Author et al. 2019, 7:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Graphical Password Based Robust Document Retrieval from Multi-Keyword Query

Nishu Kumari **Asst. Prof. Ritu Ranjani Singh**
Dept. of Computer Science & Engg.
Bansal Institute of science and Technology
Bhopal, M.P., India
nknishukumari@gmail.com

Abstract- As the digital data increases on server's different researcher have focused on this field. From last few decades document are obtained from the various set gathered data from users, researcher, authors, etc. This work focus on the spatial basis privacy method where secret password was developed with the help of image by using various combinations from same image set. So whole process is divide into two steps first is generating a password where user create password which is a combination of text as well as image. While in the next step login is done where password is generate from the same image and retrieve desired document from the set of available documents as per user query. In order to maintain the privacy encrypted query was pass to fetch relevant document. Experiment was done on real dataset and results were compared with existing methods on various evaluation parameters.

Index Terms- Supervised Classification, Text Mining,, Text Feature, Text pre-processing.

1. INTRODUCTION

Nowadays, large quantity of data is being accumulated in the data repository. Usually there is a huge gap from the stored data to the knowledge that could be constructed from the data. This transition won't occur automatically, that's where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data.

Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. Fast developing computer science and engineering techniques and methodology generates new demands to mine complex data types.

A number of Data Mining techniques (such as association, clustering, classification) are developed to mine this vast amount of data. Previous studies [18] on Data Mining focus on structured data, such as relational and transactional data. Document Retrieval (more commonly referred to as Information Retrieval by researchers in the field) is the computerized process of producing a list of documents that are relevant to an inquirer's request by comparing the user's request to an automatically produced index of the textual content of documents in the system. These documents can then be accessed for use within the same system. Nearly everyone todument Retrieval systems, although they may not refert them as such, but rather as Web-based search engines, e.g. Google, Yahoo, Alta Vista, etc.

Document Retrieval systems are based on different theoretical models, which determine how matching and ranking are conducted. The most prevalent models are Boolean, Vector Space, Probabilistic, and Language Modeling, each of which is explained below. Within the indexing aspect of each model, the system processes, represents, and weights the substantive content of documents and queries for matching. It is here in feature selection that one might expect to see linguistic theories and models used extensively, however, to date, most systems utilize only the morphological and lexical levels of language,

with some notable exceptions where full Natural Language Processing is utilized.

II.RELATED WORK

In [4] presented an approach using closest neighboring algorithm with cosine analogy to classify research papers and patents published in several fields and stored in different conferences and journals database. Experimented results proves that user get better outcomes by traversing research paper or patent in specific category. The primary advantage of presented technique is that search area become compact and waiting time for query's solution has reduced. They have calculated the threshold depending upon similarity of terms of query, patent and research paper. Threshold calculation was not numerical value based. Hence the presented technique categorize more precisely than existing approach.

In [5] examined that social media posts can analyse the personal intelligence. Primary base of human behaviour is personality. Personality tests elaborate the individual's persona that influences the relations and priorities. User reveal their opinions on social media. The text classification was exploited to predict the character and nature on the basis of their comments. Indonesian and English language were used for this test. Naïve Bayes, SVM and K-Nearest Neighbour are executed methods for classification. Naïve Bayes performed better than other techniques. The research work uses MyPersonality dataset. In this dataset used to classify the personality based-on an online ques

In [6] traversed internet for huge data to gather knowledge. It consists of huge unstructured data like text, image and video. Challenging issue is organization of big data and gathers useful knowledge that could be used in bright computer system. Ontology covers the big area of topic. To construct an ontology with specific domain, big dataset on web was used and arranging with particular domain before the completion of organization. Naïve Bayes classifier was implemented with Map reduce model to organize big dataset. Plant and animal domain articles from encyclopaedia available online were used to experiment. Proposed technique yielded robust system with high accuracy to classify data into domain specified ontology. In this research work, dat sets use plant and animal domain animals article in online encyclopedia and Wikipedia as dataset. In [7] presented a Bayesian classification technique for text categorization using class-specific characteristics. Unlike regular approaches of text categorization proposed method had chosen a particular feature subset in every class. Applying such class-dependent characteristics for classification, a Baggenstoss's PDF Projection Theorem was followed to recreate PDFs from class-specific PDFs and construct a Bayes classification rule. The importance of suggested approach is that feature selection criteria, like: MD (Maximum Discrimination), IG (Information Gain) are included easily. Evaluated the performance on several actual benchmark data set and compared with feature selection approaches. The experiments , they tested approach for texture classification on binary real time benchmarks : 20-Reuters and 20-Newgroups.

In [8] proposed a BI-LSTM (Bidirectional long short term memory) network to inscribe the short text classification with 2 settings. The short-text classification is required in applications of text mining, especially health care applications in short texts mean linguistic ambiguity bound semantic expression due to which traditional approaches fails to capture actual semantics of limited words. In health care domains, the text includes infrequent words, in which due to lack of training data embedding learning is not easy. DNN (Deep neural network) is potential to boost the performance as per their strength of representation capacity. Initially, a common attention mechanism was adopted to guide network training with domain knowledge in dictionary. Secondly, direct cases when knowledge dictionary is unavailable. They presented a multi-task model to learn domain knowledge dictionary and performing text classification task in parallel. They applied suggested technique to existing healthcare system and exclusively available ATIS dataset to get better results.

In [9] surveyed the process of text classification and existing algorithms. Large amount of data is stored as e-documents. Text mining is a technique of extracting data from these documents. Classifying text documents in specific number of pre-defined classes is Text classification. Its application consists of email routing, spam filtering, language identification, sentiment analysis, etc.

III.PROPOSED METHODOLOGY

1. Select Image- As this work focus on strong password creation method where user put his user

name and move for password creation. While creating password user has to select image of its own choice. So this selection of image as per user choice is done 5.in this step.

2.Pre-Processing- Read a image means making a matrix of the same dimension of the image then fill the matrix correspond to the pixel value of the image at the cell in the matrix.

3. Blocked Image- Image obtained after pre-processing was divide into fix size block where size of block are depend on the user it means user can pass the value of block size in form of input parameter for the image. So this act new dimension for the password robustness. This can be understand by below expiate let pre-processed image is IM[] having dimension 128x128, where block size is b={2 / 4 /8 / 16 / 32 ……}.

4. Add to Secret Key- In this step if user click on any block in the blocked image than that block position is stored in the dataset where shuffling round is also stored.

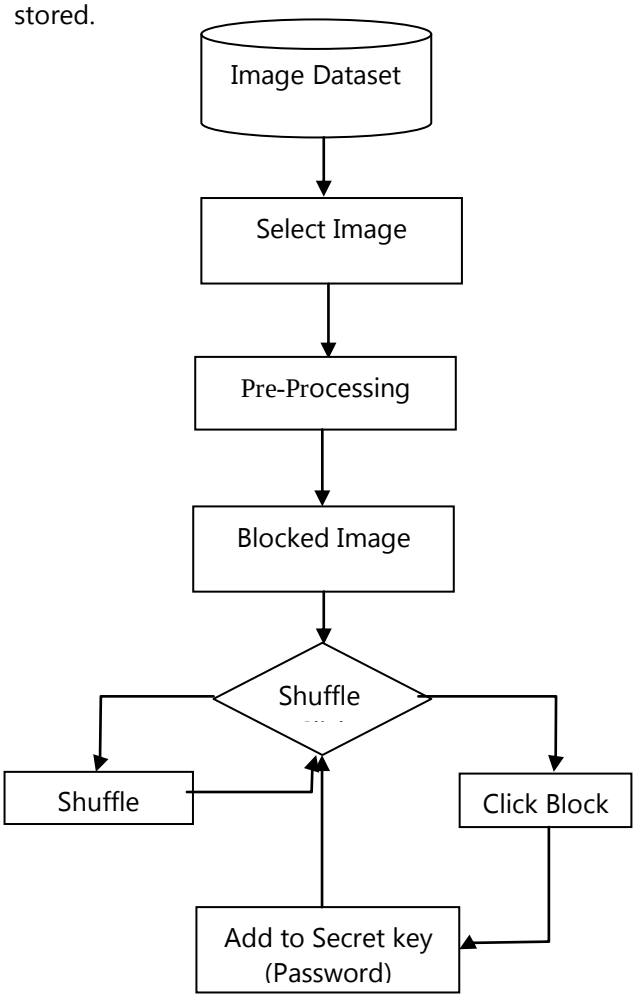

Fig.1 Flow chart of proposed password creation algorithm.

5. Shuffle- This operation increase one more dimension for the image based password creation where selection of shuffle operation at any click point sequence was include. Here user is free to shuffle image any number of time before or after selecting selection a set of click point. In order to shuffle blocks in the image proposed work utilize chaotic function for block jumbling where λ is the parameter for correct set of jumbling in the image block. For every shuffle operation new set of blocks are available. This shuffling increases the confusion in the password, simultaneously complexity of the password also raises.

6. Chaotic function- In this function one matrix is multiple by the block position of the image which generate new position for the block. Here as per the λ value different position was developed by the function. So as per chaotic function multiplying matrix is represent as:

$$Chaotic\_Matrix = \begin{bmatrix} 1 & 1 \\ \lambda & 1+\lambda \end{bmatrix}$$

7. Document Retrieval Methodology

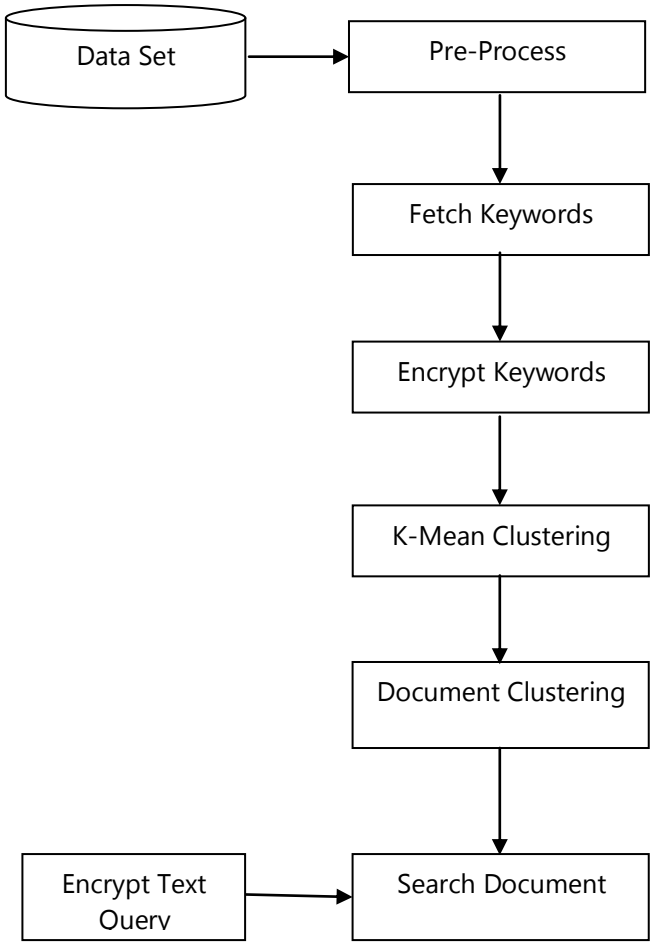

Fig.2 Block diagram of training module**.**

8. Pre-Processing- Preprocessing is a process used for conversion of document into feature vector. Just like text categorizations the preprocessing also has controversy about its division. This work utilizes text preprocessing which consist of words responsible for lowering the performance of learning models.

Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination and stemming. Here Stop-words are functional words which occur frequently in the language of the text (for example a, the, an, of etc. in English language), so that they are not useful for classification.

9. Feature Term-The vector which contains the pre-processed data is use for collecting feature of that document. This is done by comparing the vector with vector KEY (collection of keywords) of the ontology of different area. So the refined. vector will act as the feature vector for that document [11, 14]. So the lists of words which are crossing the threshold are consider as the keywords or feature of that document.

[Feature] = mini_threshold ([processed text])---(1)

In this way term feature vector is created from the document.

10. Encrypt Keywords- In this work keywords obtained from the user generated key are encrypt by AES algorithm. This algorithm is safe and fast. Here server provide this encryption to the keywords obtained from the dataset. Now common step for all kind of data is that each data need to be convert into 16 element set of input.

11. K-Mean Clustering- In this work few document keyword set are consider as the cluster center of the individual data owner. For finding difference between two document work used similarity function. Here fetch keywords from documents are compared with the cluster center keywords. As the number of similar keywords increases than fitness value is high.

12. Search Document- In this step as per the keywords (terms) from the user text query. Fig. 2 show whole steps of fetching. Now all term that are present in the text query act as key for the selecting the cluster where each document set from the matched cluster are index as per the text query term. So selected cluster find the document rank in that cluster only.

IV.EXPERIMENT RESULTS

MATLAB 2012 as is the tool use for the implementation of this work. It is use because of its rich library which has many inbuilt function that can be directly use in this work for different purpose. Out of different function few are intersection, comparing of the string, etc. Experiment was done on real as well as on artificial dataset. Here different set of dataset was use for retrieving documents.

Results

Table 1 Comparison of precision value with previous work [15].

Comparison of Precision values		
Query	Proposed Work	Previous Work
Q1	0.555556	0.444444
Q2	0.777778	0.6667
Q3	0.666667	0.555556
Q4	0.888889	0.777778

From above table 1 it is obtained that proposed work precision value is higher than previous work on different queries. As query set has good quality keywords results of proposed work is also high.

Table 2 Comparison of Recall value with previous work [15].

Comparison of F-Measure values		
Query	Proposed Work	Previous Work
Q1	0.526316	0.470588
Q2	0.608696	0.571429
Q3	0.571429	0.526316
Q4	0.64	0.608696

From above table 2 it is obtained that proposed work f-measure value is higher than previous work on different queries. As query set has good quality keywords results of proposed work is also high. From above table 3 it is obtained that proposed work execution time value is comparatively low then previous on different queries. As query set has good quality keywords results of proposed work is also high.

Table 3 Comparison of execution time in second with previous work [15].

Comparison of execution time in second		
Query	Proposed Work	Previous Work
Q1	11.3627	12.9628
Q2	8.29032	9.32459
Q3	10.8338	11.2351
Q4	10.2114	12.5823

Table 4 Comparison of proposed and previous work password creation algorithm.

Number of User	Sign-Up Successful Rate	
	Proposed Work	Previous Work
12	0.9166	0.833
15	0.8667	0.733
18	0.8333	0.777

It has been obtained from table 4 that proposed work has high sign-up rate as compared to previous password creation algorithm. Here by the use of whole block as click point user can easily click and remember that position in the image. Here freedom of creating a block size as per user choice is also helpful to make successful signup.

V.CONCLUSIONS

With the drastic increase of the digital text data on the servers, libraries it is important for researcher to work on it. Considering this fact work has focus on one of the issue of the document retrieval. Here many researchers have already done lot of work but that is focus only on the content classification where in this work document are classify. Proposed work has increase the retrieval efficiency of the work in all different evaluation parameters. So use of hash based indexing provides privacy with efficiency for document retrieval.

REFERENCES

1. Alan Díaz-Manríquez , Ana Bertha Ríos-Alvarado, José Hugo Barrón-Zambrano, Tania Yukary Guerrero-Melendez, And Juan Carlos Elizondo-Leal. "An Automatic Document Classifier System Based on Genetic Algorithm and Taxonomy". accepted March 9, 2018, date of publication March 15, 2018, date of current version May 9, 2018.
2. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana. "Relevance Feature Discovery for Text Mining". IEEE transaction knowledge and Data ENGINEERING, VOL. 27, NO. 6, JUNE.
3. Souneil Park, Jungil Kim, Kyung Soon Lee, and Junehwa Song. "Disputant Relation-Based Classification for Contrasting Opposing Views of Contentious News Issues".Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.
4. B. Gourav& R. Jindal, "Similarity Measures of Research Papers and Patents using Adaptive and Parameter Free Threshold," International Journal of Computer Applications, vol. 33, no. 5. 2011.
5. B.P.Yudha, and R. Sarrno. "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," In Data and Software Engineering (ICoDSE), in proceedings od International Conference on, pp. 170-174. IEEE, 2015.
6. J. Santoso, E. M. Yuniarno, et al., "Large Scale Text Classification Using Map Reduce and Naive Bayes Algorithm for Domain Specified Ontology Building." In Intelligent Human-Machine Systems and Cybernetics (IHMSC), in proceedings of the 7th International Conference on, vol. 1, pp. 428-432. IEEE,2015.
7. B.Tang, H. He, et al., "A Bayesian classification approach using class-specific features for text categorization." IEEE Transactions on Knowledge and Data Engineering 28, pp: 1602-1606,no. 6, 2016.
8. S. Cao, B. Qian, et al.," Knowledge Guided Short-Text Classification for Healthcare Applications", 2017 IEEE International Conference on Data Mining (ICDM) vol. 2, no. 6,pp: 234-289. 2017.
9. V. K. Vijayan, K. R. Bindu, et al., "A comprehensive study of text classification algorithms." IEEE Advances in Computing, Communications and Informatics (ICACCI),, vol 12, no. 1 pp: 42-53. 2017.
10. K.. Y. Wu, M. Zhou, et al., "A fuzzy logic-based text classification method for social media data," Systems, Man, and Cybernetics (SMC), IEEE International Conference on, vol.13,no.3 pp:23-32. 2017.
11. Wei Zhang, Yaping Lin, Sheng Xiao, Jiewu, And Siwang Zhou. "Privacy Preserving Ranked Multi-Keyword Search For Multiple Data Owners In Cloud Computing". Ieee Transactions On Computers, Vol. 65, No. 5, May 2016.