International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

# Indexing and Optimization Techniques in Biomedical Industry

**Gladiss Merlin N.R, Vigilson Prem**

***Abstract*: The unpredictable amount of data generated everyday by smart phones, social networks, health care systems etc. is really mind blowing. Smart phones alone generate 335exabytes of data ineveryyear that is really big data.Thus, the storage industry is facing several challenges in providing high magnitude of storage and retrieval devices at lowest costs which help to fulfill the requirements of big data and even technologies like de-duplication on storage devices are also becoming very important. Similarly, in recent days storing and retrieving the health care information in biomedical area is also becoming a great challenge in providing the best optimum data because of its huge amount of biomedical datasets. In order to achieve efficiency in providing highest quality health care information, an optimized index scheme is needed for big data which is based on accuracy and timelines. The existing indexing and optimization solutions are not enough to meet the emerging grow of index size and seek time. The objective of this paper is to identify better indexing solutions by investigating the basic big data requirements on indexing and optimization. This also includes a comparative study of various indexing and optimization techniques along with a taxonomy which contains Artificial Intelligence (AI) and Non Artificial Intelligence (NAI) based indexing techniques, optimization enhancement techniques which improves the performance efficiency of big data health care informatics.**

***Index Terms*: *Indexing-Optimization-BigData-Artificial Intelligence***

## I. INTRODUCTION

In recent days, decision making plays major role in health care. Health care includes enormous amount of patient information, where the traditional or existing searching and data retrieving techniques are incapable to produce the responses quickly. Due to this drawback, producing timely responses for a request plays a major role in decision making. Therefore new, fast and effective data analytics solutions are required to compete with rapidly growing Exabyte's of data. Also, performing search on big data repositories is a challenge task. For big data analysis on cloud, efficient indexing techniques should be designed.

Revised Manuscript Received on August 30, 2019.

* Correspondence Author

**Gladiss Merlin N.R*,** Assistant Professor, CSE Department, Jeppiaar Institute of Technology. Chennai, India.

**Dr. Vigilson Prem,** Professor, CSE Department, RMK College ofEngineering and Technology. Chennai, India.



The existing techniques are not up to meet the current big data analytics requirements, since the data volume has exceeded Exabyte's and now it is crossing from terabytes to Petabyte's. So, indexing becomes important in performing search on big repositories of data. It is impossible to perform manual data retrieval on a high volume and complex dataset. Instead, we go with efficient indexing techniques to access big data. The indexing techniques need to satisfy the big data requirements like volume, velocity, veracity, value, variety, variability and complexity. Researchers have used different indexing techniques on big data. For example, to improve query execution and the performance of searching[4], big data coupled with elastic search technology to satisfy the daily health care needs[5], a model with block creation module, index creation module and query creation[6], R-tree based indexing to support cloud with multi dimension data indexing, Data Med for searching biomedical datasets across repositories[7]. Similar to indexing,optimization is also one of searching techniques in big data for finding the best optimum. Researchers have used various optimization techniques. For example, PESM measure is use in retrieving the patient information[1] in case-based reasoning using similarity measure and adaptive fractional brain storm optimization, population based optimization technique is used in looking for the optima of optimization problems[2], biogeography-based optimization(BBO) is used which works under some natural phenomena, Jayaoptimization algorithm[JOA] is used for constrained and unconstrained optimal problems[3], genetic algorithm(GA) is used in optimization which works under chromosomes and genes[2]. This research includes a comparative study of different indexing and optimization techniques along with a taxonomy which includes Artificial Intelligence (AI), Non-Artificial Intelligence(NAI), and Collaborative Artificial Intelligence(CAI) based indexing techniques. This also includes the requirements of big data indexing techniques and the optimization enhancement techniques.The remaining portion of this paper is organized as follows. Section 2 analyses the existing research works on big data indexing along with taxonomy and requirements study. Section 3 analyses the existing research works done on big data optimization enhancement techniques which involves some of the applications of optimization. Section 4 includes a detailed survey on optimization and indexing in biomedical research. Some of the optimization process is tabulated along with its techniques and applications. Similarly, indexing techniques are also investigated and tabulated along with their techniques and applications. Section 5 states the conclusion which provides the significance of this survey.

## II. METHODOLOGY

The existing indexing techniques are categorized as structured in Figure 1. This organization provides a detailed view of the existing categories of indexing techniques, which is classified in to three categories [8] such as non-artificial intelligence (NAI), artificial intelligence (AI) and collaborative artificial intelligence (CAI). For efficient and fast data retrieval, NAI is developed. They comprise techniques like bitmap, hashing, B-tree and R-tree. B-tree indexing technique deals with big datasets in terms of number of objects. This technique was applied on temporal data, which allows changing values frequently. Additional to variability and volume, value is also included in B-tree indexing for easiest implementation. Bitmap indexing techniques include two datasets to perform evaluation test. This also cope up with frequent changes in data volumes and so it is more scalable. This method satisfies the big data requirements like volume, velocity, variability and complexity.

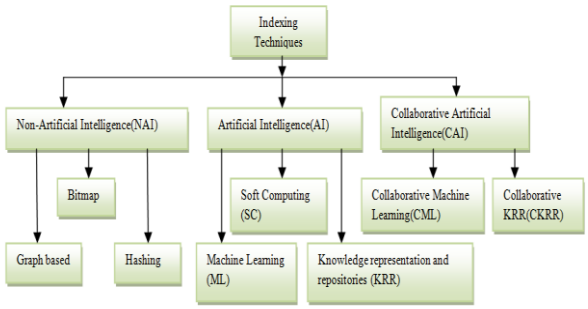


**Figure 1**: Indexing techniques taxonomy

Hash based indexing is mainly used in multimedia searches in designing sparse index solution for different kinds of data like audio, optical characters can be achieved by this method. Requirements like velocity and volume is satisfied by this method. AI utilize a knowledge base for efficient data retrieval and this is technical oriented which also apply indexing on moving objects. They comprise techniques like machine learning(ML), knowledge representation andreasoning(KRR) and soft computing(SC). Using ML, video searching index can be designed to achieve improved accuracy. Data variety can be supported by this method. KRR involves semantic method which improves the performance of large datasets.

This method satisfies the big data requirements like data variability, veracity, variety and value. CAI helps in increasing accuracy and search effectiveness. They comprise techniques like collaborative machine learning (CML) and collaborative knowledge representation (CKRR). The collaborative machine learning method allows an interpretation-based indexing which is a social learning model for large datasets. This method satisfies the big data requirement variety. The collaborative knowledge-based representation is mainly used in improving the performance of digital music data. This method achieves high fault tolerance and robustness.

The big data optimization enhancement techniques are structured in Figure 2. Process capability enhancement:This enhancement technique is used in application level optimization [9] to improve the performance of data transmission over pipelining and also supports parallelism and concurrency control. Another one application is evolutionary optimization [10], which proposes genetic operators and focuses on high dimensional optimization problems which even works on complex solution space. Memory management enhancement: This enhancement technique is used in In-memory big data optimization, which shows a detailed examination of the required technology related to memory management [11] along with the related works and also supports all the memory operations to be more effective in planning and execution.

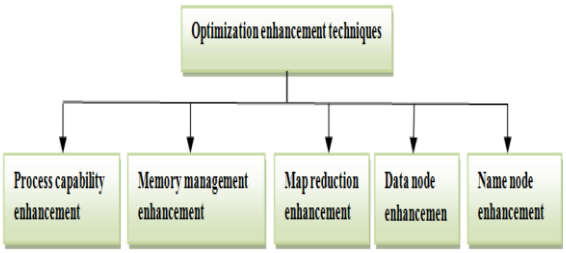


**Figure 2**: Big data optimization enhancement techniques

Map reduction enhancement: This enhancement technique is applied in platform big data optimization, which specifically uses data size and time as parameters in selecting a platform for an application. In order to attain performance efficiency, before selecting a platform [12] for an application one should gain enough knowledge about all the available existing platforms. Data node enhancement: This enhancement technique supports multi objective optimization in big data particularly used in knowledge discovery [13]. A multi objective optimizer starts with a set of entities, executes them iteratively and finally ends with an effective performance target or estimated output.Name node enhancement: This enhancement technique supports small files optimization in big data. It involves four stages, in the first stage several small files are merged, secondly the structurally related small files are prefetched, thirdly the files are grouped and finally the logically related small files are prefetched.

## III. RESULT AND DISCUSSION

Fractional brain storm optimization is used for case retrieval [1]. Here, a similarity measure is proposed known as PESM measure, which is used in retrieving the patient information. In PESM, two patient cases can be matched with four distinguished parameters with the help of probability methods which calculates the occurrence and non-occurrences. Adaptive fractional brain storm optimization (AFBSO) algorithm is a modified brain storm algorithm which replaces the mathematical theory with a fractional calculus (FC) in order to improve searching in the search space as well as for improved utilization of large datasets. The similar patient cases for the input query can be easily captured and retrieved by integratingthe PESM with AFBSO neural network. PESM measure allows accessing the similar patient cases stored and the neural network access the neighbor index value. Finally, the outputs of both PESM measure and neural network are combined which ease the identification of the similar patient's health diagnosis.

An algorithm called novel-based optimization is proposed [2] which applies sine cosine algorithm (SCA) initially and starts the process of optimization with a set of random solutions. The exploration of search space is done every time when the SCA function returns a value >1 or <-1.

The exploitation of search space is done every time when the SCA function returns value between -1 and 1 which is then pointed as promising regions. The sine and cosine function use adaptive range transformation for the SCA algorithm to move smoothly in the search space from exploration to exploitation. As a result, the global optimum is declared as the destination point. During optimization, the best region of the search space found so far is updated as the solution. A new optimization algorithm called Jaya [3] is proposed which works with a smaller number of parameters. This algorithm always focuses on the best solution ignoring the worst solution. It is named 'Jaya' because; it always tries for the victory by attaining the best solution. In this paper, a well-defined set of 24 constrained optimization problems are identified and the proposed Jaya algorithm is implemented on it. The obtained results are compared with the existing optimization algorithms, which results in improved performance efficiency. Particles worm optimization (PSO) [14] which involves image segmentation techniques is proposed with Fuzzy C-means (FCM). FCM is a clustering algorithm which can be applied on complex data sets. In health care, this clustering algorithm is applied in brain MR image segmentation. Also, an improved kernel possibilistic c-means algorithm (IKPCM) is proposed. Below Table [1] briefly list out some optimization processes used in big data analytics.

**Table 1** Optimization survey

| Author | Optimization Technique | Application |
|---|---|---|
| Joel Wolf et.al [24] | File allocation scheduler | Job scheduling |
| Bo Dong et.al [25] | Merging of several files, Prefetching for structurally related small files, Grouping of files, Perfecting for logically related small files | Small file detection |
| Esma Yildirim et.al [26] | Recursive chunk division(RCD) for optimal pipelining, optimal parallelism- concurrency pipelining(PCP) | Map reduce optimization |
| Kostas kolamvatsos et.al[27] | Two prototypes are used. One is centered on a finite horizon time optimized model and the second one is built on an infinite horizon optimally scheduled model | Improve the performance of querying big data clusters. |

A tree based indexing technique is used in performing indexing on shapes [15] with optimal embedding in medical image databases. It is a shape embedding procedure to retrieve the similarities between the complete and partially complete shape. K-tree based indexing technique is used for the study of diabetic health care system [16] along with advanced database technologies. This indexing method effectively processes the reverse k-nearest neighbors (RKNN) queries or requests. Hash based indexing data structure technique [17] is used in cloud-based health care monitoring to identify the lost data in body sensor networks. Here, the Merkle hash tree is used in detecting the lost data. Hash based indexing data structure technique is used in cloud based generic health data sharing and also to develop a secure monitoring platform [18]. Here, a hash function is used to design a system which provides rapid health care facilities and data downloads. In [19], a three-level indexing hierarchy is used in collaborative telemedicine applications as a provision of smart playback functions with a novel indexing architecture. Self-learning indexing technique [20] is applied for supporting context aware health care applications in a probabilistic ontology-based platform. In [21], a phrase-based indexing technique is used in retrieving medical text documents which is a scenario specific and a knowledge-based method. A collaborative learning technique is proposed [22] to develop a collaborative filtering based medical knowledge recommendation system so that clinicians can retrieve trust based accurate knowledge. A knowledge management technique [23] is proposed to present a method for reconstituting a medical ontology by translating a medical database in to RDF language in the context of a health care network. A virtual staff is developed where a greater number of health care members are involved for better diagnosis. The below Table[2] list a set of indexing techniques in health care big data analytics.

**Table 2** Indexing survey

| Author | Indexing Technique | Application |
|---|---|---|
| Qian et.al.[15] | Applying an optimal shape embedding method for retrieving replicas using tree based indexing structures | Medical image databases |
| Hsu et.al[16] | Efficient query processing with reverse k-nearest neighbors(RKNN) algorithm using a K-tree indexing | Diabetic health care system |
| Ali et.al[17] | Detecting lossy data in body sensor networks using a Merkle hash tree indexing | Health care monitoring |
| Thilakanathanet.al [18] | Rapidly downloading data for sharing information using a hash function indexing | Health data sharing in Cloud |
| Wang et.al[19] | A novel indexing architecture is presented with a three-level indexing hierarchy | Telemedicine |
| Ongenae et al[20] | Probabilistic, self-learning, ontology based frame work | Health care context aware applications |
| Chu et.al[21] | Scenario specific data retrieval using phrase-based indexing | Medical text documentation |
| Huang et.al[22] | Filtering based collaborative learning indexing | Medical knowledge(semantic) system |

| | | |
|---|---|---|
| Dieng Kuntz et.al[23] | Translating amedical database in to RDF language | Medical ontology |

## IV. CONCLUSION

This paper surveyed different indexing and optimization techniques on biomedical area using big data tools. The uniqueness of this paper is, it includes a tabulation which displays a concise outcome of existing research methodologies. The table also instructs the previously used algorithms along with their applications on real time, which can be utilized or applied for the future improvements and development of big data requirements. Investigation of the most important research issues and challenges are also highlighted. Importance was given to process capability enhancement and collaborative artificial intelligence technique with respect to optimization and indexing. The examined methods result in improving the utilization, performance efficiency and data retrieval in big data analytics.

## REFERENCES


1. P. Yadav, "Case Retrieval Algorithm Using Similarity Measure and Adaptive Fractional Brain Storm Optimization for Health Informaticians," *Arab. J. Sci. Eng.*, vol. 41, no. 3, pp. 829–840, 2016.
2. S. Mirjalili, "SCA: A Sine Cosine Algorithm for solving optimization problems," *Knowledge-Based Syst.*, vol. 96, pp. 120–133, 2016.
3. R. Venkata Rao, "Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems," *Int. J. Ind. Eng. Comput.*, vol. 7, no. 1, pp. 19–34, 2016.
4. Siddiqa, A. Karim, and V. Chang, "SmallClient for big data: an indexing framework towards fast data retrieval," *Cluster Comput.*, vol. 20, no. 2, 2017.
5. D. Chen *et al.*, "Real-time or near real-time persisting daily healthcare data into HDFS and elasticsearch index inside a big data platform," *IEEE Trans. Ind. Informatics*, vol. 13, no. 2, pp. 595–606, 2017.
6. Siddiqa, A. Karim, and V. Chang, "Modeling SmallClient indexing framework for big data analytics," *J. Supercomput.*, pp. 1–22, 2017.
7. X. Chen *et al.*, "DataMed - an open source discovery index for finding biomedical datasets," *J. Am. Med. Informatics Assoc.*, vol. 25, no. 3, pp. 300–308, 2018.
8. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowl. Inf. Syst.*, vol. 46, no. 2, pp. 241–284, 2016.
9. Yildirim, E., Arslan, E., Kim, J., and Kosar, T. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. IEEE Transactions on Cloud Computing 4, 1 (2016), 63-75.
10. Bhattacharya, M., Islam, R., and Abawajy, J. Evolutionary optimization: a big data perspective. Journal of network and computer applications 59 (2016), 416-426.
11. Zhang, H., Chen, G., Ooi, B. C., Tan, K.-L., and Zhang, M. In-memory big data management and processing: A survey. IEEE Transactions on Knowledge and Data Engineering 27, 7 (2015), 1920-1948.
12. Singh, D., and Reddy, C. K. A survey on platforms for big data analytics. Journal of Big Data 2, 1 (2015), 8.
13. Bandaru, S., Ng, A. H., and Deb, K. Data mining methods for knowledge discovery in multi-objective optimization: Part a-survey. Expert Systems with Applications 70 (2017), 139-159.
14. Mekhmoukh and K. Mokrani, "Improved Fuzzy C-Means based Particle Swarm Optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 266–281, 2015.
15. Qian X, Tagare HD, Fulbright RK, Long R, Antani S (2010) Optimal embedding for shape indexing inmedical image databases. Med Image Anal 14(3):243–254. doi:10.1016/j.media.2010.01.001
16. Hsu W, Lee ML, Ooi BC, Mohanty PK, Teo KL, Xia C (2002) Advanced database technologies in adiabetic healthcare system. Paper presented at the proceedings of the 28th international conference onvery large data bases, Hong Kong, China
17. Ali ST, SivaramanV,OstryD(2013) Authentication of lossy data in body-sensor networks for cloud-basedhealthcare monitoring. Future GenerComput Syst 35:80–90. doi:10.1016/j.future.2013.09.007
18. Thilakanathan D, Chen S, Nepal S, Calvo R, Alem L (2013) A platform for secure monitoring and sharingof generic health data in the Cloud. Future GenerComput Syst 35:102–113. doi:10.1016/j.future.2013.09.011
19. Wang C-H, Jiau HC, Chung P-C, Ssu K-F, Yang T-L, Tsai F-J (2010) A novel indexing architecture forthe provision of smart playback functions in collaborative telemedicine applications. Comput Biol Med40(2):138–148
20. Ongenae F, Claeys M, Dupont T, Kerckhove W, Verhoeve P, Dhaene T, De Turck F (2013) A probabilisticontology-based platform for self-learning context-aware healthcare applications. Expert Syst Appl40(18):7629–7646. doi:10.1016/j.eswa.2013.07.038
21. Partha Protim Das Priyank, Kumar Ghadai, M. Ramachandran, Kanak Kalita, Optimization of Turning Process Parameters by Taguchi-Based Six Sigma, Mechanics and Mechanical Engineering, 21(3):649-656, 2017.
22. Huang Z, Lu X, Duan H, Zhao C (2012) Collaboration-based medical knowledge recommendation. ArtifIntell Med 55(1):13–24
23. Dieng-Kuntz R, Minier D, R°uži˘cka M, Corby F, Corby O, Alamarguy L (2006) Building and using amedical ontology for knowledge management and cooperative work in a health care network. ComputBiol Med 36(7–8):871–892. doi:10.1016/j.compbiomed.2005.04.015
24. Wolf, J., Rajan, D., Hildrum, K., Khandekar, R., Kumar, V., Parekh, S., Wu, K.-L., et al. Flex: A slot allocation scheduling optimizer for mapreduce workloads. In Proceedings of the ACM/IFIP/USENIX 11th International Conference on Middleware (2010), Springer-Verlag, pp. 1-20.
25. Dong, B., Zheng, Q., Tian, F., Chao, K.-M., Ma, R., and Anane, R. An optimized approach for storing and accessing small files on cloud storage. Journal of Network and Computer Applications 35, 6 (2012), 1847-1862.
26. Yildirim, E., Arslan, E., Kim, J., and Kosar, T. Application-level optimization of big data transfers through pipelining, parallelism and concurrency. IEEE Transactions on Cloud Computing 4, 1 (2016), 63-75.
27. Kolomvatsos, K., Anagnostopoulos, C., and Hadjiefthymiades, S. An efficient time optimized scheme for progressive analytics in big data. Big Data Research 2, 4 (2015), 155-165.


## AUTHORS PROFILE

**Mrs. N.R. Gladiss Merlin**, Currently she is working as an Associate Professor in the Department of Computer Science and Engineering in Jeppiaar Institute of Technology, Chennai. India. She has presented various papers in various national and international conference.

**Dr. M. Vigilson Prem,** Currently he is working as a Professor in the Department of Computer Science and Engineering in RMK College of Engineering and Technology. Chennai, India. He has presented various papers in various national and international conference. he has session the chair in various national and international conferences