frequency is determined by counting the frequent terms occurred in a document.

c. Sorting Techniques:
Sorting technique is performed for the ranking of the articles according to their score, and the article with the highest rank will be displayed first on the output screen.

- **Searching**

a. Keyword Extraction Technique
Keyword extraction will be done on the basis of three keyword extraction techniques which are Custom build, open Calais and Yahoo API.

b. Index Search:
The Extracted keywords will passed to index and the article containing more occurrences of that keywords will be resulted as output. While searching the articles relevancy between the keywords is also checked.

V RESULTS

The below fig.2 depicts extracted keywords from the input passed. The contents from which the keywords are to be extracted are passed as input into the search box.

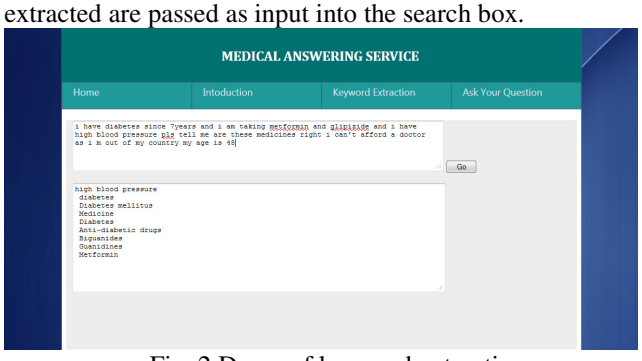

Fig. 2 Demo of keyword extraction

After clicking on "GO" button the content is passed to the external entity for the extraction of keywords. If the external entity fails to give any keyword, then the input is passed to the custom built technique for extracting keywords. This technique will extract keywords based upon the keyword length. The extracted keyword will then be displayed in the textbox just below the search textbox.

.

Ranking of the articles having different scores is done at runtime. Rank is decided considering that the all the keywords extracted from the input by the user is present in any of the article. If such is present in the database, then the particular article will be ranked first and displayed as first output to the input. The below figure shows the initial stage of the ranking database which is created at run time. Initially the rank of all the articles is zero as shown in fig 3.

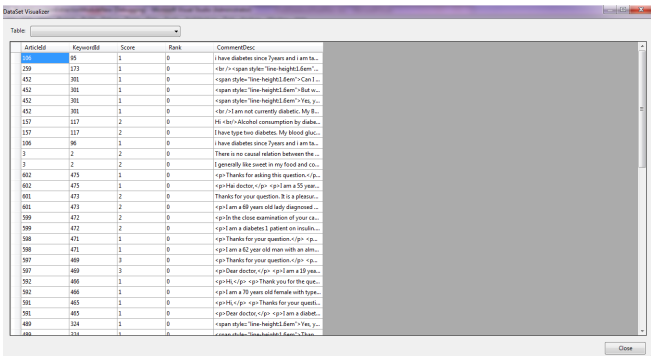

Fig.3 Runtime table before ranking of the keywords

The value of the rank column goes on changing until all the articles matching that keyword are gone through. The article containing all the extracted keywords from the input is given higher rank and is displayed as the exact solution for the input query. If there is no such article which contains all the keywords, then the ranking of the article is done on the highest score of the keyword. The article with the highest rank is given first and then accordingly

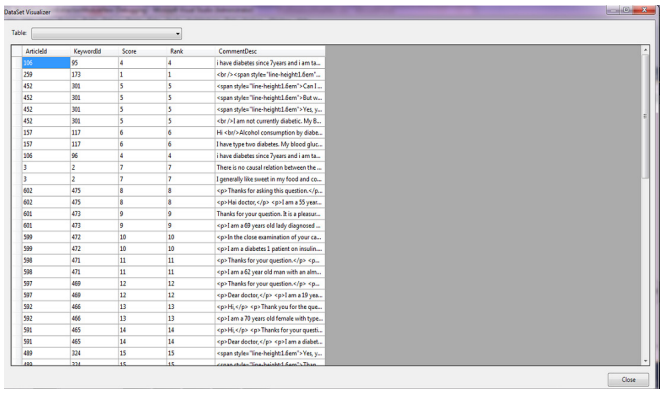

Fig: 4. Runtime table after ranking of the keywords

After the ranking of the article at runtime, the articles related to particular input are given as output. The output consists of number of comments in the decreasing order. The output consists of questions and comments related to the query. The comments are displayed systematically one after the other.

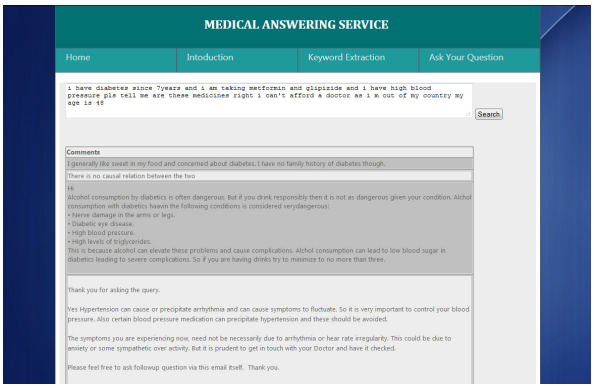

Fig. 5 User console screen

V CONCLUSION

The problem of comparison, retrieval time and extracting of precise information from the database has been efficiently solved by keyword extraction from the article, scoring that keyword and ranking the article based on score. This method is useful as the user need not to browse large amount of data to find relevant article.

.

REFERENCES

[1] Dilip Kumar Sharma, A. K. Sharma,"A Comparative Analysis of Web Page Ranking Algorithms", Dilip Kumar Sharma et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02,,2010.

[2] Vishal Gupta, Gurpreet S. Lehal,"A Survey of Text Mining Technique and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 11, AUGUST 2009.

[3] Namita Gupta,"Text Mining For Information Retrival", May 2011.

[4] Menaka S, RadhaN,"An Overview of Techniques Used for Extracting Keywords from Documents", International Journal of Computer Trends and Technology (IJCTT) – volume 4, 7– July 2013.

[5] Min Ye,"Text Mining for Building a Biomedical Knowledge Base on Diseases, Risk Factors, and Symptoms", 2011.

[6] Roberto De Virgilio ," Efficient and effective ranking in Top-K exploration for Keyword Search on RDF " Dipartimento di informatica e automazione universita RomaTre, Rome Italy.

