# A Systematic Cross Modal Multimedia Retrival Using Maui Indexing Algorithm

R.Kannan, R.Umesh

***Abstract*****—** The data sets with multiple autonomous resources are popular in now a day. The big data plays a vital role in all the science and engineering departments such as physical, biological, biomedical science. Our main goal of this project is to reduce the storage space of the cloud when storing big data into the cloud server. In our proposed concept we are going to use the Huffman coding is used to effectively store the big data in the cloud server. The Huffman encoding algorithm is an optimal compression algorithm. Huffman encoding is an algorithm for the lossless compression of files based on the frequency of occurrence of a symbol in the file that is being compressed. In our proposed system the data with larger amount is compressed and it becomes a big data which is stored in a cloud server which reduces the data storage. The data is compressed and stored into the big data into the server. Our proposed system improves the scalability at the end of result. That means a process to handle a growing amount of work in a capable manner.

***Keywords:*** big data, optimal compression algorithm, Huffman encoding algorithm, data storage.

## I. INTRODUCTION

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

HACE Theorem. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

## II. LITERATURE SURVEY

1.R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012. This paper presents a new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real world applications show that the method is efficient and scalable. We presented an algorithm for finding all maximal non-redundant evolution paths of the induced relational states in a dynamic network. This can be used to discover the transitions of the conserved relational states over time and to better understand the cause of such changes in the stable patterns in a dynamic network

2. M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012. In this paper, we study the problem of crawl scheduling that biases crawl ordering toward important pages. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known PageRank as the importance metric. In order to score URLs, the proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We conduct a large-scale experiment using publicly available data sets to examine the effect of each feature on crawl ordering and evaluate the performance of many algorithms. The experimental results verify the efficacy of our schemes. In particular, compared with the representative Rank Mass crawler, the FPR-title-host algorithm reduces computational overhead by a factor as great as three in running time while improving effectiveness by 5 % in cumulative Page Rank.

3. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012. Identifying social influence in networks is critical to understanding how behaviors spread. We present a method that uses in vivo randomized experimentation to identify influence and susceptibility in networks while avoiding the biases inherent in traditional estimates of social contagion. Estimation in a representative sample of 1.3 million Face book users showed that younger users are more susceptible to influence than older users, men are more influential than women, women influence men more than they influence other women, and married individuals are the least susceptible to influence in the decision to adopt the product offered..

4.A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

R.Kannan, P.G Scholar, Department of Computer Science & Engineering, St.Michael college of Engineering. &Technology, Kalayarkoil, Sivanganga, India. ( Email: vrs.kannan@gmail.com)

R.Umesh, Assistant Professor , Department of Computer Science & Engineering, St. Michael College of Engineering. & Technology , Kalayarkoil, Sivaganga, India. ( Email: umesh.niz@gmail.com)

In this article, we describe some general themes in research in this area, with the aim of pointing out opportunities for students. Keeping with its interdisciplinary nature, we present perspectives from both computer science and statistical science, which are our two home departments. We begin by describing research into how to define and measure the risks of confidentiality breaches. We then describe some approaches to data protection. We end with a few general areas where students can engage in research.

5. S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

The rapid growth of the availability and popularity of interpersonal and behavior-rich resources such as blogs and other social media avenues, emerging opportunities and challenges arise as people now can, and do, actively use computational intelligence to seek out and understand the opinions of others. The study of collective behavior of individuals has implications to business intelligence, predictive analytics, customer relationship management, and examining online collective action as manifested by various flash mobs, the Arab Spring () and other such events. In this article, we introduce a nature-inspired theory to model collective behavior from the observed data on blogs using swarm intelligence, where the goal is to accurately model and predict the future behavior of a large population after observing their interactions during a training phase.

## III. EXISTING SYSTEM

In our existing system the general purpose parallel program method used a weighted linear regression. It proposed a HACF (Heterogeneous Autonomous Couple x Evolving relationship) theorem. The heterogeneous was used the different collector which uses different protocols to manage system for recording. Each data is able to generate and collect information without involving any centralized control in autonomous. The value of big data was increased in the complex and evolving relationship. The existing system found the best feature from the entire feature present in the data.

In our existing system the characteristics made it an extreme challenge for discovering useful knowledge from the Big Data. The existing work considered each individual as an independent entity without considering their social connections. That was one of the most important drawbacks of our existing system. The correlations between individuals inherently complicate the whole data representation and any reasoning process on the data. The most fundamental challenge for Big Data applications were to explore the large volumes of data and it requires the larger amount of storage space to store these big data.

Big data needs large amount of space for storing their data, so cloud service providers must spent more money for memory. They want to reduce the cost by reducing the memory space. Many researchers can implement some other technology to reduce the big data file size. But it is not in effective manner and also increase the system complexity. Decentralized control is one of the problems in existing system. Storing and retrieving of big data must be secure and flexible; the above existing theorem could not meet it.

## IV. PROPOSED SYSTEM

In our proposed system the big data is compressed and stored it into a cloud server. In our proposed work we are going to use the Huffman coding algorithm. The Huffman coding is used as the compression technique in our proposed system. This compression technique which reduces the storage space of the big data in the cloud. The Huffman encoding is an lossless compression technique. In our proposed system there is no any data loss during the compression process.

In our proposed system the data with larger amount is compressed and it becomes a big data which is stored in a cloud server which reduces the data storage. The data is compressed and stored into the big data into the server. Retrieval of stored files from the server we have to use k-means clustering algorithm. It provides more flexibility and scalability in the network. Our proposed system improves the scalability at the end of result. That means a process to handle a growing amount of work in a capable manner.

### *A. MODULES:*

- Authentication
- Compression
- Storage
- Retrieval Process

## V.MODULE DESCRIPTIONS:

### *A. AUTHENTICATION:*

The main process of our concept is to register the details of the users. Every user in the cloud can enroll their details. According to the registered information’s they have to move onto the next stage in the cloud. Here, we login to the cloud with the help of registered details. After the authentication process performed successfully we can select the files to upload into the cloud. In our process we have to choose different domains file for uploading. In this stage we protect the privacy by restricting the access to the data such as adding certification or access control to the data entries, so sensitive information access to the group of authenticated users. Second, we prevent the pinpointed sensitive information in individual data. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be mis-conducted by unauthorized individuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. One of the major benefits of the data annomization-based information sharing approaches is that, once anonymized, data can be freely shared across different parties without involving restrictive access controls. This naturally leads to another research area namely privacy preserving data mining, where multiple parties, each holding some sensitive data, are trying to achieve a common data mining goal without sharing any sensitive information inside the Data. We providing authentication the user have the rights to access, alter, and delete their own data.

### *B. COMPRESSION:*

Compression is the important process in big data application. As Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, we can always carry out mining activities at each distributed site. The rise of Big Data is driven by the rapid increasing of

complex data and their changes in volumes and in nature Generally, Big data means large volume of inter related date, that have been integrated by getting information from the different location or variable sources. We must reduce the storage space for this information by doing compression techniques. Compression involves encoding information using fewer bits than the original representation. Compression can be either lossy or lossless. Popular Haffman compression technology can be used for compressing the original message. The Haffman Lossless compression reduces bits by identifying and eliminating statistical redundancy. No information is lost in lossless compression. Lossy compression reduces bits by identifying unnecessary information and removing it. The process of reducing the size of a data file is popularly referred to as data compression, although its formal name is source coding. Using haffman encoding we can efficiently reduce the total file size and it need minimum storage space.

### *C. STORAGE:*

Storage is the third process of our concept. This will leads to the capacity of the particular location. Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from "quantity" to "quality". Here, we can use different location for storing the compressed data's. Different location means we have to choose different locations from the system and stored the information's which could be compressed using Huffman technique. The retrieval process can be done with the help of clustering. For clustering we choose the process of k-means clustering algorithm.

### *D.RETREIVAL PROCESS:*

Finally, we enter into the process of retrieval. For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data. To carry out Big Data mining, having an efficient and effective data access mechanism is vital, especially for users who intend to hire a third party (such as data miners or data auditors) to process their data. Under such a circumstance, users' privacy restrictions may include 1) no local data copies or downloading, 2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many other. Retrieving information from this large data environment is complex one. At the time of retrieval we have to perform the clustering process. Because already stored information contains different domain. Here, we can cluster or grouping the domain information's. For that purpose we have to use k-means clustering algorithm. This algorithm should properly cluster the files which we downloaded from the cloud. Here, the performance of the retrieved files can be evaluated. This process can reduce the cost and also reduce data loss.

The input design of an information system must the following objectives:

The input design of the system must attempt and try reducing the data requirements. It should also avoid capturing unnecessary data such as constant and system-computable data.

The input design must avoid processing delays during data entry. Capturing automatic data can reduce this kind of delay.

The input design must avoid data entry errors. This can be achieved by checking the errors in data entry program. This technique of checking data entry program programs for errors is known as input validation technique.

The input design must keep the process simple and easy to use.

## VI. CONCLUSION

In our proposed system we have to perform the bigdata applications. In those bigdata applications we have to use the concept of HACE. The purpose of that HACE is to choose multiple file in different are3a and also different domain. The process of big data applications contains map reduce. Here, we have to uploading the files into the cloud as in compressed manner. For compression process we have to use Huffman code compression technique. According to that compression technique the files has to e compressed. The reason for choosing this compression algorithm is to reduce the data loss at the time of compression. The compressed details are to be stored in multiple cloud servers. Next, we have to done the process of retrieval at the time of retrieval we can use the concept of map reduce. The retrieval process has to be done according to those srevers. Here for grouping the same domain files we can use k-means clustering algorithm. Using this k-means clustering algorithm we can group the retrieved files from the cloud. Finally, we have to retrieve the group of data from the cloud.

## REFERENCES

[1] K. Su, H. Huang, X. Wu, and S. Zhang, “A Logical Framework for Identifying Quality Knowledge from Different Data Sources,” Decision Support Systems, vol. 42, no. 3, pp. 1673-1683, 2006.

[2] “Twitter Blog, Dispatch from the Denver Debate,” http:// blog.twitter.com/2012/10/dispatch-from-denver-debate.html, Oct. 2012.

[3] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, “Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters,” Proc. Int’l Conf. Data Mining Workshops (ICDMW ’09), pp. 296-301, 2009.

[4] Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Secure Cloud Storage” IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[5] X. Wu and X. Zhu, “Mining with Noise Knowledge: Error-Aware Data Mining,” IEEE Trans. Systems, Man and Cybernetics, Part A, vol. 38, no. 4, pp. 917-932, July 2008.

[6] A.da Silva, R. Chiky, and G. He´brail, “A Clustering Approach forSampling Data Streams in Sensor Networks,” Knowledge andInformation Systems, vol. 32, no. 1, pp. 1-23, July 2012.

[7] W. Liu and T. Wang, “Online Active Multi-Field Learning for Efficient Email Spam Filtering,” Knowledge and Information Systems, vol. 33, no. 1, pp. 117-136, Oct. 2012.

[8] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman,“Shoroud: Ensuring Private Access to Large-Scale Data in the Datacenter,” Proc. 11th USENIX Conf. File and Storage Technologies (FAST ’13), 2013.

[9] D. Luo, C. Ding, and H. Huang, “Parallelization with Multiplicative Algorithms for Big Data Mining,” Proc. IEEE 12th Int’l Conf. Data Mining, pp. 489-498, 2012.

[10] J. Mervis, “U.S. Science Policy: Agencies Rally to Tackle Big Data,” Science, vol. 336, no. 6077, p. 22, 2012.

[11] F. Michel, “How Many Photos Are Uploaded to Flickr Every Day and Month?” http://www.flickr.com /photos/ franck michel/ 6855169886/, 2012.

[12] T. Mitchell, “Mining our Reality,” Science, vol. 326, pp. 1644-1645, 2009.

[13] Nature Editorial, “Community Cleverness Required,” Nature, vol. 455, no. 7209, p. 1, Sept. 2008.

[14] S. Papadimitriou and J. Sun, “Disco: Distributed Co-Clustering with Map-Reduce: A Case Study Towards Petabyte-Scale End-to-End Mining,” Proc. IEEE Eighth Int’l Conf. Data Mining (ICDM ’08), pp. 512-521, 2008.

[15] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, “Evaluating MapReduce for Multi-Core and Multiprocessor Systems,” Proc. IEEE 13th Int’l Symp. High Performance Computer Architecture (HPCA ’07), pp. 13-24, 2007.