IJCTA, 10(9), 2017, pp. 619-624
© International Science Press

ISSN: 0974-5572

Survey on Big Data Indexing Techniques

Vaishanvi Gawande* and Ambika Pawar**

ABSTRACT

In this era of Internet of Things, Social Media, Mobile and Cloud Computing the data generation rate has gone to such an extent that has led to phrase data as '*Big Data'*. *Big data* has distinguishing characteristics like volume, variety, velocity and veracity which has created demand for new techniques to deal with *big data*. Traditional techniques which were designed to store, handle, process and query data are not efficient to store and manage big data. Thus big data has opened doors for research on design of efficient big data management techniques. This paper targets indexing techniques for big data and covers existing work on big data indexing techniques. This paper also discusses in detail about the hash indexing concepts and future extensions to improve the efficiency.

Keywords: Big Data, Big Data Management, Efficient Query Processing,Indexing Techniques, Hashing, Sparse Hashing

1. INTRODUCTION

Big Data is the term used to describe tremendous amount of data which is generated at very high speed. Big Data can be defined with the help of 3Vs that is Volume (large amount of data), Velocity (speed at which data is generated) and Variety (different types of data formats, semantics, uncertain data etc.). Later studies have pointed out that the definition of 3Vs is insufficient to explain the big data. Thus, veracity (defines abnormalities in data), validity (defines quality of data, heterogeneous nature of data), variability (refers to seasonal, customer dependent, dynamic type of data), venue (from where data is generated) and vocabulary (refers to schema, data model, semantics, taxonomies of the data) are added to make some complement explanation of big data [2]. The tremendous growth in volume, velocity, and variety of data produced by all devices like social sites, trading sites, OnLine Analytical Processing (OLAP) and cloud applications are generating big data. Available solutions for efficient data storage and management cannot fulfil the needs of big data as the amount of data is continuously increasing. Due to its distinguishing characteristics, big data need to have rethinking and redesign on existing techniques like indexing.

1.1. Introduction to Big Data Indexing

Indexing is the technique which improves the speed of the data retrieval operation on a database. The main purpose of having an index is to speed up search queries.

Basically there are three types indexing techniques (Figure 1):-

1) Artificial Intelligence Approach
2) Non-Artificial Intelligence Approach
3) Collaborative Artificial Intelligence Approach

1.2. Artificial intelligence indexing

This technique has an ability to detect unknown behaviour in Big Data and to establish relationships between data items by observing patterns and categorizing items for that it uses knowledge base. This technique

* Department of Computer Science, Symbiosis Institute of Technology, Pune (MH), India, *Emails: vaishanvi.gawande@sitpune.edu.in, ambikap@sitpune.edu.in*

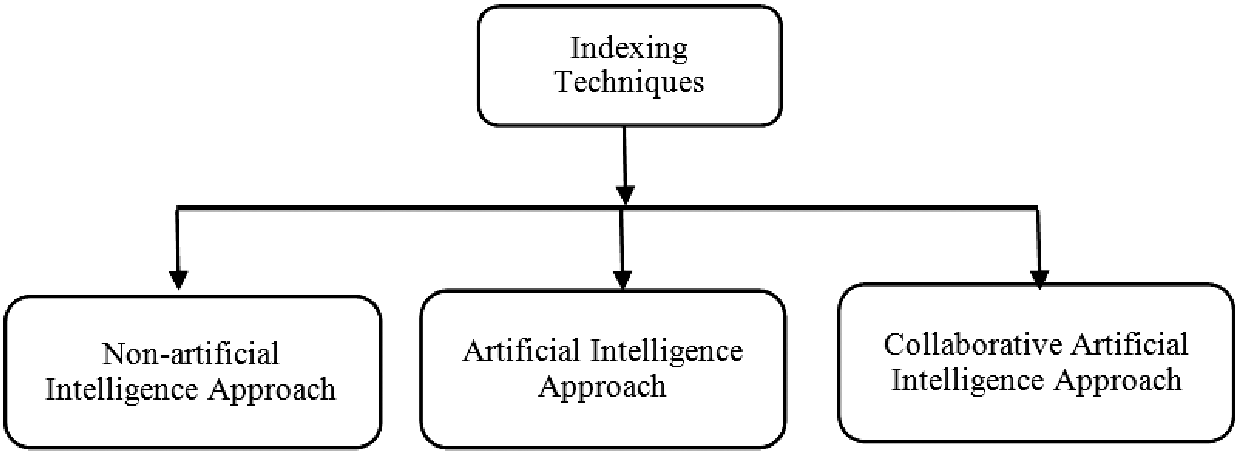
Indexing Techniques
Non-artificial Intelligence Approach
Artificial Intelligence Approach
Collaborative Artificial Intelligence Approach

Figure 1: Indexing Techniques

takes more time for information retrieval as data is continuously changing which requires frequent updates moreover it gives accurate results.

This technique is further sub-divided into two techniques as shown in figure 2 are as follows 1. Latent Semantics 2. Hidden Markov Model

1.2.1. Latent Semantic Indexing [LSI]

Latent Semantic Indexing is an indexing technique that identifies relationship between texts. The main characteristic of this technique is to extract the conceptual content of data sets and to establish relationships between terms with similar contexts. This technique supports keyword queries.Challenges faced by this technique are scalability and performance. LSI strategy demands very high computational performance as well as memory to index Big Data.

1.2.2. Hidden Markov Model [HMM]

Hidden Markov Model indexing approach is method which is developed from the Markov model. Markov model is consists of states which are associated by transitions, where future states are completely dependent on present state and independent on historical states. In this technique query results are generally predictions of future states of an item, based on thecurrent state. The present state is used to predict the future states using the dependent data which increases a good performance.

1.3. Non-Artificial Intelligence indexing

In this approach, the formation of indexes does not depend on the meaning of the data item or the relationship between texts. Preferably, indexes are formed based on how frequently the items are queried or searched in a particular data set. This approach covers most of the characteristics of Big Data like volume, velocity, variety, value, variability and complexity but they cannot detect unknown behaviour of Big Data. This approach is mostly used for fast and efficient retrieval.This technique is further subdivided into four types as shown in figure 2 are as follows.

1. Tree Based Indexing: In the Tree based indexing technique, retrieval of data is done in a sorted order which satisfies nearest neighbour queries.

   a. B Tree: B-tree works like the binary search tree but in complex manner because the nodes of B-tree have more branches than binary tree which has two branches per node. So a B-tree is more complicated than a binary tree .B-tree indexes satisfy range queries and similarity queries. This technique works only on two dimensional data.

b. R Tree: This indexing strategy is used for spatial queries. Mostly it is applied in geospatial systems with each entry having X and Y coordinates with minimum and maximum values. Advantage of using an R-tree over a B-tree is that, the R-tree satisfies multi-dimensionalor range queries.

2. Hash Indexing: Hash Indexing Works on equality operator. They find the most similar data items from data set. Hash Indexing is significant for retrieval of multidimensional data.Depending on various applications Hash Indexing technique is classified into 3 sub types as follows.

   a. Supervised Hash Indexing: This technique works on labelled data (any data having label). Supervised hash Indexing technique is more flexible for real world applications.

   b. Unsupervised Hash Indexing: This technique works on unlabelled data for similarity search for fast Retrieval of multidimensional data.

   c. Semi-supervised Hash Indexing: This technique is the combination of supervised and unsupervised technique to reduce the quantization loss.

3. Custom Indexing: Custom indexing technique is based on arbitrary or user defined indices.They generally based on indexing strategies such as B-tree, R-tree, inverted index, and hash indexing strategy.

4. Inverted Indexing: An inverted index is built on a list of all unique words which appear in documents, and a list of documents in which each word appears. Multipledocuments can have the same key as index in inverted index.

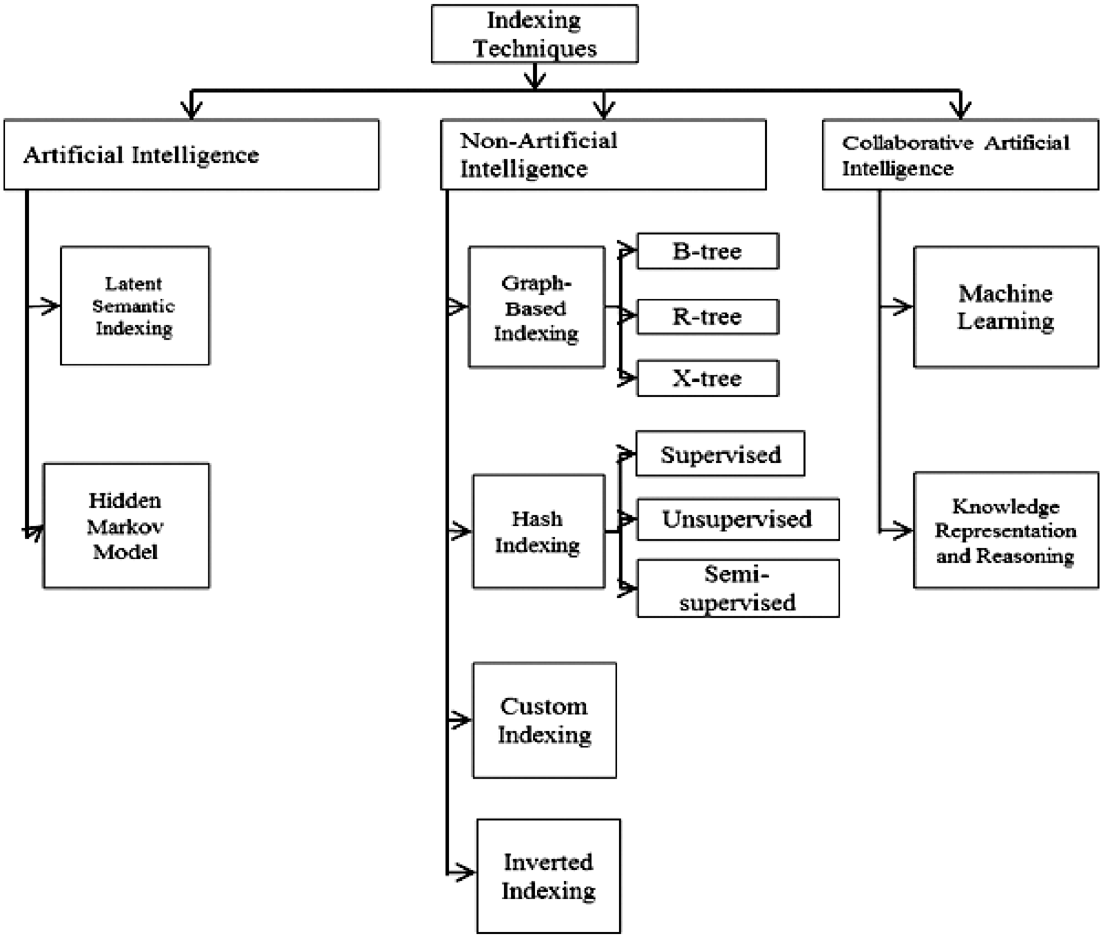

Figure 2: Taxonomy of Indexing Techniques

1.4. Collaborative Artificial Intelligence indexing

As the name suggests this technique is the combination of *artificial intelligent indexing and non-artificial intelligent indexing techniques*. The main aim of this technique is it consists of collaborative agents of each cluster to provide Machine Learning(ML) and Knowledge Representation and Reasoning(KRR) which helps indecision making for indexing which improves accuracy and search efficiency.

2. CLASSIFICATION OF INDEXING TECHNIQUES

Many researches had been carried out till date in past two decades in the area of Big Data. Table 1 discuss all the detailed description and advantages of different indexing techniques for various applications.

Table 1
Classification of Indexing Techniques

Indexing Technique	*Authors and Year*	*Description*	*Advantages*
		Artificial Intelligent Indexing	
* Latent Semantic Indexing [3]	Multi-label informed latent semantic indexing [Kai Yu et al Aug 2005]	Proposed a new technique- multi-label informed latent semantic indexing (MLSI) algorithm which preserves the information of inputs and meanwhile captures the correlations between the multiple outputs.	• Improve the prediction accuracy
* Hidden Markov Model [4]	Content-based video indexing of TV broadcast news using hidden Markov models [S. Eickeler et al. 2002]	Presents a new approach to content-based video indexing using hidden Markov models (HMMs)	• Automatic learning capabilities
		Non-Artificial Intelligent Indexing	
* Graph Based			
# B Tree [5]	Top-k queries on temporal data [Li et al. 2010]	To design efficient indexing technique for ranking queries on temporal data	• Index requires less space. • Fast Query Response
# R Tree [6]	A Framework For Efficient Spatial Web Object Retrieval [Wu et al.2012]	To design a hybrid inverted file R tree for text retrieval	• Index requires more space • Query processing cost is less • Response time is depends on buffer size
* Hashing			
# Supervised [7]	Supervised semantic indexing [Bing Bai et al 2009]	To provide an approach to easily retrieve different tasks	• More flexible for real-world applications
# Unsupervised [8]	Sparse hashing for multimedia search [Zhu et al 2013]	To develop a sparse hashing for fast approximate search	• Accurate query results • Encoding is fast
# Semi-supervised [9]	Semi-supervised hashing for large scale search [Wang et al 2012]	To develop a techniques for nearest neighbor search	• Minimizes quantization loss over both the labelled and unlabelled data
* Custom Indexing [10]	Potential-aware Automated Abstraction of Sequential Games, and Holistic Equilibrium Analysis of Texas Hold'em Poker [Andrew Gilpin et al. 2007]	Present a new algorithm for sequential imperfect information games and also presented a custom indexing scheme based on suit is omorphisms that enables one to work on significantly larger models than was possible before.	• User defined indices

(*contd...Table 1*)

Indexing Technique	*Authors and Year*	*Description*	*Advantages*
* Inverted Indexing [11]	Efficient set intersection for inverted indexing [Alistair Moffat et al. 2010]	Propose a simple hybrid method that provides both compact storage, and also faster intersection computations for conjunctive querying than is possible even with uncompressed representations.	• Investigate intersection techniques that make use of both uncompressed "integer" representations, as well as compressed arrangements.
	Collaborative Artificial Intelligent Indexing		
* Machine Learning [12]	Collaboration-based medical knowledge recommendation [Huang et al 2012]	To provide effective methods of seeking and recommending appropriate medical knowledge in order to help clinicians perform their work.	• Faster query response
* Knowledge Representation and Reasoning [13]	An Architectural Paradigm for Collaborative Semantic Indexing of Multimedia Data Objects [Clement H. C. Leung 2008]	Introduces new approach collaborative semantic indexing, which uses dynamic evolutionary approach	• Accurate query results

3. CONCLUSION

The survey of some important big data indexing techniques is highlighted in this paper. In taxonomy of indexing techniques were classified as Non-Artificial Intelligent approach, Artificial Intelligent approach and Collaborative Artificial Intelligent approach. The approach is to review the description and advantages of the various indexing strategies. In future, to improve performance efficiency important characteristics of each technique can be obtained and reviewed from this paper.

REFERENCES

[1] Gani, Abdullah, et al. "A survey on indexing techniques for big data: taxonomy and performance evaluation." *Knowledge and Information Systems* 46.2 (2016): 241-284.

[2] Adamu, Fatima Binta, et al. *A Survey On Big Data Indexing Strategies*. No. SLAC-PUB-16460. SLAC National Accelerator Laboratory (SLAC), 2016.

[3] Yu, Kai, Shipeng Yu, and Volker Tresp. "Multi-label informed latent semantic indexing." *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005.

[4] Eickeler, Stefan, and Stefan Muller. "Content-based video indexing of TV broadcast news using hidden Markov models." *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. Vol. 6. IEEE, 1999.

[5] Li, Feifei, Ke Yi, and Wangchao Le. "Top-k queries on temporal data." The VLDB Journal—The International Journal on Very Large Data Bases 19.5 (2010): 715-733.

[6] Wu, Dingming, Gao Cong, and Christian S. Jensen. "A framework for efficient spatial web object retrieval." The VLDB Journal—The International Journal on Very Large Data Bases 21.6 (2012): 797-822.

[7] Bai, Bing, et al. "Supervised semantic indexing." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.

[8] Zhu, Xiaofeng, et al. "Sparse hashing for fast multimedia search." *ACM Transactions on Information Systems (TOIS)* 31.2 (2013): 9.

[9] Wang, Jun, Sanjiv Kumar, and Shih-Fu Chang. "Semi-supervised hashing for large-scale search." IEEE Transactions on Pattern Analysis and Machine Intelligence 34.12 (2012): 2393-2406.

[10] Gilpin, Andrew, TuomasSandholm, and TroelsBjerreSørensen. "Potential-aware automated abstraction of sequential games, and holistic equilibrium analysis of Texas Hold'em poker." *Proceedings of the National Conference on Artificial Intelligence*. Vol. 22. No. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

[11] Culpepper, J. Shane, and Alistair Moffat. "Efficient set intersection for inverted indexing." *ACM Transactions on Information Systems (TOIS)* 29.1 (2010): 1.

[12] Huang, Zhengxing, et al. “Collaboration-based medical knowledge recommendation.” *Artificial intelligence in medicine* 55.1 (2012): 13-24.

[13] Leung, Clement HC, et al. “An architectural paradigm for collaborative semantic indexing of multimedia data objects.” *International Conferenceon Advances in Visual Information Systems*. Springer Berlin Heidelberg, 2008.