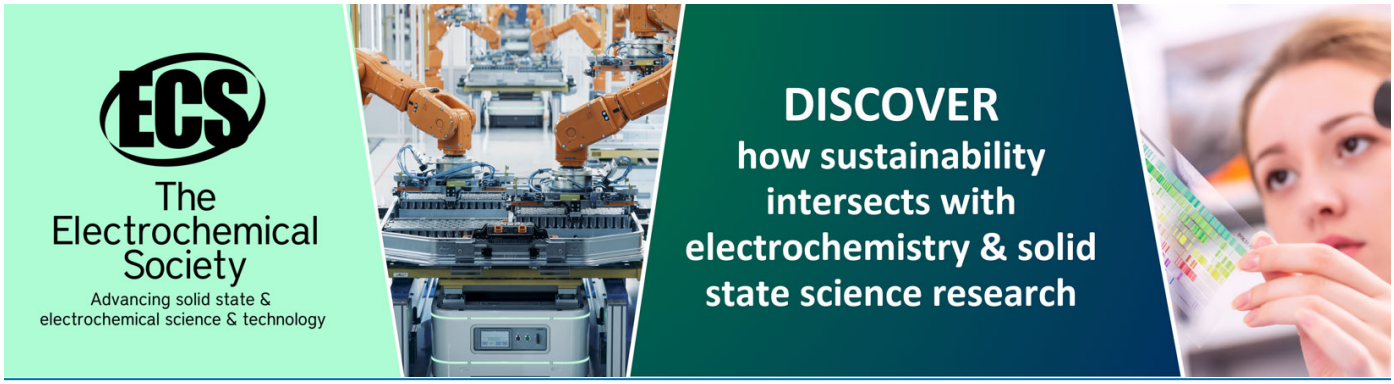
Application of Business Intelligence Based on Big Data in E-commerce Data Analysis



Journal of Physics: Conference Series **1395** (2019) 012011 doi:10.1088/1742-6596/1395/1/012011

Application of Business Intelligence Based on Big Data in E-commerce Data Analysis

Xiaoli Du, Beixiong Liu and Jiangli Zhang*

Guangdong Polytechnic of Environmental Protection Engineering, Foshan, China

duxiaoli83@qq.com; jiangli@just.edu.cn

Abstract. This paper redesigns the framework platform of business intelligence by combining big data with traditonal business intelligence. This paper focuses on the way of data acquisition.Taking the e-commerce data of an enterprise as an example, this paper uses K-Means algorithm in clustering analysis to cluster consumers, so as to realize the purpose of personalized marketing for different consumers.

1. Introduction

The concept of Business Intelligence (BI) was proposed by Gartner Group. It involves information search, management and analysis. The purpose is to enable decision makers to acquire knowledge and make more effective decisions. Business intelligence is not an independent technology, but a complete set of solutions. It integrates data warehouse, OLAP, data mining and visualization technologies into business activities, transforms complex information into auxiliary knowledge, and finally presents knowledge to users to support enterprise decision-making.

With the continuous expansion of the scale of Internet applications, the amount of data to be processed increases exponentially, and the data structure becomes more and more complex. The pressure of business operation increases sharply, which directly promotes the development of large data processing technology [1]. With the rapid development of new generation IT technologies such as e-commerce, cloud computing and mobile social media, traditional BI systems can not meet the needs of enterprise data analysis. Personalized, data-based and scientific data analysis technology has gradually made the traditional BI system need to be combined with big data technology to achieve a new platform architecture to meet the needs of big data analysis.

2. Design of Big Data Application Based on Traditional BI System

In the era of big data, traditional BI's data storage ability, data analysis ability and real-time data processing ability are not competent for the application and analysis of unstructured complex data sources. Therefore, how to make full use of existing BI and big data technology is the key to the design of new platform architecture. Traditional BI data mainly comes from internal operating system and management system; the main source of big data is the Internet, such as microblog, web pages and other data exchange. They are fundamentally different in terms of data sources, data collection, data processing, data storage and future data applications [2]. Based on the above considerations, a new architecture platform is designed as shown in Figure 1.

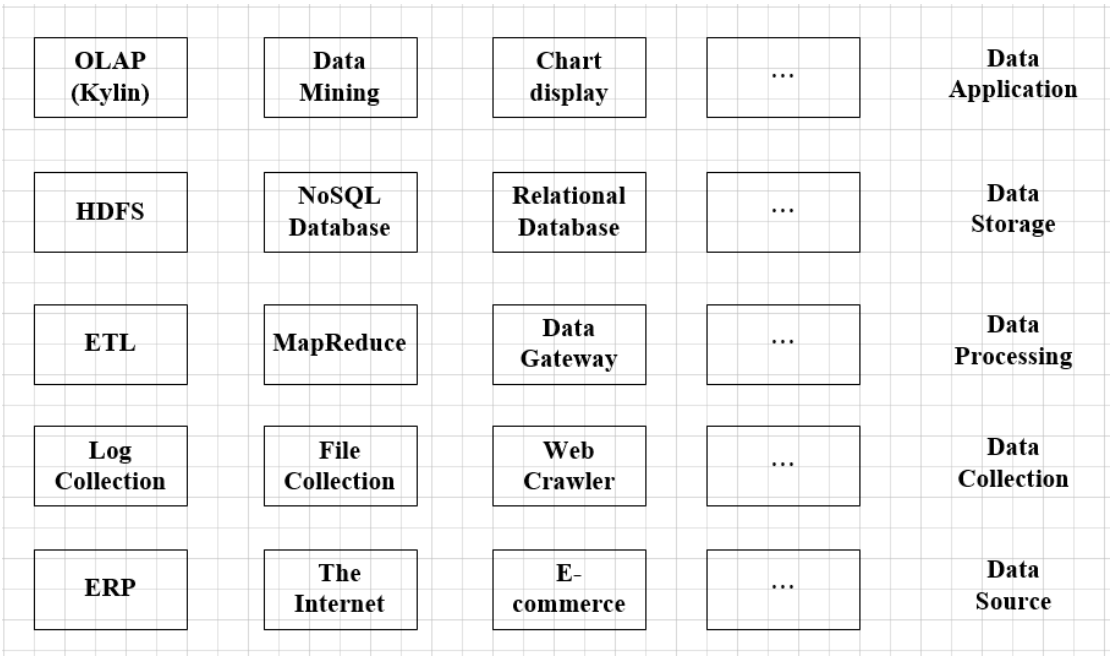


Figure 1. BI and big data combined data platform

Data sources mainly include internal data and external data of enterprises. Internal data consists of OA system, ERP system, financial statements system and other related structured data. External data includes unstructured data on the Internet, such as hypertext, images and videos. Data acquisition adds a new way of collecting Internet web crawlers to the original way of collecting data. Different processing methods are used for structured and unstructured data. Unstructured data is organized into structured data and stored in distributed structured database; traditional data is still stored in relational database. Large data is mainly stored in the form of distributed file system (HDFS) and NoSQL database. The final data is mainly used for on-line analysis and processing, data mining, data visualization and so on.

3. Data Acquisition Method

Data collection methods in the context of large data mainly include three categories: system log collection, network data collection and data interface collection. Log data acquisition is achieved through the logging subsystem in the device, which can generate log messages when necessary. Common commercial data APIs support REST APIs to obtain data information. Web crawler technology is mainly used in network data acquisition. Its core principle is to use HTTP protocol to simulate the browser to access the Web server through a unified resource locator URL address, obtain the permissions of the Web server, return to the original page and parse the data [3].

Traditional crawler technology may have problems, so focused crawler technology designed to crawl web resources emerges as the times require. Focus crawlers selectively access web-related links on the Internet to obtain the information they need based on established crawling targets (using an e-commerce sales theme). Focused crawler does not pursue full coverage of web pages. Instead, it aims at web pages related to specific topics and prepares data resources for topic-oriented user queries.

4. Application case of an enterprise e-commerce data

4.1. K-Means algorithm

K-means is a widely used clustering method that divides D entities into N clusters. This ensures that the similarities within the cluster are as high as possible and the similarities between the clusters are as low as possible. The process of the K-means algorithm is as follows:

- randomly select N data points as the centroid;
- Calculate the distance from each data point in the data set to the centroid, and aggregate all the data points in the data set into N clusters;
- According to the N sets of data points calculated in step 2, iteratively calculates a new centroid;

- Repeat steps 2-3 until the distance between the final centroid and the previous centroid is small (satisfying convergence);
- Finally, all observations are read, and each observation is classified according to the category closest to the centroid, and the classification ends.

Centroid and distance are two basic concepts of K-MEANS algorithm. Centroid can be regarded as a sample, or as a data point A in a data set, and it is defined as the center of a set of data with similarities. Centroid selection has a great impact on clustering results, because the algorithm randomly selects any object as the centroid of the initial clustering, and initially represents the clustering results. Of course, this result is usually unreasonable, just randomly partitioned data sets. In order to approximate the desired clustering results, the concrete correction of centroid requires several iterations: objects with similarities are grouped into groups, all of which have the same centroid. In addition, because of the randomness of the initial centroid selection, the final result is not necessarily expected, so it needs many iterations to get the initial centroid again randomly in each iteration until the final clustering results meet the expectations [4].

Distance is actually a measure of similarity. Common distance formulas include Manhattan distance, Euclidean distance, Minkowski distance, Chebyshev distance and so on. The most commonly used distance formula in clustering analysis is Euclidean distance, because Euclidean distance is intuitive and easy to calculate, and Euclidean distance is used for coordinate migration and change rotation of object points. Finally, the value of distance remains unchanged, so the object similarity can still be judged by the original similarity of the object. If D (x, y) is the distance between object a and object b, then d (x, y) should satisfy the following three attributes:

- Non-negative: that is, d (x, y) 30 is constant; if and only if x = y, d (x, y) =0
- Symmetry: d (x, y) = d (y, x)
- Triangle inequality: any object a, b, c has d (x, y) +d (y, z) 3d (x, z)

4.2. Application Analysis of Electronic Commerce Data in the Enterprise

In the era of big data, independent data itself is of little value. Forecasting future trends through data and discovering hidden knowledge through data are the key. Many Chinese herbal decoction enterprises keep pace with the development of the times, and there are corresponding stores selling Chinese herbal decoction on the electronic commerce website, so a large number of customers have accumulated the consumption records of purchasing Chinese herbal decoction. The analysis of these consumption records can divide consumers into groups, and different groups of consumers can personalize marketing according to their consumption behavior. Customer classification is conducive to providing differentiated services for different groups of customers. It also enables enterprises to timely detect some minor changes in the market and customers and adjust strategies for them.

RFM model is a widely used multi-factor customer classification method, R (Recency) represents the time period from the customer's recent transaction to the current time. F (Frequency) represents the number of times a customer cooperates with an enterprise (i.e. purchasing behavior) during a specified period of time. M (Monetary) represents the amount generated by the exchange between customers and enterprises within a specified period of time. RFM measures the value of customers by the absolute amount created by customers [5].

Now we crawl the relevant data from the website of a Chinese herbal medicine decoction piece e-commerce, clean and collect the original data according to certain data processing principles, and get the consumer data (3000 pieces) after processing. R here indicates the time interval between the last purchase of Chinese herbal medicines. F means the frequency of purchasing pieces of traditional Chinese medicine. M represents the total amount consumed on a certain platform. Some valid data are intercepted as shown in Table 1.

Table 1. Consumer related data

Customer serial number	R	F	M
1	27	6	334.21
2	3	5	759.19
3	4	16	1383.39
4	3	11	3280.77
5	14	7	154.65
6	19	6	501.38
7	5	2	1721.93
8	26	2	107.18
9	21	9	973.36
10	2	21	764.55
11	15	2	1251.4
12	26	3	923.28
13	17	11	1011.18
14	30	16	1847.61
15	5	7	1669.46

There are differences in numerical size and unit between different data items, so they can not be directly used to participate in the calculation. For example, the total amount of product M purchased by consumers is a large numerical attribute, the unit is generally more than 100, and the frequency of purchasing products in a certain period of time is often small, and has little effect relative to the amount of consumption. In order to make these attributes work, it is necessary to compare the attributes with their corresponding ranges to ensure that there is no difference between units and values, so that these standard data can be directly used for calculation in the later period. In this paper, data are processed by normalized processing method. Table 2 below is part of 3000 processed data.

Table 2. Consumer-related data after normalization

Customer serial number	R	F	M
1	0.000889 182	0.000329254	0.000168387
2	9.8798E-05	0.000274379	0.000382507
3	0.000131731	0.000878011	0.000697002
4	9.8798E-05	0.000603633	0.001652971
5	0.000461057	0.00038413	7.79183E-O5
6	0.00062572	0.000329254	0.000252613
7	0.000164663	0.000109751	0.000867571
8	0.000856249	0.000109751	5.40012E-05
9	0.000691586	0.000493881	0.000490414
10	6.58653E-05	0.00115239	0.000385208
11	0.00049399	0.000109751	0.000630501
12	0.000856249	0.0001 64627	0.000465182
13	0.000559855	0.000603633	0.000509469
14	0.00098798	0.000878011	0.000930893
15	0.000164663	0.00038413	0.000841134

Using the K-Means algorithm, the number of clusters is set to 3, the maximum number of iterations is 3, and the distance function uses the Euclidean distance. Since the initial centroids are random, the results for each cluster may be different. After repeated experiments, the detection clustering results are basically the same, so this clustering result can be used to analyze the cluster characteristics of the cluster users and carry out group personalized marketing. The following is a picture of the group one, two, and three generated by the K-Means algorithm clustering, as shown in the consumer group in Figure 2.

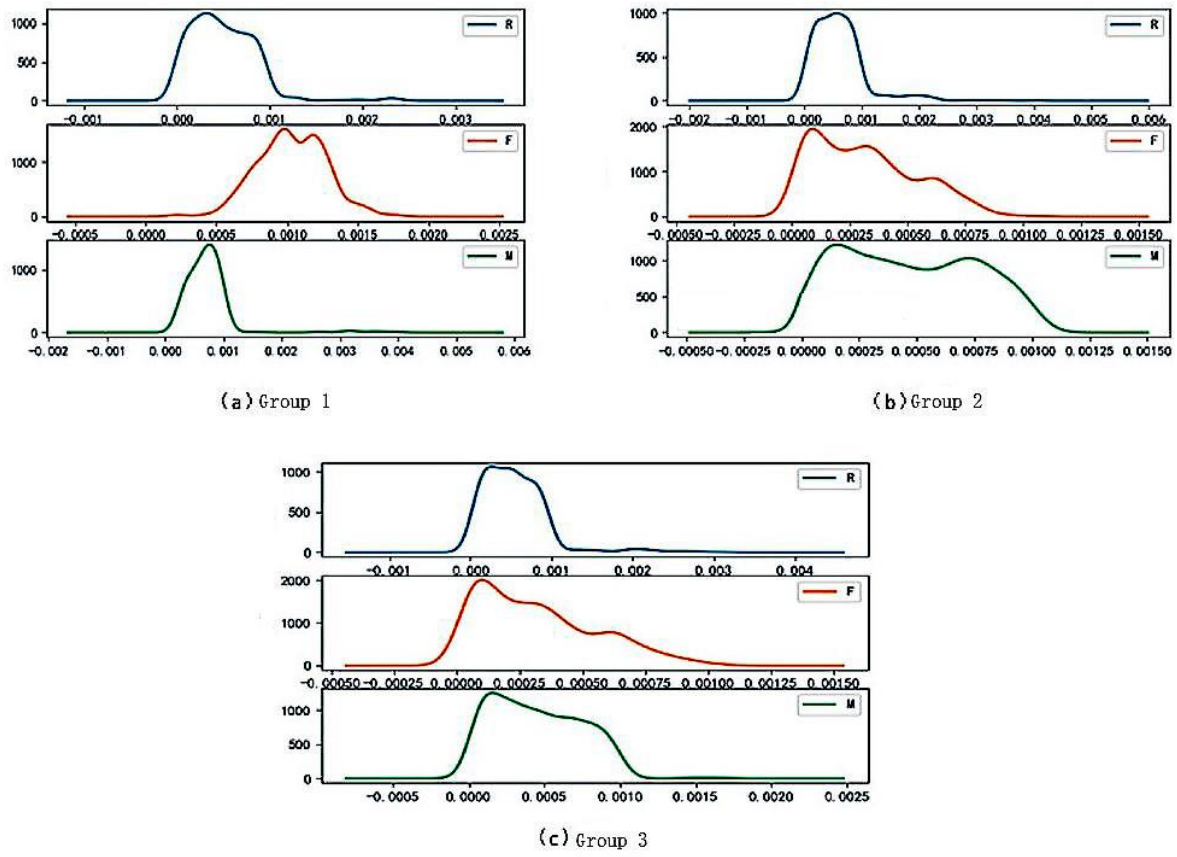


Figure 2. Consumer groups

Group 1: The last time these customers spent time interval (R) was shorter and the total amount of consumption (M) was more. They are not only the most ideal customer type, but also potential customers. They contribute a lot to the company, but their proportion is very small. Enterprises should give priority to putting resources on them to achieve differentiated management and one-to-one marketing, so as to improve the loyalty and satisfaction of such customers, and maximize the high consumption level of such customers.

Group 2: The purchase frequency (F) of these customers is general, the time interval (R) of last consumption on e-commerce website is short, and the total consumption (M) is moderate. Their customer value changes are highly uncertain, and the reasons for the decline in consumption vary. Therefore, it is particularly important to keep abreast of customer information and interact with customers in a timely manner. According to the recent consumption intervals and frequency, enterprises can predict the changes of customers'consumption behavior, focus on these customers and adopt specific marketing programs to prolong the life cycle of such customers.

Group 3: The purchase frequency (F) of these customers is general, the last time the number of days between consumption (R) is moderate, and the total amount of consumption (M) is less. They are the general users and low-value customers of traditional Chinese medicine decoction enterprises, and may only be purchased when the traditional Chinese medicine decoction pieces are discounted and promoted.

5. Conclusion

Under the background of large data, making full use of data mining information can seize market opportunities. In addition to offline sales, many enterprises also carry out online transactions with unique advantages, mining hidden information from large e-commerce data, according to these information, personalized marketing for different customer groups, thereby improving customer

satisfaction and economic benefits. This paper mainly studies the application of big data and traditional business intelligence in data analysis of electronic business enterprises, focusing on the K-Means algorithm of clustering analysis and its application in the mining of customer consumption data in e-commerce websites. Cluster analysis divides customers into three groups. According to the characteristics of different customer groups, it helps enterprises to identify customers, so as to achieve differentiated marketing objectives.

6. References

[1] Yang Chao. Research on key technologies of BI system based on big data technology [D]. South China University of Technology, 2016.
[2] cloud-based business intelligence processing technology analysis [J]. Wang Fu-min. Information and Computer (THEORY EDITION). 2017 (15)
[3] Bian Weiwei, Wang Chao, CUI really, Guo Wei, Li Hui, Zhou Miao, Xue Fu Zhong, Liu Jing based health and medical web crawler technology, big data collection and processing system [J]. Journal of Shandong University (Medical Sciences), 2017, 55 (06): 47-55.
[4] Research progress and prospects of foreign business intelligence innovation [J]. Ma Jun, Zhou Jianbo. Journal of Harbin University of Commerce (Social Science Edition). 2018(06)
[5] Lee products Rui, Xu Shou either, Xu Hui core customer relationship management research and identify the RFM model - the insurance industry as an example [J] Modern Management Science, 2015, (6): 24-26.

Acknowledgments

Thanks to the support of colleagues, especially the support of key platforms and major scientific research projects of the Guangdong Provincial Department of Education - Featured Innovation Project (Natural Science) Project "Key Technology Research of Big Data-Based Smart and Environmental Information Platform" (Fund No.: 2017GKTSCX042).