# Determining Document Relevance using Keyword Extraction

**Vinay Patil[1], Sachin Jadhav[2], Pawan Lokapur[3], Akash Mhatre[4], Prof. Tushar Ghorpade[5]**

*[1,2,3,4]Student, Dept of computer Engineering, Ramrao Adik Institute of Technology*
*[5]Assistant Professor, Dept. Of Computer Engineering, Ramrao Adik Institute of Technology*

---

**Abstract -** *This paper lies in the data analysis domain describing about the system which attempts to search for a relevant document from a large set of documents, or more specifically to fetch a summary of answer for a given query from one of the selected relevant document. In few of the organizations involving large data set in form of documents, where clients and customers need to retrieve some of these documents frequently, the process of searching this document becomes a very hectic task. For instance, the educational institute like Mumbai University which has large corpus of documents. So to overcome this manual searching approach we have proposed a system which successfully fetches the desired documents to user based on query provided to system. This is being done by priory extracting the documents at time of uploading and storing the necessary stats required for search algorithm. For document extraction we use the TF-IDF algorithm. And during search we analyse the TF-IDF weight of keywords in search algorithm to fetch the desired set of documents to user. There will also be a feedback mechanism for user to interact with system through which user can upvote or downvote for a particular document thus making the system to learn and improve its search in future. The system is supposed to deliver accurate results for every query given by user combined with less processing time. The system contains three operational elements i.e. keyword extraction, search module and topic selection.*

***General Terms:*** **Data Mining, Data Analysis, Keyword Extraction.**

***Keywords:*** **TF-IDF, QnA, Document Search, Artificial Feedback, Keyword Extraction.**

## 1. INTRODUCTION

In recent world of digitalization, data is coming in huge amounts from many sources like news, social media, banking and education. And hence because of this unregulated and unordered growth in data, there is a need of automated information retriever which will help users to retrieve relevant information by searching in piles of unstructured data. But there exists challenges to implement one, such as retrieving correct sense of information. Information retrievers is emerged as an important research area in recent past. In this regard, study of existing work is useful to carry on further research. Using keywords to predict document contents is accurate and fast method. Keywords can be used as entry points into an index which will then help to identify files, records, texts or any unstructured data. However, a large number of data is not marked using keyword and giving it to a human for tagging will be difficult and also needs large amount of time. Thus, there is need of an algorithm which will tag every document or data by its relevant keywords. Hence in this paper, a keyword extraction algorithm is presented also a faster method of searching through those extracted keyword is proposed. An algorithm named "TF-IDF (Term Frequency- Inverse Document Frequency)" is one of the simple method which serves the purpose of retrieving keywords with high accuracy. The working of TF-IDF algorithm is based on the number of times a word appeared in a sentence and the number of times it appeared in whole corpus. More the occurrences in a document, more is the weightage and contradictorily more the occurrences in corpus, less is the weightage. The keywords extracted by this algorithm can be used in searching algorithm which we proposed. This searching algorithm make use of combination concept and sets to give list of documents which contains the search query. It also discusses about different databases structures used for text extraction and a quick question answering bot which can be used in different domains. Finally, it discusses briefly about issues and research challenges faced by us along with future direction.

## 2. LITERATURE REVIEW

In LDA based Paper [1], system uses generative probabilistic model on Chinese corpus data and categories documents by different topic names, but the major drawback is that it does not remove stop-words and also the topic names or subject needs to be known in advance. hence limiting the corpus contents.

In TF-IDF based paper [2], a system is developed which will search social media platforms to get the latest trends which then can be used for advertisement purpose. Paper presented an algorithm which will monitor Instagram accounts for 50 recently posted photos and its captions. These captions are then analysed to get latest trends and fashion accessories which will be used for marketing those product effectively. But limitations of this system is it is concentrated only on 20 most followed users on Instagram and hence the scope of this system is narrow down to 20 users only.

Modified IDF paper [3] focuses more on ambiguous word sensing by using KNN approach which has less worth

from document relevance checking point of view, since word ambiguity won't affect much in large corpus of documents.

Delta TF-IDF paper [4] deals with positive and negative set of classes for efficiently weighing the word scores before pushing them into classification. So, it requires positive and negative training set to work prior to classification.

TF-IDF to Determine Word Relevance in Document Queries paper [8] where the TF-IDF algorithm was not able to equate one of the word (e.g. drum) with its plural form (drums) slightly decreasing the TF-IDF weight of that particular keyword. For larger corpus of documents this could present an escalated problem

## 3. EXPLANATION

The system contains three modules i.e. topic extraction, document search and question answering bot. All three modules will be mutually functioning with the help of global and local database schemas. So before understanding the module functioning let us dive into the structure of databases and relationship between the tables.

### 3.1 Database

Database module has two modes Private and Organization. These modes will then be used to serve different purposes.

Private Mode: This database will provide private workspace for every user. Hence users can upload their private files in this workspace which then can only be visible and accessible to its owner only. but for operations like search and QnA will work the same, just depending on the mode the respective database will be altered. Hence can be act as private Document Manager.

Organization Mode: This mode is designed for organizations to manage their files which should only be available to their employees. In this mode, files will be act as public and only authorized person can upload the data from this workspace.

Mode selection: These options will be available to users which will allow them to select mode while uploading or searching files.

Database will include small corpus of documents which contains only stop-words so as to reduce the impact of higher TF-IDF weights of those words during initial document extraction.

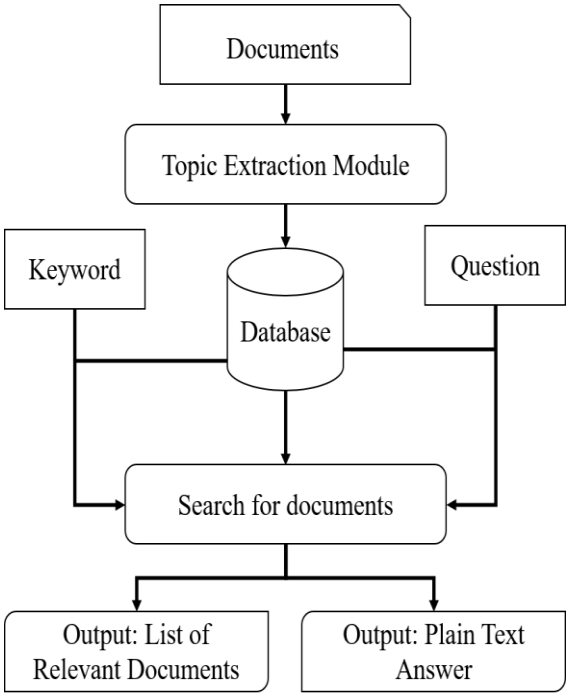


**Fig 3.1:** Database structure

### 3.2 Topic Extraction

This is the very first module with which user will interact by uploading all the documents which he/she desires. Once the documents are uploaded, document details are stored in a doc_details table. All words of a document are extracted then stop words are removed for reducing memory overhead and to improve the power of searching. Also, instead of the keywords plural form its root word is inserted into the database so that TF-IDF weight for each word is computed correctly without discriminating between singular and plural form. Term frequencies are calculated [6] for each word and send for the database operation. There are two threads for simultaneously processing the database operations. The first thread will do two tasks. It will increment a count of a total number of documents in a doc_count table after that IDF (Inverse Document Frequency) is calculated.

TF: Term Frequency, which measures how frequently a word occurs in a document. TF(t) = (Number of times word w appears in a document) / (Total number of words in the document).

IDF: Inverse Document Frequency, which measures how important a word is.

IDF(t) = log_e(Total number of documents / Number of documents with word w in it).

3.2.1 Equations

*Variables*:

**Table 1**. Symbols with Description

| Description | Symbol |
|---|---|
| Number of times word w appears in a document | $W_i$ |
| Total number of words in the document | $T_w$ |
| Total number of documents | N |
| Number of documents with word w in it | $D_i$ |

**Equation No. 1**: Term Frequency

$$tf(i) = \frac{Wi}{\text{Tw}}$$

**Equation No. 2**: Inverse Document Frequency

$$\text{idf(i)}=\log_e\left(\frac{N}{Di}\right)$$

**Equation No. 3**: Term Frequency-Inverse Document Frequency

$$\text{tf}idf(i) = tf(i) \times idf(i)$$

TF-IDF weight of each word is calculated and stored in a tfidf table before which fuzzy inference system is implemented so as to threshold number of keywords which will be inserted in database for searching purpose and hence all the stop words [7] are eliminated.. The second thread for updating search_list and search_doc tables to improve searching.

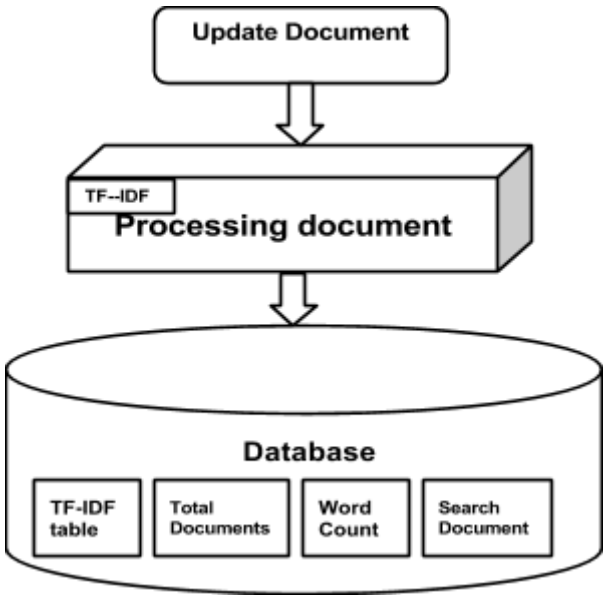


**Fig 3.2:** Keyword Extraction

## 3.3 Document Search

This is the core module of the system which runs the search algorithm and fetches the appropriate set of documents to the user depending on the query given to it. The algorithm primarily relies on the extracted data stored in the database for its processing. It will first sort the keywords in query and determine which keywords are required for searching. Accordingly, a list of such keywords is prepared and is passed as parameter to algorithm. The algorithm works on principle of "Highest priority to Largest intersection" for which the combinational logic has also been implemented to systematically form the sets of doclist of keywords. Also, the ordering in combination is taken care of on the basis of rank which is taken as a feedback from user. This artificial feedback will help to rank the documents and make the right choice in future.

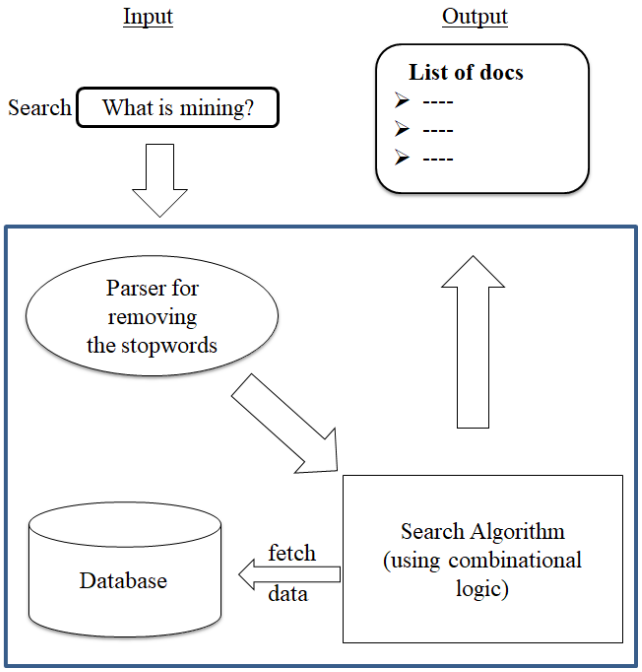


**Fig 3.3:** Searching module

## 3.4 Question and Answering bot

A natural language processing (NLP)[5] gives capability to computer which allows communication to happen between user and computer using human natural languages. There are three stages to understand natural language i.e. parsing, semantic interpretation, and summarization. In parsing phase, two actions are performed on input provided from user. First, the main linguistic words are identified. Second, using TF-IDF values of words in input query for selected document the sum is calculated. Then in next phase i.e. summarization, the sentences are collected from selected document for paragraph summarization.

Main words are retrieved from sentence by using sentence semantics. These keywords are then searched in database which in turn gives list of documents sorted by its rating taken from users which might content the answer to the question. Later, the top document is tokenized and values for tf-idf are calculated for every sentence. And those sentences having greater values than tf-idf value of the query are selected for summarization. Resultant answers are processed to get a most popular answer or most accepted answer using ranking algorithm which will then be shown as output to users. Ranking Algorithm works on feedback taken from users. This feedback asks users whether the answer was helpful or not? If user thinks it was helpful will mark yes and the document rank is upgraded otherwise it will be degraded.

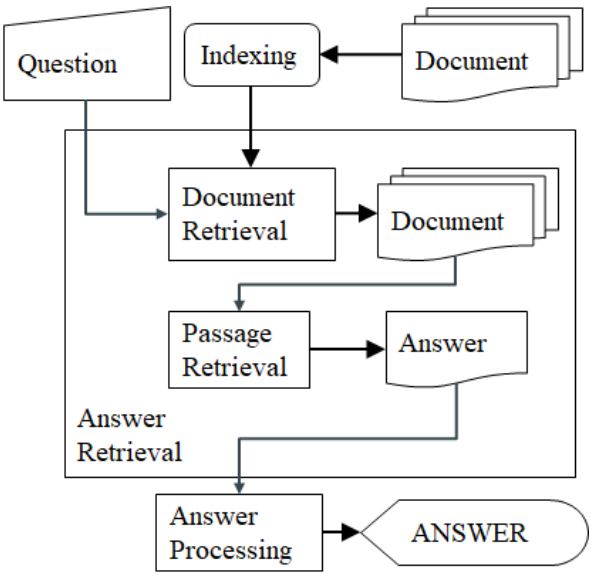


**Fig 3.4:** Question and Answering bot Module

## 4. Conclusions

In order to get a relevant document or specific answer from a corpus of a large collection of organizational documents, this paper proposes a document extraction system which has following modules first is keyword extraction in this user will upload all the documents which he/she desires. Once the documents are uploaded, the system will internally start the extraction of keywords by using the TF-IDF algorithm. From each document, significant keywords are selected these keywords will occupy their positions in database which will be used during the document searching. Second is document search is responsible for fetching the relevant document to the user by making the use of weighting of a combination of keywords and uses indexed database which contains the list of keywords along with their respective list of documents.

Further on the basis of user satisfaction, a feedback will be taken up by the user to check the correctness of the generated output. If the user is not satisfied then the next set of documents will be fetched. This system also has Question and Answering bot. which will provide user to have active conversation with system using simple human language(English). QnA also includes an artificial feedback taken from user which will help improve the answer quality and relevance regarding to question asked.

This method helps people to locate relevant documents from the corpus of a large collection of organizational documents. On the other hand, an artificial feedback mechanism can rank the documents and make the right choice in future.

## 5. Future Work

The System currently doesn't support odt and image file format. So, support for these files can be added in future. Semantic analysis can be used for getting context wise meaning of a word to improve the power of searching. Comments can be used with artificial feedback to improve searching. The correctness of result is checked by artificial feedback which currently just suggests the alternate ranked documents. But in future, this feedback technique can be replaced with views hence can be used for ranking the document. So, it will be more accurate as most of the times users don't vote up although it helped them. Users can set the expiry date of documents which they needed for short period of time. After expiry date that document is deleted hence it will improve the searching and reduce the overhead of database.

## 6. References

[1] Qihua Liu, "A novel Chinese text topic extraction method based on LDA", International Conference on Computer Science and Network Technology, Nanchang, 2015.

[2] Bernardus Ari Kuncoro, Bambang Heru Iswanto, "TF-IDF method in ranking keywords of Instagram users' image captions", TF-IDF method in ranking keywords of Instagram users' image captions Information Technology System and Innovation, Bali, November 2015.

[3] Zun May Myint, May Zin Oo, "Analysis of modified inverse document frequency Variants for word sense disambiguation", International Journal of Advanced Computational Engineering and Networking, Myanmar, Aug.-2016.

[4] Justin Martineau, Tim Finin,"Delta TFIDF: An Improved Feature Space for Sentiment Analysis",

Third AAAI International Conference on Weblogs and Social Media, San Jose CA, May 2009.

[5] Bayu Setiaji, Ferry Wahyu Wibowo, "Chatbot Using A Knowledge in Database Human-to-Machine Conversation Modeling", 2016 7th International Conference on Intelligent Systems, Modelling and Simulation, Yogyakarta, Indonesia, April 2016.

[6] Aizhang Guo,Tao Yang, "Research and Improvement of feature words weight based on TFIDF Algorithm ", International Conference on Computer Science and Network Technology,Jinan250353, China,May 2016.

[7] Rajaraman, A., Ullman J., "Data Mining: Mining of Massive Datasets", India, 2011.

[8] Juan Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries", Department of Computer Science, Rutgers University, Piscataway, NJ.