International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

A Robust Storage and Fetching of Data by maintaining the Individual Privacy

Mrs. Shivani Pandey [1], Mr. Rajneesh Pachouri[2], Mr. Anurag Jain[3]

[1]*M.Tech Research Scholar Department of Computer Science Engineering AIST, Sagar*
[2]*Assistant Professor,Department of Computer Science Engineering AIST, Sagar*
[3]*Assistant Professor,Department of Computer Science Engineering AIST, Sagar*

Abstract - *Searching of information from a huge quantity document, information mining is used which means extraction of attractive model. Bu this brings extraction of intruder and sensitive rule mining that should be kept private. So this paper focuses on providing secure access to the user and retrieves relevant data from the data source. Security of the data was maintained by introducing the image based password system. Here user need to remember a image based password to retrieve information. While data is store on server by using AES encryption system with document indexing. This document indexing improves searching efficiency of the work. Experiment was done on real dataset and results were compared with existing algorithms. It was obtained that proposed system has improved various evaluation parameters as compared to previous existing methods.*

Key Words: **Distributed Data, Data Mining, Encryption, Information Retrieval.**

1. INTRODUCTION

In today's era in data storehouse an enormous amount of knowledge is collected. There's differentiation between the knowledge and the data that is keep and also the knowledge that we have a tendency to earn from the knowledge. The change won't occur consequently, that's the rationale the term data processing appears. There must be some idea about data for data analysis but to urge deeper idea about the info , data processing can help to us. From the collected data getting idea is that the main aim of knowledge mining. for data analysis Human effort is employed for brief interval and for huge data it builds a bottle neck.

To store complicated types of data, Enhancement and development of technology leads. To store such a large amount of data, various techniques are adopted such as association, clustering classification. Actually the complete information is kept in text databases, which is nothing but collected works of text from diverse sources like news articles, books, digital libraries and web pages. Due to accessibility of data in electronics from such as electronics publication e-mail, C.D ROM and the World Wide Web. Information kept in content databases are by and large semi-structured i.e. neither they are unstructured nor they are structured. Information retrieval (IR) machine identifies the files in a set which match a consumer's question. The most well known IR machine are seek engine as Google, Which perceive those documents on the world wide web which might be applicable to a fixed of given words IR structures are often used in libraries, where the documents are usually no longer the book themselves but virtual file containing information about the books. that is however converting with the arrival of digital libraries, wherein the document being retrieved are digital model of books and journals IR device permit us to slim down the set of documents which can be relevant to a selected hassle. As textual content mining involves making use of very computationally in depth set of rules to big file series, IR can accelerate the analysis substantially by using reducing documents for evaluation. as an example if interested in mining records only about protein interplay, might restrict our analysis to files that includes the name of a protein or some form of the internet 'to engage' or certainly one of its synonymous.

1.1 Problem Statement

As big amount of text information is constantly upload on the server and this takes to new problem of the relevant records fetching. So specific search are required for the fetching of the relevant report from the group. As wide variety of intruders make non-stop attack at the system so secure system is quite demand. As this is an quick an form of kind service so execution time for passing user seek query and locating relevant records required much less time. System must accept multi-keyword searching method. So a right data structure is needed for the arrangement of the phrases from the available dataset.

2. LITERATURE SURVEY

1. S. Ramasundaram et al. [4] intended to get better the N-grams classification algorithm by applying Simulated Annealing (SA) look for method to the classifier. The hybrid classifier NGramsSA realized an improvisation to the first NGrams classifier while acquiring every one of the upsides of Ngrams approach. Highlight diminishment utilizing strategy is utilized yet its multivariate incentive among the n-grams influences the execution of the classifier.

2. Deepa B. Patila in [9], has classify image data which is extremely compound and required stochastic relations for

the creation of feature vector from images. Here various sorts of relations are consolidated where individuals from the element vector is fluffy in nature. So this connection based picture grouping is very relying on upon the sort of picture organization and in addition on the edge determination. This algorithm is easy to handle, while stochastic connection help in distinguishing the different vulnerability properties. Here to develop that stochastic relation deep study is required, on prior knowledge accuracy is depends.

3. K. Fragos et al. in [2] also close for consolidating diverse methodologies for content characterization. The strategies that author have joined have a place with same paradigm – probabilistic. Naïve Bayes and Maximum entropy classifiers are tried on the applications where the individual execution is great. The merging operators are utilized over the individual outcomes. Maximum and Harmonic mean operators have been utilized and the execution of combination is superior to the individual classifiers.

4. Dr. B. Poorna, Sudha Ramkumar in [1] has done content document bunching which was utilized to aggregate an arrangement of records in view of the data it contains and to give recovery comes about when a client browses the web. . Investigational evidences have shown that Information recovery applications can benefit from document clustering and to improve the performance of retrieval of information it has been used as a tool. An interdisciplinary field of knowledge management and text mining is Information retrieval. A typical step in many text mining problems is known as Dimensionality Reduction (DR) which includes changing meager information into a shorter and more compact one. DR can be done in 2 ways: feature reduction and feature selection. This review implements dimensionality diminished through component choice with k-means algorithm.

5. In [20] author utilizes, a hierarchical clustering technique is proposed to support more inquiry semantics and furthermore to take care of the demand for quick figure content pursuit inside a major information environment. The proposed hierarchical approach groups the archives in light of the base importance edge, and after that parcels the subsequent bunches into sub-bunches until the imperative on the greatest size of group is come to. This approach can achieve a direct computational complexity against an exponential size increment of archive collection, in the search phase. In order to confirm the legitimacy of list items, a structure called least hash sub-tree is composed in this paper.

In tree data arrangement, standard terms are accepted here, where levels increase according to the number of keywords. Similar records that are no longer time consuming, it is also important to find the current record recursive steps.

3. Proposed Methodology

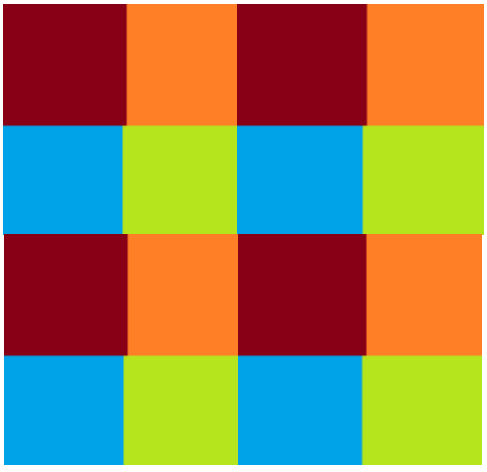
This work focus on the spatial basis privacy method where secret password was developed with the help of image by using various combinations from same image set. So whole process is divide into two steps first is generating a password where user create password which may be a combination of text also as image. Although login is done in the next step, the password is created from the same image that is obtained after password formation. Here password steps are so taken that it can save the originality identity of the carrier image. In Fig. 1 whole embedding work block diagram is explained. Proposed work has increase the retrieval efficiency of the add all different evaluation parameters. So use of hash based indexing provides privacy efficiently for document retrieval.

The following sets of possibilities are available in this work:

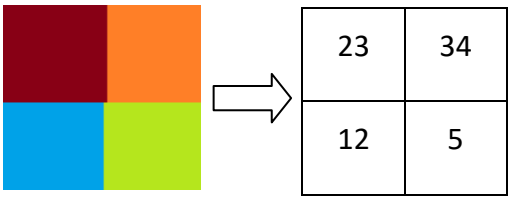
1. User enters userid.
2. For password creation user randomly select one image.
3. Now system asks for image block size and λ value for chaotic function. Here this increases the flexibility and combination of password.
4. As strength of password depends on number bits or characters. So proposed work store whole block as the secret key of the user.
5. In chaotic function as per λ value different set of jumbling is possible for the same set of blocks so if one knows the block size or someone find than λ value is again play an important role.

As this work focus on strong password creation method where user put his user name and move for password creation. The user must choose a picture of their own choosing when creating a password. So this image selection is done in this step as per user preference. There is not bound to select fix set of image over here. Once the user selected the image, transfer the image for the pre-processing stage.

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. Let us consider a 2x2 dimension image it means have four pixels. Two pixels in each row. So read an image will transform image into same 2x2 dimension matrix having four values represent color of the pixel.



(a) Consider image of 4x4 dimensions



(b) Reading Example of 2x2 block of 4x4 image.

Figure 1 Represent image color matrix to matrix conversion.

This is the flow chart of the proposed work.

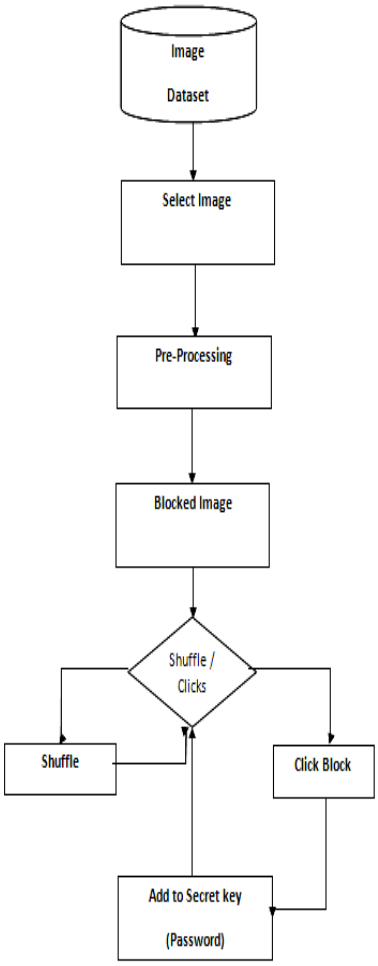


Figure 2 Flow chart of proposed password creation algorithm.

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 which is obtained after the creation of password. Here password steps are so taken that it can save the originality identity of the carrier image.

Feature Term-

The vector which contains the pre-processed data is use for collecting feature of that document. So the lists of words which are crossing the threshold are consider as the keywords or feature of that document. Here one bag of words are maintained which collect the rest of words that is not the same as in dictionary.

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.

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.

Following Step will find similarity between the selected solution and document set.

1. Fitness =0;
2. Loop n = 1:P
3. Loop x = 1:Nc
4. D[x] = S(Ds[n], x) // Here Dist is a Euclidean function
5. endLoop
6. S←Max_Index(D) // Find matximum value index position
7. C[S]←Ds(n) // Sort matrix in decreasing order
8. Fitness← Fitness +Max(D)
9. EndLoop

Now other documents are cluster into relevant cluster as per the fitness function.

4. RESULT ANALYSIS

As various techniques evolve different steps of working for classifying document into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. So following are some of the evaluation formula shown in

equation number 4,5, 6and 7 which help to judge the classification techniques ranking.

$$\text{Precision} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Positive}}$$

$$\text{Recall} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Negative}}$$

$$F\_Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

In above true positive value is obtain by the system when the ranked article is in favor of user query and system also says that article is in favor of the user query. While in case of false positive value it is obtain by the system when the input article is in favor of user query and system do not rank that article in their list.

Signup Success Rate

This is the ratio of number of successful signup done by different user to the total number of user who has tried to make password. This can be understand as if M number of user successfully create password while signup step and N number of total user who try to make password, than M/N is signup success rate.

$$SignUp\_Success\_rate = \frac{Number\_of\_Sccueccful\_SignUp}{Total\_Number\_of\_User}$$

Login Success Rate

This is the ratio of number of successful login done by different user to the total number of user who have try to login. This can be understand as if M number of user successfully login and N number of total user who try to login, than M/N is signup success rate.

$$LoginUp\_Success\_rate = \frac{Number\_of\_Sccueccful\_LoginUp}{Total\_Number\_of\_User}$$

Table 1 Comparison of precision value with previous work.

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's obtained that proposed work precision value is above previous work on different queries. As query set has good quality keywords results of proposed work is additionally high.

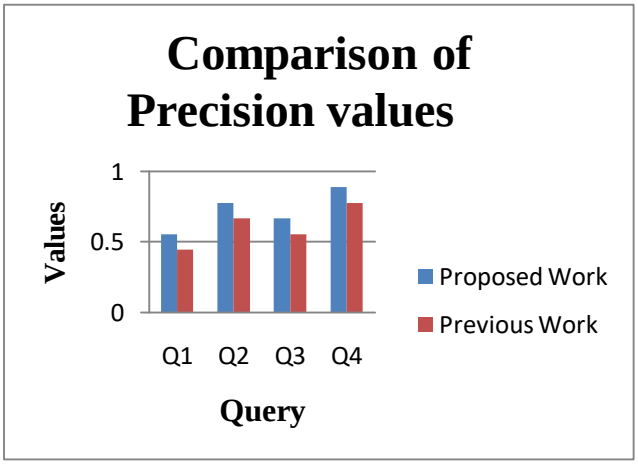
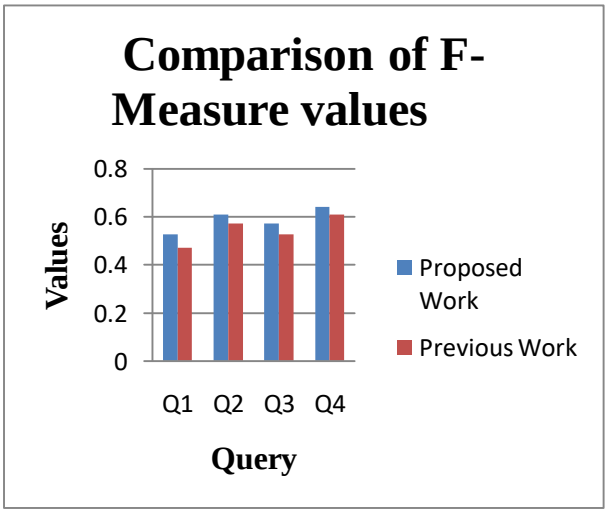


Table 2 Comparison of Recall value with previous work.

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 additionally high.



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 3. Comparison of execution time in second with previous work.

From above **table 3** it's obtained that proposed work execution value is relatively low then previous on different queries. As query set has good quality keywords results of proposed work is additionally high.

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

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

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.

Number of User	Login Successful Rate	
	Proposed Work	**Previous Work**
12	0.833	0.667
15	0.8	0.733
18	0.777	0.667

Table 5. Login rate comparison of proposed and previous work password creation algorithms.

It has been obtained that proposed work has high login 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.

5. CONCLUSIONS & FUTURE WORK

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

As this is often a replacement introduction of classification of articles by the ontology more work got to be done by including this for the opposite language also because this work is done in English language articles. One more parameter that is required to be decrease that is execution time as by introduction of ontology technique it is necessary to filter words in the field, and for comparison it take time so some kind of parallel comparator need to develop for this which will decrease the overall execution time. As there's always work remaining in every because research may be a never ending process, here one can implement similar thing for various other language.

REFERENCES

1. B. Poorna, Sudha Ramkumar. "Text Document Clustering Using Dimension Reduction Technique". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 7 (2016) pp 4770-4774.
2. K. Fragos, P.Belsis, and C. Skourlas, "Combining Probabilistic Classifiers for Text Classification",Procedia - Social and Behavioral Sciences, Volume 147 Pages 307–312, 3rd International Conference on Integrated Information(IC-ININFO), doi: 10.1016 /j.sbspro .2014.07. 098 , 2014.
3. S. Keretna, C. P. Lim and D. Creighton, "Classification Ensemble to Improve Medical Named Entity Recognition", 2014 IEEE International Conference on Systems, Man, and Cybernetics, San Diego, CA, USA, 2014.
4. S.Ramasundaram, "NGramsSA Algorithm for Text Categorization", International Journal of Information Technology & Computer Science (IJITCS), Volume 13, Issue No : 1, pp.36-44, 2014.
5. Christiane Fellbaum, editor. "WordNet An Electronic Lexical Database. MIT Press, Cambridge, Mass", 1998.
6. Blaffz Fortuna, Carolina Galleguillos, and Nello Cristianini "Detecting the bias in media with statistical learning methods". In Text Mining: Theory and Applications. Taylor and Francis Publisher, 2009.
7. Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. "Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence", 2013.
8. Anurag Sarkar1 , Saptarshi Chatterjee2 , Writayan Das3 , Debabrata Datta. "Text Classification using Support Vector Machine". International Journal of Engineering Science Invention ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 www.ijesi.org ||Volume 4 Issue 11|| November 2015 || PP.33-37.
9. Deepa B. Patila , Yashwant V. Dongre. 'A Fuzzy Approach for Text Mining". I.J. Mathematical Sciences and Computing, 2015, 4, 34-43 Published Online November 2015 in MECS (http://www.mecs-press.net) DOI: 10.5815/ijmsc.2015.04.04 Available online at http://www.mecs-press.net/ijmsc .
10. Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari and Jordan Pascual. "KNN based Machine Learning Approach for Text and Document Mining" International Journal of Database Theory and Application Vol.7, No.1 (2014), pp.61-70 http://dx.doi.org/10.14257/ijdta.2014.7.1.06 ISSN: 2005-4270 IJDTA Copyright ⓒ 2014 SERSC
11. Ghosh S, Roy S, and Bandyopadhyay S K, (2012), A tutorial review on Text Mining Algorithms, International Journal of Advanced Research in Computer and Communication Engineering,1(4).
12. Sagayam R, Srinivasan S, and Roshni S, (2012), A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques,International Journal Of Computational Engineering Research, 2(5).
13. Rada Mihalcea and Andras Csomai. "Wikify!: Linking documents to encyclopedic knowledge",pages 233, 2007.
14. Yuefeng Li, Ning Zhong, and Sheng-Tang Wu "Effective Pattern Discovery for Text Mining". IEEE transaction on knowledge and data engineering Vol. 24 no. 1 Jan 2012.
15. Wenhai Sun, Bing Wang, Ning Cao, Ming Li, Wenjing Lou, Y. Thomas Hou, and Hui Li. "Verifiable Privacy-Preserving Multi-Keyword Text Search in the Cloud Supporting Similarity-Based Ranking IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 11, NOVEMBER 2014.
16. Senja Pollak, Roel Coesemans, Walter Daelemans, and Nada Lavraffc. "Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. Pragmatic"s, 2011.
17. Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In Anne Kao and Stephen R. Poteet, editors, "Natural Language Processing and Text Mining", pages 9 (Springer London), 2007.
18. Lawrence Reeve and Hyoil Han."Survey of semantic annotation platforms ACM symposium on Applied computing", pages 1634{1638. ACM, 2005.
19. Elad Segev and Regula Miesch. "A systematic procedure for detecting news biases: The case of israel in european news sites. International Journal of Communication", 2011.
20. Chi Chen, Xiaojie Zhu, Peisong Shen, Jiankun Hu, Song Guo, Zahir Tari, Albert Y. Zomaya. "An Efficient Privacy-Preserving Ranked Keyword Search Method". IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 27, NO. 4, APRIL 2016.