Big Data Implementation of Drug Query with Disease Prediction, Side Effects and Feed Back Analysis

S. Priya Dharshini, V. S. Priya Dharshini, B. Renuka, M. Tamil Thendral

Abstract: ***Big data is an all-encompassing term for any collection of data sets which are large and complex. This project deals with the integration of big data with cloud computing. In the existing system designing question answering system requires efficient and deep analysis of natural language questions. The disease cannot be analyzed properly if the information is incomplete and noisy in nature. Here a tool is provided to assist professionals and consumers in finding and choosing drugs. An approach is developed that allows a user to query for drugs that satisfy a set of conditions based on drug properties, such as drug indications, side effects, and drug interactions, and also takes into account patient profiles. The appointment of the best doctor is scheduled for the consultation based on user feedbacks. The best drug is advised to the specific patient based on their disease through Big Data analysis. User can post a query through system or through Android Application.***

Keywords: Big Data, System Designing, Feedbacks. Android Application.

I. INTRODUCTION

The idea of using clouds for scientific applications has been around for several years, but it has not gained traction primarily due to many issues such as lower network bandwidth or poor and unstable performance. Scientific applications often rely on access to large legacy data sets and pre-tuned application software libraries. These applications today run in HPC environments with low latency interconnect and rely on parallel file systems. They often require high performance systems that have high I/O and network bandwidth. Using commercial clouds gives scientists opportunity to use the larger resources on-demand. However, there is an uncertainty about the capability and performance of clouds to run scientific applications because of their different nature. Clouds have a heterogeneous infrastructure compared with homogenous high-end computing systems (e.g. supercomputers). The design goal of the clouds was to provide shared resources to multi-tenants and optimize the cost and efficiency. On the other hand, supercomputers are designed to optimize the performance and minimize latency. we analyze the cost of the cloud computing based on different performance metrics from the previous part. Using the actual performance results provides more accurate analysis of the cost of cloud computing while being used in different scenarios and for different purposes.

The performance metrics for the experiments are based on the critical requirements of scientific applications. Different scientific applications have different priorities. We need to know about the compute performance of the instances in case of running compute intensive applications. We also need to measure the memory performance, as memory is usually being heavily used by scientific applications. We also measure the network performance which is an important factor on the performance of scientific applications.

The key contribution of this paper are as follows.

- The disease can be diagnosed based on the symptoms for the respective patient.
- Best drugs are provided for patient by analyzing their personal profile.
- The positive and negative improvements of patients should be analyzed for particular drug.
- The best doctor is suggested and also appointment is fixed for consultation.

II. RELATED WORKS

A. Medical Question Answering Translating Medical Questions Into Sparql Queries Medical Question Answering Translating Medical Questions Into Sparql Queries

Asma Ben Abacha et al. (2012) describes designing question answering systems requires efficient and deep analysis of natural language questions. A key process for this task is to translate the semantic relations expressed in the question into a machine-readable representation. In this paper tackle question analysis in the medical field. More precisely, we study how to translate a natural language question into a machine-readable representation. The underlying transformation process requires determining three key points: (i) What are the main characteristics of medical questions? (ii) Which methods are the most fitted for the extraction of these characteristics? (iii)How to translate the extracted information into a machine-understandable represented a complete question analysis approach including medical entity recognition, semantic relation extraction and automatic translation to SPARQL queries. The fact that SPARQL can represent a wide range of natural language questions in a question-answering perspective. Experiments on a corpus of real questions show that encouraging results in medical entity recognition and relation extraction is obtained. The obtained results also show that the output SPARQL queries correctly represent more than 60% of the original questions.

Revised Version Manuscript Received on March 30, 2018.

V. S. Priyadharshini, Student, Department of Computer Science and Engineering, Kingston Engineering College, Vellore (Tamil Nadu), India.

S. Priyadharshini, Student, Department of Computer Science and Engineering, Kingston Engineering College, Vellore (Tamil Nadu), India.

B. Renuka, Student, Department of Computer Science and Engineering, Kingston Engineering College, Vellore (Tamil Nadu), India.

M. Tamil Thendral, Assistant Professor, Department of Computer Science and Engineering, Kingston Engineering College, Vellore (Tamil Nadu), India.

Advantages: It presents a complete question analysis approach including medical entity recognition, semantic relation extraction and automatic translation to SPARQL queries.

Disadvantages: It has inefficient on deep analysis of natural language questions.

B. Exploring The Pharmacogenomics Knowledge Base (Pharmgkb) For Repositioning Breast Cancer Drugs By Leveraging Web Ontology Language (Owl) And Cheminformatics Approaches

Qian zhu et al. (2011) describes Computational drug repositioning leverages computational technology and high volume of biomedical data to identify new indications for existing drugs. Since it does not require costly experiments that have a high risk of failure, it has attracted increasing interest from diverse fields such as biomedical, pharmaceutical, and informatics areas. The pharmaco genomics data generated from pharmaco genomics studies,➢ applied informatics and Semantic Web technologies to address the drug repositioning problem. Specifically, PharmGKB is explored to identify pharmacogenomics related associations as pharmacogenomics profiles for US Food and Drug Administration (FDA) approved breast cancer drugs then converted and represented these profiles➢ in Semantic Web notations, which support automated semantic inference. We successfully evaluated the performance and efficacy of the breast cancer drug pharmacogenomics profiles by case studies. The results demonstrate that combination of pharmacogenomics data and Semantic Web technology/Cheminformatics approaches yields better performance of new indication and possible adverse effects prediction for breast cancer drugs.

Advantages: Pharmacogenomics data generated from pharmacogenomics studies, applied informatics and Semantic Web technologies to address the drug repositioning problem.

Disadvantages: Traditional drug development is costly and labor-intensive, and scientists are devoted to finding an alternative way to facilitate the drug discovery process.

C. A Side Effect Resource To Capture Phenotypic Effects Of Drugs

Jiahui Jinyz et al. (2010) describes the Massive information networks, such as the knowledge graph by Google, contain billions of labeled entities. Star queries, which aim to identify an entity, given a set of related entities, are common on such networks. Answering star queries can be modeled as a graph pattern matching problem. Traditional approaches apply graph indices to accelerate the query processing. Unfortunately, it is so costly that it is nearly infeasible to build indices on billion node graphs since the time or storage complexity of most indexing techniques is super-linear to the graph size. In this paper, an algorithm is proposed to identify the top-k best answers for a star query. Instead of using expensive indices, our algorithm utilizes novel bounding techniques to derive the top-k best answers efficiently. Further, the algorithm can be implemented in a distributed manner scaling to billions of entities and hundreds of machines.

Advantages: An efficient algorithm for finding the best k answers for a given query without pre-computing graph indices.

Disadvantages: Information networks are incomplete and noisy in nature it will leads difficult to discover answers that match exactly as well as answer that are similar to queries.

III. LIMITATIONS OF EXISTING SYSTEM

- Designing question answering system requires efficient and deep analysis of natural language questions.
- Higher level expense.
- People were bought drugs from pharmacy without knowledge and several factors are not considered.
- The disease cannot be analyzed properly if the information is incomplete and noisy in nature.

A. Proposed System

In the proposed system, side effects and effectiveness, depends on characteristics of patients, such as age, gender, lifestyles, and genetic profiles. Our goal is to provide a tool to assist professionals and consumers in finding and choosing drugs. The disease can be predicted based on symptoms given by patient.

We also arrange appointment to the Best Doctor for the consultation based on user feedbacks. We Integrate Big Data and Cloud computing in this project. Through this we are predicting disease by giving symptoms on application or in android application and provide the drug depends on their disease. Pre stored dataset will available on database when patient give their symptoms it will compare with the dataset and predict the type of disease.

B. Advantages of Proposed System

Easily analyze and predict the disease automatically through the system.It can be used as a training tool to diagnose the patients.Arrange the doctor appointment. People can easily communicate with doctors and time will consume.The SVM algorithm is used to identify exact and closest matches of disease based on the symptoms. Suggest the corresponding drugs based on the predicted disease.

C. Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are.

- ➢ ECONOMICAL FEASIBILITY
- ➢ TECHNICAL FEASIBILITY
- ➢ OPERATIONAL FEASIBILITY

D. Economical Feasibility

This study is carried out to check the economic impact that the system

Will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

E. Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

F. Operational Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

IV. SYSTEM DESCRIPTION MODULES

- A modular design reduces complexity, facilities change (a critical aspect of software maintainability), and results in easier implementation by encouraging parallel development of different part of system. Software with effective modularity is easier to develop because function may be compartmentalized and interfaces are simplified. Software architecture embodies modularity that is software is divided into separately named and addressable components called modules that are integrated to satisfy problem requirements.
- Modularity is the single attribute of software that allows a program to be intellectually manageable. The five important criteria that enable us to evaluate a design method with respect to its ability to define an effective modular design are: Modular decomposability, Modular Comps ability, Modular Understandability, Modular continuity, Modular Protection.
- The following are the modules of the project, which is planned in aid to complete the project with respect to the proposed system, while overcoming existing system and also providing the support for the future enhancement.

LIST OF MODULES

- SERVER DEPLOYMENT
- CONSTRUCTION OF DISEASE TRAINING SET
- DRUG AND SIDE EFFECTS TRAINING SET CONSTRUCTION
- BIG DATA BASED ANALYSIS
- BEST DRUG RECOMMENDATION
- SIDE EFFECT INTIMATION & DOCTOR APPOINMENT

A. Modules Description

SERVER DEPLOYMENT

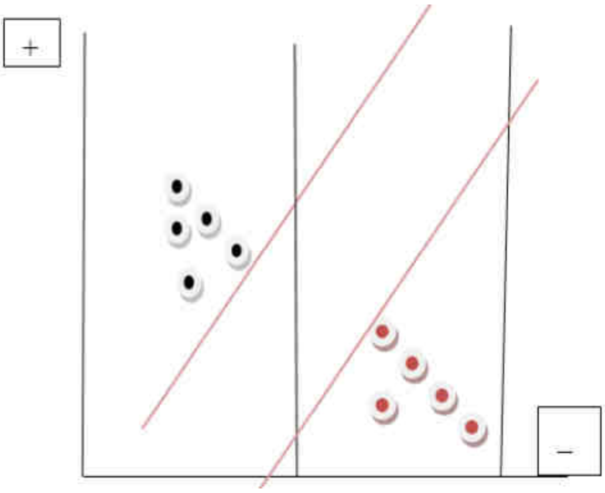
Data Service Provider will contain the large amount of data in their Data Storage. Also the Cloud Service provider will maintain the all the User information to authenticate the User when are login into their account. The drugs and disease information will be stored in the Database of the Cloud Service Provider. Also the Data Server will redirect the User requested job to the Resource Assigning Module to process the User requested Job.

B. Construction of Disease Training Set

In this module we can design and implementation of train the disease to system. Server will store a set of trained dataset and its relevant diagnosis pattern is explained below.

Computation:

We implement this by using Support Vector Machine algorithm. Let us see how this works. Here we uses linear kernel to predict the disease and best drug for the treatment. Lets we consider that we have 2 dimensional space and 2 classes of objects. We want to put the border between them. We can put the border as wide as possible and keep the objects separated. Let us take example for two objects. Hyper plane can be represented by scalar Z and normal vector A. The normal vector determines the orientation. The bios on the other hand contrivers the displacement from the origin. The margin mentioned above can be described by 2 hyper planes.



Graph Hyper Plane 1

ATx + z = 1

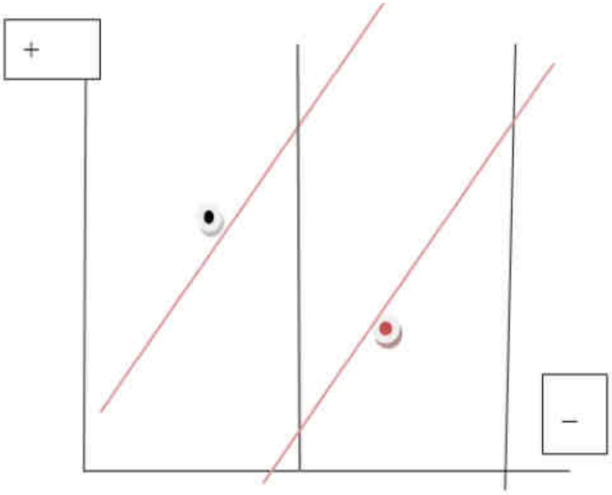
ATx + z = -1.

By changing the angle of margin we can rotate the hyper plane or if you want to shift you can increase or decrease the bios. The width of the margin can be represented as 2||A|| . Normal vector it means the width is proportional to the length of the normal vector. x1 and x2 is represented . Here y1 and y2 is used to describe the labels.

X1 = [-1,1]T

X2 = [1,-1]T

Hence we want to maximize the width of the margin.
$\therefore \min w,,,=\alpha i-12\alpha i\alpha jyiyjxiTxji,j=L(\alpha)$
This formula helps to increase the width of the margin.
By



Graph Hyper Plane 2

By this we can predict the disease and suggest for best drug and analyze the side effects with positive outcome and analyze the best doctor for appointment.

C. Drug and Side Effects Training Set Construction

In this module we will train the drugs for every disease and also train the side effects of the drugs. User will be giving their Symptoms & Diagnostic reports to the system for the diagnosis of the disease.

Predicting drug side effects is an important topic in the drug discovery. Although several machine learning methods have been proposed to predict side effects, there is still space for improvements. Firstly, the side effect prediction is a multi-label learning task, and we can adopt the multi-label learning techniques for it. Secondly, drug-related features are associated with side effects, and feature dimensions have specific biological meanings. Recognizing critical dimensions and reducing irrelevant dimensions may help to reveal the causes of side effects. Drugs can help to treat diseases, but usually come with side effects or adverse reactions. Because of unintended side effects, a great number of approved drugs were even withdrawn from the market. Therefore, recognizing potential side effects helps to reduce costs and avoid risks in the drug discovery. However, wet experiments are costly and time-consuming. Since researchers collected drug data and compile them in the public databases, computational methods were developed for the side effect prediction.

The traditional computational methods analyzed the structure-activity relationship or quantitative structure–property relationship but they are not suitable for the large-scale data. In recent years, machine learning methods were applied to the drug side effect prediction, because of their capability of dealing with complicated data. It combined drug targets, protein-protein interaction networks and gene ontology annotations, and then respectively adopted the support vector machine (SVM) and logistic regression to build prediction models. It considers chemical substructures of drug candidate molecules, and respectively adopted four machine learning methods (k-nearest neighbor, support vector machine, ordinary canonical correlation analysis and sparse canonical correlation analysis) to construct prediction models. It combined the chemical substructures and target protein information about drugs, and adopted the sparse canonical correlation analysis for prediction.

The integration of the phenotypic information, chemical information and biological information about drugs, and then built the prediction models by using different machine learning classifiers (logistic regression, naive Bayes, k-nearest neighbor, random forest and SVM). The decision trees and inductive logic programming is used to identify and characterize side-effect profiles shared by several drugs. The integrated protein-protein interaction networks and drug substructures, and built SVM-based models. It determines molecular predictors of adverse drug reactions with causality analysis.

Although several machine learning methods have been proposed to predict side effects, there is still space for improvements. Firstly, the side effect prediction is actually a multi-label learning task, but far less attention has been paid to this point. Therefore, we make efforts to solve the problem in the frame of multi-label learning. Although lots of multi-label learning methods have been proposed, they can’t be directly used for our task, which have thousands of labels and severely imbalanced data. Secondly, several drug-related features are associated with side effects, and dimensions of each feature are biological components. For example, there are 881 types of substructures. Since a drug may have specific substructures, it is represented by an 881-dimensional feature vector, in which ‘0’ or ‘1’ means the absence or presence of the corresponding substructure. However, not all substructures are necessarily related with side effects, and some may be redundant. Therefore, identifying critical feature dimensions or reducing irrelevant dimensions can help to investigate the cause of side effects, and thus probably improve predictive performances.

In the side effect prediction, prediction models are constructed on the training drugs, and are applied to the testing drugs. Formally, multi-label learning is to build a model that maps inputs to binary vectors, rather than scalar outputs of the ordinary classification. Since a drug is usually associated with multiple side effects, the work can be formulated as a multi-label classification problem.

D. Big Data Based Analysis

In this module we implement big data, in this big data we will have lot or vast amount of data that may wanted or unwanted information in simple the information in the big data are unstructured. So in this module the patient is going allow permission to access the server by the big data analyst. The big data analyst get the all the disease and drugs information which mention above and extract the information by the technique of map reducing formation to get useful information which is useful for patient.

Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

Existing analytical techniques can be applied to the vast amount of existing (but currently unanalyzed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics.

Increasingly, the data is in multimedia format and unstructured. The enormous variety of data—structured, unstructured and semi-structured—is a dimension that makes healthcare data both interesting and challenging. Structured data is data that can be easily stored, queried, recalled, analyzed and manipulated by machine. Historically, in healthcare, structured and semi-structured data includes instrument readings and data generated by the ongoing conversion of paper records to electronic health and medical records.

E. Best Drug Recommendation

In this module we develop an app that allows a user to query for drugs that satisfy a set of conditions based on drug properties, such as drug indications, side effects, and drug interactions, and also takes into account patient profiles. In this we automate the suggestion of the alternative drugs. So that we can provide the medicine to diseases i.e. a preferred medicine .so the researcher will analysis the dosage of the drug and the symptoms.

In this study data mining tools are used to identify disease for which prescriptions are written in order to evaluate the performance of the methods, the obtained results compared with Naïve method . The result shows that Support Vector Machine gives better performance than the other methods.

The results indicate that the implementation of the data mining algorithm resulted in a good performance in characterization of outpatient disease. These results can help to choose appropriate methods for the classification of symptoms in larger scale. The method firstly constructs a drug set and a disease set based on the side effects and symptoms respectively. Because similar drugs imply similar disease, then cluster the two sets to identify drug and disease modules, and connect all possible drug-disease module pairs.

F. Side Effect Intimation & Doctor Appointment

In this module, we can design and implementation of side effect of the drugs. Analyze the disease and also doctor appointment is fixed for the consultation based on user feedbacks. Side-effects are measureable behavioral or physiological changes in response to the treatment. Intuitively, if drugs treating a disease share the same side-effects, this may be manifestation of some underlying mechanism-of-action (MOA) linking the indicated disease and the side-effect. Building on this confirmation of strong correlation between drug indications and side-effects, we further compiled a list of relationships among all known drug-disease and drug-side-effect to build disease-side-effect profiles and identify statistically significant relationships between drug side-effects.

➢ Pre Stored Data Comparison

In this module doctor will import all the details about the medicine i.e. what are the symptoms, dosages and drug .And hw will store more about of the medicine so that we can make some use of it for example we can give awareness to the society .we store the all the data in the clustering format so that data can spitted and stored in the different clusters. So that it will easily to classify the data for the research.

➢ Predictive Disease Analysis

In this module we implement predictive disease analysis system in which the data will be analysis so that we can predictive the disease based on the symptoms .This module interact with server to analysis, the analysiation is done by the researchers. So they get the data from the server to make analysis to find the disease based on the symptoms

➢ Suggestive Alternative Drugs

The best result can be analyzed and they prefer the best medicine for the list of disease mentioned in the proposed system.

TESTING

TEST CASES

UNIT TESTING

Unit testing is a software testing method by which individual units of source code, sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures, are tested to determine whether they are fit for use.

In this project, all statements are executed properly. All units of program programs are tested in different computer. And the result of the project is same in all system.

G. Integration Testing

Integration testing is the phase in software testing in which individual software modules are combined and tested as a group. It occurs after unit testing and before validation testing. Integration testing takes as its input modules that have been unit tested, groups them in larger aggregates, applies tests defined in an integration test plan to those aggregates, and delivers as its output the integrated system ready for system testing.

In this project there are six modules like server deployment, construction of disease training set, drug and side effects training set construction, big data based analysis, best drug recommendation, side effect intimation & doctor appointment. Each module satisfies with other module. Every module gets an input and deliver expected output.

H. Validation Testing

The process of evaluating software during the development process or at the end of the development process to determine whether it satisfies specified business requirements.

Validation Testing ensures that the product actually meets the client's needs. It can also be defined as to demonstrate that the product fulfills its intended use when deployed on appropriate environment. In this project, client will receive their expected outcome. The motive of the project is to identify to predict the disease based symptom and provide best drug for that disease.

V. CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

We have proposed an approach for answering drug queries to support drug prescription. Our focus is on how to obtain and rank answers based on incomplete information and provide personalization. To cope with incomplete and noisy data, we allow both exact and close matches when answering queries. We also present an intuitive approach to display answers to users, which aims to help users to understand the ranked results and possibly refine their queries.

FUTURE ENHANCEMENT

It is very important that the big data research community does not repeat the same mistake. While there is clearly an important research space examining the fundamental methods and technologies for big data analytics, it is vital to acknowledge that it is also necessary to fund domain-targeted research that allows specialized solutions to be developed for specific applications. Healthcare in general and computational biomedicine in particular, seems a natural candidate for this.

REFERENCES

1. S. Khemmarat and L. Gao, “Supporting drug prescription via predictive and personalized query system,” in PervasiveHealth. IEEE, 2015.
2. Jin et al., “Querying web-scale information networks through bounding matching scores,” in WWW 2015, 2015, pp. 527–537.
3. Doulaverakis et al., “Panacea, a semantic-enabled drug recommendations discovery framework,” J. Biomed. Semant., vol. 5, p. 13, 2014.
4. Langer et al., “A text based drug query system for mobile phones,”Int. J. Mob. Commun., vol. 12, no. 4, pp. 411–429, Jul. 2014.
5. Ben Abacha and P. Zweigenbaum, “Medical question answering: translating medical questions into sparql queries,” in Proceedings of the 2nd ACM SIGHIT. ACM, 2012, pp. 41–50
6. M. Kuhn et al., “A side effect resource to capture phenotypic effects of drugs,” Molecular systems biology, vol. 6, no. 1, p. 343, 2010.
7. Dumontier and N. Villanueva-Rosales, “Towards pharmacogenomics knowledge discovery with the semantic webx, vol. 10, no. 2, pp. 153–163, 2009.
8. K. Sangkuhl et al., “Pharmgkb: understanding the effects of individual genetic variants,” Drug Metab. Rev., vol. 40, no. 4, pp. 539–551, 2008.
9. Doulaverakis et al., “Panacea, a semantic-enabled drug recommendations discovery framework,” J. Biomed. Semant., vol. 5, p. 13, 2008.
10. T. Fawcett, “An introduction to roc analysis,” Pattern recognition letters, vol. 27, no. 8, pp. 861–874, 2006.