International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 06 | June -2017 www.irjet.net p-ISSN: 2395-0072

A Robust Keywords Based Document Retrieval by Utilizing Advanced Encryption System

Ajaykumar Mourya[1], Asst. Prof. Chinmay Bhatt [2]

[1] M.Tech. Scholar Department of Computer Science & Engineering R.K.D.F. I.S.T. Bhopal, India
[2]Assistant Professor Department of Computer Science & Engineering R.K.D.F.I.S.T. Bhopal, India

---

Abstract - *As the digital data increases on servers different researcher have focused on this field. As various issues are arise on the server such as data handling, security, maintenance, etc. In this paper document retrieval was proposed that efficiently the fetch document as per query. Here hash based indexing of the dataset document was done by utilizing term features. In order to provide privacy for the terms each of this is identified by a unique number and each document has its hash index key for identification. Experiment was done on real and artificial dataset. Results shows that NDCG, precision, recall parameter of the work is better as compare to previous work on different size of datasets.*

Key Words: *Information Retrieval, Text Feature, Text Mining, Text Ontology* .

1. INTRODUCTION

Data fetching is a field that has been creating in parallel with database frameworks for a long time. Not at all like the field of database frameworks, which has concentrated on query and transaction handling of organized information, data fetching is worried with the association and fetching of data from documents. Since data fetching and database frameworks each handle various types of information, some database framework issues are normally not present in data fetching frameworks, for example, concurrency control, recovery, transaction management, etc. Additionally, some regular data fetching issues are generally not experienced in customary database frameworks, for example, unstructured reports, tentative search by use of keywords and the idea of relevance. Because of the wealth of content data, data fetching has discovered numerous applications. There exist numerous data fetching frameworks, for example, on-line library list frameworks, on-line record administration frameworks, and the web crawlers. An ordinary data fetching issue is to find relevant documents in an archive accumulation in light of a client's query, which is regularly a few keywords depicting a information need, in spite of the fact that it could likewise be an illustration important record. In such a pursuit issue, a client steps up with regards to pul the significant data out from the gathering; this is most suitable when a client has some impromptu data need, for example, discovering data to purchase an car. At the point when a client has a long data require , a fetching framework may likewise step up with regards to —push‖ any recently arrived data thing to a client if the thing is judged as being significant to the client's data require. Such a data get to process is called data filtering, and the comparing frameworks are regularly called filtering frameworks or recommender frameworks.

Text mining is a minor departure from a field called information mining that tries to discover intriguing examples from vast databases. Content databases are quickly becoming because of the expanding measure of data accessible in electronic shape, for example, electronic productions, different sorts of electronic records, email, and the World Wide Web. These days the vast majority of the data in government, industry, business, and different organizations are put away electronically, as content databases. Information put away in most content databases are semi organized information in that they are neither totally unstructured nor totally organized. For instance, an archive may contain a couple organized fields, for example, title, creators, distribution date, and class, et cetera, additionally contain some to a great extent unstructured text data, for example, summary and conclusions.

2. RELATED WORK

Yang et al., [35] proposed a new approach that is L2, 1 -norm regularized Unsupervised Discriminative Feature Selection (UDFS). The algorithm chooses the most discriminative feature subset from the entire feature set in batch mode. UDFS outclasses the existing unsupervised feature selection algorithms and selects discriminative features for data representation. The performance is sensitive to the number of selected features and is data dependent.

Cai et al., [36] presented a novel algorithm, called Graph regularized Nonnegative Matrix Factorization (GNMF) [37], which explicitly considers the local invariance. In GNMF, the geometrical information of the data space is pre-arranged by building a nearest neighbor graph and gathering parts-based representation space in which two data points are adequately close to each other, if they are connected in the graph. GNMF models the data space as a sub manifold rooted in the ambient space and achieves more discriminating power than the ordinary NMF approach.

Fan et al., [38] suggested a principled vibrational framework for unsupervised feature selection using the non Gaussian data which is subjective to several applications that range

from several diversified domains to disciplines. The vibrational frameworks provides a deterministic alternative for Bayesian approximation by the maximization of a lower bound on the marginal probability which has an advantage of computational efficiency. 2) Text summarization and Dataset: Several approaches have been developed till date for automatic summarization by identifying important topic from single document or clustered documents.

Gupta et al., [39] describes a topic representation approach that captures the topic and frequency driven approach using word probability which gives reasonable performance and conceptual simplicity.

Negi et al., [40] developed a system that summarizes the information from a clump of documents. The proposed system constructs identifiers that are useful for useful for retrieving the important the information from the given text. It achieves high accuracy but cannot calculate the relevance of the document.

3. PROPOSED METHODOLOGY

The content mining is done in this work by the proposed technique for recovery of the record or articles in the data-set without having any earlier learning of the archives. Entire work is clarified in fig 1 and 2.

3.1 Stop Word Removal

Stop Word Removal is a procedure utilized for transformation of record into feature vector. Content stop word removal is comprising of words which are in charge of bringing down the execution time of learning models. Stop word removal reduces the span of the info present in articles. Stop-words are practical words which happen as often as possible in the language of the content (for instance a, the, an, of and so forth in English language), with the goal that they are not valuable for grouping [3, 10, 13]. Let D is archive [India is an incredible nation. Its a nation of various religion and caste.], Stop-word S is [a, are, an, and, am, for, is, its, when, where, etc.]. At that point in pre-handling subtraction operation is done on these sets. Here D-S = [India, incredible, nation, diverse, religion, caste].

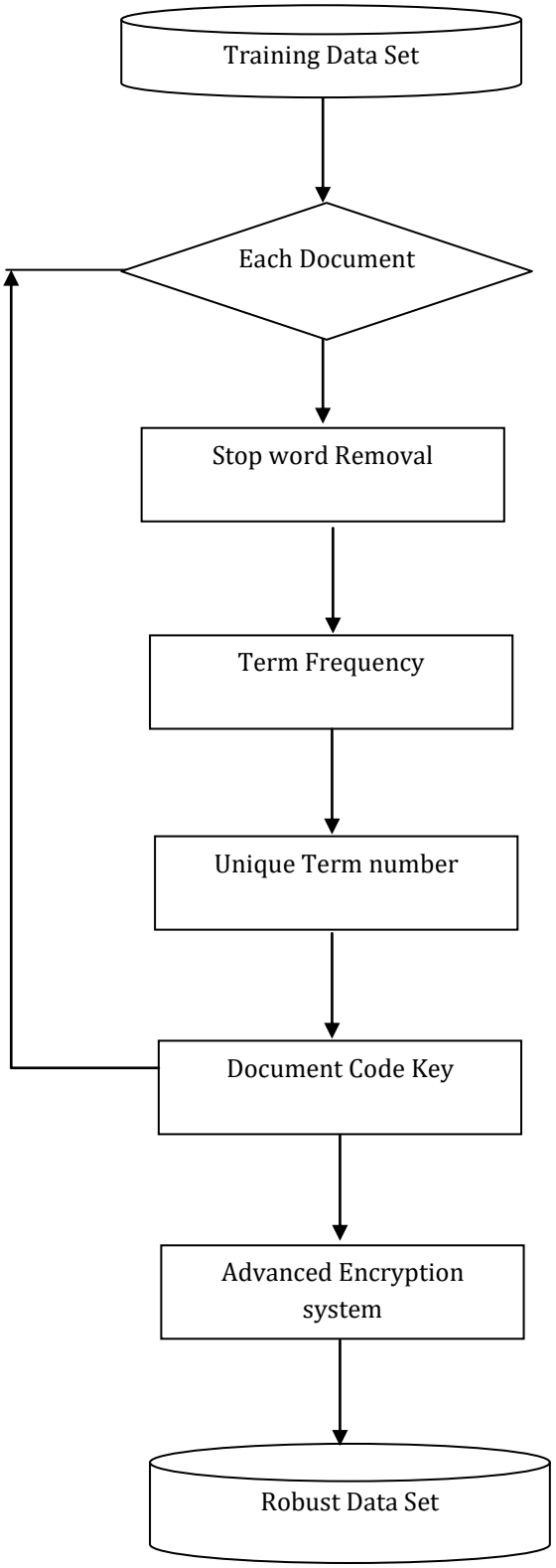

Fig.1 Block diagram of proposed Learning Model.

3.2 Term Frequency

The vector which contains the first step 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 list 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.

3.3 Unique Term Number

Now assign number to each term of the different document. So that a dictionary of words with there number is created where each text is identified by separate number. Here words coming from different document which are already present in the dictionary is not updated. So those terms which are not present in the dictionary is insert in the dictionary with unique termed.

3.4 Document Code Key

In this step document index is decide based on the terms collected from the document. Here all the term are arrange in decreasing order as per the terms frequency value of the terms in the document. So new order of the document term is 918465 this is based on the decreasing order of the term frequency. So from above table one has number is generate for selected document in similar fashion other document in the dataset get collected. Now as per the index value document is identify.

3.5 Advanced Encryption System

Now common step for all kind of data is that each data need to be convert into 16 element set of input. Here each input need to be in integer data type. In case of numeric this is ok, but in case of image gray scale will convert pixel values in integer form. While for text unique number is assign for all extracted words.

In this encryption algorithm four stages are perform in each round. While final round consist of three stages only. These steps are common in both encryption as well as decryption algorithm where decryption algorithm is inverse of the encryption one. So round consist of following four stages.

1. Substitute bytes
2. Shift rows
3. Mix Columns
4. Add Round Key

In final round simply all stages remain in same sequence except Mix Columns stage.

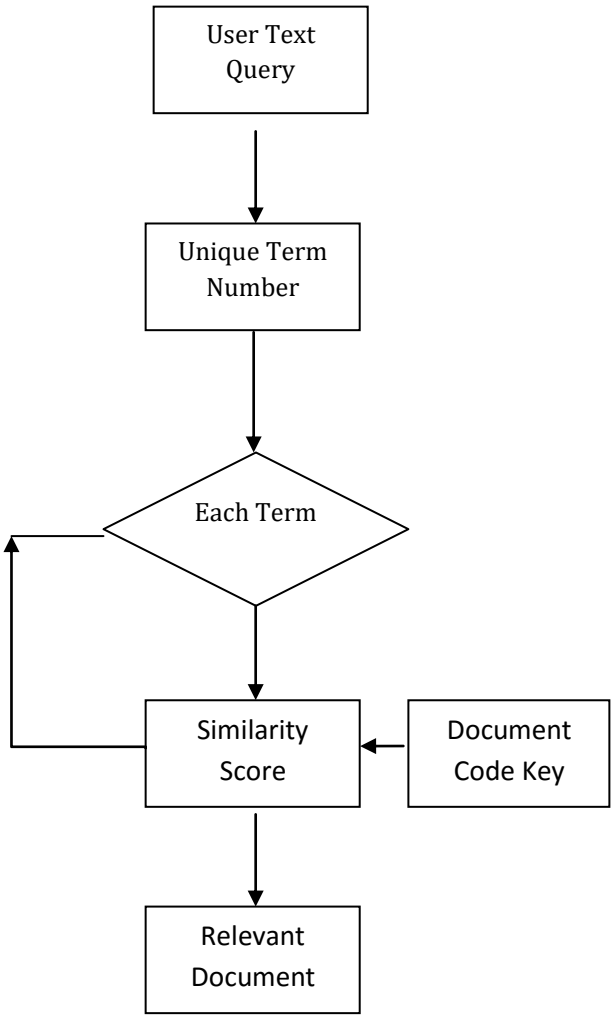

Fig.2 Block diagram of proposed Searching Model.

In this searching model whole step of assigning unique number is same as done in previous steps. So searching and retrieving related document is new for searching model. For convince words obtained from user text query is called as keywords.

3.6 Searching

In this step as per the keywords (terms) from the user text query have their own unique number while number are same as present in the dataset. Due to this unique number privacy of the user query is increases. Now all unique number that are present in the text query is compare from each document code key. So top most document code having similar codes as present in the user query is rank.

4. EXPERIMENT AND RESULT

In order to implement above algorithm for document retrieval MATLAB 2012a tool was used. Here same work can be implement on other programming language as well. But as some of the function was inbuilt in the tool which help researcher to focus on the work. Experiment was done on real as well as on artificial dataset. Here different set of dat set was use for retrieving documents.

4.1 Evaluation Parameter

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.

Precision = (True_positive / (False_positive+ True_positive))

True positive Rate = (True_positive /(False_negative+ True_positive))

F-Measure = (2xPrecisionxRecall/ (Recall + Precision))

4.2Results

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

Comparison of True Positive Rate		
Query	Proposed Work	Previous work[15]
Q1	0.5	0.44
Q2	0.6	0.5
Q3	0.44	0.4
Q4	0.5	0.44

From above table 1 it is obtained that proposed work true positive rate is higher then previous work on different queries. As query set has good quality keywords results of proposed work is also high.

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

Comparison of precision values		
Query	Proposed Work	Previous work[15]
Q1	0.7143	0.5714
Q2	0.8571	0.7143
Q3	0.5714	0.5714
Q4	0.7143	0.5714

From above table 2 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. 3. Comparison of F-Measure value with previous work [15].

Comparison of F-Measure values		
Query	Proposed Work	Previous work[15]
Q1	0.5882	0.5
Q2	0.70	0.5882
Q3	0.5	0.4706
Q4	0.588	0.5

From above table 3 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.

5. 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 has 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 provide privacy with efficiency for document retrieval. As there is always work remaining in every because research is a never ending process, here one can implement similar thing for different other language.

REFERENCES

[1]. Aparna Humad, Vikas Solanki, A New Context Based Indexing In Search Engines Using Binary Search Tree, International Journal Of Latest Trends In Engineering And Technology (Ijltet) Vol. 4 Issue 1 May 2014.
Data Engineering, Vol. 27, No. 6, June 2015.

[2]. Disputant Relation-Based Classification For Contrasting Opposing Views Of Contentious News Issues Souneil Park, Jungil Kim, Kyung Soon Lee, And Junehwa Song. Ieee Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.

[3]. Fabrizio Silvestri, Raffaele Perego And Salvatore Orlando. Assigning Document Identifiers To Enhance Compressibility Of Web Search Engines Indexes. In The Proceedings Of Sac, 2004.

[4]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man,

And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[5]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, And Ou Liu. "An Ontology-Based Text-Mining Method To Cluster Proposals For Research Project Selection". Ieee Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[6]. 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.

[7]. N. Cao, C. Wang, M. Li, K. Ren, And W. Lou, "Privacy-Preserving Multikeyword Ranked Search Over Encrypted Cloud Data," Proc. Ieee Infocom, Pp. 829-837, Apr, 2014.

[8]. N. Cao, S. Yu, Z. Yang, W. Lou, And Y. Hou, "Lt Codes-Based Secure And Reliable Cloud Storage Service," Proc. Ieee Info- Com, Pp. 693-701, 2012.

[9]. Oren Zamir And Oren Etzioni. Web Document Clustering: A Feasibility Demonstration. In The Proceedings Of Sigir, 1998.

[10]. Privacy Preserving Ranked Keyword Search Over Encrypted Cloud Data Dinesh Nepolean, I.Karthik, Mu.Preethi, Rahul Goyal And M.K. Vanethi. Volume 4, No. 11, November 2013.

[11]. Privacy-Preserving Multi-Keyword Ranked Search Over Encrypted Cloud Data Ning Cao,Cong Wang, Ming Li, Member, And Wenjing Lou, Ieee Transaction Parallel And Distributed Ssystems, Vol. 25, No. 1, January 2014

[12]. 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.

[13]. S.Ramasundaram, "Ngramssa Algorithm For Text Categorization", International Journal Of Information Technology & Computer Science (Ijitcs), Volume 13, Issue No : 1, Pp.36-44, 2014.

[14]. Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, And Moch Arif Bijaksana. "Relevance Feature Discovery For Text Mining". Ieee Transactions On Knowledge.

[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.