ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 1, January 2019

Data Extraction and Recommending Document through ASR

Shamita N. Kapote[1], Samruddhi V. Kharde[2], Gayatri R. Patil[3], Vishakha D. Kumbhar[4], Vishakha N. Pawar[5]

Diploma Student Final Year, Computer, Guru Gobind Singh Polytechnic, Nashik, Maharashtra[1, 2, 3, 4]

Guide, Computer, Guru Gobind Singh Polytechnic, Nashik, Maharashtra[5]

Abstract: In world of automation every individual needs an instant result of any query. As result of which different applications and software are coming into existence. The information is the major aspect of human life. The information is available as documents, database, multimedia resources, etc. Through this project we are extracting appropriate keyword from several documents input. Extracted keywords are matched with available documents. Finally, we recommend appropriate documents to the participants for reference. In document clustering, hundreds of thousands of files are usually examined. Much of the data in those files consists of unstructured text, whose analysis by computer examiners is difficult to be performed. In this context, automated methods of analysis are of great interest. Algorithms for clustering documents can facilitate the discovery of new and useful knowledge from the documents under analysis. We have proposed a more efficient document clustering algorithm. This will enhance the searching and analysis of the document and the best suitable results related to the query loaded will be recommended. In this software, we can add any number of documents. After that we can see all the documents, then after tokenization and clustering, we gain the extracted keywords and their frequency (count of the words) and then recommend the data sets generated to the user. Due to it the computation process of finding the data will be reduced in amount of time and efforts.

Keywords: Keyword Extraction, Stop-Words Analysis and Removal, Stemming of Clusters, Data Clustering Techniques and Document Recommendation

I. INTRODUCTION

In world of automation every individual needs an instant result of any query. As result of which different applications and software are coming into existence. The information is the major aspect of human life. The information is available as documents, database, multimedia resources, etc. with an unprecedented wealth of information, maintaining and retrieving the data or information is major issue for the organizations. The large institutions have their own data warehouses. Data warehouses are central repositories of integrated data from one or more disparate sources. They store current and historical data in one single place that are used for creating analytical reports for workers throughout the enterprise. Whenever the data is needed, the data is mined from the warehouses. The query is analysed, and results are prescribed. This results in the satisfaction of the user or developer, as he/she attains the assumed information. Relevance and diversity of documents can be modelled at three levels:

- While extracting queries
- Building one or several implicit queries
- Re-ranking the results of queries.

II. EXISTING SYSTEM

In existing system, humans are surrounded by an unprecedented wealth of information, available as documents, databases, or multimedia resources. Access to this information is conditioned by the availability of suitable search engines. The following challenges gave us the motivation to clustering of the documents:

1. The number of available articles was large.
2. Documents were being added from different sources.
3. The recommendations had to be generated and updated in real time.
4. Data retrieval is rapid, but the imposed data is not obtained, so in such conditions, users need to explore the relevant data or information which they require.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 1, January 2019

A. Examples of Existing Systems:

1. Linguakit (Keyword Extractor):

With **Linguakit**, exploring, analysing and obtaining better information from texts and written documents is possible. This multilingual web, which integrates, among other linguistic tools, a summarizer, a sentiment analyser or an extractor of keywords that give meaning to a text, is aimed at a wide range of users that make the language a professional use, educational or general. Linguakit is designed so that every person with a linguistic interest can get the most out of written texts. This platform presents its linguistic modules organized into four orientated sections: a first that deals with more generic aspects of language with modules such as the conjugator or translator; a second, for a user profile more linked to the educational field, with modules such as the morphosyntactic tagger or the parser; a third section designed for communication and marketing professionals such as the sentiment analyser or the keyword extractor; and, finally, an experimental section where Linguakit presents the new tools in project.

Linguakit is an idea of Client Language Technology that arises fruit of years of research of the company in the field of Natural Language Processing (PLN).

- **Drawbacks of Linguakit:**

1. Could not be able to upload the document form our desktop files.
2. Needs the Internet Connection.
3. Needs the user to type the data to extract the keywords.
4. It does not even take the websites URL to extract the page.

2. Pingler (SEO Tool like Google Keyword Planner):

Pingler is a web-based SEO Tool which helps you to ping your website, know number of visitors visits to your website by using the keywords you have set as your search optimization. It is same as the Google Keyword Planner which also let you know which keywords to use for your website to rank in the search of the Google first page.

- **Drawbacks of Pingler:**

1. One cannot add any document to it nor input any text.
2. It needs continuous Internet connection.
3. Require money to access the facilities.

III. PROPOSED SYSTEM

Then to overcome the existing system drawbacks propose a method to obtain multiple topically separated queries from this keyword set, in order to make best use of the chances of making at least one appropriate recommendation when using these queries to search. The proposed methods are evaluated in terms of significance with respect to conversation rated by several human judges. The scores show that our proposal improves more than previous methods that consider only word frequency or topic similarity and represents a capable solution for a document recommended system to be used in conversations. We propose just-in-time retrieval of information, based upon the keywords from the queries of the user. Firstly, an algorithm is designed to extract keywords from all documents and gather in the local repository of the system. Then, a method derives multiple topically separated queries from this keyword set, to maximize the chances of making at least one relevant recommendation to user using these queries to search over the English Wikipedia. By clustering the documents, we could reduce our domain of search for recommendations as most of the users had interest in several clusters. This improved our time and efficiency of searching the information.

IV. WORKING OF PROJECT

1. Automatic Speech Recognition (ASR): Speech recognition is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. In this phase of the process, we are going to collect the data from the voice of the speaker and then will convert it into a text document and the document will be then forwarded for the stop-words analysis and removal.

2. Stop-words Removal and Analysis: A stop word is a commonly used word (such as "the"). The list of stop-words that are not to be added is called a stop list. Stop words are deemed irrelevant for searching purposes because they occur frequently in the language. To save both space and time, these words are dropped at indexing time and then

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 1, January 2019

ignored at search time. The keyword set generated after removal of the stop-words is then advanced for the extraction of the essential keywords.

3. Keyword Extraction: Keyword extraction is tasked with the automatic identification of terms that best describe the subject of a document. Keyword Extraction involves steps such as the tokenization, stemming, extraction. The system starts with the process of tokenization of the document by breaking-down the query into keywords (means a word or concept of great significance). The benefit of diverse keyword extraction is that the coverage of the main topics of the conversation fragment is maximized. The future method for diverse keyword extraction proceeds in three steps,

1. Used to represent the division of the abstract subject for each word.
2. These topic models are used to determine weights for the abstract topics in each conversation fragment represented by
3. The keyword list W = {w1, w2……wk}. Which covers a maximum number of the most important topics is preferred by rewarding range, using a unique algorithm introduced in this part.

4. Keywords Clustering: The different sets of extracted keywords are measured to denote the possible information needs of the applicants to a discussion, in terms of the ideas and topics that are declared in the discussion. To maintain the variety of topics alive in the keyword set, and to reduce the noisy result of each data need on the others, this set must be divided into several topically-disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document recovery system. These subsets are acquired by clustering topically-similar keywords, as follows. Clusters of keywords are constructed by ranking keywords for each main topic of the fragment.

5. From Keywords to Document Recommendation: As a first impression, one implicit query can be arranged for each discussion part by using as a query all keywords special by the various keyword extraction techniques. However, to enhance the retrieval results, multiple implicit queries can be formulated for each discussion part, with the keywords of each cluster from the before fragment. In tests with only one implicit query per discussion fragment, the document results parallel to each discussion fragment were arranged by selecting the first document retrieval results of the implicit query.

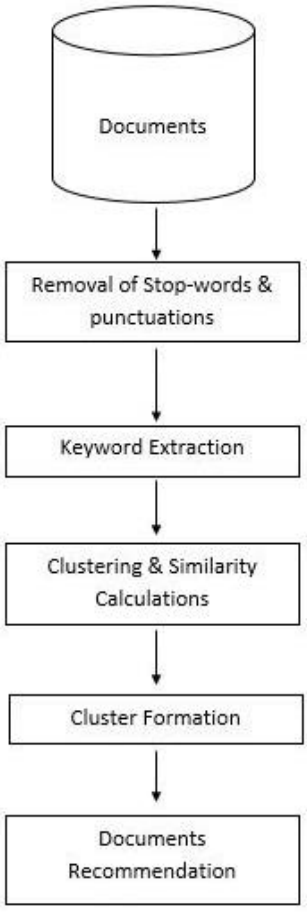

Fig.1 Flow Chart of the System

DOI 10.17148/IJARCCE.2019.8120

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 8, Issue 1, January 2019

6. Advantages of Proposed System: To maintain multiple hypotheses about user's information need. To present a small sample of recommendations based on the most likely ones. Retrieving of documents by keyword query is faster and Clustering of documents by multi-key word similarity.

1. Expert examiner's job is reduced.
2. Speeds up the computation process.
3. No need of Internet connection.

V. RESULT AND ANALYSIS

We conduct experiment on five transcripts. The results are based on the number of words in the transcripts. With different transcripts we obtained different number of keywords. So, the results depend on number of words in the transcript. Here in fig 2. we input a transcript of different word length to existing as well as proposed system and observed the keywords extracted by both the system, where we consider the keywords extracted in percentage (%). We noticed that proposed system extracts more keywords for different transcript input.

Table I Accuracy Computation

	Existing Accuracy (%)	**Proposed Accuracy (%)**
1	78	85
2	75	79
3	72	78
4	55	65
5	60	68

CONCLUSION

We have considered a form of just-in-time retrieval systems intended for conversational environments, in which they recommend to user's documents that are relevant to their information needs. We focused on modelling the user's information needs by deriving implicit queries from documents. These queries are based on sets of keywords extracted from the documents. We have proposed a novel diverse keyword extraction technique which covers the maximal number of important topics in a fragment. Then, to reduce the repetition of words, queries of the mixture of topics in a keyword set, we proposed a clustering technique to divide the set of keywords into smaller topically-independent subsets constituting queries. We compared the diverse keyword extraction technique with existing methods depended on word frequency or topical similarity in terms of representativeness of the keywords and the relevance of recommended documents. Our current goals are to process queries, and to rank document results with the objective of maximizing the coverage of all the information needs, while minimizing redundancy in a short list of documents. Integrating these techniques in a working prototype should help users to find valuable documents immediately and effortlessly.

REFERENCES

[1]. Maryam Habibi and Andrei Popescu-Belis " Keyword Extraction and Document Clustering and Recommendation"
[2]. Hsi-Cheng Chang and Chiun-Chieh Hsu "Using Topic Keyword Clusters for Automatic Document Clustering
[3]. Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin and Craig G. Nevill-Manning "KEA: Practical Automatic Keyphrase Extraction
[4]. G. Subhashini, D.Jayakumar "Document Recommendation for Conversation Based on Keyword Extraction and Clustering "
[5]. Kumodini V. Tate, Bhushan R. Nandwalkar "Document Recommendation Using Keyword Extraction for Meeting Analysis"