JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2932 - 2937
https://publishoa.com
ISSN: 1309-3452

An Efficient Top-K Relevant Data Retrieval Using Catsort Algorithm

Namratha K H,
Dept of CSE (MCA)
Visvesvaraya Technological University,
Center for Postgraduate Studies, Mysuru
Email: namrathakoluttara@gmail.com

Dr Kumar P K
Dept of CSE (MCA)
Visvesvaraya Technological University,
Center for Postgraduate Studies, Mysuru
Email : pandralli@gmail.com

Nayana K
Dept of CSE (MCA)
Visvesvaraya Technological University,
Center for Postgraduate Studies, Mysuru
Email: nayanakgowda24@gmail.com

ABSTRACT
In the current era of internet world data act as an important source that holds many entity of information about the corporates, Financial Institutions, stock exchanges, IT firms, health organisations etc. The data extraction transforming the data into the required format unloading the data into the system is typically used to retrieve the information created by the organisation. The information act as an asset for the organisation to make for the improvements and enhancements of the business form and act as an asset. The driving the data from the massive storage is an important task. However the volume of data increases its it become difficult progressively to handle the data with current methods. The downside of employing the conventional approach motivated to create a robot model for returning top k search analysis. The proposed approach is focused on creating an efficient top-k relevant data retrieval using Catsort algorithm. From the massive Collection of data the top-K relevant data alone retrieved with reduced frame of time and accuracy of 90%.
Keywords— ***Retrieve data, Data extraction, Top k-retrieval.***

I. INTRODUCTION

Data retrieval is an important task for any kind of organisation. Loss of certain data can affect the development of the organisation in spite of losing information and recordings. The problem behind the massive data is retrieving the data from the database. In Spite of massive databases, acquiring the data is the process of identifying and extracting the relevant information based on a relevant query provided by the user or application. It enables the system to search over the complete database in order to find out the relevant data and display the data into the monitor or the display used inside the application[1]. In the current era of cloud computing numerous data are stored inside the memory and due to locally shifting the Enterprises to the cloud many data are shared over the cloud during communication. The high quality storage, increase of time to complete the process is required. Owners cannot completely relay on the real time monitoring of data. The service provider needs to provide the data protection. Data that need to be retrieved from the cloud during the deployment of cloud and privacy concerns a bigger issue[2].
The data encryption makes the data user interactable only after the encryption part is being completed. On the other hand, data need to be retrieved from the cloud the data need to be decoded. The cost required for storing the data into the cloud also expensive[3]. Downloading the data from the cloud also has particular steps of process and rapidly searches the encrypted target data, become a critical problem. To achieve ranked multi keyword search in massive data including the privacy preserving concern multi keyword search process is being proposed.

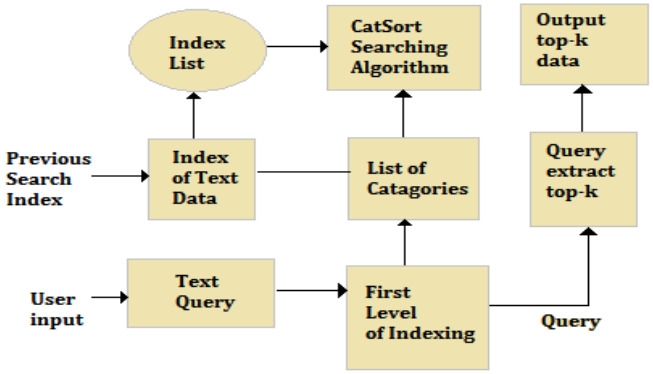

Fig. 1. General architecture of top-k retrieval model

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2932 - 2937
https://publishoa.com
ISSN: 1309-3452

Fig. 1 shows the general architecture of Top-k retrieval process. Frequently used method to secure the data in terms of privacy preserving is by providing a secure encryption and decryption process. The query act as an important factor to retrieve the data from the cloud effectively. To propose an efficient ranking process for multiple keyword search the data owner need to allow the multiple keyword access. To address the issue, the proposed system is evaluated. The top-k search practice provides the relevant top-k data alone will be displayed after the query has been posted. The launching of Keyword provided by the user need to be cypher text and trapped over the privacy sensitive keywords enters into the system to track the keywords. The clouds platform determines whether the person allowed to search the data is authenticated or not. Many existing systems provide remedy for making a multi keyword search that travel with privacy preserving for multiple data users impacted by keyword guessing attack. In order to avoid these kinds of problems randomisation process also evaluated. Every time to search the relevant data, the search process randomise the results and provide the output display in the screen. The same set of keywords is again randomized and further utilized for next search.

In order to search multi keyword based *top-k* relevant search process, the user need to be authenticated before making a search and further the search processing time need to be reduced. Using depth-first-search(DFS) algorithm and balanced binary tree(BBT)[4] algorithm the existing frameworks generate and display relevant results with a top k search data users. Data retrieval requires writing and execution of data and executes the commands recover to grab the information from the massive database. Based on the query provided, the database search for the relevant data in the database and display the relevant information alone in the display. User defined applications, Software platforms are generally used to retrieve the data from different formats[5].
In addition simple data retrieval techniques need to be focused to gather the information from the massive database without disturbing the current process. Data retrieval is a form of indexing the complete database to search for the relevant information. It is also called as reorganizing the database for a certain period of time to make the process complete. The database retrieval process consists of getting the query from the external source or user or application indexing the database based on the search query. The robust technique is required to display the top k search reliable retrieval process.

- The proposed system focused on CatSort algorithm to retrieve the top K relevant data retrieval from the massive database.
- The CatSort algorithm is nothing but a category based sorting algorithm that completely hold the massive database and categories the database into relevant data and further that fetch the relevant information then rank the data into top k process.
- The proposed approach considers a massive database of user information stored in the cloud. For the time taken to complete the process or the computation process is being reduced significantly comparing with the state of art approaches.

The rest of the paper is formulated as making detailed literature study in Section II. The system tool selection, problem identifications are discussed in Section III. The system architecture, detailed system design steps are discussed in Section IV. The rest of the paper is concluded with future enhancement.

II.BACKGROUND STUDY

C. Sha, et al., (2014) the author presented a cloud service retrieval problem from both text perspective and semantic aspect of cloud services with best answers. The presented approach considered the best query to make a relevant answer and balanced correlation between the content relevant to the topic asked. The service content and the service topic are considered to be the unique data on the asked question. The proposed approach created as a function and a sub modular function to search the process completely and make the algorithm to find the best answer using greedy search algorithm. TREC benchmark services are used to make the effective search process.

D. Wu et al., (2014) The author presented a new type of query based top-k search and textual clustering process that cluster the massive data into a number of sub blocks. The separated blocks make a top K keyword relevant request. the category based top K process completed then the final top k search process will be evaluated by comparing all the sub modules evaluated hence the number of search process increases level by level when making the cluster to be completed. The proposed approach find a good quality of clustering and capable of excellent performance in data search for that competition time will be little bit higher.

J. Yu et al., (2013) the author presented a system, addressing data privacy issues and discuss the top-k relevant search and their problem of privacy in the applications. the data acts as an important source and the privacy preserving the data

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2932 - 2937
https://publishoa.com
ISSN: 1309-3452

with a robust encryption method is important. In order to have a efficient data storage and encryption process the multi keyword retrieval process need homomorphic encryption. The proposed vector based search approach provides sufficient search accuracy and leads the system to rank over the complete database using cypher text based encryption. The Complete database needs to be decrypted while applying the search process. The presented system focus on security and performance analysis for the time taken to repack the complete database.

M. Rastegari et a., (2011) The author presented two retrieval scheme and compare the two schemes with each other in terms of efficiency and accuracy. The first experiment on classification model create a number of linear combinations of inputs to form small relevant data need to be combined to form the final search for the second process takes the retrieval scheme using ranking and vectorization method. In the ranking and vector formation method the results of the sub modules will be displayed only based on the ranks. The top K highest ranked data only displayed over the output Window for the classification of the retrieved data will be displayed.

A. Ghanbarpour et al., (2019) Presented a novels scoring function to optimize the accuracy of results provided by the search process. The presented approach focused on creating two levels of search algorithm to retrieve the top-k answers from the database for the answers need to be optimised before displaying into the final window. The results are further make a rank according to the Optimization results. Extensive experiments for conductor based on the evaluation of Framework Real time database are considered and efficiency of the database are displayed.

J. Sun et al., (2021) the author presented a novel approach on ranked multi keyword retrieval scheme to empower the cloud server system to protect the multi keyword based data. The *top-k* rank research process in terms of data leaking is discussed here. The proposed approach considers various forms of data that is used to secure the cloud storage and further show the Novelty and effectiveness of multi keyword search in the cloud and their defects in retrieving process. The efficiency of the system is for the shown by extracting the relative features of the data. Feature based extraction scheme is completely discuss to over in the presented approach and most relevant results are updated.

III. SYSTEM DESIGN

The top-K data retrieval process is identified as an important task in managing the massive database. To search the data in a faster way the system need to identify the relevant data from the database and make an indexing process to compute the data extraction from the database within a short span of time.

- In practice top K classification is typically performed by using many encryption and decryption process.
- Detection score and the metric score is determined by the number of times the correct classification has been implemented and for the cross correlation efficiency and correlation constant are helpful to determine the search accuracy.
- The proposed approach is focus on developing a robust top-k search process where the computational time is focused. The problem of searching time and effective display of time need to be focused here.

IV. METHODOLOGY

A. System architecture

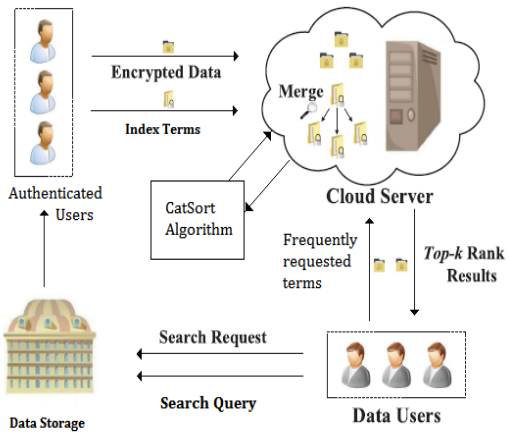

Fig. 2. System architecture of top-k retrival

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2932 - 2937
https://publishoa.com
ISSN: 1309-3452

Fig. 2. Figure shows the system architecture of top K retrieval process where the data is stored in the data authority or the agency having the complete database stored inside the memory. The data owners are the user provided with the application create a query to search the database. The data storage process takes encryption of data and storing the data into the cloud. Every time the query is applied to search the database further the database need to be encrypted before acting into the search query. The cloud server is controlled by the data user further maintained by the service provider. Every time the authenticated user access the cloud to search the database, the Cloud Service Provider provide a privacy preserving access to the user. Trustability and efficiency of the system need to be maintained in order to allow the data owner to enter to the cloud securely without affecting the feasibility and working process of the cloud.
The proposed approach considered these facts, the data categorised before applying the search process, the keyword search database indexing the history of search keywords. The implementation using category is being optimised and applied for searching process. The Optimisation process for reframe the indexing keywords while making search process is enahnced.
In terms of security, the goal is to provide relevance search without having any attack to diversify the data. The keywords cannot open the Trap door security and further the privacy preserving steps need to be improved. At this point the detailed process using CatSortr algorithm is explained below. The proposed approach consider the logical view of cloud data of a particular organisation is considered.

Algorithm

1. Get search query from user
2. Check authentication level
3. If USER = authentic then
Perform search process
Cluster DB/ N cluster
Get frequent query
Compare and fetch N relevant top-k
Move N top-k words to new DB
Apply CatSort (New_DB)
Get new_top_k data
Complete process.
End.
4. Back to Main_monitor
5. Display

The proposed algorithm work with two aspects of operations. The foremost work considers the raw database, decrypt the complete database and further apply the categorical sorting algorithm to cluster the whole database. The initial categorical process is the probabilistic random functional value of sorting the database.
Secondly the clustered database is further applied to CatSort process to make top-k relevant data. these data considers the frequently occurring query data from the secondary database connected with the search process.
The role of CatSort algorithm allow the system to search the requested query in short span of time. In the current context, if the relevant data found within the initial clusters then further the remaining database search process is suppressed.

V. RESULTS AND DISCUSSIONS

A. Word cloud

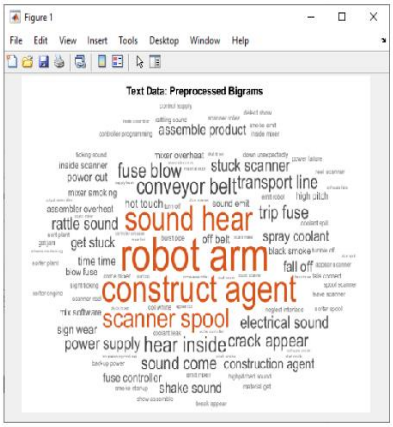

Fig. 3. Formation of wordcloud

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2932 - 2937
https://publishoa.com
ISSN: 1309-3452

Fig. 3. The figure shows the system results opening the word cloud formation of relevant data available in the complete database.

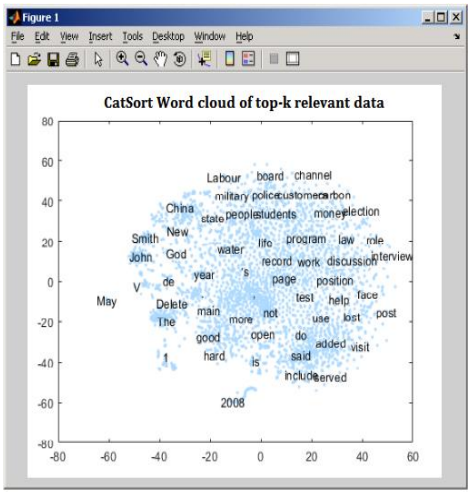

Fig. 4. CatSort word cloud of Top-K relevant data

The Fig. 4 shows there optimized results of relevant data displaying in the complete embedded word cloud after the search process.

Table 1. Computational Time of top-k query search with N iterations

Computational Time			
Iter 1	Iter 2	Iter 3	Iter 4
45.42	45.02	44.32	43.42
36.254	35.854	35.154	34.254
35.124	34.724	34.024	33.124
25.475	25.075	24.375	23.475
21.025	20.625	19.925	19.025
19.124	18.724	18.024	17.124
15.147	14.747	14.047	13.147

Table. 1 shows the computation time taken to complete the clustered search process of relevant data with respect to N number of iterations, for the whole database.

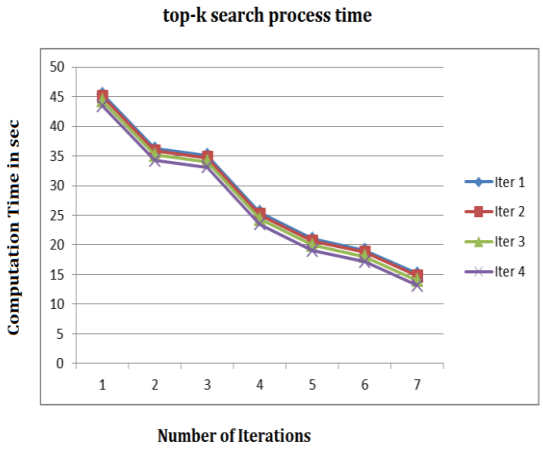

Fig. 5. Processing time of top-k relevant data search

The Fig. 5 shows the comparison of computational time with respect to the relevant search process and the separate categories of top-K retrieval process employed over the complete database.

The Major challenge persist with the implementation part is that during the clustered categorical division of data, the large categorical sorting provides similar data to get repeated. Hence removal of duplicate data is another task.
Separate function need to be implemented to remove the duplicate data.
Further the proposed approach can be improved by implementing artificial intelligence enabled data search process with systematic model for process evaluation is focused.

VI. CONCLUSION

Cloud Computing technology has been improved significantly for the effective usage of data handling and accessibility of data. Storing the data into the cloud securely and retrieving the data from the cloud is important. The proposed approach is focused on developing an efficient CatSort algorithm based top-K ranked retrieval process of data from the cloud. The scaleability and effectiveness of the search process improves in terms of computation time by making the category based search process. The two aspect of proposed approach is, first the category based search process is implemented, second the Optimization process of searched query and indexing of relevant query in the separate database. The proposed approach outperform the relevant top k search process and display the results within a short span of time and further the propose the approach need to be improved by enhancing more efficient algorithms and considering globally accessible massive data.

REFERENCES

[1] M. Armbrust, A. Fox, and R. Griffith, “A view of cloud computing,”
Commun. ACM, vol. 53, no. 4, pp. 50–58, 2010.
[2] H. Takabi, J. Joshi, and G. Ahn, “Security and privacy challenges in
cloud computing environments,” IEEE Secur. Privacy, vol. 6, pp. 24–31,
Nov./Dec. 2010.
[3] Y. Chang and M. Mitzenmacher, “Privacy preserving keyword searcheson remote encrypted data,” in Proc. Int. Conf. Appl. Cryptography Netw.Secur., 2005, pp. 442–455.
[4] C. Sha, K. Wang, K. Zhang, X. Wang and A. Zhou, "Diversifying Top-k Service Retrieval," 2014 IEEE International Conference on Services Computing, 2014, pp. 227-234, doi: 10.1109/SCC.2014.38.
[5] D. Wu et al., "Density-Based Top-K Spatial Textual Clusters Retrieval," in IEEE Transactions on Knowledge and Data Engineering, doi: 10.1109/TKDE.2021.3049785.
[6] J. Yu, P. Lu, Y. Zhu, G. Xue and M. Li, "Toward Secure Multikeyword Top-k Retrieval over Encrypted Cloud Data," in IEEE Transactions on Dependable and Secure Computing, vol. 10, no. 4, pp. 239-250, July-Aug. 2013, doi: 10.1109/TDSC.2013.9.
[7] M. Rastegari, C. Fang and L. Torresani, "Scalable object-class retrieval with approximate and top-k ranking," 2011 International Conference on Computer Vision, 2011, pp. 2659-2666, doi: 10.1109/ICCV.2011.6126556.
[8] A. Ghanbarpour and H. Naderi, "A Model-based Keyword Search Approach for Detecting Top-k Effective Answers," in The Computer Journal, vol. 62, no. 3, pp. 377-393, March 2019, doi: 10.1093/comjnl/bxy056.
[9] J. Sun, S. Hu, X. Nie and J. Walker, "Efficient Ranked Multi-Keyword Retrieval With Privacy Protection for Multiple Data Owners in Cloud Computing," in IEEE Systems Journal, vol. 14, no. 2, pp. 1728-1739, June 2020, doi: 10.1109/JSYST.2019.2933346.
[10] J. Sun, Y. Bao, X. Nie, and H. Xiong, “Attribute-hiding predicate encryption with equality test in cloud computing,” IEEE Access, vol. 6,
pp. 31621–31629, 2018.
[11] S. Sun et al., “An efficient non-interactive multi-client searchable encryption with support for boolean queries,” in Proc. Eur. Symp. Res. Comput. Secur., 2016, pp. 154–172.