# Efficient Ranked Multi-Keyword Search using Machine Learning Algorithms

**Mr. Namdeo S. Kedare[1], Neha Jha[2], Nidhi Jha[3], Meghna Chavan[4]**
*[1,2,3,4]Department of Information Technology*
*[1,2,3,4]Dhole Patil College of Engineering*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -**Now a days, classification is the process of classifying the text documents based on words, phrases and word combination with respect to set of pre-defined categories. Data classification has many applications such as mail routing, email filtering, content classification, news monitoring and narrow-casting. Keywords are extracted from documents to classify the documents. Keywords are subset of words that contains the most important information about the content of the document. Keyword extraction is a process used to take out the important keywords from documents. In this proposed system keywords are extracted from documents using TF-IDF and naïve bayes algorithm. TF-IDF algorithm is used to select the candidate words. The words which have highest similarity are taken as keywords. The experiment has been done using Naive Bayes algorithms and its performance is analyzed based on machine learning.

***Key Words*:**Keyword based search, machine learning, naïve bayes algorithm, TF-IDF algorithm, ranking.

## 1. INTRODUCTION

Over the last decade, the number of digital documents available for various purposes has grown enormously with the increasing availability of high capacity storage hardware and powerful computing platforms. The vivid increase of documents demands effectual organizing and retrieval methods mainly for large documents. Text classification is one of the key techniques in text mining to categorize the documents in a supervised manner. The processing of text classification involves two main problems are the extraction of feature terms that become effective keywords in the training phase and then the actual classification of the document using these featureterms in the test phase. Text classification can be used for document filtering and routing to topic specific processing mechanisms such as information extraction and machine translation. Various methods are used for document classification such as Naive Bayes, Support Vector Machine, K-Nearest Neighbor, Fuzzy C-means, Neural Networks, Decision trees and Rule based learning algorithms out sourcing.

## PROBLEM STATEMENT

To develop an efficient system to retrieve given data in response to user with ranking system and to provide search file based on keywords with ranked result by using machine learning algorithm.

## PROPOSED METHODOLOGY

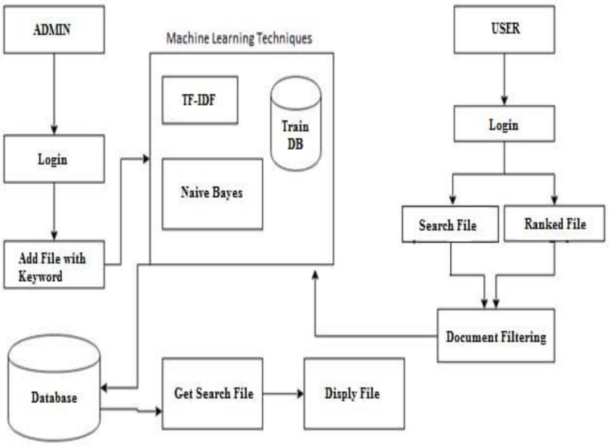


**Fig -1**: Proposed System

**User**

This module helps clients to enter their query keyword to get the most important documents from set of uploaded documents. This module recovers the documents from cloud which coordinates the query keyword.

**Data Owner**

After expansion of keywords the data owner assist data with multiple keywords the document utilizing based on machine learning Algorithm and after that upload the document to store the database.

**Ranked Results**

Clients/user can download the resultant arrangement of documents just if he/she is approved client who has allowed consent from data owner to download specific document. Here user get the ranked and mostly search records from the ranking system to get exactly data to the all user.

## ALGORITHM

The proposed architecture of four models: user interface, log pre-processing, feature clustering using of Naïve Bayes Classification, training and testing using TF-IDF for more accurate categorization of opinion. This system can solve irrelevant data and more accuracy by associating Modified K means with Naïve Bayes Classification algorithm.

### A. Naive Bayes (NB):

Naive Bayes Classifier uses Bayes Theorem, which finds the probability of an event given the probability of another event that has already occurred. Naive Bayes classifier performs extremely well for problems which are linearly separable and even for problems which are non-linearly separable it performs reasonably well.

### B. TF-IDF Algorithm:

TF_IDF stands for Term frequency-inverse document frequency. The TF-IDF weight is a weight often used in information retrieval and text mining. Variations of the TF-IDF weighting scheme are often used by search engines in scoring and ranking a document's relevance given a query. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus assumed that all the data users are trustworthy, but in practical, the dishonest data user may distribute his secure keys to unauthorized users.

## EXPEXCTED RESULTS

**Data Encryption and Decryption Result**

When is applied on the data then we get encrypted data and that encrypted data is stored on the database. User can access the data after downloading the decrypting file.

**Ranking Result**

When any user request for the data then Ranking is done on the requested data. After ranking, the user gets the expected results of the query.

## FUTURE SCOPE

The detailed analysis and studying privacy and efficiency guarantees of proposed schemes is given, and experiments on the real-world dataset show our future system introduce low overhead on both computation and communication. The system provides the accurate result ranking documents. The system provides search efficiency due to the use of efficient search algorithm. The proposed model can be enhanced in symmetric searchable encryption scheme.

## 3. CONCLUSIONS

The system is designed keyword with top-k ranked search over secure server data. The system provides the accurate result ranking documents. The system provides search efficiency due to the use of tree based index and efficient search algorithm. For future work there are many challenges in symmetric searchable encryption scheme. As it is assumed that all the data users are trustworthy, but in practical, the dishonest data user may distribute his secure keys to unauthorized users.

## OUTPUT SCREENSHOTS:

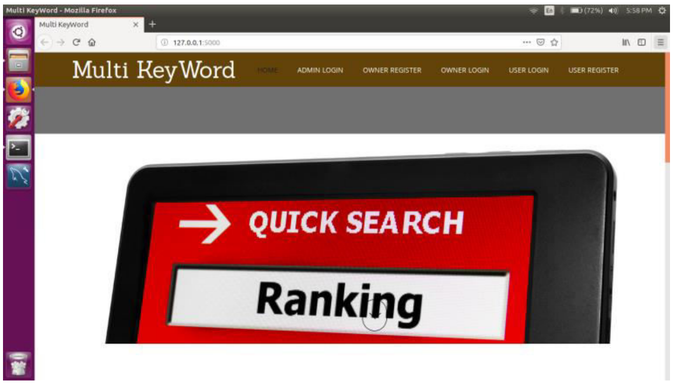
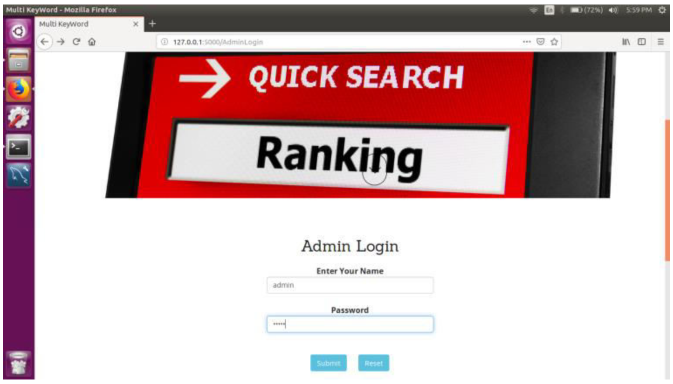


**Fig -2**: Home page



Admin Login

Enter Your Name: admin

Password: •••••

**Fig -3**: Admin Login

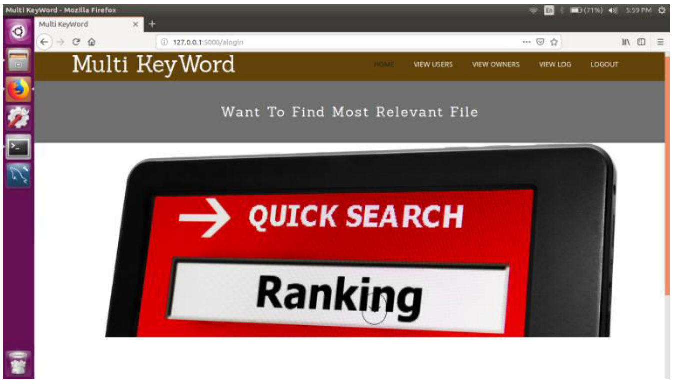
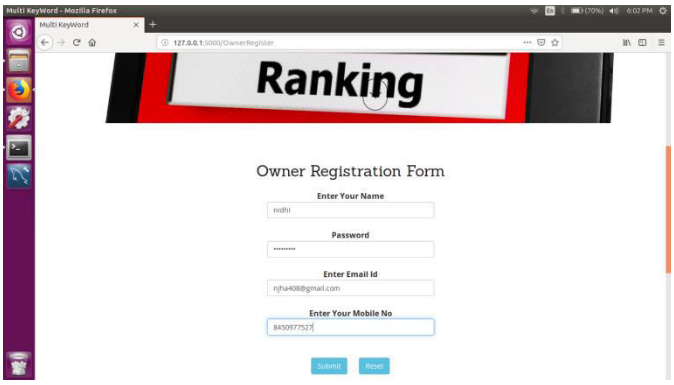


**Fig-4:** Admin Home page



Owner Registration Form

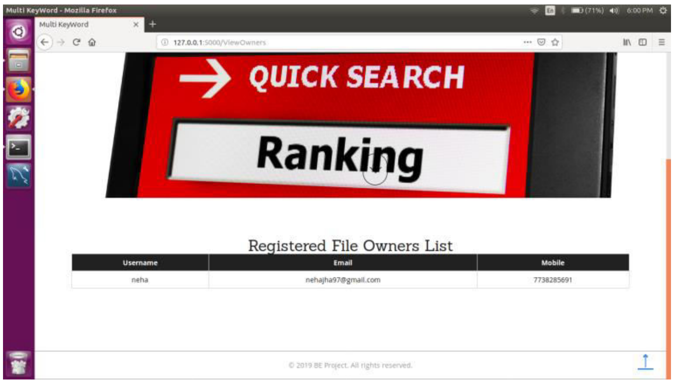
Enter Your Name: nidhi

Password: •••••••••

Enter Email Id: njha408@gmail.com

Enter Your Mobile No: 8450977527

**Fig-5:** Owner Registration

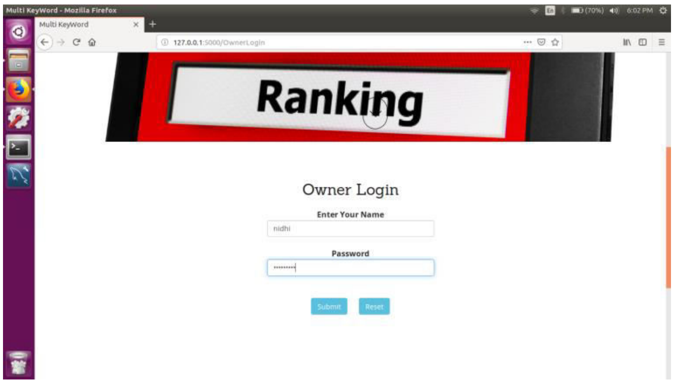


Registered File Owners List

| Username | Email | Mobile |
|---|---|---|
| neha | nehajha97@gmail.com | 7738285691 |

© 2019 BE Project. All rights reserved.

**Fig -6**: Registered Owner



Owner Login

Enter Your Name: nidhi

Password: ••••••••

**Fig -7**: Owner Login

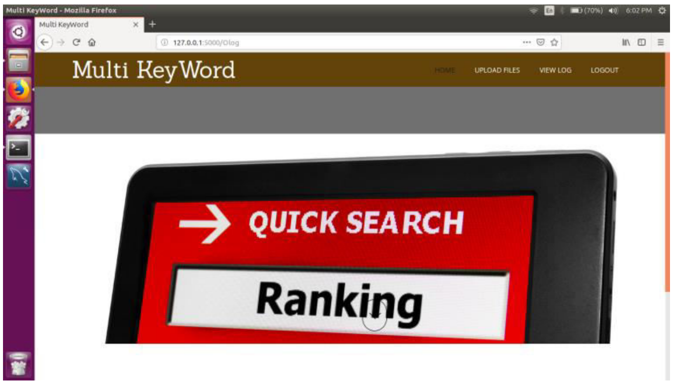
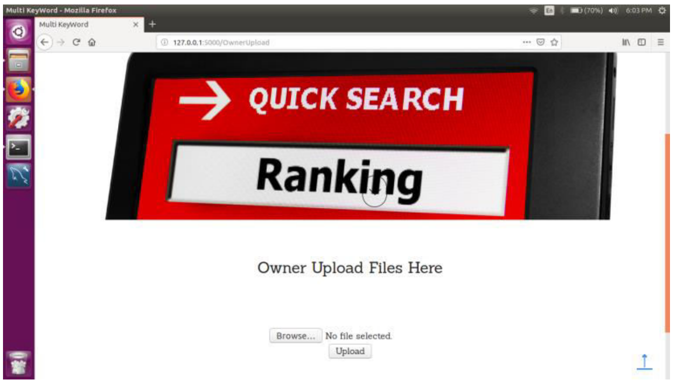


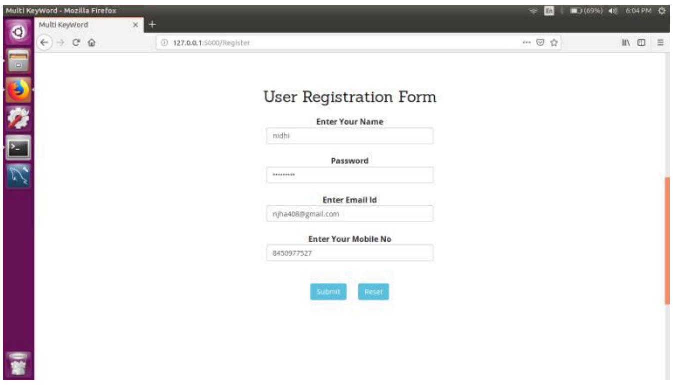
**Fig-8:**Owner Home page



Owner Upload Files Here

No file selected.

**Fig-9:** Owner Uploads



User Registration Form

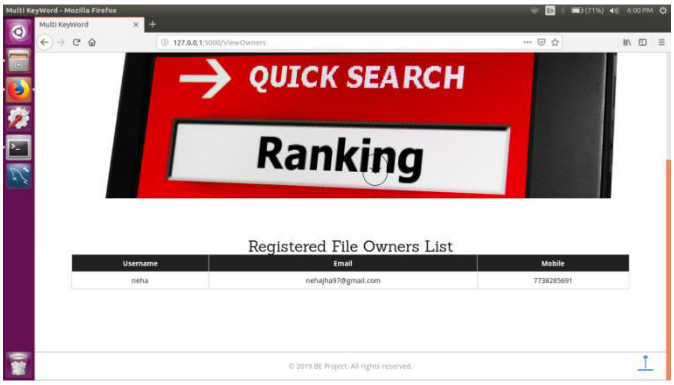
Enter Your Name: nidhi

Password: ••••••••••

Enter Email Id: njha408@gmail.com

Enter Your Mobile No: 8450977527

**Fig-10:** User Registration

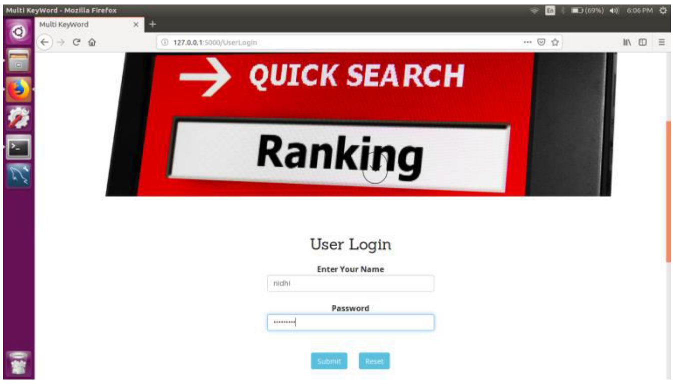


Registered File Owners List

| Username | Email | Mobile |
|---|---|---|
| neha | nehajha97@gmail.com | 7738285691 |

© 2019 BE Project. All rights reserved.

**Fig-11:** Registered User

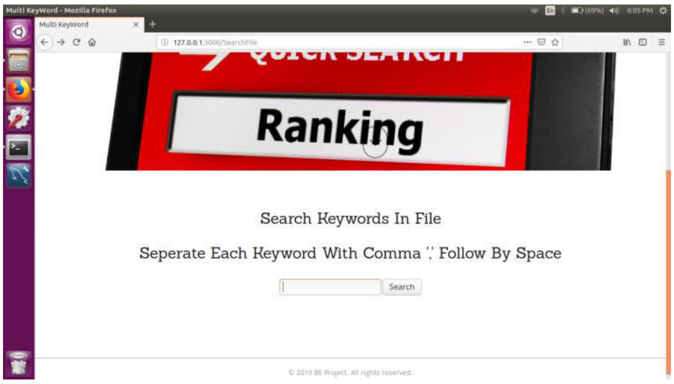


User Login

Enter Your Name: nidhi

Password: ••••••••

**Fig-12:** User Login



Search Keywords In File

Seperate Each Keyword With Comma ',' Follow By Space

© 2019 BE Project. All rights reserved.

**Fig-13:** User's Search Page

## ACKNOWLEDGEMENT

I wish to express my profound thanks to all who helped us directly or indirectly in making this paper. Finally I wish to thank to all our friends and well-wishers who supported us in completing this paper successfully I am especially grateful to our guide for him time to time, very much needed, valuable guidance. Without the full support and cheerful encouragement of my guide, the paper would not have been completed on time.

## REFERENCES

[1] A. Ghanbarpour, H. Naderi, IEEE TRANSACTIONS On Knowlegde And Data Engineering,2018.An Attribute-Specific Ranking Method Based on Language Models for Keyword Search over Graphs.

[2] Karl Severin, Swapna S. Gokhale Aldo Dagnino. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), pp: 978-1-7281-2607-4.Keyword-Based Semi-Supervised Text Classification

[3] Vidhya.K.A, G.Aghila (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 2, 2010. A Survey of Naïve Bayes Machine Learning approach in Text Document Classification

[4] Pawar Supriya, Dr. S. A. Ubale.International Journal for Research in Applied Science & Engineering Technology (IJRASET) *Volume 5 Issue VII, July 2017.* Multi-Keyword Top-K Ranked Search over Encrypted Cloud Using Parallel Processor.

[5] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré Stanford University, pages 3567–3575, 2016. Data Programming:Creating Large Training Sets, Quickly.