ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

BIG DATA ANALYTICS IN HEALTHCARE SYSTEM FOR DIVERSE PERSPECTIVES

P.Shanmuga Sundari*, Dr.M.Subaji**
*Research Scholar, School of Computing science and Engineering, VIT University, Vellore,
Tamil Nadu 632 014, India.
** Professor, Center for Industry and International Studies, VIT University, Vellore, Tamil Nadu 632 014, India.
Email:sundari.sigamani@vit.ac.in

Received on 06-08-2016 *Accepted on 10-09-2016*

Abstract

In the era of advanced technology in healthcare system, data are produced in huge amount from varied sources like clinical, e-health record, prescription, tests report and test image records. The greatest challenging is to extract knowledge from raw data and provide useful information to the medical researchers and practitioners, who in turn put it to work in real life scenario which will benefit a common man. Big Data analytics in healthcare is used to predict disease from historical data such that it predict epidemics disease, adverse drug reaction from social media improves quality of life and avoid preventable decease in earlier. In traditional system handling varied health care data for collection, storage and processing is complex, and this complication leads more opportunities for big data analytics. The main objective of this paper is to focus on various types of analysis and techniques to solve the big data problem in healthcare system.

Keyword: Big data, Healthcare, Big data storage, Text mining, Web mining.

1. Introduction

During the last decades substantial flow of data is perceived in Healthcare domain representing patients' health states in the form of laboratory results, treatment plans, and medical reports. Report from NCBI says that data from the U.S. healthcare system alone reached 150 Exabyte in 2011. Moreover at this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes) . Kaiser Permanenteis the California-based health network, which has more than 9 million members, is believed to have between 26.5 and 44 petabytes of potentially rich data from Electronic Health Records (EHR), including images and annotations .Significant clinical knowledge and a deeper understanding of patient disease pattern can be gleaned from such collection[1] .

Recent studies shows Pew Internet & American Life project says that, 81% of U.S. adults use the Internet and 59% looked online for health information regarding diseases, diagnoses and different treatments. As increasing information available in the internet the patients becomes a decision maker. Many patients' use the Internet to obtain health-related information. It is anticipated that health-related Internet information will reduce cost without consultation practice of physicians .Digital information in the form of news sites and web forums are freely available for patient-oriented decision making that has increased drastically. The biggest challenge for the user is the overloaded information that is irrelevant for drawing conclusions on the personal health status and taking satisfactory actions. Faced with a large amount of medical information on different channels, users often get lost or feel uncertain when investigating on their own. In addition, a manifold and heterogeneous medical vocabulary poses another barrier for laymen. Therefore improved personalized delivery of medical content can support users in finding relevant information.

Big data analytics can play a major role in real-time alerting system. Big data machine learning libraries faithfully predict the course of a patient over time while providing above mentioned analysis in leveraging historical data from other patients with similar conditions, predictive algorithms and that too in timely and cost effective manner. From such way that using claims data the system Medco can extract patient indications, treatments, dates of treatment, and outcomes like whether the patient was hospitalized or not can be extracted. Putting this multi-layered data together, Medco can search for associations between drug use, patient characteristics, and clinical impact such as good, bad or indifferent, in order to determine whether a drug works the way it should . Increasing patient access to medical records could encourage patient participation in improving the accuracy of medical records [2]. IBM rightly said that, " rising rates of chronic disease, aging populations, changing consumer expectations about how they want to purchase and receive care, and increasing access to social media and mobile technologies are transforming the way healthcare is obtained and delivered".

Health social media sites such as Daily Strength and Patients Like Me provide unique research opportunities in healthcare decision support and patient empowerment especially for chronic diseases such as diabetes, Parkinson's, Alzheimer's, and cancer [1]. Association rule mining and clustering, health social media monitoring and analysis, health text analytics, health ontologies, patient network analysis, and adverse drug side-effect analysis are promising areas of research in health-related system. This paper illustrated leveraging the big data analytics in healthcare systems. Section 2 brings out impact of big data in healthcare system. Section 3 conveys varied data. Section 4 expressed big data processing stages and it discusses on different approaches and methodologies for predictive analytics in following sections.

2. Varied Data In Health Care

IBM stated that the big data are featured with 4V's namely volume, velocity, variety and veracity. Volume refers to large size of data such as petabytes to zettabyte. These data is the number of customer on Facebook, YouTube, twitter and other domains. Velocity refers to the swift data processing, such as data generated from wireless sensor device and video surveillance camera. Variety refers to different data types such as Structured, Unstructured and semi structured data. Data may be in text, audio, video, image, internet data or click stream data. Unclear, incomplete, inconsistent and doubtful data is called as Veracity. Following Table 1 shows that how these four characteristics which relate healthcare data.

Table 1. Characteristics of big data.

Dimensions	Data types
Volume	Digitizing existing patient data, human genetics, population data, genomic sequences, 3D imaging, genomics and biometric sensor readings.
Variety	Office medical records, handwritten nurse and doctor notes, hospital admission and discharge records, MRI, CT and other images.
Velocity	Real-time data (trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors and Electrocardiograms (ECG).
Veracity	Veracity in healthcare data faces diagnoses, treatments, prescriptions, procedures.

3. Big Data Processing

Value chain of big data can be divided into four stages. Figure 1 shows the four stages in big data processing. Data generation and acquisition is the initial stages of value chain. The raw data from various sources will generate the huge amount of data and it can be stored into proper storage devices. Finally data analysis is the way to analyze the data in different data format. According [3] analysis consists of three types prescriptive, predictive and descriptive. This paper focuses on predictive analysis.

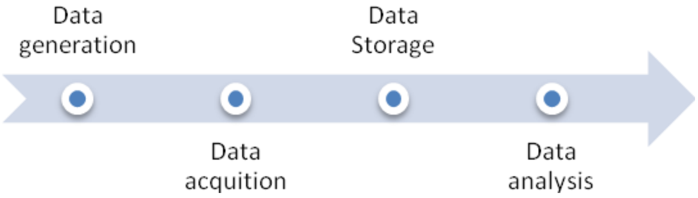

Figure1. Big data processing [3].

3.1. Data generation in healthcare

Advanced technologies such as capturing devices, sensors and mobile applications play a major role in data generation. Today social media network, world wide web plays the major role in data generation. The internet of things in healthcare encompasses heterogeneous computing and wireless communication systems, apps and devices that help patients and providers alike to monitor, track and store patients' vital statistics or medical information. Examples of such systems are smart meters, RFID, wearable health monitoring sensors, and smart video cameras. Smart phones, intellectual machines, and robotics are also considered to be the part of IoT. Table 2 shows some available data sources.

3.2. Data acquisition in healthcare

Data acquisition is carried out in three distinct technique namely data collection, data transfer and data pre-processing. Electronic devices play a major role in data collection. The collected raw data through these devices is usually heterogeneous, noisy, severely distorted and contains many relevant, sometimes associated features. This data resides in multiple databases such as individual EMRs, lab and imaging systems, physician notes, medical correspondences, claims, CRM systems, and hospital finance department servers. The collection, integration, and analysis of such big, complex, and noisy data in healthcare are a challenging task. For this reason, healthcare data can be considered as a form of big data not only for its complete volume, but also for its complexity and diversity that makes traditional data warehousing (ETL) solutions prohibitively unwieldy and unsuited for large scale data exploration and modeling.

Hadoop is a perfect platform to run Extract Transform Load (ETL). It has tools that can extract the data from diverse data sources such that log files, machine generated data or online databases and loads them in best possible time. It is probable to transform on the fly as well, although more elaborate processing is better done after the data is loaded into Hadoop. Programming and scripting frameworks allow complex ETL jobs to be deployed and executed in a distributed manner. An open source platform namely Hive [4] is used to advance SQL type scripts to do MapReduce operations that extract relevant information from the Hive schema, run queries using an interface similar to standard SQL. This actually gets transformed to programmatic constructs and performs as multiple MapReduce jobs. MapReduce is another programming model intended to process large volumes of data in parallel by dividing the work into a set of independent tasks. Another widespread open source distributed data management tool that influence the data extraction process faster is Cassandra which can handle very huge quantities of data spread out across many commodity servers with no single point of failure (due to replication). Cassandra distributes a controlled key-value store with tunable consistency when compared with Hive. It is immensely scalable due to its ring architecture. To protect from loss during node failure, data is replicated to multiple nodes. By offering the organization of a traditional RDBMS table layout combined with the flexibility and power of no stringent structure requirements, Cassandra also provides flexible schema-less data modeling.

3.3. Data storage in healthcare

An efficient way of data storage is a big challenge in the healthcare domain because of varied data format. Modern CPUs are becoming highly parallel to achieve scalability and performance improvement in large-scale data processing. Traditional data storage like DBMS is inadequate to tackle and manage these varieties of data. Distributed and parallel database systems are needed to handle large scale data to achieve flexibility, reliability and fault tolerance [5]. Storing big data include handling very large amounts of data and keep scaling to keep up with growth. It also make available the input/output operations per second (IOPS) necessary to deliver data to analytics tools. Amdahl's Law stated that improvements in processing speed must be matched by improvements in I/O. The table below shows some data bases that support large scale data processing in parallel and distributed computing and storage. Following Table 3 shows different databases that support large scale data storage.

Table 2. List of databases that support large scale data storage.

Data base type	Database name	Data storage	Consistency	Map reduce support	Purpose
Column-oriented data base	HBase [6] Cassandra [7] HyperTable [8]	HDFS Disk Plug-in Google file	Immediate consistency Eventually consistency Immediate consistency Immediate consistency Eventually	Yes Yes Yes Yes	Column-family databases store data in column families as rows that have many columns associated with a row key

	Bigtable [9]	system	consistency		
Document based	SimpleDB [10] MongoDB[11] Couch DB [12]	S3(Simple storage solutions) Disk Disk	Immediate consistency Eventually consistency Eventually consistency Immediate consistency Eventually consistency	No Yes Yes	A database record consists of a collection of key value pairs plus a payload.
Graphic database	Neo4J[13] Infinite graph[14] OrientDB [15]	- Disk -	- Eventually consistency -	No - -	To store entities and relationships between these entities. Entities are also known as nodes that have properties.

Mapreduce Framework:

MapReduce is an information preparing or parallel programming model presented by Google. In this model, a client indicates the calculation by two capacities, Map and Reduce. In the mapping stage, MapReduce takes the information and encourages every information component to the mapper. In the decreasing stage, the reducer forms every one of the yields from the mapper and touches base at a last result. In straightforward terms, the mapper is intended to channel and change the contribution to something that the reducer can total over. The fundamental MapReduce library consequently parallelizes the calculation, and handles confounded issues like information appropriation, load adjusting and adaptation to internal failure. Gigantic information, spread crosswise over numerous machines, need to parallelize. Moves the information, and gives planning, adaptation to non-critical failure. The first MapReduce execution by Google, and its open-source partner, Hadoop, is gone for parallelizing figuring in substantial groups of ware machines. Map Reduce has picked up an incredible fame as it nimbly and naturally accomplishes adaptation to internal failure. It consequently handles the social event of results over the numerous hubs and returns a solitary result or set. MapReduce model point of interest is the simple scaling of information preparing over various processing hubs. Figure 2 shows working flow of MapReduce framework. With the MapReduce programming model, software engineers just need to determine two functions: “Map” and “Reduce”. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

Using MapReduce, a programmer defines the with only Map and Reduce functions, without having to specify physical distribution of the job across nodes. Advantages [16] of MapReduce framework is Flexible, fault tolerance, High scalability.

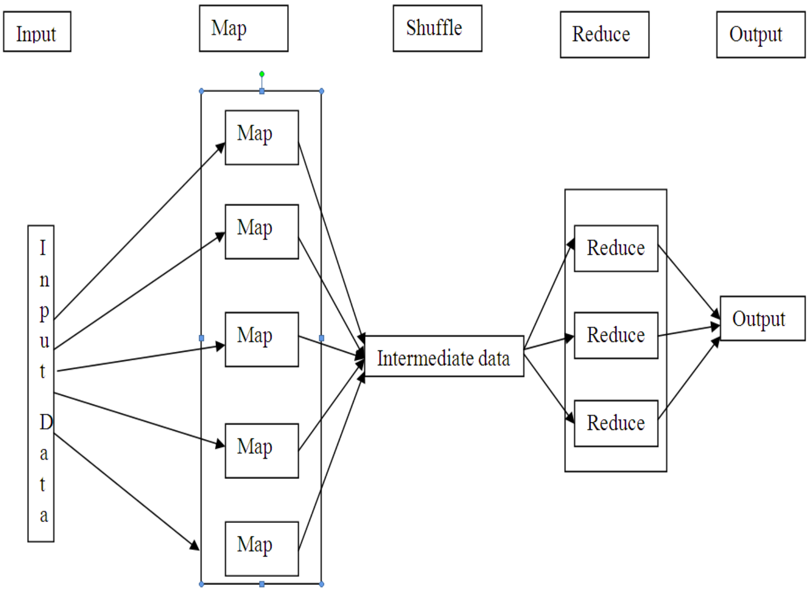

Figure 2: Workflow of MapReduce Freamework [16].

5. Predictive Analytics

An advanced analytics which is used to make predictions about unknown future events comprise a variety of techniques used to predict future outcomes based on historical and current data. Predictive analytics can be applied to any disciplines. Predicting the failure of jet engines based on the stream of data from several thousand sensors is an example best suited or predicting customers’ next guess based on what they usually buy, when they buy and even what they comment on social media [17]. Predictive analytics techniques are sectioned into two groups. One of the techniques includes moving averages that attempt to discover the historical patterns in the outcome variable(s) and extrapolate them to the future. Other technique includes linear regression, aim to capture the interdependencies between outcome variable(s) and explanatory variables, and exploit them to make predictions. Based on the fundamental approach, methods can also be categorized into two groups: regression techniques and machine learning techniques such that neural network.

4.1 Technique that support predictive analytics

Data mining [18] refers to "extracting knowledge from large amounts of data". Raw data from healthcare organizations are huge and assorted. They need to be composed and stowed in the controlled forms, and their integration supports creating of hospital information system. Healthcare data mining provides countless possibilities for hidden pattern investigation from these data sets. These patterns can be used by general practitioner to define, analyzes, diagnoses and treatments for patients. Classification and prediction are the Data mining techniques that assist most common forming depending on the modeling objective and purpose. Classification models predict categorical labels such as discrete and unordered data and these algorithms are used by Decision Trees and Neural Networks. Prediction models predict continuous-valued functions and their algorithms are used by Association Rules and Clustering [19].

Healthcare industry incorporates genomic and clinical data to enable advance personalized medicine. Medical Data related to treatments, genomes, body characteristics and care preferences of earlier patients who had the same condition of the current patients is made available for Practitioners and Biologist to determine what course to prescribe to get the best outcomes.

Viral transmittal of flu and sickness has been a recurrent difficult especially in mass transit situations in airports, railway stations and places of religious gathering, where crowds of strangers intermingle in close proximity from and to different destinations. Remote thermal sensor data anticipate techniques to analyze the spread of viral or bacterial infections in a thick population [20]. The method comprises recognizing a body temperature of passengers crossing the remote thermal sensor placed on their path. This method store the body temperatures of the passerby in a database, conveying at least one geographic characteristic to the stored body temperatures in the database, and associating the body temperatures to a known standard body temperature of individuals.

Web application model [21] for disease diagnosis will describe the disease to the patients and also provide a preventive measure. It is based upon the Naïve Bayes classification algorithm that runs on Apache Mahout, a machine learning based algorithm library that provides data analytic tasks to archive better scalability in term of time and resources. This model advocates diseases very competently resulting in cost effective treatments.

4.2. Machine learning

Machine learning is an artificial intelligence method of discovering knowledge for making intelligent decisions [22]. Big Data has great impacts on scientific discoveries and value creation. Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available to a learning system. These are supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.

Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

5. Text Analytics

Text analytics facilitate healthcare system to convert huge volumes of human generated text into significant reviews, which support proof based policy making. A adaptive information extraction technique “Multiply sectioned Bayesian Network (MSBNs)”, a choice support system for evidence-based administration that enhances the quality and recommends best practice for medical prescription [23]. Evidence-based study delivers the basis for sound clinical practice guidelines and recommendations. It will find and verify evidence from multiple sources, leading to cost-effective use of drugs, improve patients’ quality of life and optimize drug-related health outcomes. Table 3 shows different text analytics techniques from various authors suggested in reliable journals.

Table 3. List of text analytics techniques and methods.

Types of analysis	**Author and Year**	**Techniques**	**Method**	**Result analysis**
Information Retrieval	Nick, *et al*, (2011) [23]	Text mining	Classification	n-gram term frequency
	Harshavardhan,*et al*,(2012) [24]	Data mining	ARX model	Clustering relevance
	Kathy,*et al*,(2013) [25-15]	Natural language processing	Bags of word model	Classification
	Abeed and Gonzalez (2015) [26]	Natural language processing	LDA	Topic model classification

Text summarization	Inderjeet and Bloedorn (1999) [27]	Natural language processing	Information extraction	Information relevance
	Dragomir R., *et al*, (2004) [28]	Data mining	Clustering	Centroid based summarization
	Klaus (2002) [29]	Statistical	Maximum marginal relevance ranking algorithm	Maximum marginal relevance
	Ioannis, *et al*, (2014) [30]	Natural language processing	Named entity mining	Document frequency
Opinion mining	Greg, *et al*,(2003) [31]	Data mining	Clustering	Deviation from mean
	Hoill and Chung (2015) [32]	Data mining	Clustering	Deviation from mean
	Assad, *et al*, (2015) [33]	Link analysis	Classification	Rank order centroid

6. Web Analytics

Web analytics refers to optimizing web usage during the process of collection, analyzing and reporting data from the web for varied purpose. In recent days internet-enabled self-diagnosis for diseases by using search engines and other web tools are just a click away for people around the world who prefers self-medication without consulting a physician.

Most of the times people with internet accessibility surf for information before stepping in to the consulting room. In web application, customer's feedback plays the important role in opining mining. From emergency to non-emergency to everyday preventative health care, location tracking technologies could make a big impact on our health and well-being in the future.

"Knowledge-based dietary nutrition recommendation for obese Management system" as proposed [32] patients based personalized recommendations system. It filters and recommends the dietary nutritional menu gathering with a high relationship that applied to knowledge based collaborative method. The problem of Spares is resolved while using the knowledge-based context-aware modeling.

Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM) [34] gives recommendation to coronary heart disease patients in the form of content. Fuzzy logic and decision tree techniques are used for Coronary Heart Disease prediction model. Consumers' evaluation of hypothetical health recommendation systems [35] designed and provides personalized nutrition advice to consumers, managerial guidance to organizations and public policy makers and whoever wishes to promote the use of complex knowledge based recommendation systems.

7. Multimedia Analytics

Multimedia analytics investigate the audio, video and image data and extract information from unstructured data. Patient's disturbed by communication patterns such as depression, schizophrenia, and cancer are treated with the support of Audio analytics [36]. In addition to this it plays a major role to analyze an infant's health and emotional status [37].

Magnetic resonance imaging (MRI) has become important in brain tumor diagnosis [38]. The Support Vector Machine (SVM) classification integrated with a selection of the features in a kernel space is proposed. It is a successful pattern recognition method especially in the case of high-dimensional data.

8. Social Media Analytics

The practice of gathering data from blogs and social media websites of respective domains and analyzing the collected data is called as Social media analytics. For instance, epidemic outbursts of a disease caused by any viruses or bacteria in any corner of the globe can be predicted and shared to the society using Social media. People share their opinions, experiences and their point of view on various aspects of health related issues and tips related to symptoms, treatments, side effects, experts and other related information. An enormous amount of data on an extraordinary scale is transferred during every share. This social media data openly available is a precious reserve for mining and actionable healthcare perceptions. Social media analysis has two approaches namely content based analysis and structure based analysis. Data posted by users on social media platforms such as customer feedback, product reviews, images and videos are focused by content based analysis. Structure based analysis is a set of nodes and edges signifying contributors and associations.

The perfect source for early-stage flu detection due to its real- time nature [39] was provided by Twitter one among the most prevalent micro blogging service. When flu breaks out in a particular zone, people affected by the flu may tweet the information about the flu which enable the detection of the flu breakout quickly. Social Network Enabled

Flu Trends (SNEFT) framework [40] monitors these messages posted on Twitter with a remark of flu indicators to track and envisage the emergence and how an influenza epidemic in a inhabitants spread. Moreover by observing the tweets related to health, health experts ranked them by using the concept of hubs and authorities, Twitter data offer users, an opportunity to interact with the health experts for consultation.

A feasible collaborative health communities where developed for the patients can acquire health information and pursue assistance from the professionals without any cost by merging the predictive modeling approaches and the social media networks [23]. Probabilistic clustering [41] is an effective way to filter a large chunk of outliers in the feature space and select high authority users on which ranking can be applied more robustly.

9. Discussions

Today industries of any domain have loads of data that is either too big or unstructured to be managed and scrutinized through traditional methods. Healthcare escalating sources are from call centre voice data to genomic and proteomic data from biological research, medical transcription records in the form of audio data and medical histories of patients. "Big data" is a slogan for cleverer, more perceptive data analysis that pave way to treatments and cures for threatening diseases. Big data tools saves time and enhance the performance of the output as required by varied stakeholders. Ability to understand high scale data sets that are generated by emergent technologies in biomedical research dealing with the cumulative amounts of omics data, combined with clinical information will depend on scalable structure.

Semantic data obsessed method is the need of the hour to create a system that marks big data vital and keen for healthcare benefactors and patients. This technique can lead to more functioning clinical decision-making, amended health upshots, and eventually reduces the costs involved in healthcare. Geographic location based healthcare facilities will be much more helpful in healthcare system in near future. This is relatively possible with the help of mobile; sensors and social media generated data related to biological patterns providing new proportions of situation, geolocation, behavior pattern.

Pharmaceutical firms have been accumulating years of research and development into medical databases to formulate effective replacements in drugs and medical consultants and pathology labs have digitized their patient records. Last decade physicians traditionally used their successive case histories while making treatment decisions, but in recent years there has been a shift toward evidence-based medicine, which involves scientifically swotting clinical data and generating treatment decisions based on the preeminent obtainable evidence.

Huge number of structure and unstructured data are affected by context in many ways [42]. A context aware research issue becomes more challenging task while handling big data processing.

Stream processing is a big challenge even among big data users. The data generated by real time event handles parallel processing and will stream at a rate of millions of events per seconds. To perform this, event correlation Complex Event Processing engine is used to extract the meaningful information from this moving stream.

10. Conclusion

The computational complexity in healthcare system becomes high as it generates voluminous, variety of data from different data sources like EHR, Internet and social media to predict epidemics, cure disease, improve quality of life and avoid preventable decease. As traditional data processing becomes inadequate to tackle these data dimensions, big data, a much less expensive to own and operate technique replaced the traditional relational database. But in near future our biggest challenge will be reductions in the cost to capture and store data. Effectively enhancing the features of Hadoop will be a huge success to efficiently store and process large quantities of data. Ample research questions, testing and application are yet to be carried out in realizing this future needs.

References

1. Chen H, Chiang RH, Storey VC. Business Intelligence and Analytics: From Big Data to Big Impact. MIS quarterly. 2012 Dec 1;36(4):1165-88
2. Hanauer DA, Preib R, Zheng K, Choi SW. Patient-initiated electronic health record amendment requests. Journal of the American Medical Informatics Association. 2014 Nov 1;21(6):992-1000.
3. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Networks and Applications. 2014 Apr 1;19(2):171-209..
4. Thusoo A, Sarma JS, Jain N, Shao Z, Chakka P, Anthony S, Liu H, Wyckoff P, Murthy R. Hive: a warehousing solution over a map-reduce framework. Proceedings of the VLDB Endowment. 2009 Aug 1;2(2):1626-9.
5. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and Challenges of Big Data Computing in Health Sciences. Big Data Research. 2015 Mar 31;2(1):2-11.
6. http :/ /hbase .apache .org/
7. http :/ /cassandra .apache .org/
8. http://hypertable.org/
9. https://cloud.google.com/bigtable/
10. https://aws.amazon.com/simpledb/

11. http :/ /www.mongodb .com/

12. http:// couchdb.apache.org/

13. http://neo4j.com/

14. http://www.objectivity.com/products/infinitegraph/

15. www.orientechnologies.com/orientdb/

16. Padhy RP. Big data processing with Hadoop-MapReduce in cloud systems. International Journal of Cloud Computing and Services Science. 2013 Jan 1;2(1):16.

17. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015 Apr 30;35(2):137-44

18. Han J, Kamber M. Data mining: concepts and techniques.

19. Kleissner C. Data mining for the enterprise. InSystem Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on 1998 Jan 6 (Vol. 7, pp. 295-304). IEEE.

20. Reinpoldt MA, inventor; Thermal Matrix USA, Inc., assignee. Method and system for the acquisition, transmission and assessment of remote sensor data for trend analysis, prediction and remediation. United States patent US 8,519,850. 2013 Aug 27.

21. Yu WD, Pratiksha C, Swati S, Akhil S, Sarath M. A Modeling Approach to Big Data Based Recommendation Engine in Modern Health Care Environment. InComputer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual 2015 Jul 1 (Vol. 1, pp. 75-86). IEEE.

22. Wang L. Machine Learning in Big Data. International Journal of Advances in Applied Sciences. 2016 Apr 1;4(4).

23. Cercone N, An X, Li J, Gu Z, An A. Finding best evidence for evidence-based best practice recommendations in health care: the initial decision support system design. Knowledge and information systems. 2011 Oct 1;29(1):159-201

24. Achrekar H, Gandhe A, Lazarus R, Yu SH, Liu B. Twitter Improves Seasonal Influenza Prediction. InHEALTHINF 2012 (pp. 61-70).

25. Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013 Aug 11 (pp. 1474-1477). ACM.

26. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. Journal of biomedical informatics. 2015 Feb 28;53:196-207.

27. Mani I, Bloedorn E. Summarizing similarities and differences among related documents. Information Retrieval. 1999 Apr 1;1(1-2):35-67.

28. Radev DR, Jing H, Styś M, Tam D. Centroid-based summarization of multiple documents. Information Processing & Management. 2004 Nov 30;40(6):919-38.

29. Zechner K. Automatic summarization of open-domain multiparty dialogues in diverse genres. Computational Linguistics. 2002 Dec;28(4):447-85.

30. Kitsos I, Magoutis K, Tzitzikas Y. Scalable entity-based summarization of web search results using MapReduce. Distributed and Parallel Databases. 2014 Sep 1;32(3):405-46

31. Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering. Internet Computing, IEEE. 2003 Jan;7(1):76-80.

32. Jung H, Chung K. Knowledge-based dietary nutrition recommendation for obese management. Information Technology and Management.:1-4.

33. Abbas A, Ali M, Khan MU, Khan SU. Personalized healthcare cloud services for disease risk assessment and wellness management using social media. Pervasive and Mobile Computing. 2015 Nov 11.

34. Kim JK, Lee JS, Park DK, Lim YS, Lee YH, Jung EY. Adaptive mining prediction model for content recommendation to coronary heart disease patients. Cluster Computing. 2014 Sep 1;17(3):881-91.

35. Wendel S, Dellaert BG, Ronteltap A, van Trijp HC. Consumers' intention to use health recommendation systems to receive personalized nutrition advice. BMC health services research. 2013 Apr 4;13(1):126.

36. Hirschberg J, Hjalmarsson A, Elhadad N. "You're as Sick as You Sound": Using Computational Approaches for Modeling Speaker State to Gauge Illness and Recovery. InAdvances in Speech Recognition 2010 Jan 1 (pp. 305-322). Springer US

37. Patil HA. "Cry Baby": Using Spectrographic Analysis to Assess Neonatal Health Status from an Infant's Cry. InAdvances in Speech Recognition 2010 Jan 1 (pp. 323-348). Springer US.

38. Hsieh TM, Liu YM, Liao CC, Xiao F, Chiang IJ, Wong JM. Automatic segmentation of meningioma from non-contrasted brain MRI integrating fuzzy clustering and region growing. BMC medical informatics and decision making. 2011 Aug 26;11(1):54.

39. Li J, Cardie C. Early stage influenza detection from twitter. arXiv preprint arXiv:1309.7340. 2013 Sep 27.

40. Lee K, Agrawal A, Choudhary A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining 2013 Aug 11 (pp. 1474-1477). ACM.

41. Pal A, Counts S. Identifying topical authorities in microblogs. InProceedings of the fourth ACM international conference on Web search and data mining 2011 Feb 9 (pp. 45-54). ACM.

42. Bagheri H, Shaltooki AA. Big Data: challenges, opportunities and Cloud based solutions. International Journal of Electrical and Computer Engineering. 2015 Apr 1;5(2):340.

Corresponding Author:

P.Shanmuga Sundari*,

Email: *sundari.sigamani@vit.ac.in*