Research of Postal Data mining system based on big data

Xia Hu[1], Yanfeng Jin[1], Fan Wang[1]

[1]Shi Jiazhuang Post & Telecommunication Technical College, 050021, Shijiazhuang, China

[1]huxia-068@163.com, [1] yanfengjin@163.com

KEYWORDS: big data, postal data, data mining, precision marketing

ABSTRACT: With the rapid developing of cloud computing and big data technology, enterprises have accumulated vast amounts of data and all of them have got more effective treatment, so companies get greater value of market. This issue analyzed the data China Post has and proposed an architecture of China Post's data-digging system in large data environments and carried out a detailed design of precise process for China Post based on big data. Finally, we used China Post's data-digging system on the precise database marketing and personalized product recommendations and other aspects, and achieved good results.

Introduction

With the rising of mobile Internet, network, social network technologies and applications, the amount of data is growing rapidly in global scope. The time of big data has coming. With the rapid development of Direct Mail business, the demand of name and address data is increasing. China Post has experienced 10 years of information technology, a large number of postal service production data has accumulated, and is in urgent need of development.

At the end of 2010, China Post established a data analysis team, which is responsible for the postal internal business data analysis and integration, to support the headquarters of marketing projects. Since its inception, the extraction of data samples for a variety of business is researched, the postal existing data resources was basically understanding, and the team also summarized some data analysis methods, constantly trying to build some models to produce some data products, for marketing application, and verification. But no integrated platform to store the full amount of the relevant business data, the data can not be processed, reconstructed, the early results can not be transformed into products. So a postal data mining platform based on big data structures need to be established, which establish business system data acquisition mechanism, and extract the full amount of data. And package part of common data preprocessing and data cleansing rules to the platform, and provides manufacturing capabilities and data platform for data analysis team. Postal enterprises are facing a critical period of transition, a large data processing system will be a key factor in the transformation of China Post.

Postal data characteristics analysis of big data environments

With Chinese postal deeply reform continued, postal services, courier logistics and finance and insurance business are booming. Integration within the plate, and the convergence between the plate become more and more prominent. The platform need a access to data sources, including postal class, express logistics, finance and so on. as follows:

Feature analysis of postal data. Postal service class data includes: Newspaper subscription data, savings SMS and delivering SMS data, customer data from the Post family, air ticket membership

data, Ule membership data, care package data, philately membership data, mail acceptance data, international packet customer data, packet acceptance data, account management system data, name and address date from system, organization data system.

Express Data Characteristics. Express data contains courier SMS information and express acceptance data. Courier send about 4 million SMS data per month, just include two fields: e-mail and telephone numbers. Express acceptance produces 1.8 million data a day. The quality is not high before the full name and address data entry work carried out.

Feature analysis of financial data. Financial data includes the Postal Savings Bank customer data, transaction data and electronic exchange savings SMS data. Electronic exchange transactions daily increasing about one million, mainly used for supporting Direct Mail. Postal Savings SMS data has 80 million customers a year, the size of transaction data about 3TB.

Data mining process and methods under big data environment

Postal big data mining system framework. Due to the characteristics of postal big data post, postal corporations can not apply the traditional data analysis techniques Both traditional data analysis and big data mining extract great useful information and show leanings from the data And they are the process of in-depth analysis and value-added exploitation of the data. However, there are essentially differences between them and mainly reflected in: ①The size of the data is different. The traditional data analysis and process the data which is usually stored in a database or file. The size of data is GB level or below, but the size of big data mining is generally PB level, even more massive stage. ②The types of data for these two analysis method are different. The traditional data analysis focused on static and structured data. Big data mining, on the other hand, is able to process either the structured data or semi-structured and unstructured data, often in real-time based. ③The analysis methods are also different. The main algorithm for traditional data analysis is based on the statistics. As shown in figure1, classification and Prediction are two common methods of data analysis. Big data mining need not only statistical method, but also machine learning and artificial intelligence algorithm for lots of situations.

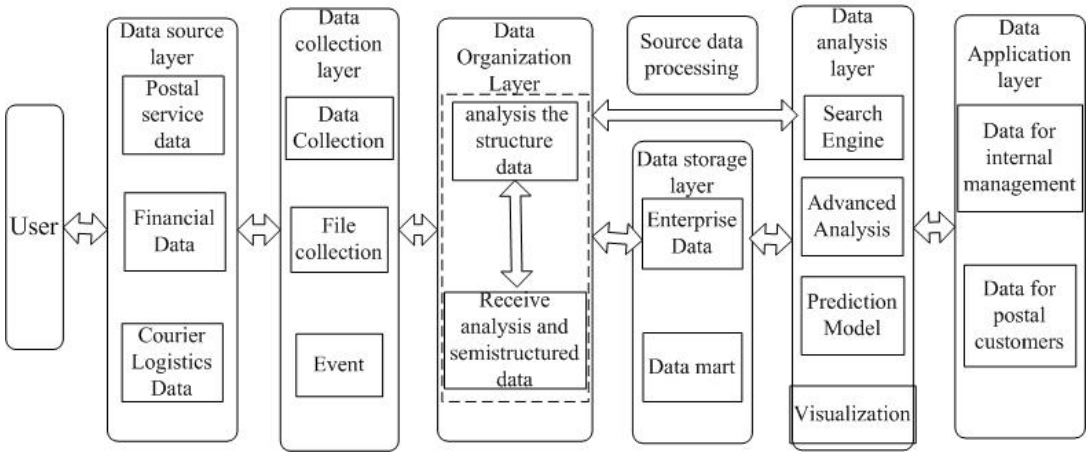

Fig.1 Big data mining framework of the postal corporations

Postal big data mining process design. For the process of analysis, the traditional data analysis is relative simple. The data is usually is organized in a file or metadata in database, and then is carried on the sample selection. Meanwhile, we use the classification algorithm and prediction algorithm to predict the type of discrete and continuous values of the data object.

Different from traditional data analysis, big data mining is a process of knowledge discovery automatically. In the absence of defined goals, we get data from different data sources, preprocess data, and greatly use machine learning and artificial intelligence algorithm to analyze the observational data. The postal data mining focus on solving such a problem: in big data from different the postal services, valuable knowledge is from analysis of the characteristics of the user groups and users' personal characteristics further, to obtain commercial value. As shown in figure 2, the data mining process includes: data collection, data preparation, data conversion, data extraction, data mining and mining applications.

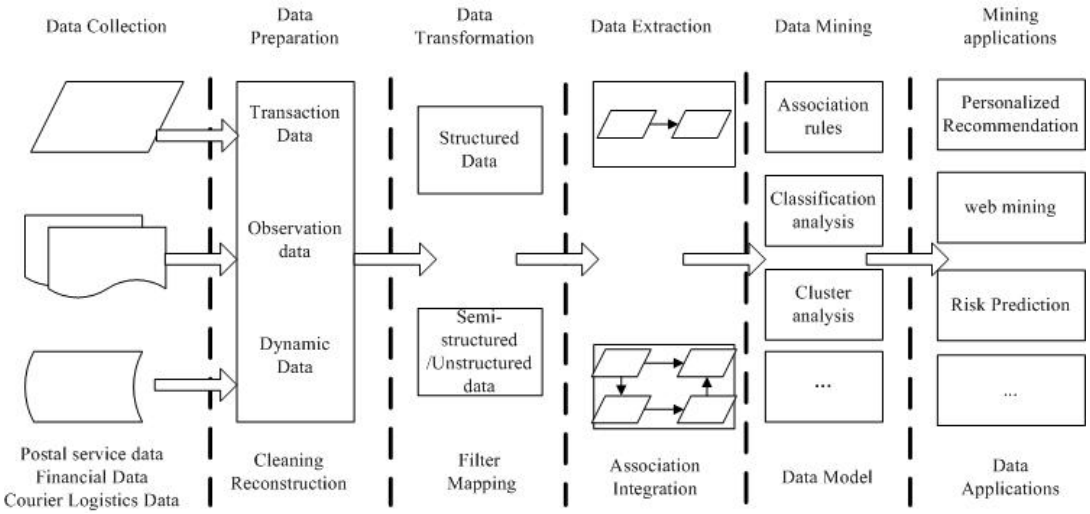

Fig.2 Postal big data mining process

•Data collection. The postal data is collected from the postal service data, financial business data and express delivery and logistics business data, stored in the postal system platform. For the press data as an example, there is interaction(a overlap) between press data and customized address list data, and also in the data content. Therefore data should be classified on the basis of data categories and properties.

•Data preprocessing. Data preprocessing includes data preparation, data transformation and data extraction. Data preprocessing determines the quality of the mining results. To some extent, data preprocessing tend to affect the validity of data mining. Due to the noise data, redundant data and the missing value and so on, during the process of data preparation, data analysis, cleaning, refactoring and filling the missing value can improve the quality of the data mining. Then the data which is unstructured, semi-structured is processed into machine language or index, for example, converting natural language user reviews, log data into weighted term logic or fuzzy logic, and different words mapped to a standard of value. Structured data is filtered to extract meaningful data and eliminate the invalid data in order to improve the efficiency of analysis.

The last step is data extraction, which detects the correlation and relevance of data. The relevant data showed more specific user activity characteristic. And those data itself can also be used for personalized services.

•Data mining and application. In the process of data mining, we select the mining model according to different application requirements for the depth of mining data. The main models include: Association Analysis, classification analysis, clustering analysis and so on. There are also some user models for data mining. These user models will group people with gender, race, age and

interest to get data mining results to interpreted and applied. The general mining applications, include the ranking and personalized recommendation, anomaly detection, Web data mining and search, and visual calculation and analysis of big data.

The Postal data mining application under big data environment

The postal data mining application is with in-depth analysis of data to dig out the user's behavioral characteristics, consumption habits and interest focus, so that postal clients can target the audiences more accurately to obtain greater business value. Postal Data mining can help postal clients to develop more accurate and effective marketing strategy. As shown in figure 3:

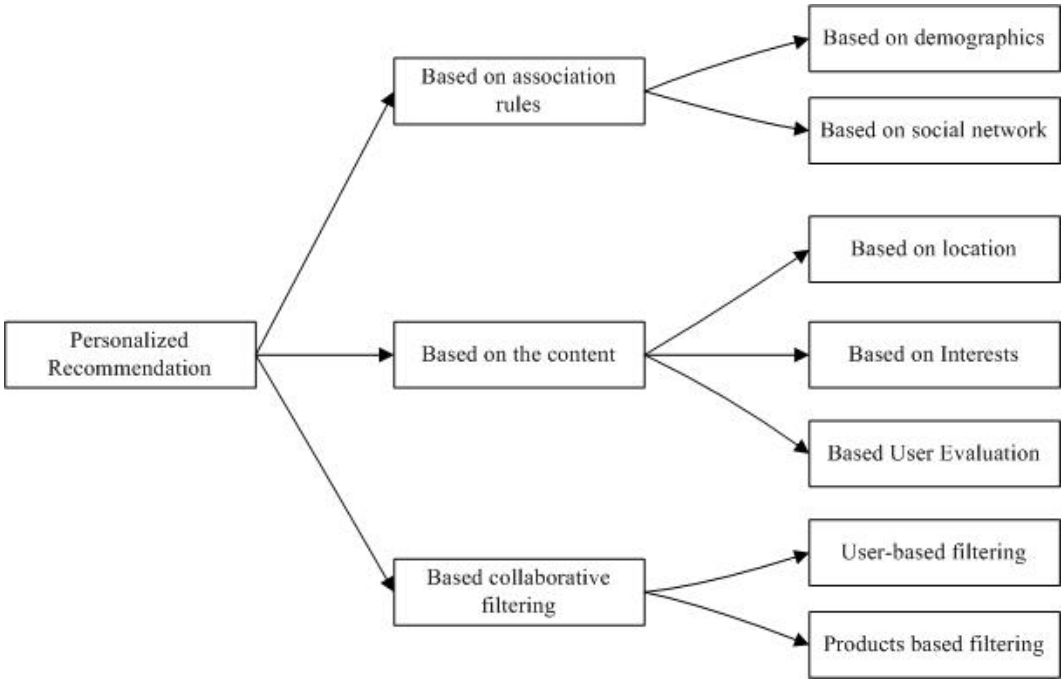

Fig. 3 Personalized Recommendations based on Postal Data Mining

Precise Marketing Based on Postal Data Mining. Postal Data Mining represents a more refined market, a more accurate prediction of user behaviors and more precise needs. By collecting, processing and handling large amounts of information related to the user's consumption behaviors, it determines interests, consumption habits, consumption trends and consumption demand for a particular user groups or individuals, and then to infer consumption behaviors for next step. And based on this, specific marketing content will target to the identified user group. Compared with the traditional mass marketing tool without distinguishing characteristics of the user targets, it saves marketing costs and improves marketing effectiveness.

The Data Mining Application for Postal Clients. Understanding users' key nodes in different consumer behaviors by postal data mining can provide a reference for online advertising strategies and targeted advertising to achieve personalized marketing for clients. Based on uses data, data mining can help on establishing the probabilistic knowledge base and fuzzy knowledge base, to do probability analysis on real-time online information. Through analyzing and dividing Ads visitors' potential information features, marketers could decide who are the real customer to the business; through analyzing customer response to certain Ads, marketers could decide next delivery channels and timing; through cluster analysis, marketers could send out targeted ads to the targeted

audiences. When the accumulated data to a certain size, marketers can accurately calculate ROI of each keyword of ads for businesses through data mining and optimize advertising content.

Conclusion

With the development of cloud computing and data mining techniques, the inherent value of the data will be increasingly likely to be explored. With the rapid development of China's postal database Direct Mail business, the demand for lists is increasing. China Post has experienced 10 years of information technology construction and accumulated a large number of postal service data, which is need to be analyzed and used urgently. First, we analyzed China Postal existing data. Based on that we proposed the framework of postal data mining system under big data environment. Besides, the postal data mining process is designed and implemented in detail. Finally, postal big data mining can be applied to precise marketing and personalized product recommendations(referrals), etc. Applications of big data technologies will be very important for the development and transformation of postal services, And it will also provide a tremendous commercial value for China post.

References

[1] Qian Y, Liang J, Pedrycz W, et al. Positive approximation: An accelerator for attribute reduction in roughest theory[J].Artificial Intelligence,2010,174(9-10):597-618.

[2] Niu S, Guo J. Top-k learning to rank: Labeling, ranking and evaluation[C]//Proceeding of the 35th international ACM SIGIR conference on Research and development in information retrieval, New York: ACM, 2012:751-760.

[3] Mahony M W. Randomized algorithms for ,matrices and data[J]. Foundations and Trends in machine Learning,2010,3(2):123-224.

[4] Gasso G, Pappaioannou A, Spivak M, el al. Batch and online learning algorithms for non convex Ney man-Pearson classification[J].ACM Transaction on Intelligent System and Technologies, 2011,2(3):Article No 28.

[5] Cheng X Q, The application and Scientific Problems of Big Data(Scientific Forum on math and Big Data)[R].Beijing: Chinese Academy of Sciences,2013(Ch).

[6] Wang Y Z, Jin X L, Cheng X Q. Network big data: Present and future[J].Chinese Journal of Computers,2013,36(6):1125-1138(Ch).

[7] Gasso G, Pappaioannou A, Spivak M, et al. Batch and online learning algorithms for non convex Ney man-Pearson classification[J].ACM Transaction on Intelligent System and Technologies, 2011,2(3):Article No 28.