Data Mining Recognition and Promotion under Online Questionnaire Survey

Ping Yin[1], Min Li[1], Qingpeng Li[2]
[1]Shaanxi Industrial Vocational and Technical College
Xianyang, Shaanxi
712000, China

[2]Northwest University of Technology
Xi'an, Shaanxi
710000, China
vrwf7774591@163.com

ABSTRACT: *This article explores the precautions for online questionnaire surveys in the recognition and promotion of information data mining. For example, attention should be paid to the quality and rationality of questionnaire design to avoid overly complex or ambiguous questions. Pay attention to protecting the privacy and personal information of the respondents. We should fully consider the applicability and accuracy of data mining technology to avoid misleading or erroneous conclusions. Summarized the application value and advantages of online questionnaire surveys in information data mining recognition and promotion. The combination of online questionnaire surveys and data mining technology can help enterprises better understand market demand and consumer behavior, develop more accurate and personalized promotion plans, and improve the effectiveness and efficiency of market promotion. At the same time, this application can also help enterprises reduce the cost and risk of market promotion, and improve their competitiveness in the market.*

Keywords: Data Mining, Communication Data, Algorithm

Received: 19 March 2023, Revised 13 June 2023, Accepted 22 June 2023

DOI: 10.6025/pms/2023/12/2/29-37

1. Introduction

To meet the WTO challenge, the government's diversified carrier strategy makes the competition in the telecom market increasingly fierce [1]. The potential profitability of mobile communications is greater than that of other telecommunications businesses. Therefore, after China's entry into WTO, the most competitive communication business is mobile business besides Internet business [2]. From the development trend of world communication, one of the trends of communication network development is wireless [3]. According to IDC, ITU and *M*, in 1990-1999, the average growth rate of telecom in the world was 0%. The telephone line is increased by an average of 17%. The average growth rate of mobile phones was 49.2%. In 1990-1999, the revenue of China's telecom business increased by an average of 39.2%, and the telephone line increased

by 36% on average, and the average growth rate of mobile phones in 1992-1999 was 137%. Mobile phones and Internet users are growing much faster than landline phones [4]. Mobile communication has gradually become the main growth point of China's telecom business [5]. The proportion of long distance business income in China has been declining continuously. Revenue from local telephone service increased rapidly in 1990-1998, but has declined since 1999. And the share of mobile revenue has gone up since 1995. By 1999 its share was close to that of long-distance and local businesses [6]. As of the first quarter of this year, the total number of telephone users in China reached 350 million. Among them, the number of fixed telephone users increased by 9.61 million to 190 million. The number of mobile phone users increased by 16.88 million, bringing the total to 160 million. Mobile phones have grown faster than landline phones. The number of mobile phone users in China is expected to reach 280 million by 2005 [7].

2. State of the Art

Telecommunications is the world's fastest growing industry. As the industry grows, so does the challenge and competition. In order to meet the challenge, telecom operators explore ways to better understand the customers' needs, and the use of information has become the key to survival [8]. For a relatively mature mobile operator, the vast amount of historical data accumulated by operating and supporting systems is undoubtedly a valuable asset. Data mining system is one of the most effective methods, and means to make full use of these valuable resources to achieve the above three goals [9]. However, we still believe that there is some research on the application of data mining in China's mobile communications industry. This is based on the following gaps in previous studies: A lot of the research that was done in the past was for telecoms, and not so much for the telecommunications industry. Therefore, based on the predecessors, this article will make the data mining more in-depth and targeted in mobile communication operation. In the past, more research was conducted on data warehouse, but the data mining based on data warehouse was not discussed enough. Our domestic research on this aspect is just at the beginning stage, and the foreign countries have been greatly advanced in this field. We hope that this research can be combined with China's national conditions to strengthen, and improve the application of data mining technology in mobile communication operation in China [10]. To sum up, this article will focus on how to carry out and use data mining technology, to improve the competitiveness of mobile communication operation enterprises in China.

3.Methodology

3.1. Design of Communication Data Mining Model

Value is generally regarded as the result of the product (service) cost performance, or the control of the acquisition cost. In the field of value theory, some scholars have considered the two attributes of value emotion and cognition, and put forward the general standard model. The model consists of three dimensions: external attributes; intrinsic properties; System properties. External attributes reflect the special purpose of being a value item (service). Intrinsic attributes represent an emotional assessment of value items (services). System attributes represent the rational and logical aspects of value items (services). Mattsson (1991) introduced Hartman's model into service marketing, and proposed three new dimensions: emotional attribute (E), practical attribute (P), and logical attribute (L). The emotional dimension represents a complete physical experience of value in psychology. The utility dimension reflects the functional and logical aspects of value. The logical dimension contains the rational aspects of value.

In the whole customer life cycle, it involves the interaction between the enterprise and the customer, and the customer's contribution to the profits and expenses of the enterprise. Based on this, many studies have put forward the concept of Lifetime Value (Lifetime Value) LTV. To compare the value of the full lifecycle of a definition is: in terms of customer relationship, enterprise start to the end of the cycle of the whole customer life cycle, a single customer direct contribution to enterprise expenses and costs (trading), and indirect contribution (recommend, recommendation of new products, etc.) the full value of combined. It can be said that LTV includes potential customer value and existing customer value. It includes both historical and future data. The definition of LTV in jannyc. Hoekstra (1999) contains two parts to calculate the total value of a customer. The first part refers to the direct financial benefit of a supplier, which is the total amount purchased by a customer. The second part refers to the influence of a customer's non-purchase behavior on the profit of the supplier. The effect may be positive or negative. Positive effects such as recommended supplier, provides information about the supplier service or product, to participate in new product development (e.g., put forward new product ideas, innovative USES of information sharing products, testing of new products, etc.). An example of the negative impact of customer behavior on supplier profits, it is to complain about the company's products or services in front of other customers or potential customers. At any time throughout the life cycle, LTV consists of two components, which is the historical and future value. The historical

part is the discounted present value of all past sales; the future is the net present value of all future sales. LTV now used for the following six kinds of decision-making of an enterprise, customer segmentation, measure the strength of customer relationship and evaluation and the desired selection, quality of customer resources, customer communication media options and loyalty program. LTV measurement usually takes two dimensions: time dimension (past and future) and data source dimension (supplier and customer). The following is a data table for measuring LTV values.

	Supplier	**Customer**
Past times	I customer qualityA period of time as a customerThe number of products sold at a certain periodSales volume for different products of the same customerSales per periodStart the total sales from the first transactionProfit contribution in each periodProfit contribution from the first transaction	IICustomer satisfaction with product related servicesCustomers' atisfaction with the purchase of products last yearCustomer budgetCustomer recommendation to the companyThe ratio of the customer to the budget of the companyConversion cost(perceived by the customer)
Future	III potential customersSales forecastPrediction of customer life cycleSales trendProfit forecast	IV potential suppliersRepeat purchase intentionThe degree of willingness to recommend companiesChanges in the proportion of the customer's expenditure on the budgetChanges in customer related budgets

Table 1. The Two Dimensions of LTV

Therefore, it can be concluded that the LTV of a customer *j* at time *p* is:

$$LTV_j = \sum_{t=0}^{p} CQ_{jt}(1+r)\,p - t + \sum_{t=\mathrm{p}+1}^{n} (CS_{jt} \mathrm{x}\, CP_{jt})(1+r)^{p-t}\ t = 0,...p, ...\, n \qquad (1)$$

CQ_{jt} is the customer quality, *f* is the unit time sales, profit contribution, different product quantity. CS_{jt} is the customer's share, expressed by *f* (*SQjt, SPjt*). SQ_{jt} is the supplier quality, with *f* (customer satisfaction, identity, trust). SP_{jt} is a potential supplier, using f(purchase intention, expected customer share, budget line). CP_{jt} is a potential customer*; R* is the discount rate; *P* is the time from the first transaction.

According to the "80/20" rule of marketing, 20% of the customers generate 80% of the profits for the enterprise. In fact, the distribution of profit contribution in most industries is even worse. Perhaps more than 100% of the profits are in less than 10% of the customer base. A study of 35 companies, including financial services, telecommunications and retail, conducted by John McKean, found that: The highest percentage of customers with a profit value is 25%, with a minimum of 2% and an average of 15%. To maintain a competitive edge in the competition, companies have to hold on to a very small number of high-value customers. Generally speaking, there are three main ways to distinguish high-value customers from many customers: According to previous trading records, the most profitable customers are found. A potential assumption of this approach is that "the customer will repeat the past behavior", that is, the past generation of high-margin customers will continue to generate high profits. The advantage of this approach is that it is easy to understand and calculate the data easily available, but it does not reflect the future of the customer. Perhaps the customer's consumption capacity has been developed and the future consumption is a level or a decline. The increased marketing input did not generate a simultaneous increase in profits. So it's not economical. Methods for customer potential value. This approach assumes that the value created by the customer in the past belongs to the past and that the enterprise cannot change. What companies really care about is how much value the customer can create in the future, and the value of the customer that is easily overlooked in the enterprise is the potential part. The potential value of the customer is defined as the profit and value of the customer's future behavior. Based on the potential value and the current value of the customer, it can be represented by a 2-by-2 matrix of the following table.

Customer value V = potential value PV+ current value CV. The first quadrant of the table has the lowest total value, which is not attractive to the enterprise. The fourth quadrant has the highest total customer value, with the highest priority in customer relationship management. In the second quadrant, the current value is lower and the potential value is higher. This kind of customer cannot make the enterprise profit now, but it should become the target customer of the enterprise customer relationship management development. The third quadrant has a high current value and low potential value as an indicator of development potential. This kind of customer has become the loyal customer of the enterprise, try to maintain this kind of customer relationship and make the enterprise gain more profit.

		Current value	
potential value		**low**	**high**
	high	II	IV
	low	I	III

Table 2. Customer Value Matrix Diagram

The current value mainly focuses on factors such as the profit growth and the change of cost that the customer brings directly to the enterprise. Assume that the customer and enterprise keep trading time for N years, initially used to attract customers cost (mainly refers to the marketing costs) for C, customer purchase price of the product for the first time for P_1, firms expect every year to get income of R from the customers, can get the current value of the customer CV to:

$$CV = P_1 - C + R \cdot \left[\frac{(1+i)^N - 1}{i\ (1+i)^N} \right] \tag{2}$$

The value of customer potential value (PV) is based on customer history and current behavior. The potential value of a single customer can be predicted by linear regression model. According to the matrix of the above table, we classify customers and predict their potential value by using probability model. Suppose customer I buys product j, then:

$$Y_{ij}^{*} = \beta_j X_i + \sum_{k=1}^{j} \gamma_{jk} Z_{ik} + \varepsilon_{ij} \tag{3}$$

$$\begin{cases} Y_{ij} = 1 & \text{When} \quad Y_{ij}^{*} > 0 \\ Y_{ij} = 0 & \text{When} \quad Y_{ij}^{*} < 0 \end{cases} \tag{4}$$

Among them, Y_{ij} represents whether customer I owns product j. Y_{ij}^{*} is a variable, Xi is the demographic index of customer I (such as age, income, etc.). Z_{ik} for customer I already has the purchased product or service k. ε_{ij} is the error term. The probability model of the potential value of customer I is as follows:

$$PV_i = \sum_{k=1}^{k} Prob\,(Y_{ik} = 1) \times Profit_k \tag{5}$$

3.2. Hybrid Data Mining Algorithm

The filling algorithm in mixed state mainly includes EM algorithm, Ml algorithm and RE algorithm. The following is a detailed introduction. This paper mainly USES the EM algorithm. For the EM algorithm, let's say $Y = \chi\beta + \varepsilon = \beta_0 + X\beta\sum + \varepsilon$. Where Y is n×1 response variable, χ is n x (P + 1) order matrix. So the first column has all 1's, so it's χ minus the first column. $\beta = (\beta_1, \beta_2, ..., \beta_p)$ is the regression coefficient vector, ε is n ×1 random error vector, and its distribution is $N(0, \sigma^2 I)$. There is a data deletion

in the matrix $Z=(Y, X)$, that is, partial data of some variables are incomplete. In 1977, the algorithm proposed by Dempster, Larid and Rubin was an iterative method which was widely used to deal with missing data problems. Each iteration of the EM algorithm consists of two steps: expected value (E step) and maximum value (M step). For mixed deletion, the missing mode of matrix $Z=(Y, X)$ can be used to establish a similar EM algorithm. Let $Z=(Y, X)$ obey the multivariate normal distribution, namely:

$$Z_i(y_i, x_i) \sim N_p(\mu, \Sigma) \tag{6}$$

Among them:

$$\mu=\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \Sigma=\begin{pmatrix}\Sigma_{yy} & \Sigma_{yx}\\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix} \tag{7}$$

In the matrix $Z=(Y, X)$, Z_{obs} is known, and the logarithmic maximum likelihood function of parameter $\theta=(\mu, \Sigma)$ is:

$$\begin{aligned} Q(\theta/\theta^{(t-1)}) &= E\left[logf(Z_{obs}, Z_{mis}/\theta)/Z_{obs}, \theta^{(t-1)}\right] \\ &= \int_{Z_{mis}} logf(Z_{obs}, Z_{mis}/\theta)/f(Z_{mis}, Z_{obs}/\theta^{(t-1)})\, dZ_{mis} \\ &= \int_{Z_{mis}} logf(X,Y/\theta)/f(Z_{mis}, Z_{obs}/\theta^{(t-1)}) dZ_{mis} \end{aligned} \tag{8}$$

Where, $\theta^{(t-1)}$ is the estimation of parameter θ obtained by the t--l step iteration. $Q(\theta/\theta^{(t-1)})$ represents the logarithmic maximum likelihood function of parameter θ under $\theta^{(t+1)}$, which is the conditional expectation of $logf(Z_{obs}, Z_{mis}/\theta)$ in $\theta^{(t-1)}$ and Z_{obs}, and $f(*)$ is the probability distribution function. In the following M steps, the expectation function $Q(\theta/\theta^{(t-1)})$ is maximized to obtain the estimation $\theta^{(t)}$ of the parameter θ of the t step iteration.

$$\theta^{(t)}=argmax\, Q(\theta/\theta^{(t-1)}) \tag{9}$$

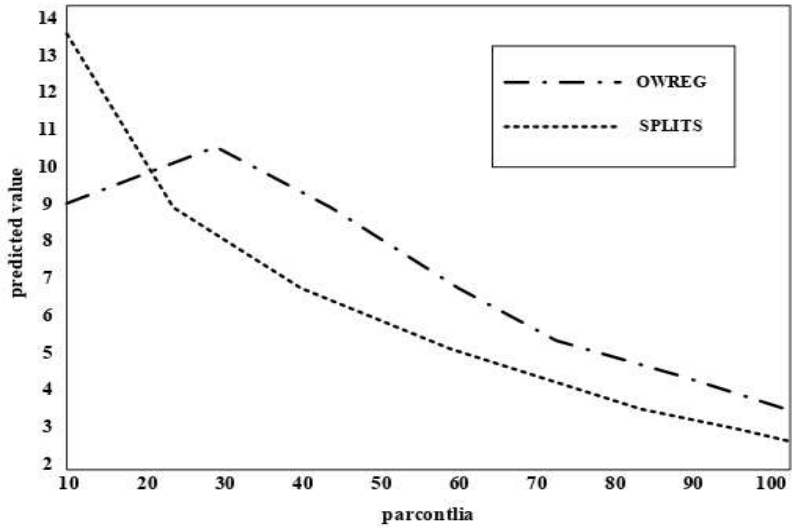

Figure 1. The predictability of the CVM model on a test set

And then, use the new maximum value of θ^t to update the $\theta^{(t+1)}$ in the prediction distribution of the bristles. Through (5)-(9), the estimation $\theta^{(t+1)}$ of the parameter θ of the t+l step iteration is obtained, and the algorithm is repeated until convergence. After convergence, we will converge θ as the estimate of parameter θ. As for the convergence of EM algorithm, there is a lot of literature on it.

4. Analysis and Discussion

In this study, the revenue generated by the customer, because of the application of SAS to the relevant data of the product cost, is only considered in this case without considering its cost. The model is applied to the test set, and its prediction accuracy is as shown in the figure 1.

The error statistics of CVM model are shown in the following table:

Average Squared Error	0.2267529213	0.2286816926
Average Error Function	0.6433798709	0.6433780419
Degrees of Freedom for Error	978	
Model Degrees of Freedom	5	
Total Degrees of Freedom	983	
Divisor for ASE	1966	982
Error Function	1264.8848262	631.79723714
Maximum Absolute Error	0.9842061437	0.8999839094
Mean Square Error	0.2279121898	0.2286816926
Sum of Frequencies	983	491
Number of Estimate Weights	5	

Table 3. CVM Model Error Statistics

As can be seen from the above table, the error classification rate in the training set and the test set is 14.2% and 15.5% respectively, which is 85.8% and 84.5% respectively. This indicates that the accuracy of the CVM model in this study is ok. From this, we can find the following patterns: customer satisfaction is positively correlated with loyalty, and the higher the degree of satisfaction, the higher the loyalty. The customer is the easiest to lose in the first year of the network, and after this easy exit, the latter years are relatively stable. The possibility of customer churn in some highly valued customer segments is particularly high. Easily lost customers are often sensitive to product prices. The mining results in SAS EM are as follows:

Then, as shown in figure 3, the clustering analysis is applied when mining the initial subset. The observation points are divided into several classes as objects to excavate the forward mining algorithm, so that the results of comprehensive search can be achieved by using very few searches. If the observation value contained in one class is less than or equal to the initial subset number p, all the selected points are selected from the random points of each class. If the observed value in a class is greater than the initial subset number p, then a point can be selected randomly from each class. After this selection, the number of points is s, satisfying: $s < m\ p < n$ (because in most cases, s is much less than n, therefore, it is possible to combine clustering with forward search to simplify the initial subset of search).

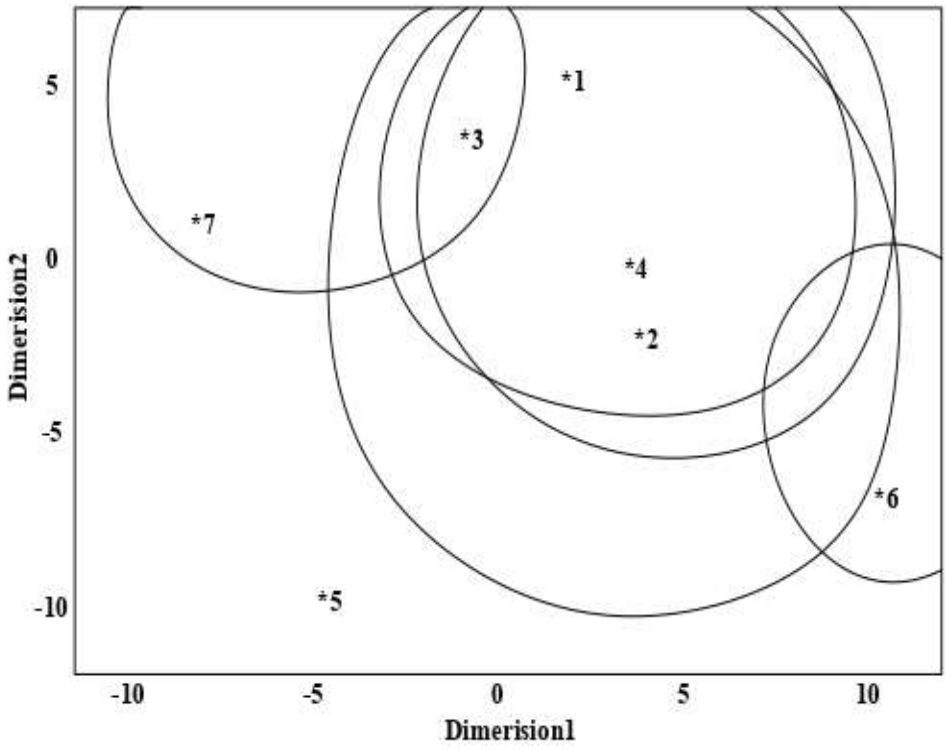

Figure 2. Clustering results of customer subdivision model

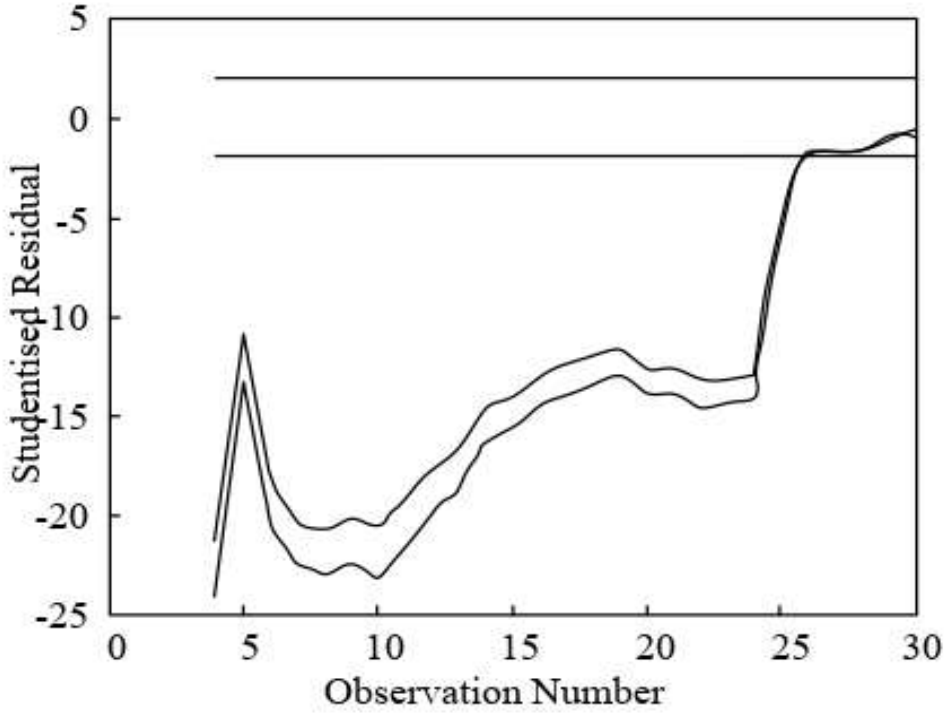

Figure 3. Observation studentised residual

According to consumer preference, we can roughly divide the customer into these types: The first is pure consumption, generally for the fashion youth, they are not sensitive to the fees, they care about the quality of the call and the service, and they are very enthusiastic about the new business. The second is the business group, which has a large number of users, but is sensitive to the unit price, and the new business demand is more practical, and the demand for roaming calls is higher. Third, the government's civil servants, such as party and government, public inspection law, news and other important department users, such users are sensitive to the cost sensitivity, new business needs are not high, and the service quality requirements are higher. Fourth, the average consumer, they are sensitive to the charge, the call quality, service and new business requirements are not high, the monthly call fee is very low, generally for ordinary pay. The first three are big customers, the fourth is small customers. Since there is no data on fraud identification, we only use the collected data to do the data mining of customer's arrears. Find out which customers are liable to owe, and what are the characteristics of these customers, and how will they behave? From the above excavation results, it can be seen that the mobile phone bill is mainly related to who pays the arrears. The errors in the training set and test set were 9.0% and 22.5%, respectively, with 81.0% and 77.5% respectively. This explains the accuracy of the customer segmentation model in this study.

5.Conclusion

Based on the research of data mining theory, this paper draws on some foreign research experience and conclusions. The application of data mining technology in mobile communication operation in China is systematically studied. Firstly, the modeling of mobile communication data warehouse and decision support system based on data mining is introduced. Several data mining topics are urgently needed in China mobile communication operation. Secondly, five data mining models, such as customer value, customer retention and customer segmentation, are established. Again using the data collected by the questionnaire. Based on the verification and evaluation of the model in SAS Enterpirsei Mner, the research results show that the accuracy of the customer segmentation model in this study is good. Due to the lack of relevant research materials and the commercial secrets of many mobile operating enterprises, the research in some aspects is not satisfactory. For example, due to the lack of data in the enterprise, we can only collect research data through questionnaire survey. This results in a small amount of data for model training and testing, and the correctness of the model is affected.

References

[1] Shi, Guangren, Zhu, Yixiang, Mi, Shiyun, Ma, Jinshan, and Wan, Jun. (2017). A Big Data Mining in Petroleum Exploration and Development. *Advances in Petroleum Exploration and Development, 7*(2), 125.

[2] Liu, Guomin, Knight, James D. R., Zhang, Jian Ping, Tsou, Chih-Chiang, Wang, Jian, Lambert, Jean-Philippe, Nesvizhskii, Alexey I. (2016). Data Independent Acquisition analysis in ProHits 4.0. *Journal of Proteomics, 149*(1), 30.

[3] Tug, Emine, Sakiroglu., Merve, and Arslan, Ahmet. (2015). Automatic discovery of the sequential accesses from web log data files via a genetic algorithm. *Knowledge-Based Systems, 19*(3), 45-47.

[4] Zhang, Wei, Kiyonami, Reiko, Jiang, Zheng, and Chen, Wei. (2016). Quantitative Analysis of Targeted Proteins in Complex Sample Using Novel Data Independent Acquisition. *Chinese Journal of Analytical Chemistry, 42*(12), 325.

[5] Rushing, John, Ramachandran, Rahul, Nair, Udaysankar, Graves, Sara, Welch, Ron, and Lin, Hong. (2017). ADaM: a data mining toolkit for scientists and engineers. *Computers and Geosciences, 31*(5), 168.

[6] Rathnamala, K. S., and Wahida Banu, R. S. D. (2016). Analysis of Data Mining Visualization Techniques Using ICA and SOM Concepts. *International Journal of Computer Science and Information Security, 9*(1), 1196.

[7] McQueen, Peter, Spicer, Vic, Schellenberg, John, Krokhin, Oleg, Sparling, Richard, Levin, David, & Wilkins, John A. (2015). Whole cell, label free protein quantitation with data independent acquisition: Quantitation at the MS2 level. *Proteomics, 15*(1), 56.

[8] Bilbao, Aivett, Varesio, Emmanuel, Luban, Jeremy, Strambio De Castillia, Caterina, Hopfgartner, Gérard, Müller, Markus, and Lisacek, Frédérique. (2015). Processing strategies and software solutions for data independent acquisition in mass spectrometry. *Proteomics, 15*(5-6), 454-457.

etry. *Proteomics*, 15(5-6), 454-457.

[9] Jörg Kuharev., Pedro Navarro., Ute Distler., Olaf Jahn., Stefan Tenzer. (2015). In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics*, 15(18), 911-912.

[10] Chih-Chiang Tsou., Chia-Feng Tsai., Guo Ci Teo., Yu-Ju Chen., Alexey I. Nesvizhskii. Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics*, 16(15-16), 303