International Journal of Advanced Culture Technology Vol.1 No.2 13-17 (2013)
IJACT.2013.1.2.13

IJACT 13-2-3

Customer Behavior Data Model using User Profile Analysis

Yong Gyu Jung[1], Agatha Lee[2], Jeong Chan Lee[3], Young Dae Lee[4]

[1]*Department of Medical IT Marketing, Eulji University, Korea*
ygjung@eulji.ac.kr

[2]*Trulogic, Unit 4, 191 Parramatta Rd. Auburn NSW 2144 Australia*
aga246@gmail.com

[3]*National DB Division, National Information Society Agency, Korea*
jclee@nia.or.kr

[4]*International Promotion Agency of Culture Technology*
youngday77@daum.net

Abstract

Today, most of the companies have numerous issues to take advantage of the data within the organization. Modeling techniques could be described using profile and historical log data as a tool of data mining techniques. It is covered increasingly with data entry, research, processing, modeling and reporting components of the icon in the form of easy-to-use in many datamining tools. Visual data mining process can create a data stream. In this paper, customer behavior is predicted in pages or products, using the history profile analysis and the navigation items are necessary to predict unknown features.

Keywords: *Bayesian Network, K2, TAN, Apriori algorithm*

1. INTRODUCTION

Recently, data mining techniques are utilized in the decision-making process for new and meaningful information from large amounts of data by extracting data. As exponential increase in each sector is highlighted, data mining is used in each company's customer management, the bank's personal and business credit score calculation, risk management, health care for the treatment of patients in clinical data analysis, such as DNA sequencing analysis in biotechnology. Mainly Logistic Regression, Neural Networks, Support Vector Machine techniques and variety of other algorithms are used. Customer data, transaction data, contact data, and the data about the process as the other side of the multiple data are hidden valuable knowledge that can be used to improve the company's business. However, they do not include most companies extract value from these data. As a solution to these problems, Clementine provides Predictive Analytics. Data mining is the process for finding hidden patterns and trends in large amounts of raw data that is made useful for decision-making information. Data mining is also powerful analytical technique with a tremendous insight on the business issues and to find new business opportunities and can lead. And more informed decisions and solve complex problems is to take advantage of the results obtained in this way. In this paper, by making site

Manuscript Received: Sept. 16, 2013 / Revised: Dec. 1, 2013 / Accepted: Dec. 6, 2013
Corresponding Author*: youngday77@daum.net*_
Tel: +82-2-407-7718, Fax: +82-2-407-7716
International Promotion Agency of Culture Technology

visits profiling and forecasting models to analyze visit data with a wide variety of data sources to obtain data using data mining, we use visiting customers and visit types for behavioral patterns on the Web site to find and navigate products with recommended model to deploy.

2. RELATED RESEARCH

2.1 k-means clustering

In data mining, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. k-means technique, how to choose what possible about the problem in the following way. Choose k to minimize cluster the center of the square of the distance cross-validated the square of the penalize distance of the training data, and use. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes. Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$) S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where μi is the mean of points in Si.

2.2 k-means clustering

The minimum description length (MDL) principle is a formalization of Occam's Razor in which the best hypothesis for a given set of data is the one that leads to the best compression of the data. MDL was introduced by Jorma Rissanen in 1978.[1] It is an important concept in information theory and learning theory. It is used MDL(Minimum Description Length) criteria in k-means clustering. Criteria used to recursively apply the k-means (For example, MDL-based) seed for the subcluster along the direction of the biggest variables in the cluster, the appointed seed selection is possible. However, the standard deviation of the parent cluster is estranged from each direction from the center of the cluster. Such information is referred to as the X-means algorithm is implemented in

To select the hypothesis that captures the most regularity in the data, scientists look for the hypothesis with which the best compression can be achieved. In order to do this, a code is fixed to compress the data, most generally with a (Turing-complete) computer language. A program to output the data is written in that language; thus the program effectively represents the data. The length of the shortest program that outputs the data is called the Kolmogorov complexity of the data. This is the central idea of Ray Solomonoff's idealized theory of inductive inference.

However, this mathematical theory does not provide a practical way of reaching an inference. The most important reasons for this are: Kolmogorov complexity is incomputable: there exists no algorithm that, when input an arbitrary sequence of data, outputs the shortest program that produces the data. Kolmogorov complexity depends on what computer language is used. This is an arbitrary choice, but it does influence the complexity up to some constant additive term. For that reason, constant terms tend to be disregarded in Kolmogorov complexity theory. In practice, however, where often only a small amount of data is available, such constants may have a very large influence on the inference results: good results cannot be guaranteed when one is working with limited data.

MDL attempts to remedy these, by: Restricting the set of allowed codes in such a way that it becomes possible (computable) to find the shortest codelength of the data, relative to the allowed codes, and Choosing a code that is reasonably efficient, whatever the data at hand. This point is somewhat elusive and much research

is still going on in this area.

Rather than "programs", in MDL theory one usually speaks of candidate hypotheses, models or codes. The set of allowed codes is then called the model class. (Some authors refer to the model class as the model.) The code is then selected for which the sum of the description of the code and the description of the data using the code is minimal. One of the important properties of MDL methods is that they provide a natural safeguard against over fitting, because they implement a tradeoff between the complexity of the hypothesis (model class) and the complexity of the data given the hypothesis

3. EXPERIMENTS

The Clementine is a very powerful and useful tool in order to find the appropriate Data Mining techniques. In a particular field, CAT(Clementine Application Template) as a library of application techniques can be applied directly to the user industries, and can be applied in other industries. Web-Mining CAT ensure the quality of Web-logs data, and then Make site visits profiling and forecasting models to a variety of data sources, and Visit data analysis to detect Visit customer and Visit Type, Of behavioral patterns on the Web site to find out then, To deploy products to the recommended model includes.

Modules: the history and modeling (complex)

History, profile analysis

Predict customer behavior, pages or products recommended

Initial behavior prediction

Use navigation items necessary to predict

In this paper, a tool for experimental use as a tool SPSS Clementine, Make site visits profiling and forecasting models to a variety of data sources, and Visit data analysis to detect Visit customer and Visit Type, Of behavioral patterns on the Web site to find out then, To deploy products to the recommended model includes that is web-mining. Based on this experiment, analyzed the history and profile analysis, and predict customer behavior, pages or products, the initial behavior predicted. Clementine Stream, as shown in Figure 1 after writing a variable assignment will try to analyze the data.

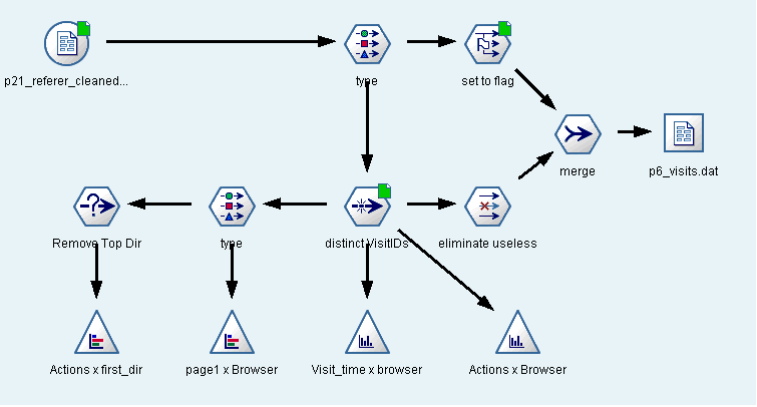

Figure 1 DATA Steam

Stream analysis is shown in the following Figure 2 and Figure 3. They represent the histogram of each visit record and the initial behavioral predictions.

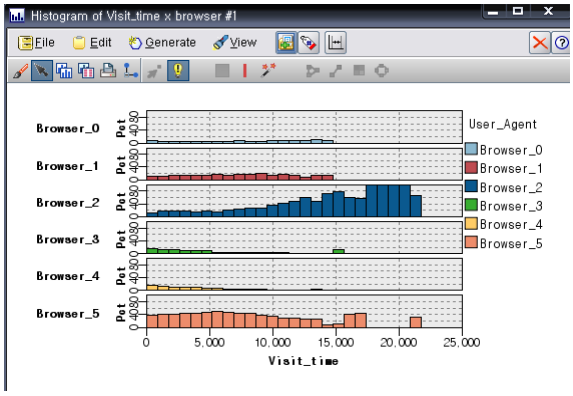

Figure 2 visit records

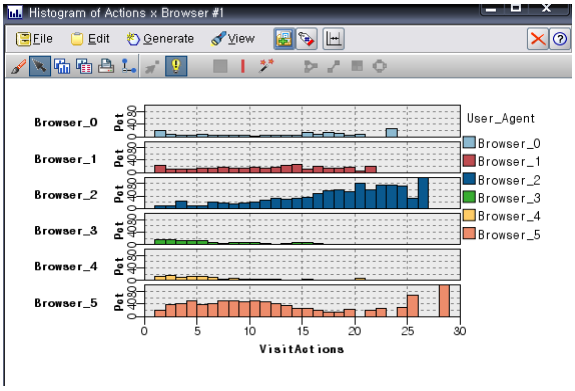

Figure 3. Initial behavioral predictions

It is used for any user with step-by-step using some time to analyze. Profile analysis is used to predict customer behavior through experimental results in Figure 4. The data were extracted and analyzed by classifying the items necessary to predict.

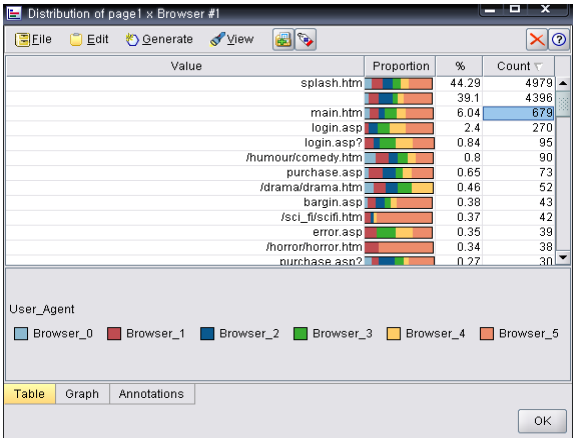

Figure 4. Use navigation items necessary to predict

The experimental results in Figure 4 shows to navigate through the necessary items needed for the prediction. Figure 5 shows to analyze and predict about customer behavior data.

Figure 5. Customer behavior prediction

4. Conclusions

Data mining techniques are used to obtain data, which visits to a variety of data sources to create profiling and forecasting model. In this paper, it could be obtained the behavior patterns by analyzing customer data and detecting the type on the websites. It appears recommendation model to expand to include the contents. The current applications of these techniques are libraries of techniques to apply directly to the industries. It could be possible applications in other industries proposed methods. In future experiments, performance will be evaluated more accurate to build real and similar environmental factors with techniques of Clementine. Web-mining techniques for modeling are analyzed for the user's browsing history.

References

[1] Stuart Moran, Yulan Hey, Kecheng Liu, “An Empirical Framework for Automatically Selecting the Best Bayesian Classifier”, Proceedings of the World Congress on Engineering 2009 Vol I, WCE 2009, July 1 - 3, 2009.
[2] Carolina Ruiz, “Illustration of the K2 Algorithm for Learning Bayes Net Structures”, Department of Computer Science, WPI, 2005.
[3] Evelina Lamma, Fabrizio Riguzzi, Sergio Storari, “Improving the K2 Algorithm Using Association Rule Parameters”, Modern Information Processing: From Theory to Applications B, 2006.
[4] http:// www.spss.co.kr - spss korea (Data solution)
[5] N. H. Kim, “Rotor fault detection system for inverter driven induction motor using current signals and an Encoder,” Journal of IWIT, Vol. 10, No. 5, pp. 128-135, Oct. 2010.
[6] F. A. Huliehel, F. C. Lee, and B. H. Cho, “Small-signal modeling of the single-phase boost high power factor converter with constant frequency control,” in Proc. PESC, pp. 475-482, 2011.
[7] T. J. E. Miller, Reactive Power Control in Electric Systems, John Willey&Sons, 2009
[8] Myway labs -PSIM- http://www.myway-labs.co.jp/psim, May 13th 2011.
[9] S. K. Dwivedi, "Power Quality Improvements and Sensor Reductions in Permanent Magnet Synchronous Drives," PhD. Thesis, IIT Delhi, 2006.
[10] Yong-Gyu Jung, Seung-Ho Lee and Ho Joong Sung, Effective Diagnostic Method Of Breast Cancer Data Using Decision Tree, Journal of IWIT, Vol.10 No. 5 pp.57-62, 2010
[11] I.C Kim, Y.G Jung, Using Baysian Network to analyze Medical Data, LNAI2734, Springer-Verlag, pp.317-327, 2003
[12] Yong Gyu Jung, Ki Young Lee and Myung Jae Lim, Discharge Decision for Post-Operative Patients, Proceedings of ICHIT, pp.195-199, 2010
[13] Shmueli, G., Patel, N. R., and Bruce, P., Data Mining for Business Intelligence, 2009
[14] lan H. Witten and Eibe Frank, Data Mining, Addison