Astha Jain, 2022, 10:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Butterfly Particle Swarm Optimization Algorithm for Cloud Data Retrieval

Astha Jain, Prof. Rajesh Nigam
Computer Science Department
Technocrates Institute of Technology Science
Bhopal India

Abstract- Availability of digital data on cloud increases day by day. Out of different type of content arrangement and fetching of document files or text files are tough. This paper has developed a model that fetch document file from the clustered data structure. Arrangement of document in cluster data structure was done by Butterfly Particle Swarm Optimization algorithm. As hybrid nature of this algorithm has improve the work performance of cluster document arrangement. Term features were extracted from the document for finding the fitness of the chromosome in the population. This paper has fetch document file in encrypted form. Experiment was done on real dataset and result shows that proposed model content relevancy is high as compared to existing works of text file fetching.

***Index Terms*-Data Fetching, Data Storage, Feature Extraction, Genetic algorithm.**

I. INTRODUCTION

Compute and storage can be expanded separately in public clouds. Modern data warehouses such as Snowflake, Google BigQuery, Amazon Redshift, and Presto [1, 2] are built on this foundation. Many features of those systems, such as the number of nodes, cores, and memory size, may be easily tweaked by users. Users can upgrade their clusters to improve performance. Users can decrease their clusters to save money. This capacity to scale up and down clusters quickly has been well-received by Google BigQuery, which effectively handles this series of operations. They are, however, invisible to consumers users are charged based on the cost of each enquiry [3].

It's worth noting that computational resources (such as cores and memory) account for a bigger share of cloud expenditures than storage. As a result, scaling down compute nodes can result in greater cost reductions. Search engines are software programs that retrieve relevant documents based on keywords entered by the user. To quickly discover relevant content, all commonly used search engines use indexes (e.g., skip lists , tree, learned indexes) [4]. They need their compute and storage to be near together on the same system for quick index traversals. This architecture has the drawback of requiring a large cluster to remain operational as more documents are indexed, even if query workloads are moderate or some pages are queried infrequently.Text classification, text clustering, concept/entity extraction, generation of granular taxonomies, sentiment analysis, document summarizing, and entity relation modeling are all common text mining activities (i.e., learning relations between named entities). Information retrieval, lexical analysis to investigate word frequency distributions, pattern recognition, tagging/ annotation, information extraction, data mining techniques such as link and association analysis, visualization, and predictive analytics are all part of text analysis [5].

Text mining is commonly used to scan a batch of natural language documents and either model them for predictive classification or populate a database or search index with the information gathered. Text (or Document) classification is an ongoing text mining research area in which documents are sorted into predetermined categories. Supervised Document Classification and Unsupervised Text Classification are two types of text classification tasks. In Supervised Document Classification (also known as document clustering), an external mechanism (such as human feedback) provides information on the correct classification of documents or defines classes for the classifier, whereas in Unsupervised Document Classification (also known as document clustering), the classification must be done without any external reference and the system has no predefined classes [6].

Developing text classification models is currently a complex process that includes not only model training but also various other procedures such as data pre-processing, transformation, and dimensionality reduction. The second section introduces the key topics directly relevant to the researchers' proposed text document retrieval algorithms. This section offers a discussion of various text representation models. The paper goes on to explain the proposed notion of document retrieval while also adopting privacy precautions while surfing and indexing. Finally, the research compares the proposed model to different document retrieval approaches currently in use.

II. REVIEW OF THE LITERATURE

Pereira et al. [7] (2018) provide a thorough review and categorization of FS approaches, with a focus on multi-label classification. However, the different methods for dealing with the high dimensionality of the feature space, the different text representation formats such as bag of words and word embedding, and the power of the features' semantics for choosing the most efficient set of features were not considered in these surveys' analyses.

Dang et. al in [8] Used fuzzy set theory, provide a fresh principled approach to passage-based (document) retrieval. The method creates a mix of passage scores based on broad relevance decision rules. Our technique supports the usual heuristic of utilising the maximum constituent passage score as the overall document score by operationalizing these ideas using fuzzy set theory aggregation operations. In [9] Akhter et. al. authors developed a big multi-purpose and multi-format dataset with over ten thousand papers organised into six types in this study. We evaluate the performance of the Single-layer Multisize Filters Convolutional Neural Network (SMFCNN) with sixteen ML baseline models on three imbalanced datasets of varied sizes.

Ying, Y., et al. [10] (2017) used abstractive text summarization to extract essential terms. During text extraction, they used a graph-based approach to prioritise key terms. They chose an abstracted text extraction strategy to connect the keywords while ignoring the sentencing implications. They recommended specific metadata links between standard terms and their associated sentences. They started by grouping several publications into clusters, which then displayed the key issues of the articles. They devised three distinct criteria and demonstrated that their work was superior to simply extracting important phrases.

To address the earlier limitations of manual text processing, Gupta, A. et al. [11] (2018) examined radiological reports using abstractive summarization and Clustering. They enhanced keyword extraction and text grouping in dictionary- and rule-based techniques. The cluster link analysis was missed by the named entity recognition procedure in previous processing methodologies. They used an unsupervised approach to extract named entity relationships with no prior knowledge. They used parse trees to cover text processing requirements and then used distributed semantics to connect them.

Moradi, M. [12] (2018) employed extractive text summarization to condense the text by removing unnecessary information. Clustering & Itemset-mining Biomedical Summarizer was the term given to this technology, which used text summarization to summarise biomedical data (CIBS). This method extracted biomedical concepts from the text input. By applying the itemset mining algorithm to reduced text, the ideas represented the major subjects, and CIBS placed sentences into the clusters' relevant set.

KUSH is a text summary approach introduced by Uçkan et al. [13] (2020). This method determines the largest number of non-overlapping abstractive summarization sets possible. To determine the context of various paragraphs in unstructured text texts, they labelled these sets as nodes. They concentrated their efforts on creating a logical text visualisation. Abstractive summarization was combined with set theory and graph visualisation tools in their proposed KUSH technology.

Sanchez-Gomez et al. [14] (2020) created a context analyzer based on multi-objective abstractive and extractive summarization. Critical objectives and operational usability are defined by context. The keywords-based technique identified the important sections of context in the paragraphs for the functional usability examination. For text selection, the summarizer uses pre-defined criteria. The goal of this project is to address the growing demand for artificial text summarising technologies.

The lack of unknown terms and fragmented phrases was highlighted by Deng et al. [15] (2020). They provided an abstractive-text-summarizing strategy as an alternative to sequence-to-sequence text summarization models in Chinese text assessment, which combined text-summarization and sequence-to-sequence models. In their suggested text summarizer, they included adversarial learning. The addition of abstractive text summarization to the suggested method improved text assessment, according to their comparison results.

III.PROPOSED DOCUMENT RETRIEVAL MODEL

Cloud storage provides flexibility to data owners for easy access of any stored data. But to resolve the security issue and relevant content fetching this paper has proposed a encrypted search model. Proposed DRBAPSO (Document Retrieval by BAPSO) model has improved data security by storing data in encrypted form and searching of data was also done with query encryption. Proposed model explanation was done in two part first brief the document storing where Butterfly Leaping algorithm [10] cluster documents for storage on cloud. In second part document retrieval was done from stored data. Model flow was shown in fig. 1 and detail of each block was done in this section of paper.

Users: In this work three type of user do activity first is Admin, second is data owner and third is data user. Admin can create data owner and provide keys for encryption. Data owner Do can create its data user and upload document files in encrypted format. Data user Du can access data files of its data owner by passing a query.

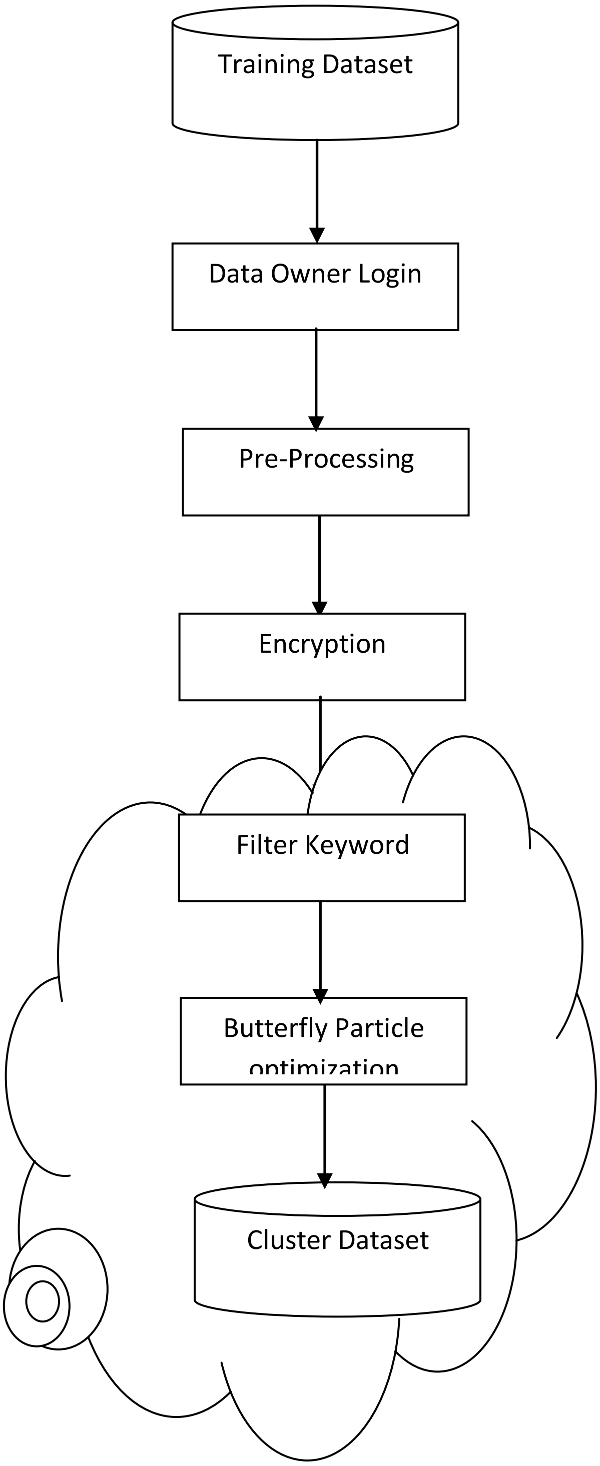

Fig. 1 Block diagram of DRBPSO document storing and clustering.

1.Pre-Processing: Text file was transform into set of word vectors Wv where each word is store in separate position. Document file may have different size of text word vector as per content. In this step space is remove from the Wv, as it was assume that all position word or special character have space separator. Each word was transform into its ASCII number range of 0 to 255. This transformation increase the security of content. Each row in the Am have 16 position to store character ASCII number, for short words position have default value 0. So set of Wv is now set of Am (ACCII matrix).

2.Encryption: Am of document file process separately for each row in matrix. Advanced Encryption algorithm [11] was used in the work for securing data from the intruders. This algorithm decrypts data without any losses. 16 digit row of Am pass in the algorithm for encryption. Output of AES is 16 digit different sequence numbers depends on encryption keys. So encrypted matrix Em is store on cloud for a data owner Do. A data owner can store any number of document files in encrypted format.

3.Filter Keyword: In order to improve the cloud performance retrieval relevancy need to be improved. For this clustering of encrypted data need to done at cloud side, this was performed by applying Butterfly Leaping Algorithm. Each row in Em is representing a word but some words are important act as keyword in the document, so as per frequency such words are filter. Stop words are identified by encrypted Sw dictionary. So rows having similar set of numbers are actually same word and counting of such word was done to get term frequency Tf [12].
Terms which have minimum number of counts are filter as keyword Fk from the document and rest words are not taken in butterfly leaping algorithm for document representation.

4.Butterfly Genetic Algorithm: This is a genetic algorithm proposed by [10, 13]. In this algorithm butterfly move towards nectar as per cognitive and social feature set values. This paper has done randomly select some of document for c number of cluster center. Output of this algorithm is cluster center representative documents.

5.Generate Butterfly: A Butterfly is set of random cluster center representative documents in the algorithm. Collection of butterfly is population in the genetic algorithm. This can be understand by let f number of butterflies present in population having c number of cluster. So Fp shown in eq. 1 is matrix of fxc. Selection of cluster representative document is done randomly by Gaussian distribution function.

Fp ←Generate_butterfly(f, c)

6.Fitness Function: Butterfly food searching ability was evaluate by fitness function. Distance between cluster center Fkc and non cluster center Fk were summed. Summation of this difference is fitness value of the butterfly. Eq. 2 shows the fitness evalution formula.

$$F_{v,f} = \sum_1^d Min_{c=1}^{c}(F_{kc} - F_k)\text{---------Eq. 2}$$

In above eq. d is number of documents for clustering.

7.Evaluate L-Best and G-Best

This step finds best chromosome from the population and fitness value of this best solution act as Local best and Global best value. Here it was obtained by evaluating the fitness value of each probable solution in the population. After this iteration of the algorithm starts where L-Best and G-Best update regularly.

8.Iteration Steps This involved calculation of Sensitivity of Butterfly by eq. 3 then cognitive values with constriction factor and inertia weight were evaluated using eq. 4 to 9 obtained from [65]. Here velocity and position of the butterfly also get update which is parameters of BA-PSO. So as per position matrix crossover is done to update population.

9.Sensitivity of Butterfly

$$S = e^{-(M_r - C_r)/M_r}\text{----3}$$

Where S is sensitivity of r^{th} iteration where M_r is maximum number of iterations takes place and C_r is current iteration of this BA-PSO algorithm.

10.Cognitive and Social parameters

$$C_1 = y * \left(\frac{C_r}{M_r} + x\right)\text{---------------4}$$

$$C_2 = x * \left(\frac{C_r}{M_r}\right)\text{--------------5}$$

Where x, y areconstant range between 0 to 1.

11.Constriction Factor C_{eq}

$$\alpha = C_1 + C_2$$

$$C_{eq} = 1 - \alpha - \sqrt{\alpha^2 - 4\alpha}\text{----6}$$

12.Inertia Weight

$$W_t = y + \frac{(M_r - C_r)}{M_r}\text{----- 7}$$

13.Update velocity V and position X of each probable solution

$$V_{i+1} = C_{eq} * (W_t * V_i + S \quad * (1 - P) * R * C_1 * (L_{best} - C_r) + P * R' * C_2 * (G_{best} - C_r)\text{-------------8}$$

$$X = R*P*V_{i+1}\text{-----9}$$

In above equation V is velocity, X is position while R and R′ are random number whose values range between 0-1. P is probability of nectar for the butterfly selection. So as per X and V values crossover operation were performed.

14.Crossover

This genetic algorithm is a hybrid combination of Butterfly and Particle swarm optimization. Here work has applied crossover operation as per butterfly sensitivity, congnitive and social feature values. In this work population P is updated as per X column wise and V values update P row wise. Change in column help to assign new position for the cluster center in same probable solution. While changes in row value as per L_{best}(where is L-best)solution increased the chance of generation of better fitness probable. Solution.

15.Update G-Best

After each iteration values of G-Best get optimized if new solution probable solution fitness function values are better than previous G-Best values. Hence if two iteration shows same values than iteration will break or if N number of iteration complete.

.

After t number of iteration of fitness function, butterfly movement and crossover operation butterfly algorithm gives final population. Best fitted butterfly is consider as final cluster center in the model.

16.Document Retrieval

Data user login to access its data owner document files. This paper has provides security of user query as well, as text query send to server in encrypted format. For encryption data owner keys were used. As per data user query each word is transform into 16 numeric values. These transformed or encrypted words were match with cluster center keywords to get the desired cluster data. Once cluster center highly matching keywords are found then files present in that cluster were list inn the index and pop out to the user. User can select any of index file and decrypt as per its requirement. This retrival increase data security, user privacy in all steps of data flow.

IV.EVALUATION PARAMETER

Experimental work was done on real dataset obtained from [21]. Implementation of proposed model was done on MATLAB software 2016a, having machine configuration of I3 processor 4GB RAM. Comparison of proposed DRBPSO model was done with existing model proposed in [22].

Results

Table 1 Precision value based comparison of document retrieval.

User Query	DRBPSO	FMRMS[15]
Query 1	0.8	0.375
Query 2	0.72	0.375
Query 3	0.933333	0.3125
Query 4	0.8	0.4375
Query 5	0.72	0.3125

Table 1 shows precision values of the document retrieval comparing models. It was obtained that proposed model butterfly particle swarm optimization based clustering of document has improved the work performance. This table shows that average precision value 54.38% was enhanced by the work.

Table 2 Recall value based comparison of document retrieval.

User Query	DRBPSO	FMRMS[15]
Query 1	0.888889	0.4
Query 2	0.782609	0.428571
Query 3	1	0.333333
Query 4	0.888889	0.466667
Query 5	0.782609	0.357143

Table 2 shows that proposed model DRBPSO has improved the retrieval relevancy of documents. It was obtained that proposed model clustering algorithm has increases the work performance as documents are arrange as per keyword similarity.

Table 3 F-measure value based comparison of document retrieval.

User Query	DRBPSO	FMRMS[22]
Query 1	0.842105	0.387097
Query 2	0.75	0.4
Query 3	0.965517	0.322581
Query 4	0.842105	0.451613
Query 5	0.75	0.333333

Table 3 shows that different query as per dataset text files has high retrieval value as compared to previous model FMRMS. This improvement is obtained by genetic based clustering.

Table 4 Accuracy value based comparison of document retrieval.

User Query	DRBPSO	FMRMS[15]
Query 1	0.842105	0.366667
Query 2	0.75	0.4
Query 3	0.964286	0.3
Query 4	0.842105	0.433333
Query 5	0.75	0.333333

Table 4 shows that accuracy of query results with actual output. It was obtained that average accuracy of proposed modle is 55.8% high as compared to previous model proposed in [22]. Storage of data improved the=is relevancy work.

Table 5 Execution time (seconds) based comparison of document retrieval.

User Query	DRBPSO	FMRMS[15]
Query 1	3.10317	5.55336
Query 2	2.07024	4.28152
Query 3	1.53882	3.29575
Query 4	2.05796	3.76114
Query 5	2.0798	3.23877

Execution time for the query based document retrieval is shown in table 5. It was obtained that proposed model DRBPSO takes less time as compared to previous model. As data arrange in cluster structure hence less comparison need for indexing the document file as per query.

V.CONCLUSION

This paper has proposed a secured document retrieval technique. Model has apply encryption algorithm on input file, so no party need to trust each other. Further paper has applied the file clustering algorithm particle swarm optimization with hybrid butterfly algorithm. This clustering increases the strength of retrieval of files. In order to increase the security of model paper has implement the user encryption as well. This encrypted query search relevant document in encrypted form, so cloud not aware of any information accept calculation. Experiment was done on real dataset and results were compared with other existing model. It was obtained that proposed model has increases the file retrieval relevancy. In future scholars can apply some block chain concept to increase the validity of data files.

REFERENCES

1. Rasool A, Tao R, Kamyab A (2020) GAWA - a feature selection method for hybrid sentiment classifcation. IEEE Access 8:191850–19186.
2. Shahid R, Javed ST, Zafar K (2017) Feature selection based classifcation of sentiment analysis using biogeography optimization algorithm. In: Proceedings of the international conference on innovations in electrical engineering and computational technologies.
3. Q. Liu, P. Hou, G. Wang, T. Peng, and S. Zhang, ''Intelligent route planning on large road networks with efficiency and privacy,'' J. Parallel Distrib. Comput., vol. 133, pp. 93–106, Nov. 2019.
4. H. Kumar, A. E. Manoli, and I. R. Hodgkinson, ''Sport participation: From policy, through facilities, to users' health, well-being, and social capital,'' Sport Manage. Rev., vol. 21, no. 5, pp. 549–562, 2018,
5. J. Kang, D. Steiert, D. Lin, and Y. Fu, ''MoveWithMe: Location privacy preservation for smartphone users,'' IEEE Trans. Inf. Forensics Security, vol. 15, pp. 711–724, 2020.
6. F. Abbas and H. Oh, ''A step towards user privacy while using locationbased services,'' J. Inf. Process. Syst., vol. 10, no. 4, pp. 618–627, 2014.
7. Pereira RB, Plastino A, Zadrozny B, Merschmann LHC (2018) Categorizing feature selection methods for multi-label classifcation. Artif Intell Rev 49(1):57–78.
8. E. K. F. Dang, R. W. P. Luk and J. Allan, "A Principled Approach Using Fuzzy Set Theory for Passage-Based Document Retrieval," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 7, pp. 1967-1977, July 2021.

9. M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood and M. T. Sadiq, "Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network," in IEEE Access, vol. 8, pp. 42689-42707, 2020.
10. Y. Ying, T. Qingping, X. Qinzheng, Z. Ping, and L. Panpan, ''A graphbased approach of automatic keyphrase extraction,'' Procedia Comput. Sci., vol. 107, pp. 248–255, 2017.
11. AGupta, I. Banerjee, and D. L. Rubin, ''Automatic information extraction from unstructured mammography reports using distributed semantics,'' J. Biomed. Informat., vol. 78, pp. 78–86, Feb. 2018.
12. [4] M. Moradi, ''CIBS: A biomedical text summarizer using topic-based sentence clustering,'' J. Biomed. Informat., vol. 88, pp. 53–61, Dec. 2018.
13. T. Uçkan and A. Karci, ''Extractive multi-document text summarization based on graph independent sets,'' Egyptian Informat. J., vol. 21, no. 3, pp. 145–157, Sep. 2020.
14. J. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, ''Experimental analysis of multiple criteria for extractive multi-document text summarization,'' Expert Syst. Appl., vol. 140, Feb. 2020, Art. no. 112904.
15. Z. Deng, F. Ma, R. Lan, W. Huang, and X. Luo, ''A two-stage Chinese text summarization algorithm using keyword information and adversarial learning,'' Neurocomputing, 2020.
16. H. Chiranjeevi and K. S. Manjula, "An Text Document Retrieval System for University Support Service on a High Performance Distributed Information System," 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2019.
17. *M. A. Alam et al., "Faster Image Compression Technique Based on LZW Algorithm Using GPU Parallel Processing," 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 2018, pp. 272-275*
18. Nishtha Mathur, Rajesh Bansode, AES Based Text Encryption Using 12 Rounds with Dynamic Key Selection, Procedia Computer Science, Volume 79, 2016.
19. Eusuff MM and K.E Lansey; Optimization of water distribution network design using SFLA(2003)
20. Xingyu Tao, Heng Li, Chao Mao, Chen Wang, Jeffrey Boon Hui Yap, Samad Sepasgozar, Sara Shirowzhan, Timothy Rose, "Developing Shuffled Butterfly-Leaping Algorithm (SFLA) Method to Solve Power Load-Constrained TCRTO Problems in Civil Engineering", Advances in Civil Engineering, vol. 2019.
21. https://ijsret.com/2017/12/14/computer-science/
22. Jiayi Li, Jianfeng Ma, Yinbin Miao, Ruikang Yang, Ximeng Liu, and Kim-Kwang Raymond Choo.. "Practical Multi-keyword Ranked Search with Access Control over Encrypted Cloud Data". IEEE TRANSACTIONS ON CLOUD COMPUTING, 2020.