The Industry Data Analysis Processing Model Design

—The Regional Health Disease Trend Analysis Model

Li Anying
School of Computer Science and Technology,
Northwestern Polytechnical University, 710072, Xi'an, China
lay18@163.com

Chen Ke
Xi'an future international information co.ltd,
710075, Xi'an ,China
chenke@ourfuture.cn

Song He
Xi'an future international information co.ltd,
710075, Xi'an ,China
songhe@ourfuture.cn

Lei Yu
Northwestern Polytechnical University Ming De College,
710072, Xi'an, China
724944758@qq.com

Abstract—In recent years, Chinese society increase higher degree of informatization, all aspects of social life more and more can not live without information. The medical profession information provide people seeking medical advice of hitherto unknown convenience, this paper mainly studies the prediction method of regional medical disease. With the help of big data processing platform, the original data suppliers provide data, we gather and collect the data, do data mining and analysis in big data computing environment, finally get the disease trends forecast report.

Keywords—disease;big data ;data mining;data analysis

I. INTRODUCTION

Informatization is the irreversible trend of economic and social development,informatization technology impact on population health service system, is far from a simple informatization technology application or the problem of efficiency, and understanding through the use of informatization technology, the traditional major renovation and reshape the population health service system. Population health service at present, the emerging informatization technology and accelerated convergence trends, the role of promoting economic growth is increasingly significant, related areas has become the focus of global competition. Strengthen population health informatization construction is not only the comprehensive deepening health reform,improve the ability of industry management, promote the implementation of scientific development is an urgent need, is also adhere to the people-oriented,implement scientific and accurate health family planning policy,effectively improve the intrinsic demand of the people's health.

China is the world's most populous developing countries. At present, as the population ages accelerated, the propulsion of urbanization and disease change, China is faced with both developed and developing countries diseases and health problems, solve the problem of good hospital of 1.3 billion people, there is no ready-made model can follow, this family planning service system and to the health of the governance ability put forward severe challenges. And informatization provides a rare opportunity.to improve and solve the problems in the medical and health services, to achieve leapfrog development.

II. BUSINESS DESIGN

A.Medical foundation

Xi'an is located in the guanzhong plain in shaanxi, is the capital of shaanxi province, 13 dynasty set up in this, now seven area, six counties and two management committee, the city resident population of about 8.2 million. By 2013, the city's total of 1350 medical and health institutions, including three level of first-class hospital (including 1) military hospital, secondary hospital 14 (including 1 military hospitals, private seven), CDC seven, seven health supervision institutions, four women's and children's organization, two skin disease prevention and control institutions, mental health institutions, four rehabilitation facility, emergency center, a bloody battle, medical, one each in towns and townships, 41, eight community health service center, 967 village clinics. In 2013 the total clinical medical institutions at all levels 8.8056 million person-time, of which this 8.5687 million, 340300 people in hospital.

B.Requirements describe

Regional health medical disease trend analysis system application subject to the centers for disease control and prevention, the centers for disease control and hope to be able to disease outbreaks occur again, to a pharmacy, hospital, physician medical institutions such as the end, timely collect medical information, and provide decision support auxiliary analysis.

1) Timely predict the spread of disease. When a disease outbreak, through pharmacies, hospitals, community health agencies related yao piano sales, accepts and regular physical examination, and so on and so forth, the real-time analysis of the spread of the disease trends and movements, collection of infectious factor, to form 8 hours, 1, 3, 1 week trend analysis and forecasts of volume with pricing.

2) Effective trend forecasting disease. According to the drug sales in recent 5 years, medical treatment, regular physical examination data, such as the formation of

978-1-4799-6621-9/14 $31.00 © 2014 IEEE
DOI 10.1109/CCBD.2014.11

calendar year during the same period of diseases, and according to the real-time updating data, to track revision anticipation.

C.Processing model

Regional trend analysis model including the public health disease, pharmacy, hospital, community health posts, the centers for disease control and prevention, research institutions, pharmaceutical companies, as well as the health development planning commission, emergency management and the provincial government body.

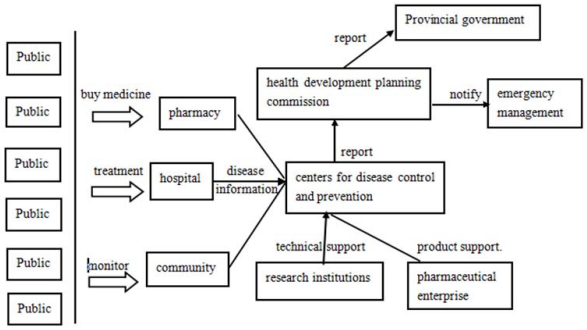

Figure 1. Processing model.

Public basis condition choose pharmacy to buy medicine, hospital and so on active way, and by the community for routine physical examination way of passive, disease information collection to the centers for disease control and prevention, the centers for disease control and prevention on the basis of disease outbreaks reported in time health development planning commission, the health development planning commission reported to the provincial government, and notify the emergency management, the centers for disease control and prevention on the basis of diseases between research institutions to provide technical support and pharmaceutical enterprise product support.

III. SYSTEM ARCHITECTURE

Regional health medical diseases trend analysis model framework is provided by the original data, data extraction, data processing and control all the parts, etc.

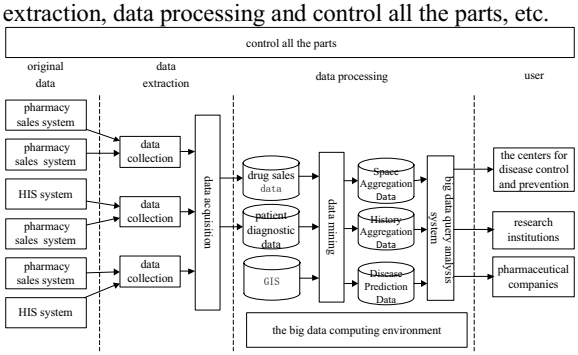

Figure 2. system architecture figure

A.Data extraction

Mainly including data acquisition, data acquisition two links.

1) Data collection. Through HIS system in hospital and pharmacy sales data interface provided by the system can obtain drug sales and patient diagnostic data. According to the different area and different data source type development of data acquisition.

2) Data acquisition. Data acquisition for receiving data acquisition plugin for different data and preliminary data cleaning and data preprocessing. Format conversion is used to data from the data acquisition component acquisition upward buffer data extraction, conversion, quality cleaning processing, and the standardization of the processed data loaded into the big data processing platform.

B.Large data processing platform

Mainly includes the big data computing environment, data mining platform and data query analysis system.

1) The big data computing environment. As the calculation basis of regional health medical disease trend analysis system, adopt cloud computing mode, build high elasticity, high redundancy and high reliability of parallel computing environment.

2) Data mining platform. Implementation of geographic information data fusion and medical information, and provide data based on fusion data according to different classification of drilling, digging, aggregation, and other services.

3) The big data query analysis system. Provide real-time online query service, batch, service, sharing programming interface, etc.

C.The whole control system

This system is used to control from raw data, to data extraction, to big data processing, and user queries all data in the process of data acquisition, process flow, calculation, dig and report.

IV. ALGORITHM ANALYSIS

A.Aggregation analysis

Large data platform aggregate data analysis, using Mahout clustering algorithm. The algorithm running as below.

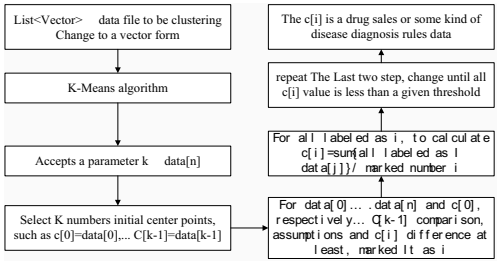

Figure 3. aggregation analysis figure

B.Machine learning

1) Pattern recognition

Construction of pattern recognition system based on statistics, the basic principle is: the similarity of the samples in the pattern space close to each other, and form a group, namely Like attracts like.The analysis method is based on the feature vectors of the Xi= (Xi1, XI2 ,..., XID) T (i=1, 2,... , N), put the pattern into numbers of C class of $\omega 1$, $\omega 2$,..., ωc, then according to the distance function between model to classification. Among them, T as the

transpose; N as sample points; D as the sample characteristic. Using the nearest neighbor classification, characteristic analysis. The drug sales and disease diagnosis data sets into small set by pattern recognition.

2) Neural network

BP neural network is a kind of three or more than three layers of structure without feedback, layer without interconnection structure to the network, which are two layers which are input layer and output layer, other layer is the hidden layer. The training process is repeat "forward to calculate the output process, back propagation error", until the error is reduced to an acceptable level, training process of BP neural network becomes the end.

In this project, the output layer is the historical law of certain diseases.

C.The trend analysis

Analysis the drug and disease history trend analysis, get the history of drug sales and disease diagnosis by machine learning, get the history trend of epidemics occur through BP neural network algorithm.

Prediction analysis methods are as follows.

Define a temporal information system, S= (n, a U {d, t}, B), where n denotes the set of objects (epidemic trend of History), a as attributes (features, variables, conditions), D as the decision attribute, t as order properties, B represents an order relation on attribute t sequence.

Time series analysis has the following three kinds of changes in decomposition form:

- Trend.
- Cycle change.
- Random variable.

Describe the prediction algorithm of time series analysis as:

- Selection method for linear trend, square tendency, three trends and index trend fitting, and follow the estimation criterion in the sequence, get an appropriate term M.
- Development method which estimated seasonal item S.
- Using autoregressive moving average model to predict the random Y.
- According to formula X=M+S+Y to get the final prediction value.

X is a set of future disease for a period of time occurrence and development forecast data.

V. DATA ANALYSIS

A.The raw data

Mainly from the pharmacy sales system, the hospital HIS system, etc.

1) Pharmacy sales system. Now the drug sales in the city mainly from several large pharmaceutical chain brand, their sales data has been connected to the Internet, access to the sales data of several major pharmaceutical retailers, you can get most drug sales.

2) The hospital HIS system. Obtained from the hospital HIS system of real-time data, pharmacy can obtain by hospital opened a prescription drug sales. HIS system from the hospital diagnosis can obtain the real-time diagnosis data in the data.

B.benchmark data

1) Drug sales data. Raw drug sales data after cleaning, extraction, format conversion loaded into the big data processing platform of Hbase and Hive.

2) The disease history data. From hospital HIS system diagnostic data original disease history data, after dealing with the same load to the big data processing platform.

3) The geographical spatial data. Have to deal with and loaded into the big data platform of geospatial data.

C.theme analysis data

Theme analysis data can be stored according to the needs of the business by analyzing large data mining technology depth analysis of the theme of the in-depth analysis and query data for users. Provided based on the types of drugs and disease diagnosis of historical data query. Space the aggregated data and historical data aggregation, can for users to use of machine learning and neural network for deep analysis.

1) Spatial data aggregation. Based on benchmark data aggregation analysis in the spatial scale, form related to the geographical spatial data aggregation analysis.

2) The history of the aggregated data. Through the study of the time series analysis of different periods of history data, form the analysis of the historical trend data.

3) The disease forecast data. For space and historical aggregation analysis situation in trend forecast analysis of the disease.

VI. BIG DATA PROCESSING DESIGN

A.Big data computing environment

Big data computing environment provide basic big data storage and big data computing environment for parallel computing processing platform, the environment is a massive data parallel processing platform building with core technique of Hadoop2.X and other big data processing.

1) Distributed parallel storage system. Distributed parallel storage system is a distributed file system running in general x86 server cluster. Use the server's local disk to provide a massive data storage solution of high fault tolerance and high throughput. Distributed parallel storage system uses an efficient distributed algorithm to distribute data access and storage in a large number of servers, also mutual backups the data in the server clusters according to the need.

2) Parallel computing framework. Parallel computing framework is a high performance batch distributed computing framework, used for parallel processing and analysis massive data. Compared with the traditional data warehouse and analysis technology, this framework is suitable for processing various types of data, including structured, semi-structured and unstructured data.

3) Distributed and parallel database. Distributed and parallel database is a data storage system for a distributed and faced to column. Distributed and parallel database is designed of TB to PB level massive data storage and high speed to read and write from the beginning, these data can be distributed in thousands of ordinary server, and can be high-speed accessed by a large number of concurrent users.

B.Data mining

1) The data preparation. The data after pretreatment and conversion store in a big data database or data warehouse.

2) Searching the rule. Association rules mining used in this project, it is divided into two stages, the first stage is to identify all from the data sets, the second stage is to create association rules by high-frequency project teams.

3) Show the rule. Interpretation and evaluation the results, use the visualization technology, and integrate the analysis data of drug sales and disease into the polymerization database.

C.Big data analysis and query system

1) The on-line real-time query. The on-line real-time query engine can query fast from population data warehouse and relational database in core data collection area, the query engine can query big data, query the relational database, query related the big data and relational database tables.

2) Batch analysis. Batch analysis engine provides data batch processing task scheduling mechanism and operating environment, use big data parallel computing framework to achieve a variety of complex data processing with the task scheduling mechanism by population data warehouse and relational database in collection area.

3) Share programming interface. Share programming interface engine provides open data access interface, to query, modification, analysis and scheduling the data etc for upper applications.

VII. SUMMARY

Modern medical and health services is the inherent requirement of standardization and normalization; Sees a doctor, examination, treatment process in general is linear, producing huge amounts of information in the process of medical services, with the aid of the big data processing platform for regional health medical disease, trend analysis, provided by the raw data to provide the data, data collection and gathering, we in the big data mining and analysis of the data computing environment prediction, finally get the trend forecast of the disease, make the most of population health information technology, to realize efficient use of medical resources, reduce the medical cost and improve service quality and efficiency, to solve the problem of the people's hospital, provide strong support.

REFERENCES

[1] Elements of Large-Sample Theory, E.L.Lehmann

[2] The Creative Destruction of Medicine, Eric Topol

[3] Big Data A Revolution That Will Transform How We Live, Work and Think, Viktor Mayer and Kenneth Cukier

[4] Hadoop in Action, Jia heng Lu

[5] Hadoop Internals: in-depth study of YARN, Xi cheng Dong

[6] Data Mining, Jiawei Han

[7] Beautiful Data, Toby

[8] Mathematical Sratistics and Data Analysis, John A.Rice

[9] Introdution To Data Mining, Pang-Ning Tan

[10] Hadoop: The Definitive Guide, Tom Wbite