ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Evidence Based Disease Analysis using Big Data.

RM Gomathi[1]*, Obin Joseph[2], and Vikash Kumar Gupta[2]

[1]Assistant Professor, [2]UG Scholar, Dept of Information Technology, Faculty of Computing, Sathyabama University, Chennai, Tamil Nadu, India

ABSTRACT

Data mining based disease analysis is usually done for a structured data. Currently there is no drug analysis based on evidence gathered and Analysis by using Big data is yet not achieved. In this paper, achieving evidence based drugs analysis is done using big data. This procedure is possible by gathering of medical evidences, grouping of data, Mapping of disease data set and Medicines, and report generation. Machine learning technique is used for the disease discovery and Medicine analysis is achieved by appropriate evidence.

Keywords: Big data, Healthcare, Disease based Data Grouping.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Big Data has ability to grasp the increasing volume of data. With the help of this we can analyze and collect data that was very strenuous few years ago. Big Data is not only used for amplifying data but also the swiftness by which it is created and the types of data we examine. It generates values from dataset of enormous volume (of both unstructured and structured data) that cannot be analyzed using conventional computing techniques. Report by Ricardo Baeza-Yates entitled "BigData: Promises & Problems"[1] issued in March. 2015. Additional advancement now propose big data complication are recognize by "5V": Volume (size of data), Variety (heterogeneous data), Velocity (high data production), Veracity (secure data provenance), and Value (in the data). Big data analytics has a promising future when it comes to healthcare and well being. It can provide information from an enormous data set and can also reduce cost spent on healthcare. Healthcare centre have collected a huge amount of data. Most of them are stored in the form manuscript or typewritten document [9]. The digitization in today's world has made them to store in data base, which has the potential to refine the current standard of healthcare and also reduce the cost. This huge size of data can support a vast variety of healthcare purpose and responsibility [10]. A survey report says that US has reached 150EB in 2011 and soon it will reach zettabyte.

In 2008, Google.org created a web service called Google flu trends which was used to estimate influenza rate in more than 25 countries. It used Google queries to accurately predict about flu illness rate. It had an accuracy of 97% on which some analysts claimed that all the modern healthcare problems can be solved by using big data.

In 2005, a framework was created by using certain methods and technologies and was termed as Virtual Physiological Human (VPH). Once it is initiated it can describe human body as a unit complex system. The plan was to sub-divide the complex human body into many parts (body part, organs) and examine it separately from others. Though the system worked well for individual parts, it was very difficult to deal with multi organ illness.

RELATED WORKS

Doug laney, In his paper"3D Data Management"[2] proposed the importance of big data in today's business world. In his paper he has shown e-commerce is benefited by the use of big data. He has also shown how e-commerce has exploded data management challenges among variety, velocity and volume.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, In their paper "Detecting influenza epidemics using search engine query data"[3] have tried to detect influenza disease using google search engine. They analyzed search query to find the current population of user suffering from influenza disease.

J. W. Fenner, B. Brook, G. Clapworthy, In their paper "STEP and a roadmap for the VPH"[4] have created a framework which can describe human beings as a silico. They have shown that with this paper they can improve the current computational method for investigating the human body.

PROPOSED SYSTEM

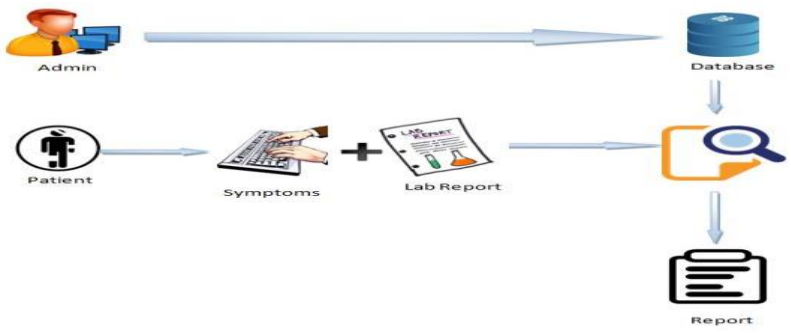

Figure 1 System Architecture

ISSN: 0975-8585

In the proposed structure, achieving evidence based drugs analysis is done using big data. This procedure is possible by gathering of medical evidences, grouping of data, Mapping of disease data set and Medicines, and report generation. Machine learning technique is used for the disease discovery and Medicine analysis is achieved by appropriate evidence [5]. The modules used are:

1. Patient Data Gathering
2. Multi Access Control
3. Disease based Data Grouping
4. Mapping of disease data set
5. Providing optimal medicines

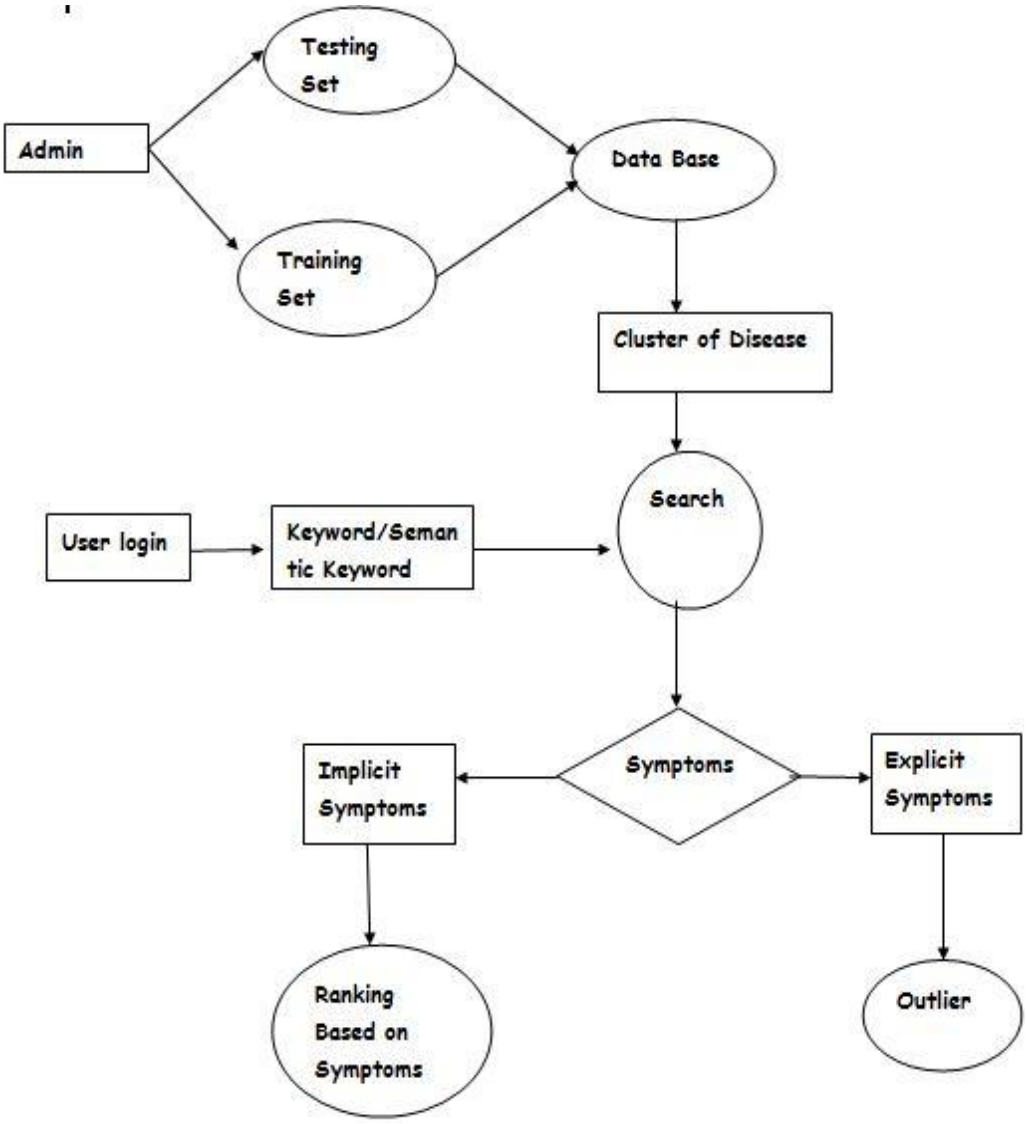

Figure 2 Data flow diagram

Patient Data Gathering

Collecting medical data is the key part of standard health-care for patients. Treatment data is collected from any of the healthcare center and is kept in the database. The administrator has the authority to manage the entire database and supervises the proper application and retrieving of data. Medical data comprises of information such as the given treatment [6-8], symptom of the disease, and laboratory report.

Multi Access Control

In this module we design to implement the two different type of account. First one is controlled by the administrator and the second is for the user.

Administrator

In the administrator's account, we have provided the authority to manage the patient data. This includes adding of information such as the given treatment, symptom of the disease, and drugs for the disease. The administrator can also update new drugs for the treatment of the disease.

Patient Account

In the patient's account, the user has to sign-up first providing all the necessary information. Now the patient can login using his/her user name and password. Now the user can provide all the symptoms of the disease with the laboratory blood report.

Disease based Data Grouping

In this module we will categories the large amount of data, based on the symptoms of disease. By doing this we can easily get the disease detail and disease can be then identified and they can be reviewed later for the future medical use.

Mapping of disease data set

In this module, disease data is mapped using user's given disease symptoms and related laboratory blood report. Machine learning algorithm is used for the diagnosis of the disease with reference to the user's symptoms of disease and laboratory blood report. System will automatically identify the disease details using machine learning algorithm.

Providing optimal medicines

In this module, we can recommend the best medicine for the identified disease. Medicines are optimized on the basis of previous patients' feedback. Feedback from the patients work as a evidence.

PROPOSED ALGORITHM

K-means clustering algorithm

K-means is used to solve clustering problem and is one of the effortless unsupervised learning algorithm. The methodology follows a simple way to categorize a specified dataset by a fixed no. of clusters (k-clusters). The idea is to provide each cluster a center (i.e. k-center), one for every cluster. Next, the step is to take all the point related to a specified dataset and link it to the closest center. If none of the point is pending, then first step is finished. Now we have to recalculate baby centers (k-new centroids) of the cluster developed from the earlier step. Since now we have the k-new centroids, binding has to be done once again among the identical dataset point and the closest recently developed center. Due to this a loop has been formed. Now we can notice the change in the location of k-centers at every step all changes are over. Finally, k-means clustering algorithm is used to reduce the objective function know as squared-error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

ISSN: 0975-8585

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

Algorithmic steps for k-means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'ci' represents the number of data points in ith cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

EXPERIMENTAL RESULT AND OUTPUT

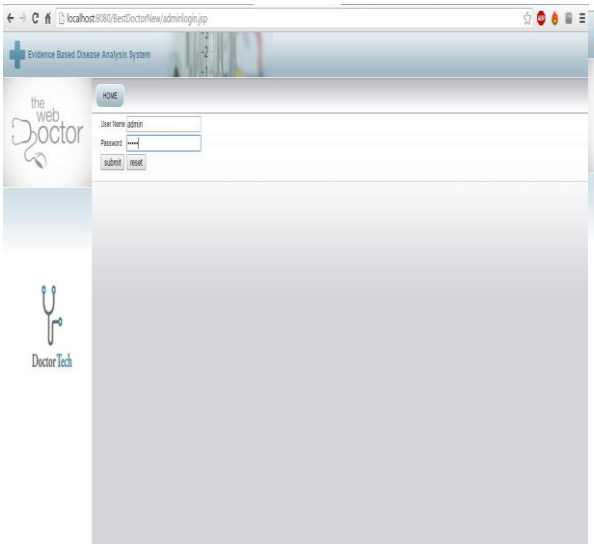

Figure 3 Admin Login

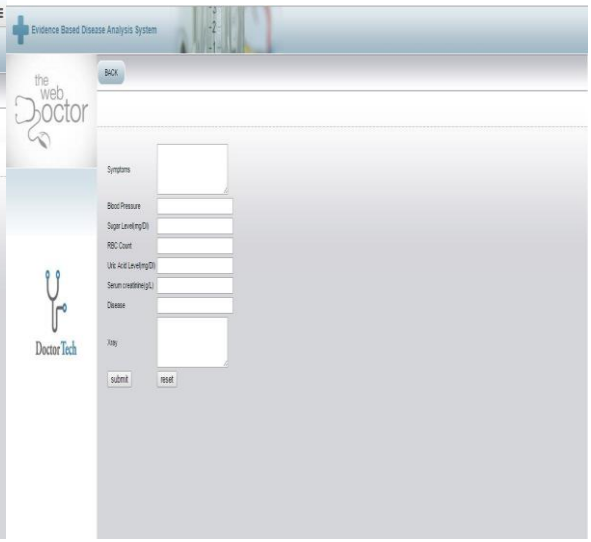

Figure 4 Adding disease symptoms

ISSN: 0975-8585

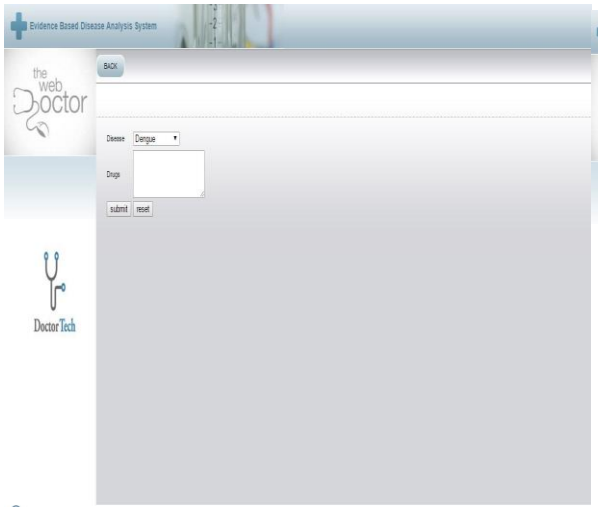

Figure 5 Adding Drugs

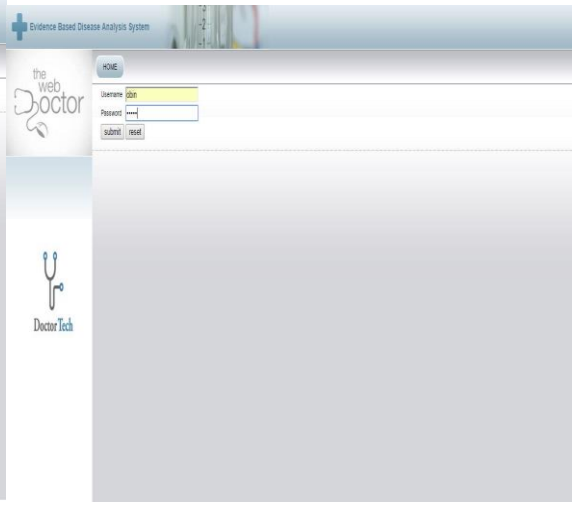

Figure 6 Patient login

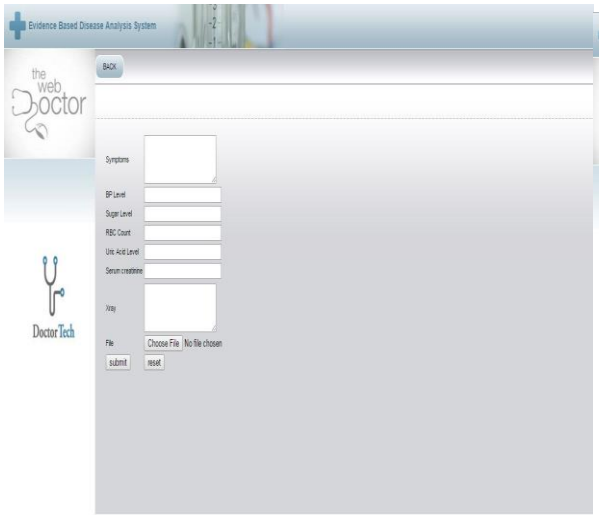

Figure 7 Finding Disease

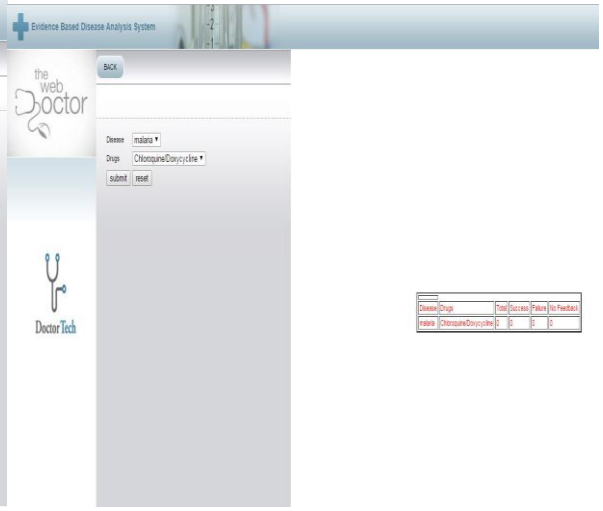

Figure 8 Finding Drugs

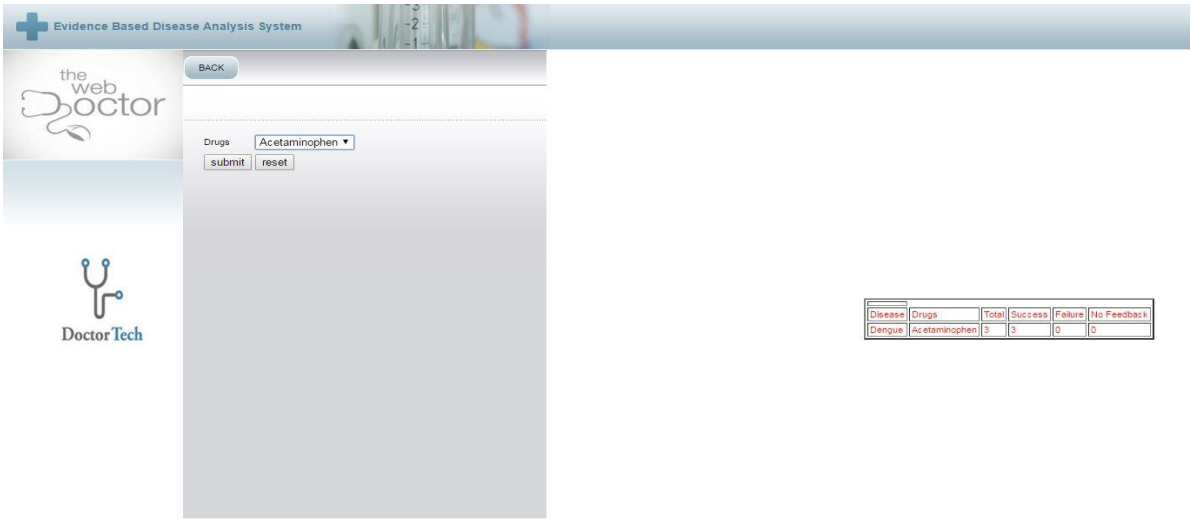

Figure 9 Drug Feedback

ISSN: 0975-8585

CONCLUSION

Although big data technologies are sometimes overrated, they have a great future in the field of biomedicine, though the development should be done with the blend of various strategies and not for the sake of rivalry. Today big data provides a crucial role where it is used to make better health-profile and models which can be used to diagnose and treat disease. By using clustering in this project, the given dataset is separated into identical groups such that same types of data are kept in a group and dissimilar data in a different group. This helps in the accurate disease analysis and medicine discovery. Also the paper is implemented on the crown of Hadoop technologies which sustain the java language.

ACKNOWLEDGEMENT

I thank Sathyabama University for providing us with various resources and unconditional support for carrying out this work.

REFERENCES

[1] VN Gudivada, R Baeza-Yates, VV Raghavan. Computer 2015; 48(3): 20 – 23.
[2] D Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Reference publication, META Group, 2001, 949.
[3] J Ginsberg, MH Mohebbi, RS Patel, L Brammer, et al. Nature 2009; 457(7232): 1012-1014.
[4] JW Fenner, B Brook, G Clapworthy, PV Coveney, et al. Philos Trans A Math Phys Eng Sci 2008; 366(1878): 2979-2999.
[5] O Terzo, P Ruiu, E Bucci, and F Xhafa. IEEE International Conference on Complex, Intelligent, and Software Intensive Systems 2013; 475-480.
[6] B Wixom, T Ariyachandra, D Douglas, et al. Communications of the Association for Information Systems 2014; 34(1): 1-13.
[7] A Wright. Communications of the ACM 2014; 57(7): 13-15.
[8] J Manyika, M Chui, B Brown, J Bughin, et al. Big data: The next frontier for innovation, competition, and productivity. Reference publication, Mckinsey, 2011, pp. 1-143.
[9] K Priya. Journal of Theoretical and Applied Information Technology 2015; 73(2): 296-300.
[10] RM Gomathi, J Martin Leo Manickam, T Madhukumar. IEEE International Conference on Innovation, Information in Computing Technology-ICIICT'15 2015; 1-8.