ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

BIG DATA: A GOLDMINE IN HEALTHCARE

[1]Hiren.T.Modi, [2]Priya.S
[1]Third year, B.Tech / Department of CSE, SRM University, Kattankulathur-603 203, Tamil Nadu, India.
[2]Assistant Professor, Department of CSE, SRM University, Kattankulathur-603 203, Tamil Nadu, India.
Email: hirentmodi@gmail.com

Received on: 18.10.2016 *Accepted on: 11.11.2016*

Abstract

Healthcare and predictive analysis is an emerging goldmine of information. With the current digitization of medical records, case histories and outcomes of the numerous treatments so provided, big data has paved way for predictive analysis in the field of medicine. The idea is to extract knowledge from large amounts of medical records and data using complex algorithms and data analytics so as to come up with the best possible treatment, given a medical condition. Using advanced recommendation systems and big datasets of medical records as input parameters, a suitable treatment strategy can be adopted. Highest threshold of impact on account of such an analysis can be achieved by integration various forms of medical data, ranging from scans to tests. Using big data ensures intercompatibility of unstructured data. Data in the size of even petabytes has become a strategic asset.

Keywords: Predictive analysis, digitization, medical records, complex algorithms, data analytics, data sets.

1. Introduction

The field of healthcare has been craving for predictive analysis for decades. Big data gives us just the right set of tools and factors to do so. The main idea of this paper is to present a brief understanding of how big data gathering and its predictive analytics can revolutionise the way our current medical world works. The basis lies in understanding the needs of a medical professional and implementing those needs to extract petabytes of medical records [1]. Data analytics then comes into the picture, bringing along with it the ability to provide the best possible treatment strategy.

The key is in gathering sufficient medical records. E-records are now the go to solution in the medical domain. Sources range from government medical databases to NGO's maintaining medical portfolios. What we are trying to achieve is not merely data extraction but gather information from the mined data. Say we were to plot a patient illness to treatment graph for millions of patients from the mined data; it would throw light on the most preferred treatment a

medical professional would choose corresponding to the illness. A similar extension may be through the analysis of treatment is to outcome graph corresponding to a particular illness. This is what will give us an idea of the true potential of big data. What we would receive here is the ability to gauge using prediction the most effective treatment, given an illness.

Various other factors including a patient's duration of exposure to the illness, geographic location, age group, ethnicity, genetic information my act as further filters to the above plots. Doctors can then develop case studies as to why a specific treatment is proving to be efficient than the other and so on. The ideas seem to be endless.

The remainder of this paper is organized as follows. Section II describes about understanding the implementation workflow, whereas Section III describes the data analytics. Section IV describes the data report/visualization, whereas Section V describes the indicative example, then Section VI about conclusion and Section VII describes the future implementation.

2. Understanding The Implementation Workflow

Medical domains:

The field of medicine has an overwhelming number of domains which need attention. New diseases, unknown infections; the list is never ending. Some areas have been a pivot of medical research from centuries. Unlocking the hidden mystery behind the effectiveness of various drugs and treatments seems to be a flattering option.

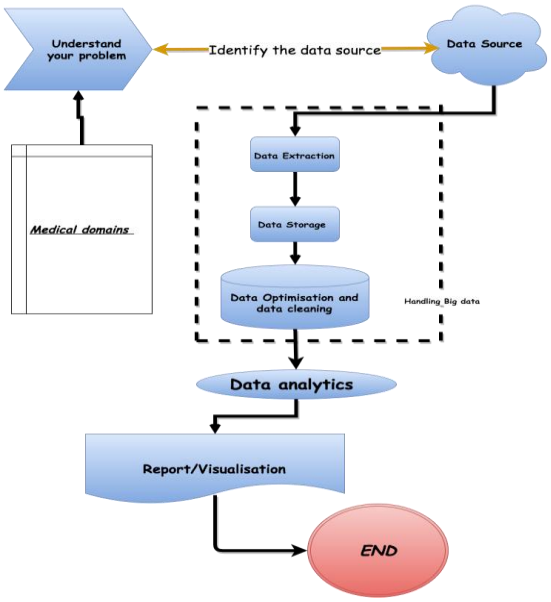

Fig. 1: Implementation Workflow.

A. Understand your problem

With an astounding 120 plus specialties (aamc.org) in the medical realm the requirement of predictive analysis becomes more evident.

1. Doctor must isolate and classify the patient's case.
2. He/She must then understand the cause/reason of the case.
3. Making a comprehensive case study is key to understanding the problem better.
4. Determine what has to achieved at the end of it all.
5. The right data source has to be chosen then.

B. Data source

This procedural step will come across as the biggest challenge to the medical professionals. The lack of abundant sources of E-records can be the most critical drawback of such a scheme. However the fields of health science are emerging rapidly. E-medical records are no longer unheard of. Predictive analytics would reach its pinnacle if governments and human services globally, join hand for the cumulative good of mankind. A global database of E-records and a regulating body to maintain its functioning and define its norms is crucial [2]. The ability to have petabytes of medical data for analysis and input for recommendation systems would take Big data's implementation in the field of healthcare to whole new level.

C. Data extraction and storage :-

Petabytes of data becomes difficult to both store and extract. The likes of open source software HADOOP and its data extraction implementation by MapReduce algorithm, have created and new spur in the field of data mining. The requirement of multi-million dollar work frames, processors and databases has been completely eliminated. Hadoop along with its ecosystem now has the ability to meet the needs of anyone having the most minimalistic amount knowledge. Hadoop works on the basis of distributed processing across clusters. With the upcoming era of cloud computing the applications seem limitless. Both the patients and medical professionals can access this data on the go.

D. Data optimization and cleaning :-

HDFS:

When we deal with the task of accessing big data that is distributedly stored on a cluster, a cluster file system is the goto solution. The cluster file system provides location access to data files to the servers on the cluster. Hadoop Distributed File System (HDFS) is a popular type of cluster file system which is designed for storing large amount of data across machines in a large scale cluster [3]. HDFS was originally originated from Google Files System (GFS) paper [4]. It provides an open source cluster file system similar to GFS. HDFS makes use of a master-slave architecture. HDFS logically separates the file system metadata and application data. The metadata is stored on a delicate computer named as NameNode (known as mater node in GFS) in HDFS. Application data was stored on

other computers named as DataNodes (known as slave node in GFS). A data file is divided into one or more blocks and these divided blocks are replicated and stored across several DataNodes. All of the nodes contained with a Hadoop cluster are full connected and communicate with each other using TCP-based protocols. The NameNode maintains the file system namespace and the mapping of file blocks to DataNodes. When an HDFS client read a file, it first contacts the NameNode to get the locations of data blocks comprising the file and then reads these data blocks from closest DataNode. Compared to traditional distributed file systems, HDFS have two important advantages:

(1) Highly fault-tolerant. Unlike traditional distributed file systems that use data protection mechanisms to make the data durable, HDFS provides replicated storage for massive data across multiple DataNodes and heartbeat for failure detection. When a DataNode fails, the NameNode will detect it and re-assign the work to other DataNodes;

(2) Large Scale Data. The Hadoop clusters today can store Petabyte (PB) data. It supports high aggregate data bandwidth and scale to hundreds of nodes in a cluster.

Map Reduces:

Map Reduce [5] is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A popular open-source implementation of MapReduce framework is Apache Hadoop. The MapReduce is inspired by the map and reduce functions that are commonly used in functional programming. In a MapReduce framework, a programming consists of two primitive steps: Map() and Reduce(). In the Map() step, the master node takes the input and divides it into smaller sub-problems, and distributes them to slave nodes. A slave node may do this Map () procedure again for further dividing the problem. The slave node processes smaller problem and passes answer back to its master node. In the Reduce() step, the master nodes collects the answers and combines them together to form the final answer to the original problem that it wants to solve. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved when processing massive data on a large cluster. MapReduce greatly simplifies the task of writing a large-scale analysis on distributed data for many types of analysis.

3. Data Analytics

Different types of Big Data require different analysis methods. We choose three widely used analysis methods in computer science and biomedicine to share with the readers:

A. Recommendation System;

B. Deep Learning and

C. Network Analysis.

A. Recommendation system

Recommendation systems have become extremely common in recently years. They are being intuitively applied in various applications, such as recommending products on Amazon.com [6], movies recommendation by Netflix [7] and MovieLens [8], music recommendation by Last.fm and Pandora Radio, and news recommendation by VERSIFI Technologies. In a recommendation system, there are two classes of entities: users and items. Let U be the set of all users and I be the set of all available items. Let R be the rating matrix, where r(u,i) denotes the rating of an user u to an item i. The value of r(u,i) indicates how a particular user likes a particular item. Usually, the rating matrix R (also known as utility matrix) is sparse, meaning that most entries in the matrix are "unknown". An "unknown" entry implies that we do not explicitly know a user's rating (i.e. preference) for an item. The goal of a recommendation system is to estimate the values of these "unknown" entries.

Once the rating matrix is estimated, for each user, we can select top N items with highest ratings and recommended them to the user. There are several ways to estimate the "unknown" entries in the rating matrix. Usually, recommendation systems are classified into three different categories based on their approaches to estimate the "unknown" ratings: (1) Collaborative filtering approaches, (2) Content-based filtering [9] and (3) Hybrid approaches. Collaborative filtering approaches are based on analysing a large amount of data about users' rating, behaviours or activities over items. The items recommended to a user are those preferred by other similar users. Content-based filtering approaches are based on a description of the item and a profile of the user's preference. They recommend the items that are similar to the items that a user liked in the past. The hybrid approaches combine collaborative and content-based methods.

B. Deep learning

Since 1980s, many machine learning methods, such as Neural Network (NN), Random Forest, Support Vector Machine (SVM), have been constructed to build classification models [10]. Then it was found that without better features, the performance of classifiers is difficult to improve. Actually, the features determine the upper bound of a classifier's performance. Since 2000, Transfer Learning [11], Manifold Learning [12] etc. were developed to learn the feature structure.

However all these methods need hand craft and require experiences of tricks in practice. In 2006, Hinton et al. published the historical paper "Reducing the dimensionality of data with neural networks" which is considered as the beginning of deep learning [13]. Unlike shallow machine learning models, deep learning uses Neural Network with many layers of hidden variables to automatically do feature learning, including pre-processing, feature extraction and

feature selection. The unsupervised feature learning models in deep learning include Autoencoder [14], Restricted Boltzmann Machine [15] and Deep Boltzmann Machine [16]. Deep learning has been quickly applied to many industrial projects and created big value, such as the speech recognition and Xbox from Microsoft, image recognition, Natural Language Processing (NLP) from Google.

C. Network analysis

Many unstructured complex data can be organized as a graph. The graph theory provided solid theoretical foundation for network analysis[17]. Several widely used network analysis methods are available.

4. Data Report/Visualization

Graphical presentation is the best way to diagrammatically get the meaning of the data and uncover the hidden truth by analysis. There are many tools to visualize the Big Data. As a generalized programming language, R has about six thousand high quality packages [18] that could achieve even the most complex functions. Its excellent help system and power functions make it the most widely used language in DataScience. For visualization R packages, ggplot2 and igraph provide general plot functions and network specific plot functions. For network specific visualization, other choices include Cytoscape [19] which can save the network layout and provide rich visualization styles, Circos (http://circos.ca/) which becomes the standard of visualizing genome chromosomes, Gephi [20] which is a java based application to create dynamic and hierarchical network graphs and GraphViz [21] which is a powerful command-line tool that can layout big network with massive numbers of nodes and edges. R provided wrappers for most visualization software, for example, RCytoscape to Cytoscape, RCircos to Circos, RGraphViz to GraphViz. For general visualization which does not require programming skills, Tableau (http://www.tableausoftware.com/) is a good choice. Unlike previous ones, it is commercial software, but has try-it-free version. Its highlight feature is that it can visualize Google map style location data.

5. An indicative example

Consider we have a patient A.

- Patient A has been suffering from, say blood cancer.
- Now patient A was obviously diagnosed first and hence has a scan report proving his case.
- Patient A also has a medical professionals report and prescription for the scan.
- Before the professional jumps to a conclusion he/she would have definitely made up a rigorous case study of patient A along with all the symptoms, signs and vitals.

- Now if the medical professional was to provide patient A with the best possible treatment then he will have to carry out extensive research in constraints of the patient's parameters.
- Assume there is a presence of a comprehensive database of medical records in a globally accessible database.
- The doctor with suitable input parameters will then obtain the respective dataset.
- Now data from the dataset will have to be filtered out. This is where the patient's portfolio comes into play. Given the patient's age, genetic background, type of cancer, stage of cancer, symptoms etc as input parameters, we will have a fine tuned big dataset that will act as an input for analytics and data representation.
- Let's say we have the ability to produce a graphical representation of the dataset, with treatment and effectiveness being the parameters of the plot, and then a simple analysis would give us a comprehensive insight as to which treatment is best suitable corresponding to the ven patient considering all the factors which essentially have to be taken into account.
- The scope is endless. Assume a plot between the patient's treatment and duration or even treatment and success rates; the analysis of such plots by overlapping reveals extensive magnitude of information, digging out the hidden truth behind how powerful big data analytics can be and why it is a hidden goldmine.

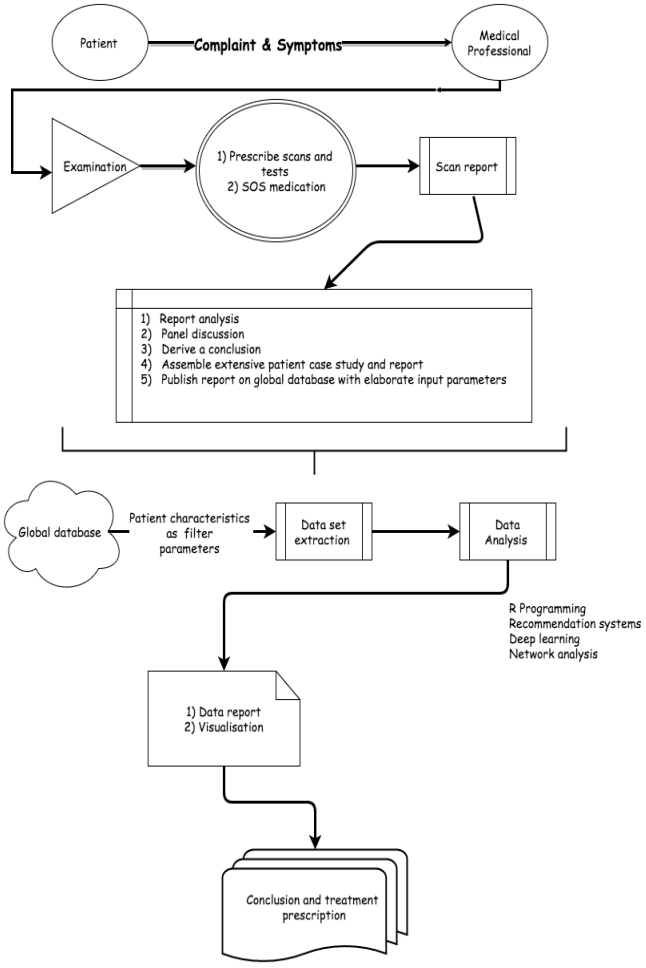

Fig. 2: **Flow diagram for the indicative example.**

6. Conclusion

Big data is truly a hidden goldmine. It consists of a wealth of information that may completely change the way the world works. The scope is not restricted to the field of medicine. Finances, stock market, gambling etc are just a few other domains where such a predictive analysis and implementation will be a game changer. Imagine having the ability to gauge the growth in value of particular stocks with the confidence of petabytes of factual data supporting your analysis. The list is endless we just need the ability to understand our problem, obtain the relevant dataset and carry out the corresponding implementation

7. Future Implementation

With improvement of big data handling, mining and analytics techniques the data report and visualization can be expanded to a whole new level. More number of input parameters and filters on the basis of patient's unique traits will make this entire process more efficient and more personalized. Overlapping graphs or visual representations will give an extensive ability to carry out comparative study. The impact on research will bring out new medication, which is not only more effective but also caters to patient's exact requirements if need be. Medical professionals will gain the confidence to prescribe a treatment which is almost guaranteed to work.

References

1. L. Fernandez-Luque, R. Karlsen, L.K. Vognild, Challenges and opportunities of using recommender systems for personalized health education, Stud. Health Technol. Inform. 150 (2009) 903–907.
2. L. Duan, W.N. Street, E. Xu, Healthcare information systems: data mining methods in the creation of a clinical recommender system, Enterp. Inf. Syst. 5 (2011) 169–181.
3. S. Ghemawat, H. Gobioff, S.-T. Leung, The Googlefile system, SIGOPS Oper. Syst. Rev. 37 (2003) 29–43.
4. K. Shvachko, K. Hairong, S. Radia, R. Chansler, The hadoop distributed file system, in: 2010 IEEE 26th Symposium on Mass Storage Systems and Tech-nologies (MSST), 2010, pp.1–10.
5. J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clus-ters, Commun. ACM 51 (2008) 107–113
6. G. Linden, B. Smith, J. York, Amazon.com recommendations: item-to-item col-laborative filtering, IEEE Internet Comput. 7 (2003) 76–80.
7. Y. Koren, Tutorial on recent progress in collaborative filtering, in: Proceed-ings of the 2008 ACM Conference on Recommender Systems, ACM, Lausanne, Switzerland, 2008, pp.333–334.

8. B.N. Miller, I. Albert, S.K. Lam, J.A. Konstan, J. Riedl, MovieLens unplugged: experiences with an occasionally connected recommender system, in: Pro-ceedings of the 8th International Conference on Intelligent User Interfaces, ACM, Miami, Florida, USA, 2003, pp.263–266.
9. M. Balabanovi´c, Y. Shoham, Fab: content-based, collaborative recommenda-tion, Commun. ACM 40 (1997) 66–72.
10. N.M. Seel, Encyclopedia of the Sciences of Learning, Springer-Verlag, New York, 2012.
11. P. Sinno Jialin, Y. Qiang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (2010) 1345–1359.
12. M. Balasubramanian, E.L. Schwartz, The isomap algorithm and topological stability, Science 295 (2002).
13. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
14. Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
15. G. Hinton, A practical guide to training restricted Boltzmannmachines, in: G. Montavon, G. Orr, K.-R. Müller (Eds.), Neural Netw., Tricks Trade, vol.7700, Springer, Berlin, Heidelberg, 2012, pp.599–619.
16. R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), 2009, pp.448–455.
17. L.C. Freeman, Centrality in social networks: conceptual clarification, Soc. Netw. (1979) 215–239.
18. R. Ihakaa, R. Gentleman, R:language for data analysis and graphics, J. Com-put. Graph. Stat. 5 (1996) 299–314.
19. P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for inte-grated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
20. M. Bastian, S. Heymann, M. Jacomy, Gephi: an open source software for ex-ploring and manipulating networks, in: International AAAI Conference on Weblogs and Social Media, 2009, http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.
21. P.T. Shannon, M. Grimes, B. Kutlu, J.J. Bot, D.J. Galas, RCytoscape: tools for exploratory network analysis, BMC Bioinform. 14 (2013) 217.