International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-9, Issue-2, March 2021
https://doi.org/10.21276/ijircst.2021.9.2.1
Article ID D10965, Pages 1-5
www.ijircst.org

A Robust Multi-Keyword Text Content Retrieval by Utilizing Hash Indexing

Mohamed Manzoor Ul Hassan

Business Analyst, ATOS, Briggs & Stratton University, University of the Cumberlands, Milwaukee, Wisconsin, USA

Correspondence should be addressed to Mohamed Manzoor Ul Hassan; mohammedmanzoor@gmail.com

ABSTRACT— Digital content on servers increase the storage and fetching issues. So, researcher works in this field to organize content for fast retrieval with data security. This paper has worked on text digital content retrieval available in form of documents, files. User can search a desired file by test query and relevant list of files get appeared. Keywords were fetched from the text content by removing noisy data during pre-processing. Pre-processed keywords are identified by the number known as term ID. As per the term-ID each text content got a Hash Index which was termed as key numbers in document index. Each term or word has its own identification number known as term Id , so privacy of comparing content terms and user query maintain by hash based searching. As document identification done by hash index key, so storage of text content was done in encrypted numbers once document select for reading then decryption of document applied for a particular user. Experiment was done on real and artificial text content dataset files on different topics. It was obtained that proposed model of Hash indexing and tem based retrieval has improved the privacy with relevancy of as per query.

KEYWORDS- Information Retrieval, Text Feature, Text Mining, Text Ontology.

I. INTRODUCTION

With the increase of digital text data on the servers. Text mining importance is increasing as this decrease lot of labor work for different use of text data. In this text mining research field classification of information and retrieval of documentation is highly required. So combination of various data mining techniques is done while gathering information from the text document [1]. As various researchers are working for improving accuracy of the work, but there is lot of improvement in the work for further increasing the parameters. Text analytics converts text into numbers, and numbers successively bring structure to the information and facilitate to spot patterns. The additional structured information, the higher the analysis, and eventually the higher the choices would be. It is conjointly tough to process the information manually and categorize them clearly. This leads to the emergence of intelligent tools in text processing, within the field of linguistic communication process, to investigate lexical and linguistic patterns. Clustering, categorization, and labeling are major techniques pursue in text analytics [2]. It is the method of distribution, as an example, a document to a selected category label among different out there category labels like “Education”, “Medicine” and “Biology”. Thus, text categorization could be a necessary introduce data discovery [3, 4]. Some of statistical and machine learning approach was already developed for text classification [5]. The aim of this text is to investigate various text classification techniques utilized in monitoring, their clarification in various application domains, strengths, weaknesses, and current analysis trends to supply improved awareness relating to data extraction potentialities.

II. RELATED WORK

In [6] inspected that social media posts will analyse the personal intelligence. Key base of human behaviour is nature. Nature tests detailed the individual’s persona that influences the relations and main concern. Users share their opinions on social media. The text categorization was demoralized to forecast the character and nature on the idea of their comments. Indonesian and West Germanic language were used for this take a look at. Naïve bayes, SVM and K-Nearest Neighbor are performed method for arrangement. Naïve bayes performed higher than different techniques. The analysis work uses Personality dataset. During this dataset used classify the nature based-on an internet queue. In [7] author navigate web for vast information to collect data. It comprises of big unstructured information like text, image and video. Tricky issue is organization of massive information and gathers helpful data that would be utilized in bright computer system. Ontology covers the massive space of topic. To build associate degree ontology with specific domain, massive dataset on net was used and arrangement with specific domain before the completion of organization. Naïve bayes classifier was enforced with Map reduce model to arrange massive dataset. Plant and animal domain articles from

encyclopaedia are online easily available for experiment. Planned technique yielded robust system with high accurateness to classify information into domain specified ontology. During this analysis work, datasets use plant and animal domain animal's article in online encyclopedia and Wikipedia as dataset. In [8] projected a Bayesian categorization technique for text categorization utilizing class-specific characteristics. In contrast to regular approaches of text classification planned methodology chosen a selected feature set in each category. Applying such class-dependent characteristics for classification, a Baggenstoss's PDF Projection Theorem was pursued to recreate PDFs from class-specific PDFs and construct a Bayes classification rule. The significance of instructed approach is that feature choice criteria, like: MD (Maximum Discrimination), IG (Information Gain) are enclosed simply. Estimating the performance on much actual benchmark information set and compared with feature choice approaches. The experiments, they tested approach for texture categorization on binary real time benchmarks: 20- Reuters and 20-Newgroups. In [9], authors projected a similarity computation approach that is based on understood links extracted from the query-log and utilized with K-Nearest Neighbors (KNN) in net web page class. The fresh computed similarity based totally on clicks frequencies facilitates increase KNN for web page category. This similarity utilizes neighborhood information and facilitates lessen the effect of the trouble of dimensionality confronted when using KNN based totally on the text alone. To categorize an internet page , KNN calculates similarities among pi and each internet page in the training set. Then, it ranks web pages within the training set based on the ones similarities. In our case, KNN does a two-stage ranking. First, it ranks net pages the usage of the implicit hyperlinks-based similarity. Then, internet pages having this similarity equal to 0 are ranked again using the cosine resemblance. In [10], they aimed to increase a classifier that may categorize internet pages primarily based on their capacity to draw random surfers. Web pages are categorized into "bad" and "good" types, where the "bad" class implies poor interest drawing ability. In the proposed approach, the net page content material is split into objects. The location occupied by using these items served as the attribute of the classifier. The experiments with various classification algorithms supported via the WEKA device prove that of the ones, particularly the random subspace and the RBF networks, gives excessive accuracy (83.33%) with high accuracy and recall. In [11] proposed a schemes to deal with privacy preserving ranked multi-keyword search in a multi-owner model (PRMSM). Rank the search results and preserve the privacy of relevance scores between keywords and files, we propose a novel additive order and privacy preserving function family. To prevent the attackers from eavesdropping secret keys and pretending to be legal data users submitting searches, we propose a novel dynamic secret key generation protocol and a new data user authentication protocol.

III. PROPOSED METHODOLOGY

This section explains proposed model of text content retrieval in encrypted format whereas per user query hash index was scroll and relevant content files are list for retrieval. Working of proposed work of inserting the text file into hash index was shown in Fig. 1 block diagram. Fetching of relevant files as per user query was shown in Fig. 2 block diagram.

A. Text Content Pre-Processing

In this step of proposed model input text files were arrange in a single dimension vector of words. Each word was compared with Stopword ontology and matching words were removed from the vector. Stopwords {a, for, at, the, in, etc.} are set of words which help in sentence framing. Output of this step is a BagofWord vector, has words appeared in input text content file with word frequency counter.

Let text content TC is pre-process by stopword SW and output BOW is obtained.

$$BOW \leftarrow Pre\_Process(TC, SW)$$

B. Feature Collection

Each word in BOW may act as feature term if term frequency of the word cross minimum frequent term threshold (FTT). Frequent term threshold is an integer value {2, 3, 4, 5, ….}. As per length of document file in terms of words threshold value can be increase or decrease. Features were collect from the below algorithm.

Input: BOW, FTT
Output: FS

1. Loop 1:BOW
2. If BOW[] greater or equal to FTT
3. FS← BOW[]
4. EndIf
5. EndLoop

Once feature set obtained from the text content then add text content in the feature collection matrix which has unique list of universal terms in column and text content hash key as rows. This feature collection matrix shows the presence and absence of the term in the Hash key position. If term present then 1 was there otherwise 0.

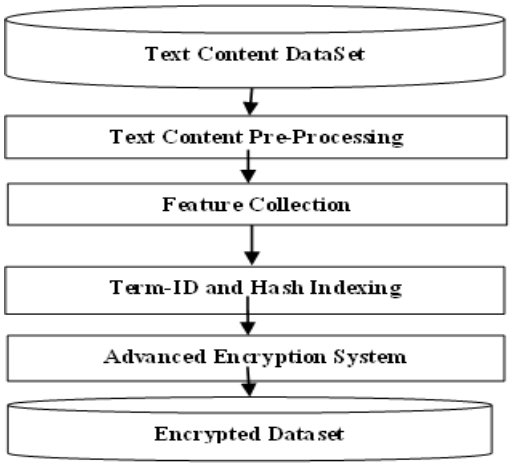

Fig.1: Block diagram of proposed Learning Model

C. Term ID and Hash Indexing

Feature set obtained from input text content file were process further to assign a unique number of the word. So a ontology of term Id was maintained and update with every new word in the ontology. Each term has its own identification in the dataset. As each document has own set of words and frequency in the text content. Based on this term a sequence of termed was prepared either in increasing order of term frequency or decreasing order of term frequency. This work has adopted increasing order term frequency. This can be understand as let text content has FS{t_1, t_2, t_3, t_8} and term ids are {5, 9, 2, 3} then as per hash key generation text content hash value is 2359. In similar way other set of text content hash key was obtained. It is possible that same set of terms were found in a content but frequency count of terms may vary which give a different index position. This paper has use modulus operator to get the hash index for a document where remainder act as position in the dataset for a document to store.

D. Advanced Encryption Algorithm

Each word of a text content was transform into numeric number of sixteen digit and each digit represent a word ASCII number. So if text content has 100 words and each word transform into 1x16 digit then 100x16 digit vector was pass in the advanced encryption algorithm. In AES algorithm four common steps were Add round Key, SubBytes, Shift Rows, Mix Column. As per key number of rounds were performed where each round has all four steps in sequence. Generally 9, 11 or 13 rounds were place for encrypting the input vector. Size of input and output vector is same, means 16 digit vector. This work use AES algorithm for encryption because decryption of encrypted data is losses. It a term has less than 16 chracter then blank character were replace by 0 ASCII number.

E. Text Content Extraction

Testing of proposed model was performed by passing a text query which is a collection of words. So a multi keyword text query were pass into the system and it gives a text content sequence as per similarity from the feature collection matrix. Steps of text content retrieval were shown in fig. 2. Each term is transform into corresponding term id at user side this increase the privacy of the user search behavior and finds the highest matching terms in the feature collection matrix. It was found that highest matching term count as per query were rank one for the text content further other matching text content were place in the output list.

F. Hash Searching

Each selected document in the list can be decrypted by finding the hash index in the dataset, using hash key. Feature collection matrix provide hash key and modulus operator provide corresponding hash index position in the dataset.

G. Decryption

Stored sixteen digit encrypted values of the selected index position text content were pass in AES algorithm which transform back to the original ASCII character sequence.

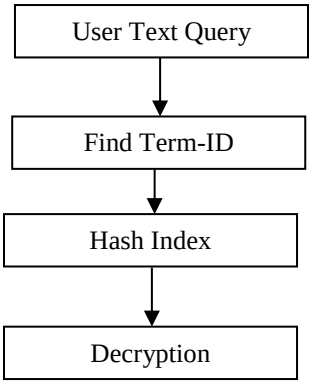

Fig. 2: Block diagram of proposed Searching Model

IV. EXPERIMENT AND RESULTS

Implementation of proposed text content retrieval model was done on MATLAB 2016 version. Experimental machine has 4 GB RAM, with I3 processor 6th generation. Dataset was taken from 13, which have set of text document files with different field. Comparison of proposed model was done with existing algorithm proposed in [12]. Different user query sets were passed in the testing phase of text content retrieval. Evaluation parameter precision, accuracy and NDCG (Normalized Distribution Cumulative Gain) was consider for comparison.

Results:

Table 1: Accuracy based Text content retrieval model comparison

Multi-Word Query	Proposed model	Previous work[12]
MWQ1	0.722	0.667
MWQ 2	0.667	0.611
MWQ3	0.722	0.389
MWQ4	0.667	0.556

Accuracy table 1 shows that proposed model retrieval of text content was more relevant as compared to previous model in [12]. This improvement was achieved by using termed and hash index concept in the work. As term Id reduce the confusion in the comparison and hash index increase data security.

Table 2: Precision based Text content retrieval model comparison

Query	Proposed model	Previous work[12]
MWQ1	0.86	0.72
MWQ 2	0.72	0.72
MWQ3	0.72	0.43
MWQ4	0.86	0.72

Precision evaluation parameter table 2 shows that proposed model finds the more relevant top ten document list as per user different query as compared to previous model in [12]. Term based document arrangement in a feature collection matrix with hashing techniques has improved the work performance.

Table 3: Recall based Text content retrieval model comparison

Query	Proposed model	Previous work[12]
MWQ1	0.6	0.556
MWQ 2	0.556	0.5
MWQ3	0.6	0.3
MWQ4	0.556	0.445

Accuracy table 3 shows that proposed model retrieval of text content was more relevant as compared to previous model in [12]. This improvement was achieved by using termed and hash index concept in the work. As term Id reduce the confusion in the comparison and hash index increase data security.

Table 4: NDCG based Text content retrieval model comparison

Comparison of NDCG Values @7		
Query	Proposed model	Previous work[12]
MWQ1	0.91	0.776
MWQ 2	0.91	0.81
MWQ3	0.91	0.485
MWQ4	0.81	0.796

NDCG evaluation parameter table 4 shows that proposed model finds the more relevant top ten document list as per user different query as compared to previous model in [12]. Term based document arrangement in a feature collection matrix with hashing techniques has improved the work performance.

Table 4: Execution time (Seconds) based Text content retrieval model comparison

Query	Proposed model	Previous work[12]
MWQ1	0.0360	3.3556
MWQ 2	0.0497	3.3034
MWQ3	0.0232	3.2214
MWQ4	0.03036	4.2755

Execution time of fetching a document list as per user query shows that use of hash key for finding the document reduces the comparison time of work. In previous work terms based fetching model uses text character wise comparison, so execution time is very high. In proposed model term Id reduce the time in the comparison and hash index increase data security.

V. CONCLUSIONS

Digital text content management and retrieval model was proposed by this paper. Paper has work on privacy of user search query as well. A secured environment for content safety was developed by the proposed work by using AES algorithm. Paper has reduced the execution time of finding the relevant file list as per user query by using the term Id based comparison, as numeric values are easy and fast to match as compared to text characters. Experiment was done on real dataset and result shows that proposed model has increase the work accuracy with NDCG values. In future researcher can train the ontology to a machine for further reducing the comparison time.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for textdocuments classification. Journal of Advances in Information Technology, 1, 4-20.
[2] Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for textdocuments classification. Journal of Advances in Information Technology, 1, 4-20.
[3] Brindha, S., Sukumaran, S., & Prabha, K. (2016). A survey on classification techniques for text mining. Proceedings of the 3rd International Conference on Advanced Computing and Communication Systems. IEEE. Coimbatore, India.

[4] K. Sarkar and R. Law, ``A novel approach to document classi_cation using WordNet," CoRR, vol. 1, pp. 259_267, Oct. 2015. [Online].
[5] Vasa, K. (2016). Text classification through statistical and machine learning methods: A survey. International Journal of Engineering Development and Research, 4, 655-658.
[6] 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.
[7] 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.
[8] 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.
[9] A. Belmouhcine et M. Benkhalifa. “Implicit Links-Based Techniques to Enrich K-Nearest Neighbors and Naive Bayes Algorithms for Web Page Classification”. Springer International Publishing, 2016, vol. 403,.
[10] G. Khade, S. Kumar, et S. Bhattacharya. “Classification of web pages on attractiveness: A supervised learning approach”. Intelligent Human Computer Interaction (IHCI), 2012.
[11] 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.
[12] 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.
[13] https://ijsret.com/2017/12/14/computer-science/