Big Data Retrieval: Taxonomy, Techniques and Feature Analysis

Israr Haneef[†], Ehsan Ullah Munir[††], Ghazia Qaiser[†††], Hafiz Gulfam Ahmad Umar[††††]

[†]Department of Computer Science1,4 COMSATS University, Wah Campus, Pakistan
[††]Department of Computer Science1,4 COMSATS University, Wah Campus, Pakistan
[†††]Bahauddin Zakariya University, Multan Pakistan
[††††]Ghazi University D.G..Khan, Department of Computer Science & IT Pakistan 32200

Summary
In recent years, Information retrieval in big data has become more popular research field. Big data is collection of heterogeneous structured and unstructured data. The heterogeneity, volume and the speed in which data is generating makes it problematic to process and analyze big data. The traditional databases system, warehouses and analyses tools are failed to process this type of data. Big data in IR system is an emerging approach not just because of the volume of data but also unstructured type of nature. The data that is related to the user query must be retrieved in IR system. Big data includes all type data like images, audio and video and from all resources like database, social media posts, and web blogs. In this paper, authors tried to provide and broad overview on different revival techniques in big data with the help of categorization of different techniques from existing literature.

Key words:
Big data, information retrieval, Feature analysis, retrieval techniques.

1. Introduction

In recent years, rapid progress of IT and the growing demand of networks, big data is increasing so fast. Big data can be describe as massive amount of data usually in terabytes [11]. The WWW generated a large amount of data and the IR techniques from large amount of data sets, play an important role. According to [4] the size of WWW surpassed from 800 million in 1999 and 11.5 billion in 2005 [8] and probably more than 30 billion these days. This fast and overwhelming change in Web make it a distinct source of information and a huge data set, and therefore the rational IR techniques cannot probably apply in this huge data set, because it can handle this amount of data. For an effective decision making, the result should be analyzed and extract properly from this unstructured form of data [10]. Hence, more effective and sophisticated IR techniques for efficient retrieval for decision making is required. These techniques should be suitable user profiles and quires. This simplify in the re-claim and organization of the data for creating knowledge required for the recommender systems. Its preferable, if the proficient features of the retrieval model are supportive for decision making and can discover the patterns, trends and correlations buried in big data.

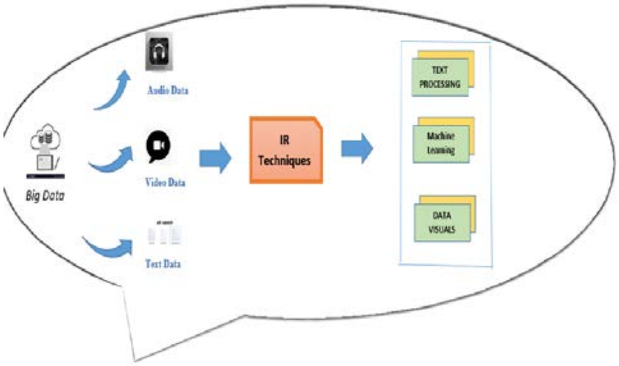

Fig. 1 IR System in Big Data

However, traditional (IR) information retrieval techniques are not capable of combining complex data to retrieve relevant information to provide better recommender systems for in time effective decision-making.With different information retrieval techniques and effective use of big data will provide us the different benefit in large networks. In these benefits, efficiency is including that will provide in fast information retrieval. Researchers still trying to get a more efficient technique in term of fast information retrieval and accurate information that are as per user's query as previous ones have some problems that are discussed previously.

The term big data is originated from the web companies who used to handle loosely structured or unstructured data. By meaning, big data in information retrieval are the states to electronic data sets so large and complex that they are problematic (nearly impossible) to handle with traditional software or hardware; they cannot easily managed with traditional data management techniques and methods [12]. "Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone. Lots of data is being collected and warehoused [14]. The big data originates from universally and everybody: different machines like sensors that are used to capture environmental information, social media

Manuscript received November 5, 2018
Manuscript revised November 20, 2018

posts, digital still images from and videos, purchase transactions record, and signals of mobile GPS to name a few. Big data discusses combinations of data whose volume, complex, and rate of growth or velocity make them problematic to be gathered, accomplished, analyzed or processed by predictable technologies and methods, like traditional databases and desktop statistics tools or visualization systems, within the time necessary to make them useful. While the size used to determine whether a data set is considered big data is not firmly defined and continues to change over time, and it is expected that the rate of growth of big data will continue to increase for the foreseeable future[9]. When big data is efficiently and professionally gathered, analyzed and processed it is considered very useful. It is persuading the way of doing business and go with real business value.

In this section a table of comparative analysis is provided in which different retrieval techniques are categorized on the bases of identified structure. From existing literature, some of parameters are recognized and labeled according to their existence. Here we have a comparative analysis with identified parameters, the techniques are from table 1. Parameters are indented from that literature, if we found that parameter in that literature we mention those in the table. Some of parameters are not identified by us, so they are unknown to us. If we see Architectural base information retrieval that contains Hadoop retrieval and Hadoop architecture for storage in which some listed parameters are discussed these architectural databases support these some of parameter and did not support NLP and security. And Hadoop storage architecture provide only classification, clustering, data analytics and load balancing. From all provided information retrieval Indexed based information retrieval will provide more parameters.

2. Taxonomical Diversification of IR Techniques for Big Data

2.1 Architectural Base

Architectural base information retrieval is based on different architectures like Hadoop architecture. Hadoop architecture is powerful architecture that is designed to explore complex type of data, transform big data and for analyzing of data [11]. In the era of big data Hadoop architecture is used for less retrieval time, huge storage capacity and high availability. In [11], authors describe the approach that can extract information from huge amount of data. In this defined approach uses word count method and inverted index with small testing data in a single environment. This architecture is flexible that allows new users to get benefits of this elaborated architecture. As it is described that this approach uses word count method so a lot of time will be consumed is this approach. This issues can resolved in future research to identify most frequently used words in data set. In [7], a distributed storage and cluster approach is used for better retrieval of information and efficient storage capacity. In this research a framework is proposed by merging Hadoop base cloud computing platforms and the storage features of Hadoop base disturbed database. This framework is implemented on Linux OS. When there is large number of users, then it will be more efficient to use. The explained framework is based on Hadoop architecture and on Linux platform. Thus, expert people are required to understand this architecture.

An idea is proposed [14] in which a web server is designed in map phase using jetty web server that can fast and efficient for searching data in map reduce standard. A searchable mechanism is implemented for real time processing by creating multiple index in web server with the help of multiple search keys and index data node.

Table 1: Architectural base Information

Type	Category	Subcategory	
Architectural base	Information retrieval using Hadoop [11]	Word count method	Big data using Hadoop Fault tolerance & availability[7]
	Hadoop architecture for storage [7]	Hadoop cloud computing framework. [7]	Distributed storage Server clustering [7]
	Meta data database [2]	User's location User's current time Common occasion [2]	

By using clustering technology, we can handle efficiently traffic and distribute the load on different servers. In future it can be more enhanced by using real time applications for large data sets.

2.2 Content Based

Content based information retrieval retrieves the data on some sort of information like meta data database, image The model has the capability to narrow down the user's preferences and need and in real life implementation this model is highly effective when data is so large. But for this model metadata is only gathered when user's GPS (global position system) is active to determine the user's exact location and place.

In [1] authors proposed a fast image retrieval system that is designed for big data. For image retrieval, firstly all features of image are extracted and it will take time for large image extraction so it is required to reduce the feature dimensions for optimizing structure of features. Finally, the similarity matching will determine the retrieved results. The proposed technique for image retrieval works with three main contributions that are feature extraction method, the reason able element ranking and efficient distance metrics that can make better the performance of algorithms. Result shows that the proposed technique can make the more effective performance and retrieved better matching results.

Table 2: Content base Information Retrieval

Type	Category	Subcategory
Content based	Probabilistic models [3]	Bayesian method [3]
	Symbolic learning and rule induction [3]	Rote learning Learning detection Learning instruction Learning analogy examples [3]
	Evolution-based models [3]	Rely on analogies [3]

In this proposed technique, feature extraction is very important but still there is need to extract more fractures for more accurate results. The proposed method in [12] improves textual by using map reduce technique. In this map reduce mechanism pattern of text is examined from different data files of big data. The text pattern recognition improves the performance by decreasing number of access to data files of big data. In future, this work can be expanding for distributed system for retrieving texts from different data nodes.From this section, input the body of your manuscript according to the constitution that you had. For detailed information for authors, please refer to [1].

Table 3: Content base information retrieval

Type	Category	Subcategory	
Content Base	Meta data database [2]	User's location User's current time Common occasion.[2]	
	Image retrieval [16]	Content base image retrieval	Reasonable elements ranking method Feature extraction method Appropriate distance metrics [16]

2.3 Machine Learning

Machine learning in information retrieval playing a really important role specially in web search engine, online advertising and recommendation systems. The idea presented in [2] is to accelerate the search operation in big data using neural networks. This cross-correlation technique is used between the user's query and the big data. Moreover, neural networks are used for retrieving of big data even the data is noisy and distorted. The process of information retrieval by using neural network are divided in two parts, first neural networks recognize the input pattern and the second to match it with the given big data. The aim of this work is to manipulate huge amount of data with less time The proposed model [13] that integrates the scattered data and organize them from multiple heterogeneous data source. This system will retrieve and integrate the non-structured or semi structured data. This model will help the IT business for finding solutions and get information from multiple resources and integrate them. Proposed model is built on Hadoop platform and collect big data using J2EE, the collected data is in the form of XML In [15] authors describe a scheme based on fuzzy similarity for color image retrieval from color library. In image retrieval from big data, color feature is the most important feature. For measuring color similarity in images direct membership value, of histogram from gray levels a gamma histogram plays an important role. For finding membership function from gamma distribution has been proposed. There are several models for these Processes established and practice the basis of implementation. For examples the vector space and probabilistic for information retrieval and 2 positon model.

Table 4: Parameters addressed by Type

		Learning base				Targeted data				Extra Features	
Type		Classification	Clustering	Data analytics	NLP	Texts	Visuals	Audio	Hybrid	Load balancing	Security
Architectural base	Information retrieval using Hadoop	✓	✓	✓	–	✓	✓	–	–	✓	–
	Hadoop architecture for storage	✓	✓	✓	–	–	–	–	–	✓	–
Content Base	Meta data database	✓	–	–	–	✓	✓	✓	–	_	✓
	Image retrieval	✓	–	–	–	✓	✓	✓	✓	_	–
	Textual retrieval	✓	–	–	–	✓	✓	✓	✓	_	–
Machine learning	Neural networks	✓	✓	✓	✓	✓	✓	–	–	_	–
	Probabilistic models	✓	–	✓	✓	✓	✓	✓	–	_	–
	Symbolic learning and rule induction	✓	✓	–	✓	✓	✓	✓	–	_	–
	Evolution-based models	–	✓	✓	✓	✓	✓	✓	–	_	–
	Analytic learning and fuzzy logic	✓	✓	✓	✓	✓	–	–	–	_	–
Indexed base	Tree pattern framework	✓	–	✓		✓	–	✓	–	✓	✓
	Artificial intelligence approach	✓	✓	✓	✓	✓	✓	✓	–	✓	✓
	Non-artificial intelligence approach	✓	–	✓	–	✓	✓	✓	–	✓	✓
Rule based	Ontology based	✓	–	✓	–	✓	✓	–	–	✓	–

This retrieval scheme with vector distance measuring and gamma membership function can effectively fine the color similarities from image database.

2.4 Rule Based

Rule base retrieval of big data is based on some rules like ontology based information retrieval and Grammar rule based information retrieval system. In [5] author proposed a model that is based on ontology for improving the search over huge document files. This model is made by adoption of vector space model, annotation weighting algorithm and ranking algorithm. This proposed model can improve keyword based search and provide the benefits of semantic based search for improving retrieval system. In future, the improved version of this model can be made for large documents and large amount of data.

Table :5 Rule based Information retrieval

Type	Category	Subcategory
Rule Based	Ontology based [13]	Knowledge base [13]

3. Challenges

Big data analytics in IR systems must provide some of key functions for processing of data. Criteria for the evolution of platform should include the continuity, scalability, availability, ease of use and ability of the manipulation at different levels. Moreover, different platforms are available open source and they have different advantage

and because they are openly available they have limitation too. IR system for big data needs to be a full fledge system like user friendly, menu driven and transparent. Information retrieval system has a key requirement that is real time big data analytics that are used in sensor networks. The dynamically processed data is also need in an information retrieval system. The most important issues that are related to management are required to be handle in an information revival system. These issues can be related to the privacy and confidentiality of the data standards and ownership of the data. Big data is an emerging filed so these issues must be handled for the upcoming IT industry. Data must be standardized and legal format of data. This great challenge needs to be addressed as well.

4. Conclusion

Big data analytics has the feature to mold the way IR systems technologies to gain their data and other sophisticated repositories and on the base of these make a decision. In the coming years, we will see the use and implementation of big data analytics in IT industries. To that end, different analytical and retrieval techniques are described above. With that technologies, we described some of challenges that must be addressed in the future. Big data has gain more attention so some of issues like security, privacy, standards governance and the sophisticated tools for the analytics purpose gain the attention of the researchers. Information retrieval and analytics in big data are at the emerging stage of the development but still there is a need for more efficient and sophisticate

Reference

[1] Adamu, Fatima Binta, Habbal, Adib, Hassan, Suhaidi, Malaysia, U Utara, Cottrell, R Les, White, Bebo, . . . Malaysia, U Utara. (2016). A Survey On Big Data Indexing Strategies: SLAC National Accelerator Laboratory (SLAC).

[2] Al-Drees, Asma, Bin-Hezam, Reem, & Al-Muwayshir, Ruba. (2016). Unified Retrieval Model of Big Data. Paper presented at the INNS Conference on Big Data.

[3] Bhosale Jaykumar P, Dr. Bhoite S D. (2016). Use of Machine Learning Techniques for Information Retrieval / Extraction from Web. International Journal of Advanced Computer Technology & Management, I(May-2016), 7.

[4] Bian, Jiang, Topaloglu, Umit, & Yu, Fan. (2012). Towards large-scale twitter mining for drug-related adverse events. Paper presented at the Proceedings of the 2012 international workshop on Smart health and wellbeing.

[5] Chandrasekar, Muthukumar. R C. (2012). Information retrieval using indexing scheme for tree pattern framework. Journal of Global Research in Computer Science, Volume 3, 6.

[6] El-Bakry, Hazem M, Mastorakis, Nikos E, & Fafalios, Michael E. (2014). Fast information retrieval from big data by using neural networks implemented in the frequency domain. sensors, 1, 6.

[7] Jie, Chen, Dongjie, Chen, & Bangming, Huang. (2014). Research on big data information retrieval based on hadoop architecture. Paper presented at the Electronics, Computer and Applications, 2014 IEEE Workshop on.

[8] Lawrence, Steve, & Giles, C Lee. (2000). Accessibility of information on the web. intelligence, 11(1), 32-39.

[9] M.M. Kodabagi, Deepa Sarashetti and Vilas Naik. A Text Information Retrieval Technique for Big Data Using Map Reduce. Bonfring International Journal of Software Engineering and Soft Computing, Vol. 6(Special Issue), 22-26. doi: 10.9756/BIJSESC.8236

[10] Milovic, Boris. (2012). Prediction and decision making in health care using data mining. Kuwait chapter of arabian journal of business and management review, 1(12), 126-136.

[11] Motwani, Deepak, & Madan, Madan Lal. (2015). Information Retrieval Using Hadoop Big Data Analysis Advances in Optical Science and Engineering (pp. 409-415): Springer.

[12] Raghupathi, Wullianallur, & Raghupathi, Viju. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

[13] Vallet, David, Fernández, Miriam, & Castells, Pablo. (2005). An ontology-based information retrieval model. Paper presented at the European Semantic Web Conference.

[14] Venkatesh, H, Perur, Shrivatsa D, & Jalihal, Nivedita. (2015). A study on use of big data in cloud computing environment. Int. J. Comput. Sci. Inf. Technol.(IJCSIT), 6(3), 2076-2078.

[15] Yang, Jiachen, Jiang, Bin, Li, Baihua, Tian, Kun, & Lv, Zhihan. (2017). A fast image retrieval method designed for network big data. IEEE Transactions on Industrial Informatics.