International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
SKRGC Since 2012

http://www.ijcsjournal.com Volume 5, Issue 1, No 27, 2017 ISSN: 2348-6600
Reference ID: IJCS-286 PAGE NO: 1870-1879

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

FEEDING BIG DATA FOR ENHANCED SEARCH ENGINE APPLICATIONS

L.Rasikannan1 **Dr.P.Alli2**
1 Assistant Professor in CSE Dept, ACCET, Karaikudi
2 Professor, CSE Dept, Velammal College of Engg. & Tech, Madurai

Abstract- Big Data has its own popularity since it surrounds huge, complicated, growing data sets with many, autonomous sources. They are analyzed computationally to reveal patterns, trends and associations, more specifically relating to human behavior and interactions. Many of the IT investments turn towards Big Data for their better survival. Search Engine plays a predominant role in presenting information to all those who need it, but still struggles to satisfy the drastic changes in user's requirements. So the search engine should be refined and fine tuned in all scopes, it is required to depend on some other emerging technologies to expedite the necessary issue in Search Engines. Search Engines request the robust services of Big Data for improving its essential functions like web crawling, indexing and summarization of data from the large repositories irrespective of the domain. The repository, in nature, may be dissimilar, independent sometimes and seems to be irrelevant with complex, evolving relationships, and keeps growing. Now a day we are brilliant enough to create quintillion websites for various applications. Maintaining and exploring the huge websites seems to be tedious. The ultimate challenge for Big Data applications is to explore the huge volumes of data and extract suitable information or knowledge for future actions. One of the toughest processes of search engine is to uncover the entire information or knowledge existing in the universe. User expects A to Z from the engine, but storing and extracting all observed data is nearly infeasible. In this paper, we are going to study and exercise by combining the

power of two giant technologies, to satisfy the need and requirements of the knowledge seekers around the globe.

Index Terms— Autonomous sources, dimension, complex data, data mining.

1. INTRODUCTION

This world that we live in today much desire for data and novel patterns of information. To manage search engines and to address the user needs and keep them in comfort is a challenging business now-a-days. The current improvements in data retrieval technology allow us to have scalable and flexible data capture, storage, processing and analytics. We live in a data driven world, the direct result of advents in information and communication technologies. Millions of resources for knowledge and information processing are made possible through the Internet and Web 2.0 collaboration. No longer do we live in isolation from vast amounts of data. Want of information from these massive amounts of data is even more demanding for enterprises to survive in competition world. Mining information from raw data is an extremely vital and tedious process in today’s information driven world. Enterprises today rely on a set of automated tools for knowledge discovery to gain business insight and intelligence.

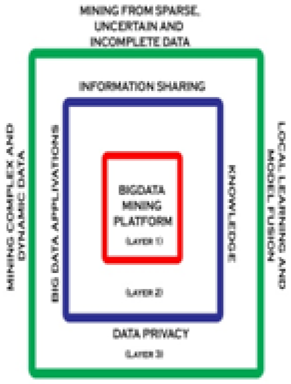
2. FEATURES OF BIG DATA

Big Data features can be described with large volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. Fundamental characteristics of the Big Data are the massive volume of data represented by heterogeneous and dissimilar dimensionalities. Practically, in a heterogeneous system, sites may run different DBMS products, which need not be based on the same underlying data model, and so the system may be composed of relational, network, hierarchical and object-oriented DBMSs. This is because different information accumulators prefer their own representations or protocols for data recording, and the nature of various applications also effects diverse data representations. Again the complexity increases with different dimensions involved in text mining, that is, simple sales data has multi-dimensional product analysis like:

Time vs. location vs. cost vs. color vs. size and so on. Time again divided into year, month, week, day, hour, minutes and seconds. Likewise a data can be represented in many dimensions. This is because user's search interest may be in any point in the data dimensionality.

Being autonomous, each data source is able to generate and collect information without comprising any of the centralized control. For major Big Data-related applications, [5] such as Google, Twitter, Facebook, and Instagram, a large number of servers are installed independently in a de-centralized manner all over the world to ensure nonstop services and quick responses for local markets.[1]

While the volume of the Big Data increases, the complexity and the relationships among the data also increase naturally. For example, social network sites, like Facebook or Twitter, are mainly characterized by social functions among friend-connections and followers. The correlations between individuals inherently complicate the whole data representation.


MINING FROM SPARSE, UNCERTAIN AND INCOMPLETE DATA
INFORMATION SHARING
MINING COMPLEX AND DYNAMIC DATA
BIG DATA APPLICATIONS
BIGDATA MINING PLATFORM
(LAYER 1)
KNOWLEDGE
LOCAL LEARNING AND MODEL FUSION
(LAYER 2)
DATA PRIVACY
(LAYER 3)

Big Data processing structure and Background

Fig 1

- Layer1 focuses on low-level data accessing and computing.

- Layer2 focuses on information sharing and data privacy application domains and it includes the knowledge of high-level semantics

- Layer3 focuses on actual mining algorithms.

Before we use Big Data platform for search engine applications, we should know about the working methodology and structure of the Big Data architecture, as well as learning the functionalities about the three layers, will help us to know how meticulously Big Data can apt for efficient and extensive search engine applications.[6]

The challenges of Layer I focus on data accessing and arithmetic computing procedures, Big Data is often stored at diverse locations and data volumes may endlessly grow. An active computing platform must have to take distributed large-scale data storage into consideration for better computing. For example, typical data mining algorithms require all data to be loaded into the primary memory. This creates a technical hurdle for Big Data because transferring data across different locations is expensive and complex, even if we have a super large main memory to hold all data for computing. To tackle such problems cloud computing may be used to support on the economic grounds by sharing and utilizing the existing idle resources across the globe.

The challenges at Layer II revolve around data semantics and domain knowledge for different Big Data application domains. Such information can provide additional benefits to the data mining process, by providing relevant meaning to the Layer I applications. Ensuring data privacy, in the information sharing scenario between data producers and data consumers can be a sensible one.

At Layer III, the data mining challenges concentrate on some of the algorithm designs in handling the difficulties raised by the Big Data volumes and distributed data distributions, because of its complex and dynamic data characteristics. Layer III contains three stages. First, sparse, heterogeneous, uncertain, missed, and multisource data are preprocessed by data fusion techniques. Second, complex and dynamic data are mined after the completion of preprocessing. Third, the knowledge obtained by learning and model fusion is tested and relevant information is fed back to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promise of smooth development of each stage, but also a purpose of Big Data processing.

3. CHALLENGES OF DATA MINING FOR SEARCH ENGINES

In typical data mining systems, the data mining procedures need computational intensive computing units for data analysis and comparisons. Parallel computing or collective mining to sample and aggregate data from different sources by applying parallel computing paradigms to carry out the mining process successfully. When the size of the data increases exhaustively, we cannot go for a single processor for the processing. Inevitably, parallel programming techniques have to be implemented to save the computing time and ensure the clarity of results to users.

Data mining task can be deployed by running some parallel programming tools, such as MapReduce or Enterprise Control Language, on a large number of computing nodes or clusters. The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes called as cluster computing.

Big Data mining offers opportunities to go beyond traditional relational databases to rely on less structured data: weblogs, social media, e-mail, sensors, and photographs that can be mined for useful information.

4. INFORMATION SHARING AND DATA PRIVACY ISSUES

In search engines, searching and sharing the information is clear; a real-world concern is that search engine results are related to sensitive and valid information it depends on like time series analysis and medical records.[4] Simple data exchanges or transmissions do not resolve privacy concerns. To preserve privacy, the two common approaches are to consider i) restrict access to the data, such as adding certification or access control to the data entries, sensitive information is accessible to an authorized group of users only, and ii) anonymize data fields such that sensitive information cannot be pinpointed to an individual record. Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of encrypting or removing personally identifiable information from data sets, so that the people who the data describes remain unknown and the privacy is preserved. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconduct by unauthorized individuals.

One of the major benefits of the data anonymization-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,[11] 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.

4.1. Identifying fake websites

Fake websites are imaginary, misrepresentative sites posing as legitimate providers of information, goods, or services used to gather illegitimate revenues by deceiving search engines or exploiting unsuspecting Internet users. Fraudsters have created several types of fake websites, including web spam, concocted, and spoof sites. Web spam sites attempt to deceive search engines to boost their rankings. Leveraging link and content spamming methods, these websites employ in black hat search engine optimization in which extremely ranked web spam sites are more visible and can be sold for a greater profit. Concocted websites are deceptive sites attempting to appear as legitimate commercial entities. Fake websites are often professional looking and difficult to identify as spurious. In response to increasing user awareness, fraudsters are also becoming more sophisticated, while current security tools are unsuitable for handling fake websites' increasing complexity. Accordingly, a need has arisen for more refined fake website detection techniques.

4.2. Importance of Domain and Application Knowledge

Domain and application knowledge provides essential information for designing Big Data mining algorithms for improving the search engine's effectiveness and efficiency. The market continuously evolves and is impacted by different factors, such as local and international news, government reports, and natural disasters, and web sites must be updated and the changing values should be maintained and stored in the index table.

Without correct domain awareness, it is a clear challenge to find effective matrices/measures to describe the market movement, and such knowledge is often beyond the mind of the data miners, although some recent research says that using social networks, such as Twitter, it is possible to predict the stock market upward/downward trends with good accuracies. [9]

4.3. Local Learning and Model Fusion for Multiple Information Sources

The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values. While most modern data mining algorithms have in-built solutions to handle missing values, data imputation is an established research field that seeks to impute missing values to produce improved models. Imputation is the process of replacing missing data with substituted values.[7] Many imputation methods exist for this purpose, and the major approaches are to fill most frequently observed values or to build learning models to predict possible values for each data field, based on the observed values of a given instance. For example, researchers have successfully used Facebook and Twitter, a well-known social networking site, to detect events such as earthquakes and major social activities, with nearly real time speed and very high accuracy [8]. In addition, by summarizing the queries users submitted to the search engines, which are all over the world, it is now possible to build an early warning system for detecting fast spreading diseases.

4.4. Impact of different data formats

In Big Data, data types include structured data, unstructured data, and semi structured data, and so on. Specifically, there are tabular data relational databases, text, hyper-text, image, audio and video data, and so on. The existing data models include key-value stores, big table clones, document databases, and graph databases, which are listed in an ascending order of the complexity of these data models. The size or complexity of the Big Data, including transaction and interaction data sets, exceeds a regular technical capability in capturing, managing, and processing these data within reasonable cost and time limits. In the context of Big Data, real-time processing for complex data is a very challenging task. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

4.5. MapReduce

MapReduce is a batch-oriented parallel computing model. The MapReduce algorithm is divided into two important tasks, they are Map and Reduce. Mapping function takes a set of data and converts or maps it onto another set of data, where individual elements are broken down into records or rows (key/value pairs). Second one is to reduce the task, which takes the output from a map as an input and combines those data into a smaller set of values. The basic Divide and conquer algorithm is used.[3] The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. In MapReduce, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is simply a configuration change.

4.6 Preserving Quality of Search

Search quality is the critical factor of organizations because almost all the references and ideas are generated by online web search engines. Employees in the firm for every change and updates they must rely on the Search Engines, no other sources can’t help instantly like SEs. Organization search assignments are mostly managed by administrators who are domain experts but not by search experts. In general search engine plays to serve two main tasks (i) to evaluate and recognize user queries and interpreting them into queries appropriate for the indexing (ii) organizing and ranking the results and present them efficiently to the user.

An important and complex query type in enterprise search is expertise search, which will search for people, presents added challenges compared with retrieving a precise document. Search algorithm should take into account the social information like locality, designation, and any nick names or alias of the expert, Inter personal skill matrix also taken in to consideration. Result representation is important for enterprise search. Another interesting job of the Search Engine is to frame the results according to the user's comfort relevant with the domain. The results may be from various sources, different structures according to the queries of different types informational, navigational and so on[2].Indexing is an important work after crawling Lucene index engine struggles for enterprise search by the excessive usage rich set of jargons, metadata and context that is absent in general natural language.

5. HIDDEN WEB

Some of the web sites are not easy to navigate by the crawlers such web sites commonly called as Hidden websites. Some studies have estimated that the deep Web is over a hundred times larger than the traditionally indexed Web, although it is very difficult to measure this accurately. Hidden web are basically divided in to three

broad categories they are:

• Private sites: Naturally they are private web sites. Such web sites are blocked access from the crawlers and there is no incoming links, otherwise you have to log in with a valid account before using the rest of the site.

• Form results: Such kinds of websites can be reached only after entering some data into a form. Such pre information or registration is compulsorily to be filled for accessing the above web sites.

• Scripted pages: pages that can be developed from the client-side language in the web page. If a link is not in the raw HTML source of the web page, but is instead generated by JavaScript code running on the browser, the crawler will need to execute the JavaScript on the page in order to find the link. But, executing JavaScript can slow down the performance of thee crawler significantly and adds complexity to the system.[10]

5.1 **static pages and dynamic pages**

Static pages are obvious that they stored on a web server and displayed in a web browser unmodified, in case of dynamic pages are the result of code executing on the web server or the client. Static pages are easy to crawl while dynamic pages are hard.[12] This is not quite true, however. Many websites have dynamically generated web pages that are easy to crawl; wikis are a good example of this. Other websites have static pages that are impossible to crawl because they can be accessed only through web forms. But the owner of the websites they want their sites to be indexed properly.

5.2 **Site Map**

Sitemaps make relationships between pages and other content components. It shows shape of information space in overview. It is a model of a website's content designed to help both users and search engines crawlers navigate the site. A site map can be a hierarchical list of pages (with links) organized by topic, an organization chart, or an XML document that provides instructions to search engine crawl bots. Sitemaps can demonstrate organization, navigation, and classification system.[10] An XML Sitemap is a structured format not for a user's concern for search engine about the pages in a site, their relative importance to each other, and how often they are updated. HTML sitemaps are designed for the user to help them find content on the page

5.3 **Deep Learning**

Deep Learning algorithms are one promising avenue of research into the automated extraction of complex data representations or features at high levels of abstraction. Such algorithms develop a layered, hierarchical architecture of learning and representing data, where higher-level (more abstract) features are defined in terms of lower-level (less abstract) features. Deep Learning algorithms are quite beneficial when dealing with learning from large amounts of unsupervised data, and typically learn data representations in a greedy layer-wise fashion [13].Deep Learning solutions have yielded outstanding results in different machine learning applications, including speech recognition, computer vision and natural language processing

6. CONCLUSION

Big data the term clearly gives a clear idea that big mind is required for managing and mining big data. The several challenges one must overcome to explore big data are at the data, model, and system levels. High performance computing platforms are required to manage big data mining. The complicated situation in big data normally arrives from autonomous information sources and the variety of data collection environments. Missing uncertain values also becomes a challenging one. Developing a safe and sound information sharing protocol is a major challenge. An Enterprise or individual now a day's heavily depend upon the information, the loss of information at the right moment is a loss to the organization and individual. Search Engine is suffered to give better performance due to the various factors. Whatever the hurdles the Search Engines may have. It must strengthen them with the robust technological boosting. In this scenario Big Data plays a vital role in supporting the SE for un interrupted and to provide value added services.

REFERENCES

[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.

[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

[3] D.Gillick, A.Faria, and J. Denero, map reduce: distributed computing for machine learning, Berkley, Dec 2006

[4] G.Duncan, "Privacy by design, "science, vol 317,pp.1178-1179,2007

[5] D.Centola,"The spread of behavior in an online social network experiment,"science,vol 329,pp.1194-1197,2010

[6] E.Birney,"The making of encode:lessons for big data projects", nature vol.489 pp 49-51

[7] B.Efron, "Missing data, imputation and the bootstrap".

[8] M.Helft,"Google uses searches to track flu's spread".

[9] S.Aral and D.walker, "Identifying influential and susceptible members of social networks.Enterprise Search in the Big Data Era:

[10] Recent Developments and Open Challenges

[11] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[12] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence,I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002

[13] Hinton GE, Osindero S, Teh Y-W(2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554