Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):562-571

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research of discovering high-value passengers of airline based on PNR data

Weidong Cao, Liang Bai* and Xiaoying Nie

College of Computer Science and Technology, Civil Aviation University of China, Tianjin, China

ABSTRACT

In the case of solving the problem of that the evaluation index of frequent flyer program is single, cannot identify high-value passengers accurately, present a method of discovering high-value passengers combines Map/Reduce and data mining. Processing gigabytes of PNR data on Hadoop by Map/Reduce parallelly, according to the improved RFD model and analytic hierarchy process, determine the customer value indexes and the weight of each index, identify the high-value passengers by data mining, and make an experiment on a real PNR dataset. The experimental result shows that, the method can effectively identify the high-value passengers of airline and provide a favorable basis for airlines to make effective decisions.

Key words: Map/Reduce; data mining; RFD model; AHP; customer value

INTRODUCTION

With the deepening of civil aviation information, a large number of PNR (Passenger Name Record) data accumulated in the booking system of every airline[1]. In the face of these valuable data resources, many airlines have not doing effective data analysis and data mining. Currently, many airlines have launched the frequent flyer program. However, the frequent flyer program just develops VIP clients according to the mileage, and improves the loyalty of passengers to the airline by integral exchange. Obviously, it cannot accurately identify high-value passengers according to the single data index. A same client may participate many airlines' frequent flyer programs at the same time. Therefore, the frequent flyer membership program in VIP system cannot form an effective attraction to passengers.

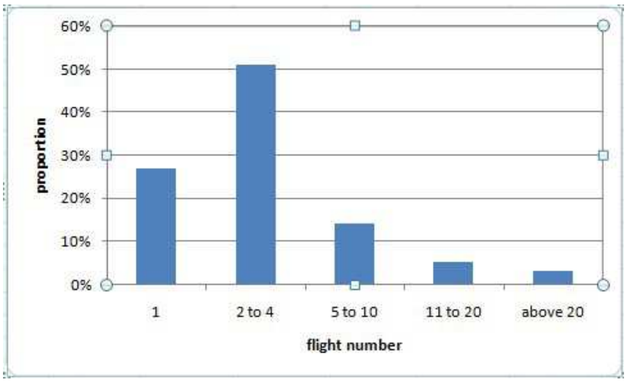

Fig. 1: The percentages of passengers of different flight number

According to the flight number of passengers, counting the booking data of passengers which is in the reservation system of civil aviation of China from 2010 to 2011, the statistic result is as follow figure 1. In figure 1, the passengers who only takes one flight in two years account for 27% of total passengers, the passengers who takes two to four flights in two years account for 51% of total passengers, the passengers who takes five to ten flights in two years account for 14% of total passengers, the passengers who takes eleven to twenty flights in two years account for 5% of total passengers, the passengers who takes above twenty flights in two years account for 3% of total passengers. It can be seen that non-frequent passengers account for above 90% of total passengers, compare to the number of frequent passengers, the number of non-frequent passengers is more massive.

In order to further investigate the ticket attributes of non-frequent, according to the ticket discount information of passengers, analyzing the booking data of non-frequent passengers which is in the reservation system of civil aviation of China from 2010 to 2011, the result is as follows figure 2. In figure 2, it can be seen that a lot of passengers had booked high-value tickets, more than 100 millions passengers would buy much more than 50 percentages off ticket, more than 8 millions passengers would buy full price ticket or first-class cabin ticket.

Therefore, to the airlines, it is an urgent question that how to use the PNR data of non-frequent flyer to identify the high-value passengers quickly and accurately in the fierce market competition, and create greater benefit in the short-term with limited resources.

The high-value passengers often choose air travel because of work needs, time, high-quality flight service or comfortable cabin environment, and the price concessions are not often important to them. Therefore, a simple definition of high-value passengers is just that[2][3], the customers who often choose air travel, and their booking class is higher than the others.

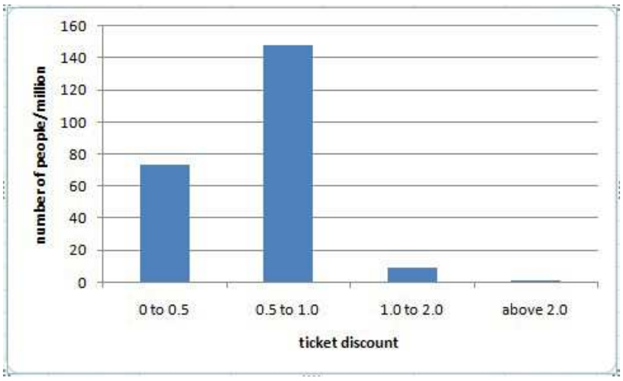

Fig. 2: The record of non-frequent book discount ticket

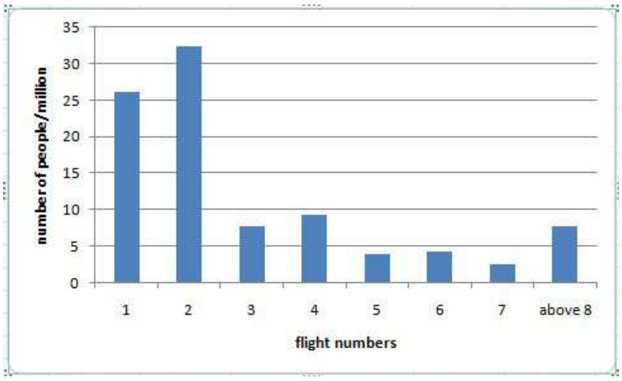

Fig. 3: The statistic of flight numbers of passengers

The routes which high-value passengers chosen are often fixed in a period of time, therefore, analyzing the short-term value of passengers in a specific route, mining the high-value passengers, and making some targeted service to them before they travel again in last route so that attracting them to fly with this airline. To the airlines, it undoubtedly has great help.

Currently, in the aspect of mining high-value airline customers, most domestic and foreign studies are combined with customer life cycle, and make data mining in the frequent flyer datasets with small amount. The result of data mining often needs long-term investment for airlines, and the utilization rate of resources is very low. Even though the precious PNR dataset of non-frequent flyer of airlines has great potential, but because of the excessive amount of data, it is difficult to handle by conventional methods[4]. Now, as an much advanced data processing model, Map/Reduce could cluster computers, parallel processing the computing task, deal with mass data resources quickly and accurately.

For this reason, the paper presents a method to analyze high-value passengers of airlines based on PNR data. At first, deal with mass PNR data resources by distributed processing mode in Map/Reduce[5][6][7][8][9][10], filtering the data which does not meet the requirements quickly and effectively. Then, according to the improved RFD customer value model, transforming and identifying the customer value indexes, determining the weight of each index by AHP(Analytic Hierarchy Process) and the experience of experts. After that, calculating customer value based on customer value indexes and weight, make clustering in the processed dataset by the optimized k-means clustering algorithm, and comparing the clustering results with the dataset mean, identifying the customer value so that discovering the group of high-value passengers, then analyzing the characteristics of the passenger groups. At last, make experiments on a real PNR dataset. The experimental results show that, the method proposed in this paper is accurate and effective, and it can identify the high-value passengers of airline quickly, so it provides a favorable basis for airlines to make effective decisions with the limited resources in short period and make better push service to the high-value passengers.

MAP/REDUCE AND THE CERTAIN OF CUSTOMER VALUE INDEXES AND WEIGHT

DISTRIBUTED PROCESSING ON MAP/REDUCE

Hadoop is an open source computing platform which is developed by apache software foundation. It has the data processing model of Map/Reduce as its main concern, and provides a distributed infrastructure which has transparent underlying details for users. As the core computing model of google, Map/Reduce could schedule and calculate different computational tasks efficiently and accurately. It highly abstracts the parallel computing process which runs on the cluster of computers as two functions: the function of map and the function of reduce. The function of map takes an input such as <key, value>, then forms an intermediate output such as <key, value-list>, all values which has the same key form a set and the set is passed to the function of reduce. The function of reduce takes an input such as <key, value>, then opens the set of values and deal with the values inside. At last, the function of reduce produces an output such as <key, value> as the processing result. Because of Hadoop has Map/Reduce which is as an efficient task scheduling model, so it allows users to develop parallel applications when they don't understand the underlying details of distributed system, and organize computer resources, build their own distributed computing platform, make full use of the computing power of clusters, deal with the mountains of data[11][12][13][14].

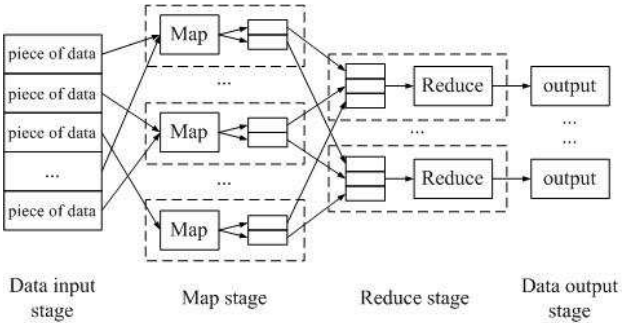

Fig. 4: The operation mechanism of Map/Reduce

THE CERTAIN OF CUSTOMER VALUE INDEXES

In customer relationship management, the RFM model is a classical model to measure customer value. It has the following three kinds of customer value indexes: Recency (recent time of consumption), Frequency (frequency of consumption) and Monetary (amount of consumption). According to the model of RFM, Goodman et al put forward to apply the limited resource of enterprises to high-value customers, enhance the use efficiency of resource, Hughes et al divide consumers into five categories, and adopt different marketing strategies to the five types of customers. However, because of the special nature of the airline industry, the traditional model of RFM is not suitable for the analysis of airline's customer value completely. Therefore, the paper modifies the three indexes of RFM as follows, and make full use of the improved model of RFD to analyze the data of high-value passengers of airlines[15-22].

R: the days between the date of a customer last takes flights in a route of this airline and the date of statistics;
F: the cumulative numbers of a customer takes flights in a route of this airline in a period of time;
D: the average discount of a customer takes flights in a route of this airline in a period of time;

THE CERTAIN OF THE WEIGHT OF CUSTOMER VALUE INDEXES

AHP(Analytic Hierarchy Process) is a simple method of combing quantitative analysis and qualitative analysis together to deal with the decision problem which is fuzzy or complex, it is put forward by A.L.Saaty, a professor of University of Pittsburgh in 1980. It makes full use of the experience and judgment of experts, and embodies the basic characteristics and development process of thinking decision. Moreover, it has the advantages of clear thinking, simple, mature, systematic and so on. Therefore, using AHP to determine the weight of each index in customer value analysis[23][24][25].

Here are the specific steps to calculate the weight by analytic hierarchy process.
(1)Constructing judgment matrix
Constructing judgment matrix of *n* orders for *n* elements $C=(c_{ij})_{n*n}$, the c_{ij} mean factors of *i* and *j* relative to the importance of target.

(2)Consistency of judgment matrix
Assume the eigenvalues of judgment matrix are $\lambda_1,\lambda_2 \dots,\lambda_n$, so $\sum_{i=1}^{n}\lambda_i = n$. When the matrix is complete consistency, $\lambda_1 = \lambda_{max} = n$, the other eigenvalues are zeros. When the judgment matrix has full consistency cannot be guaranteed, the corresponding eigenvalues of judgment matrix will change. Therefore, judge the consistency of matrix deviation by negative mean value CI of eigenvalues except maximum of judgment matrix. The calculation of CI is as follows formula (1),

$CI = (\lambda_{max} - n)/(n-1)(1)$

To give a measure indicator of consistency, it needs to introduce the mean random consistency index RI of judgment matrix of different orders. When the order is greater than 2, the ratio of consistency index of judgment matrix with consistency index of the same order mean random is called random consistency ratio, as CR, as follows formula (2),

$CR = CI/RI(2)$
When $CR < 0.1$, the judgment matrix is consistent, or it needs to be adjusted.

(1)The calculation of weight distribution
At first, calculating the weight of every effective judgment matrix, it could be attributed to a problem that calculating the maximum eigenvalue and characteristic vector of judgment matrix, here is the root method.

Calculating the product of elements of each row of judgment matrix M_i, as follows formula (3), $i = 1,2 \dots, n$

$M_i = \prod_{j=1}^{n} c_{ij}(3)$
Calculating n root of M_i, as follows formula (4),

$\overrightarrow{W} = \sqrt[n]{M_i}(4)$

Normalizing the vector $\overrightarrow{W} = (\overrightarrow{W_1}, \overrightarrow{W_2}, \dots, \overrightarrow{W_n})^T$, as follows formula (5),

$W_i = \overrightarrow{W_\iota}/\sum_{j=1}^{n}\overrightarrow{W_J}(5)$
$W = (W_1, W_2, \dots, W_n)$ is just the vector what wants, is also the corresponding weight coefficient.

After calculating the weight of each effective judgment matrix, the arithmetical average of weights of above effective judgment matrixes is the weight of each factor.

CLUSTER ANALYSIS BASED ON THE K-MEANS CLUSTERING ALGORITHM OF OPTIMIZING THE INITIAL CLUSTER CENTER

DATA STANDARDIZATION

Because of the magnitudes of different consumer value indexes vary widely, the effects of three factors are imbalance obviously, so that data values will make a big difference. In order to eliminate the influence of imbalanced distribution and different magnitudes, dataset requires standardized data processing before the clustering analysis. In general, there are two kinds of data standardization mode, the first is to convert the dataset, make all values of the dataset in the 0 ~ 1; The second is to standard the dataset, make it obey the standardized normal distribution of the average number is 0 and the standard deviation is 1; In this study, taking the first data standardization method.

Suppose X is R, F or D variable, X^L is the maximum value of R, F or D variable in the dataset, X^S is the minimum value of R, F or D variable in the dataset, X' is X variable after standardization.

To R variable, its value is more higher, the value of consumer is more lower. Therefore, R variable has a negative effect on customer value, it is a negative correlation index, standardizing it by formula (6),

$$X' = (X^L - X)/(X^L - X^S)(6)$$

To F or D variable, its value is more higher, the value of consumer is more higher. Therefore, F or D variable has a positive effect on customer value, it is a positive correlation index, standardizing it by formula (7),

$$X' = (X - X^S)/(X^L - X^S)(7)$$

K-MEANS CLUSTERING ALGORITHM OF OPTIMIZING THE INITIAL CLUSTER CENTER

K-means clustering algorithm is a kind of much classical and mature clustering algorithm, the greatest characteristic of this algorithm is be able to make the data which is in the same cluster have high similarity, but the data which is in the different clusters have low similarity. Moreover, k-means clustering algorithm has the advantages of a small amount of calculation and rapid constringency speed. When deal with big datasets, compared with other clustering algorithms, k-means clustering algorithm occupies much smaller memory space and computing time[26][27]. The specific application procedure of K-means clustering algorithm as follows:

(1)Inputting a dataset that has N pieces of data, and
appointing the amount of cluster is K. Make $I = 1$, select k nodes in the dataset as the initial cluster center $Z_j(I)$, $j = 1, 2, \ldots, k$;

(2)Computing the distance between each data in the
dataset and the k initial cluster centers $D(x_i, Z_j(I))$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, if meet $D\left(x_i, Z_m(I)\right) = min$ $\{D\left(x_i, Z_j(I)\right), j = 1, 2, \ldots, k\}$, putting x_i into cluster m;

(3)According the formula (8) to compute the sum
Squared error criterion function J_c, and judging that if meet $|J_c(I) - J_c(I-1)| < \xi$, the algorithm terminates.

$$J_c(I) = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| {x_i}^{(j)} - Z_j(I) \right\|^2 (8)$$

Or make $I = I + 1$, according to the formula (9) to compute the new cluster centers, and return (2)

$$Z_j(I) = \frac{1}{n_j} \sum_{i=1}^{n} {x_i}^{(j)} (9)$$

However, it is inevitable that the dataset has some isolated data nodes, which are far from the data intensive area. Because of that it is random to select initial clustering center, it will make the clustering result produce deviation. Therefore, improve the classic k-means clustering algorithm. At first, exclude isolated data nodes according to the thought of sum of distance, optimize the selection of initial clustering centers, then make clustering on the dataset which excludes isolated data nodes, the isolated data nodes are clustered at last[28].

According to the thought of sum of distance, compute the distance between data nodes in the dataset, list the matrix of sum of distance (see table 1, $d(i,j)$ is euclidean distance; $D(i,j)$ is the sum of distance; $d = sqrt((x_2 - x_1)^2 + (y_2 - y_1)^2 + \cdots + (z_2 - z_1)^2)$;), screen the data node which has the biggest sum of distance with the other data nodes. According to the accuracy requirements, screen M data nodes, make the isolated data nodes not participate the selection of initial clustering centers, so avoid the clustering results emerge big deviation.
After screen M isolated data nodes, make another matrix of sum of distance with the dataset which has left $\mathrm{N} - \mathrm{M}$ pieces of data, so that find out the two data nodes that one is the farthest away from the other one. In this study, the data which is in the dataset is three-dimensional, make a line segment with above two data nodes, set the center of line segment as the centre of sphere, set the line segment as the diameter, draw a sphere. Then, set the center of line segment as the centre of sphere, set half of the line segment as the diameter, draw an endosphere. Set the centre of sphere as origin, build a three-dimensional Cartesian coordinate system, in the eight quadrants, take the center in each arc as the initial cluster center. After selecting the initial cluster center, make clustering according to the classic k-means clustering algorithm. At last, compute the distance between each isolated data node and cluster centers, and sort out the isolated data nodes.

Tab. 1: Martix of sum of distance

	node 1	*node 2*	…	*node N*	*D(i, N)*
node 1	0	*d(2,1)*	…	*d(N,1)*	$\sum_{i\geq 1}^{N} d(i,1)$
node 2	*d(1,2)*	0	…	*d(N,2)*	$\sum_{i\geq 1}^{N} d(i,2)$
…	…	…	…	…	…
node N	*d(1,N)*	*d(2,N)*	…	0	$\sum_{i\geq 1}^{N} d(i,n)$

EXPERIMENTAL ANALYSIS

DATA PREPROCESSING

In this study, make an experimental analysis on a real PNR dataset, which is provided by an information company that includes all domestic passengers' travel data from January 1, 2010 to February 28, 2011, the size of dataset is 48.6G, processed by five PCs which are clustered, the time of processing is four hours. Suppose the statistical date is January 1, 2011, so R variable represents the distant days between statistical date and the date of the passenger last takes flights. At first, make full use of the data processing model of Map/Reduce to preprocess the big original dataset. Then, compute the PNR data of passengers who have same ID and same route as follows:

(1)Get the maximum value of departure date;
(2)Count the number of the PNR data which has same identity ID and same route;
(3)Get the average value of flight discount;

Tab. 2: The processed PNR dataset

	ID	*Airl*	*Orig*	*Dest*	*Rece*	*Freq*	*Disc*
1	265	290	SZX	PEK	23	4	1.021
2	318	155	PEK	SHA	25	5	0.979
...	...	...	...	...	...	...	...
n	411	315	PEK	CAN	3	4	1.047

Because of the three basic conditions of identifying high-value passengers are travelling frequently, choosing high classes and having travel records in short term. Therefore, the data which not meet above conditions is not in the range of this research. So, filtering these data according to the following conditions:

(1)The travel number is greater than or equal to four in recent two years;
(2)The average travel discount is greater than or equal to four;
(3)Have travel records in the last sixty days;

DETERMINING THE WEIGHT

In the determination of airline's customer value indexes, invited many experienced experts of civil aviation area to participate, use questions and AHP to analysis the relative importance of customer value indexes.

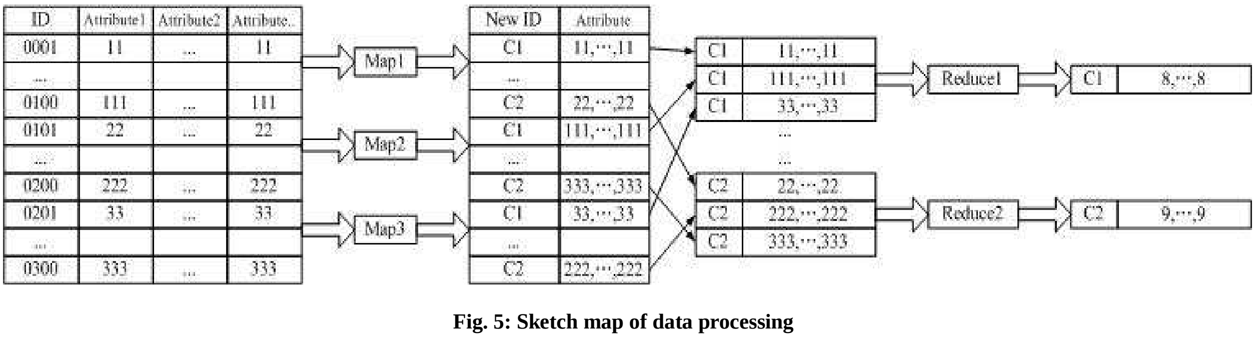

Fig. 5: Sketch map of data processing

At first, let the experts compare the relative importance of the three customer value indexes by nine scales method and make judgment matrix. Then, test the consistent of each expert's judgment matrix, as a result, there are eight judgment matrices conform to the consistency requirement, these judgment matrices are effective judgment matrices. At last, compute each effective judgment matrix's weight, take the arithmetic mean among these weights as each customer value index's weight[29]. The final judgment result is : $w_R = 0.1036,\ \ w_F = 0.3705,\ w_D = 0.5259$.The result shows that, the weight of D variable is bigger than the other two. Hence, the experts agree that the average ticket discount is the most important factor that affects the customer value.

CLUSTER ANALYSIS

In order to make the analysis more representative, in this cluster analysis, choose three golden routes which is from Beijing to Shanghai, from Beijing to Guangzhou and from Beijing to Shenzhen as the study object.

Before the cluster analysis, the dataset need to be standardized. Then, according to the determined customer value indexes' weight and the normalized PNR data, compute each consumer's value by formula (10):

$Value = w_R * V_R + w_F * V_F + w_D * V_D(10)$

Among them, V_R, V_F, V_D is separately the customer value index of *R, F, D* after standardization.

According to the three customer value indexes of distance days, travel numbers, average ticket discount, make cluster analysis on the dataset by k-means cluster algorithm of optimizing initial cluster center. The cluster algorithm must give the number k of clusters in advance, because of the average of customer value index of each cluster in the cluster result needed to compared with the average of customer value index of full dataset, and there are only two conditions of comparing each customer value index, bigger or smaller, therefore, in the case of the dataset is uniform distributed, the number of clusters is 2*2*2=8, so k=8.

After get the result of clustering, compare the average of customer value index of each cluster in the cluster result with the average of customer value index of full dataset, so that more clearly show customer feature of each cluster, moreover, provide more favorable reference for airlines

Tab. 3: The result of clustering in the route which is from Beijing to Shanghai

Cluster	Rate	*R*	*F*	*D*	*Value*
1	5%	7.46	9.31	1.97	0.28
2	16%	27.58	6.20	0.98	0.13
3	16%	39.97	5.89	1.05	0.11
4	17%	17.66	6.59	1.00	0.15
5	3%	26.21	6.86	1.97	0.24
6	16%	2.96	7.85	1.00	0.18
7	15%	9.89	7.02	0.99	0.16
8	13%	53.53	5.65	1.08	0.09
Dataset	100%	23.57	6.70	1.09	0.15

Tab. 4: The result of clustering in the route which is from Beijing to Guangzhou

Cluster	Rate	*R*	*F*	*D*	*Value*
1	11%	53.97	5.63	0.97	0.16
2	13%	29.15	6.12	0.95	0.20
3	16%	10.95	6.16	0.95	0.23
4	17%	20.05	6.31	0.95	0.22
5	3%	40.55	6.40	1.78	0.38
6	18%	3.38	7.60	0.96	0.26
7	7%	11.80	7.61	1.78	0.44
8	14%	40.39	5.83	0.94	0.18
Dataset	100%	23.68	6.43	1.04	0.23

Tab. 5: The result of clustering in the route which is from Beijing to Shenzhen

Cluster	Rate	*R*	*F*	*D*	*Value*
1	5%	9.20	15.27	1.06	0.29
2	19%	6.24	5.68	0.95	0.18
3	15%	53.91	5.34	1.01	0.10
4	17%	29.23	5.33	0.98	0.14
5	16%	41.71	5.52	1.02	0.12
6	20%	17.33	5.56	0.93	0.16
7	4%	12.69	6.07	1.87	0.28
8	3%	29.55	13.10	1.06	0.23
Dataset	100%	26.62	6.23	1.02	0.16

Tab. 6: The result of comparing the average of each cluster with full dataset in the route which is from Beijing to Shanghai

Cluster	*R*	*F*	*D*	*Value*
1	↓	↑	↑	↑
2	↑	↓	↓	↓
3	↑	↓	↓	↓
4	↓	↓	↓	↓
5	↑	↑	↑	↑
6	↓	↑	↓	↑
7	↓	↑	↓	↑
8	↑	↓	↓	↓

Tab. 7: The result of comparing the average of each cluster with full dataset in the route which is from Beijing to Guangzhou

Cluster	*R*	*F*	*D*	*Value*
1	↑	↓	↓	↓
2	↑	↓	↓	↓
3	↓	↓	↓	↓
4	↓	↓	↓	↓
5	↑	↓	↑	↑
6	↓	↑	↓	↑
7	↓	↑	↑	↑
8	↑	↓	↓	↓

Tab. 8: The result of comparing the average of each cluster with full dataset in the route which is from Beijing to Shenzhen

Cluster	*R*	*F*	*D*	*Value*
1	↓	↑	↑	↑
2	↓	↓	↓	↑
3	↑	↓	↓	↓
4	↑	↓	↓	↓
5	↑	↓	↓	↓
6	↓	↓	↓	↓
7	↓	↓	↑	↑
8	↑	↑	↑	↑

CLUSTERING FEATURE ANALYSIS

From table 3 and table 6, it can be seen that, the customers in first cluster are $R \downarrow F \uparrow D \uparrow V \uparrow$ and the customer value is the highest in full dataset. Therefore, to the airlines, these customers are the most valuable commercial passengers. They often take the airline's flights between two cities, and the average ticket price is high, more concentrated in the first-class cabin. This part of passengers creates more considerable profit than others, airlines should focus their limited resources to serve this part of passengers, so that retain them, and establish long-term friendly relationship with them.

The customers in fifth cluster are $R \uparrow F \uparrow D \uparrow V \uparrow$ and the customer value is higher than others. R variable of this part of passengers is greater than the average, it is possible that they have selected other airlines when they travel again in this route in recent. This part of passengers have no fixed airline's flights when they travel, therefore, each of airlines has the same chance to establish close relationship with them. Airlines should pay special attention to further development of these customers, and carry out some targeted market promotions to them, so that improve airlines' attractive.

The customers in sixth and seventh are $R \downarrow F \uparrow D \downarrow V \uparrow$ and the customer value is a little higher than others. This part of passengers travel frequently in this route but take a little lower ticket price than others. Based on their consumption behavior, airlines should launch more flights service activity, for example, if take the ticket discount up to 120%, just enjoy the first-class cabin service of the ticket discount up to 150%, so that stimulate their consumption.

The average of customer value index of other clusters is close to or lower than the average of customer value index of full dataset. To the airlines, their commercial value is low, airlines should not put the limited resources to attract these passengers.

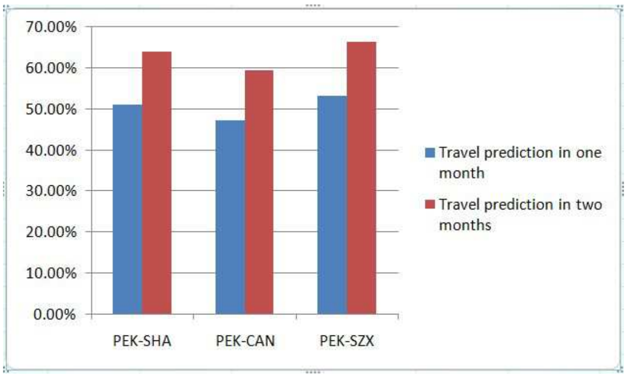

Fig. 6: The result of test accuracy

PREDICTION TEST

Test the probability of high-value passengers keep on travelling in a route in future short time which is unearthed by above method with the PNR dataset of the previous two months of 2011, the result of test accuracy is as above figure 6, the probability of high-value passengers keep on travelling in the route of Beijing to Shanghai in January, 2011is 51.1%, in January and February, 2011is 63.9%. The probability of high-value passengers keep on travelling in the route of Beijing to Guangzhou in January, 2011is 47.2%, in January and February, 2011is 59.5%. The probability of high-value passengers keep on travelling in the route of Beijing to Shenzhen in January, 2011is 53.1%, in January and February, 2011is 66.4%. It can be seen, the forecasting accuracy is high, this method is fast and effective.

CONCLUSION

The PNR data in reservation system of airlines based on non-frequent flier is very valuable, airlines could dig out high-value passengers in it. The majority of these customers travel frequently between two cities in a certain period of time because of work, and take higher ticket price than others. They are not constraint by specific airlines, therefore, which airline could find this part of passengers quickly and accurately and establish a long-term friendly relationship with them, which airline will be outstanding in the fierce competition. The paper presents a method of discovering high-value passengers combines Map/Reduce and data mining, and make experimental analysis on a real PNR dataset. The result shows that, although in the face of a PNR dataset which has massive data, the method also could deal with easily, and it could effectively identify high-value passengers, provide reference for airlines to focus limited resources in a short period of time and make effective decisions.

Acknowledgments

The authors wish to thank anonymous reviewers for their valuable comments to improve the quality of the paper. The work is supported by the Project of the Nation Natural Science Foundation, China (No.60879015) and the Civil Aviation Administration of China Research Foundation, China (No.MHRD201130).

REFERENCES

[1] J. Manyika, M. Chui, B. Brown and J. Bughin, *Big Data: The Next Frontier for Information, Competition and Productivity*, http://www.fujitsu.com/downloads/svc/Fla/03_Michael.Chui. pdf, November **2012**.

[2] P. Liu, *Modeling of Aviation Customer Value Based on Data Mining*, Guangzhou, South China University of Technology, July **2010**.

[3] A. Walter, T. Riter, H. G. Gemilnden, *Value Creation in Buyer-Seller Relationship*, Industrial Marketing Management, no. 30, **2001**.

[4]J. Gantz, D. Reinsel, *The Digital Universe Decade Are You Ready*, http://viewer. media.bitpipe.com/9380448859_264/1287663101_75/Digital_Universe.pdf, November **2012**.

[5] M. Dou, L. J. Wen, J. M. Wang, *Parallel Algorithm to Convert Big Event Log Based on MapReduce*. Computer Integrated Manufacturing System, vol.19, no. 8, pp. 1784-1793, August **2013**.

[6] W. M. P. Van. Der. Aalst, K. Van. Gee, J. M. Wang, *Workflow Management: Models, Methods, And Systems*, Cambridge, Mass, USA: MIT Press, **2004**.

[7] C. W. Gunther, *Process Mining in Flexible Environments*. Eindhoven, the Netherlands: Eindhoven University of Technology, **2009**.

[8] C. Bratosin. *Grid Architecture for Distributed Process Mining*. Eindhoven, the Netherlands: Eindhoven University of Technology, **2006**.

[9] H. Geguieg, F Toumani, H. R. Motahari-Nezhad. *Using MapReduce to Scale Events Correlation Discovery for Business Processes Mining.* Business Process Management, Berlin, Germany: Springer-Verlag, pp:279-284, **2012**.
[10] J. Dean, S. Ghemawat, *Communication of the ACM*, vol. 1, no. 51, pp. 107-113, January **2008**.
[11] T. White, *Hadoop: The Definitive Guide.* O'Reilly Media, lnc, **2009**.
[12] J. H. Lu, *Hadoop in Action*, Beijing, China Machine Press, **2011**.
[13] Y. Yang, X. Long, B. Jiang, *Journal of Computer*, vol 8, no. 10, **2013**
[14] R. Li, J. H. Luo, D. Yang, H. B. Hu, L. Chen. *Journal of Computer*, vol 8, no. 9, **2013**
[15] X. B. Xu, J. Q. Wang, H. Tu, M. Mu, *Journal of Computer Applications*, vol. 5, no. 32, pp. 1439-1442, May **2012**.
[16] J. W. Han, M. Kamber, *Data Mining: Concepts and Techniques*. Beijing, China Machine Press, **2002**.
[17] M. Maia, J. Almedia, V. Almedia, *Identifying User Behavior in Online Social Networks.* Proceedings of the 1st Workshop on Social Network Systems, New York: ACM, **2008**.
[18] J. Caverlee, S. Webb, *A Large-scale Study of Myspace: Observations and Implications for Online Social Networks*, http://faculty.cs.tamu.edu/caverlee/pubs/caverlee08alarge.pdf, October **2011**.
[19] C. H. Liu, Q. Mei, S. Q. Cai, *Technoeconomics & Management Research*, vol. 5, pp. 33-36, May **2012**.
[20] H. W. Shin, S. Y. Sohn, *Experts Systems with Applications*, vol. 1, no. 27, pp. 27-33, January **2004**.
[21] N. C. Hsieh, *Experts Systems with Applications*, vol. 4, no. 27, pp. 623-633, April **2004**.
[22] K. N. Lemon, T. B. White, R. S. Winer. *Journal of Marketing*, vol. 1, no. 66, pp. 1-14, January **2002**.
[23] X. F. Luo, A. H. Ren, M. Li, H. Y. Ren, *Computer Engineering and Design*, vol. 12, no. 31, pp. 2749-2753, December **2010**.
[24] D. F. Liu, *Computer Engineering and Design*, vol. 3, no. 34, pp. 894-898, March **2013**.
[25] F. Wang, J. G. Liu, Y. H. Chen, *Computer Systems & Applications*, vol. 1, pp. 26-28, January **2009**.
[26] W. X. Zhang, *Research on Frequent Flyer Segmentation of Airlines,* Nanjing, University of Aeronautics and Astronautics, **2009**.
[27] F. D. Wang, Y. F. Ma, *Computers Engineering and Applications*, vol. 4, no. 47, pp. 215-218, April **2011**.
[28] A. W. Zhou, Y. F. Yu, *Computer Technology and Development*, vol. 2, no. 21, pp. 61-65, February **2011**.
[29] L. S. Luo, W. X. Zhang, *Research of Method Customer Segment of Airlines Based on Database of Frequent Flyer*, Modernbusiness, **2008**.