Using Data Mining to Investigate the Behavior of Video Rental Customers

Tai-Chang Hsia[1] and An-Jin Shie[2]
[1]*Department of Industrial Engineering and Management, Chienkuo Technology University*
[2]*Department of Industrial Engineering and Management, Yuan Ze University*
[1]*Changhua, Taiwan, R.O.C.*
[2]*Taoyuan, Taiwan, R.O.C.*

1. Introduction

Living standards have been steadily rising in Taiwan, and indoor recreational activities are now receiving much attention. As a consequence, the market for video rentals has been flourishing in recent years. Renting videotapes, VCDs, and DVDs, to watch at home has become a common consumer entertainment activity. With the arrival of the US video store chain Blockbuster in Taiwan, stagnant personal incomes, and inflation, competition among video rental stores has grown, and their management is becoming more difficult. Like many retail businesses, video rental stores must be guided by the pattern of customer demand. Therefore, the purpose of this study is to investigate how to help rental stores using information in their databases to understand each customer's preferences and demand, so as to increase the rental ratio.

Using data mining theory, this study investigated the customer database of a single store of a local chain of video rental stores in a medium-sized city in central Taiwan, for the period January to March, 2007. First, the records were explored and analyzed in detail by decision tree algorithm. We determined the relationships among customer gender, occupational, favorites leisure activities, and video categories. Second, using the Apriori association rule algorithm for a rougher analysis, we explored and analyzed customers' personal preferences and video categories. We determined favorite videos and personal preferences, and developed rules for predicted which video types will be rented next time. Using these results, video rental stores can recommend personal favorites to each customer and invite customers to rent videos so that rental stores can increase their operating achievements.

2. Literature review

2.1 Consumer behavior

A survey by UDNJOB.Com CO. LTD (2006) showed that 27.8% of people in Taiwan watch TV videos to relax, the most popular activity, followed by travel, physical activities, dining out, shopping, playing on-line games, and enjoying culture. Obviously video entertainment is the most common and most basic form of consumer entertainment.

Scholars have formulated different definitions of consumer behavior. Schiffman and Kanuk (1991) categorized consumer behavior into searching, purchasing, using, estimation, and

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

handling products or services that consumers desire in order to meet their own needs. Engel et al. (1982) said that consumer behavior means the decision-making and actual realization processes the consumer engages in to meet his demand when he seeks, purchases, uses, appraises, and handles the product or the service. Wilkie (1994) observed that consumer behavior refers to the activities produced in the mind, the emotion and the body of consumers to meet their demand and desire while they choose, purchase, use, and handle products and services.

Based on the foregoing, we know that consumer behavior is the consumption processes and transaction activities that consumers carry out to satisfying their desires. Understanding the factors that shape consumer behavior can help video rental stores to design appropriate marketing strategies. Bazerman (2001) argued that the focus of research into consumer behavior should be changed from sales and marketing personnel to the customer himself. Above all, Bazerman says that if you can follow the consumer, you can follow the market. Given these understandings, this study uses the records of customer gender, occupational categories, personal preferences, and video rental categories in a video rental database to discover customer preferences and behaviors, for use in designing marketing strategies for video rentals.

2.2 Data mining

Scholars use a variety of different definitions for data mining. Frawley et al. (1991) declared that data mining is actually a process of discovering of nonobvious, unprecedented, and potentially useful information. Curt (1995) defined data mining as a database transformation process, in which the information is transformed from unorganized vocabulary and numbers to organized data, and later turned into knowledge from which a decision can be made. Fayyad et al. (1996) stated that data mining is an uncomplicated process of discovering valid, brand new, potentially useful, and comprehensive patterns from data. Hui and Jha (2000) defined data mining as an analysis of automation and semi-automation for the discovery of meaningful relationships and rules from a large amount of data in a database. They further categorized the data mining process into seven steps: 1. Establishing the mining goals; 2. Selection of data; 3. Data pre-processing; 4. Data transformation; 5. Data warehousing; 6. Data mining; 7. Evaluating the data mining results. Peacock (1998) declared that data mining can be categorized as Narrow and Broad. The narrow definition is limited by the methodology of mechanical learning, which emphasizes the discovery process and uses artificial intelligence such as Neural Networks, Correlation Rule, Decision Tree Algorithms and Genetic Algorithms. By contrast, the broad definition emphasizes the knowledge discovery in database (KDD), the process of obtaining, transforming, clarifying, analyzing, confirming and enduring the meaning of the data within existing customers or outside of the cooperation, which then results in a backup system of decision making that is continuously modified and maintained. Hand et al. (2000) stated that data mining is a process that discovers interesting and valuable information from a database. Berson et al. (2001) argued that the appeal of data mining lies in its forecasting competence instead of merely in its ability to trace back.

To summarize the foregoing definitions, data mining is a process of obtaining knowledge. The key to the process is comprehension of the research application, and then construction of a data set by collecting data relevant to the research field, purifying the data in the targeted database to eliminate erroneous data, supplementing missing data, simplifying and transforming the data, and then discovering the patterns among the data and presenting them as useful knowledge.

The range of data mining is extensive as artificial intelligence systems have made remarkable progress. A diversity of data mining algorithms have been developed for application to different types of data. When data mining needs to be performed, data features, research goals and predicted results should be considered first so that the most useful algorithms may be applied to the data.

Decision tree algorithm uses the technique of Information Gain and the theory of classification standards to automatically discover the correlation of targeted and forecasted variables. Based on a preset significance standard, the data is classified and clustered automatically. The Apriori association rule algorithm can find correlations among the attributes of different kinds of variables.

Because of their usefulness to this research, the above-mentioned two algorithms were used to classify the video rental database in order to find the correlation between consumer favorite activities and video categories. When the video rental stores have new videos, they can recommend videos to customers based on consumer preference and the correlation.

3. Methodology

This study used information contained in a database of 778 customers and their video rental data from a single store of a local chain of video rental stores in Changhua City, Taiwan, collected from January to March, 2007. Using the Chi-squared Automatic Interaction Detection (CHAID) decision tree data mining algorithm and the Apriori association rule, the behavior of the consumers was explored and analyzed. Prior to data mining, based on the purpose of the study, four variables were defined, i.e., customer gender, professional category, personal preference category, and video category. The four variables and their attributes are as follows:

1. Customer gender category

In the study the attributes of customer gender variable are male and the female. Each customer is defined as one or the other.

2. Occupational category

There are 6 items in the attributes of this category: students, manufacturing, engineering and commerce, government officials and teachers, freelancers, and finance and insurance. Each customer is assigned to one category.

3. Favorite leisure activities category

Based on the survey of UDNJOB.Com CO. LTD (2006), excluding watching videos, the favorite recreational activities of people in Taiwan are: travel, physical activities, dining out, shopping, on-line games, and enjoying culture. Each customer may be assigned to more than one choice.

4. Video category

Based on the classifications of the video rental store, there are nine attributes: comedy, action, horror, science fiction, crime, soap operas, romance, animation and adventure. Each customer may be assigned to more than one choice.

3.1 Decision tree algorithm

Prior to the application of CHAID, targeted and forecasted variables should be defined. CHAID can automatically find the correlation between forecasted and targeted variable. Therefore, the study defined the three categories of customer occupation, favorite leisure activities and video rentals, and used their characteristic data as the forecasted variables. Customer gender and its characteristic data were used as the targeted variables.

The calculation process of CHAID in decision-making tree was developed using the classification of the attributes of male and female. The, in light of the gender attributes, the attributes of customer occupational category were classified. Similarly, the attributes of male and female, favorite leisure activity, and video were next clustered. The classification of forecasted variables continued until the patterns in the data were discovered. This data mining approach creates a detailed analysis of the data.

One of the advantages of Decision Tree Algorithm is that the data can be automatically split and clustered based on the preset significance standard, and then be built up into a tree structure from the clustering event. Based on the tree structure, certain rules can be obtained, and the correlation between events found for further forecasting.

Each internal node in the tree structure is tested by a preset significance level. Branches are processed by the values of groups or multiple values of groups. This means that the branch of each internal node may have another branch that may at the same time be the internal node of another branch. They are tested in the order of the significance level until the branch cannot be split and comes to an end. The terminal node of the branch is called the leaf node. The path of each leaf node explains the reasons for and the results of each event. The study used the CHAID of the decision tree algorithm to carry out data mining and to build the tree correlation figure of the targeted and forecasted variables, as depicted in Figure 1.

Based on the tree correlation structure of Figure 1, we find there are differences in the video choices among the customers from different occupations and favorite leisure activities. Customer occupation may be divided into three categories: students (62.2%)(Node 1); engineering and commerce, finance and insurance, freelancers, government officials and teachers, (23.7%)(Node 8); and, manufacturing (14.1%)(Node 15).

First we carried out an analysis of students (Node 1). We found two kinds of student customers: those who like playing on-line games (44.5%) and those who dislike playing on-line games (17.7%), on two nodes. The former were called Node 3, the latter, Node 2. Because the p-value preset significance level is less than 0.05, Node 2 and Node 3 may each be divided into two nodes. Node 3 can be divided into Node 6 and Node 7, representing the proportion of student customers who like playing on-line games and choose action movies (31.6%) of which 95.1% are male and 4.9% female. Similarly, 12.9% of those students who like playing on-line games and choose non-action movies. In that group, 78% are male, while 22% are female. Nodes 6 and Node 7 have a p-value greater than the preset significance standard of 0.05 and thus cannot be further subdivided. On Nodes 4 and 5, subdivided from Node 2, no choice of any movie appeared. Hence, these nodes will not be considered in the analysis.

Next, we analyzed customers of the second sort of professional category (Node 8) working in engineering and commerce, finance and insurance, as freelancers, and as government officials and teachers. We found they choose crime movies (6.4%, Node 10) and non-crime movies (17.9%, Node 9), two nodes in total. With the p-value preset significance level of less than 0.05, Nodes 9 and 10 may each be further subdivided into two nodes. Node 10 can be subdivided into Nodes 13 and 14, meaning that 2.1% of the group of professional occupations choose both crime movies and romance movies. In this group 50% are male and 50% female. Non-romance movies are chosen by 4.4% of this occupational grouping, with males constituting 82.4%. At the preset significance level of less than 0.05, Nodes 13 and 14 cannot be further subdivided. Nodes 11 and 12, subdivided from Node 9, contain no individuals choosing movies and will not be considered in this study.

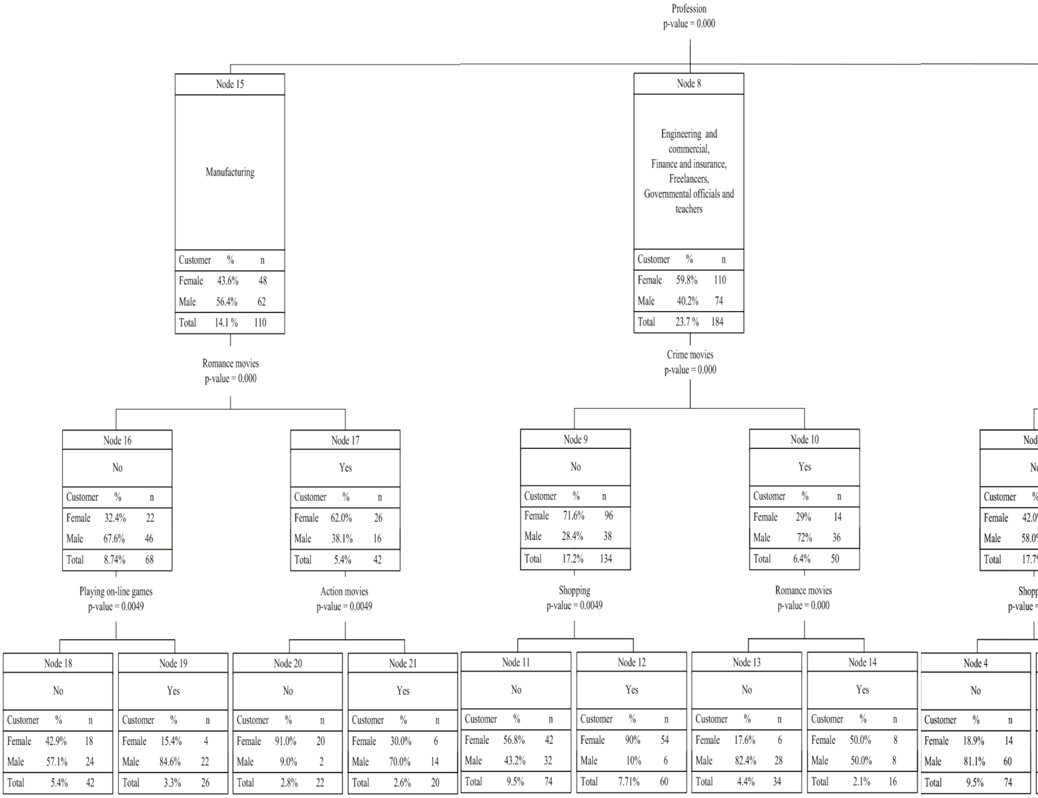


Fig. 1. Structure tree relation between video rentals and customer gender, profession, and favori

Finally we analyzed customers (Node 15) of the third occupational category working in the manufacturing industry. We found the proportion of customers working in the manufacturing industry and choosing romance movies is 5.4% (Node 17), while Node 16, manufacturing industry and non-romance movies, comprises 8.7%. At the preset significance level of less than 0.05, Nodes 16 and 17 may each be further subdivided. Node 17 can be subdivided into Nodes 20 and 21, representing manufacturing industry customers who choose romance and action movies (2.6%, 70% male). Among such customers choosing non-action movies (2.8%) 9% are male. Nodes 20 and 21 have a p-value greater than the preset significance standard of 0.05 and thus cannot be further divided. Nodes 18 and 19, subdivided from Node 16, do not contain anyone choosing movies and will not be considered.

3.2 Apriori algorithm

The Apriori algorithm was first published by Agrawal and Srikant in 1994. During the application of the Apriori association rule algorithm, no defined targeted variables or forecasted variables need be defined, but the attributes of two or more than two variables which meet the following two conditions are necessary. The first condition is that the attributes of variables must correspond with Boolean type ("1"/"0"). The second condition is that the attributes of the variables must be multiple-choice. With those two conditions fulfilled, the Apriori algorithm can be used to find the common correlation among the attributes.

Among the four defined variables of the study, both favorite leisure activities and video category met the above-mentioned conditions. The Apriori algorithm could thus be applied to find correlations among the attributes. These might include discovering what leisure activities are correlated with what video preferences, or what videos will be rented by individuals with a given personal favorite when they rent the next time. Aprior algorithm can find a portion of the correlations among the variables that need to be explored and does not need to discover the entire set of correlations among all defined variables. This data mining method provides a rougher analysis of the data than the Decision Tree approach.

The Apriori association rule algorithm explores the attributes of various variables in which 0 and 1 coexist. This coexistence relationship causes some level of correlation. Rules which rely on one another can be discovered and the proper conclusion stated at last. For instance, imagine that customers bought hamburgers, fried chicken, and potato chips at McDonald's. It could then be deduced that such consumers also tend to buy Coca Cola. Similarly, when individuals have high cholesterol, high triglycerides, and high uric acid, Apriori would deduce that they would also have high blood pressure. Of course, the strength of such correlations may vary. The strength of a given correction is called confidence and is shown by a percentage (%). The higher the percentage is, the higher the degree of correlation between Antecedent and Consequent.

In this study Apriori algorithm was used to analyze favorite leisure activities and video category. After application of the Apriori algorithm, 13 rules were produced with confidence ranging from 85.4% to 90.2%. Five examples are given below.

Rule 1: With a preference for crime movies, customers whose favorite activity is physical activities and like renting science fiction videos tend to rent action movies next. The confidence is 90.2%.

Rule 2: With a preference for crime movies, those customers whose favorites are physical activities and playing on-line games tend to rent action movies next (confidence = 87.8%).

Rule 3: With a preference for crime movies, those customers whose favorite activity is enjoying culture and like renting science fiction movies tend to rent action movies next (confidence = 87.5%).

Rule 4: Customers whose favorite activities are shopping and enjoying culture and who have a preference for renting action movies tend to rent comedy next (confidence = 87.2%).

Rule 5: Customers whose favorite activities are physical activities, travel, and enjoying culture tend to rent comedy next (confidence = 85.4%).

In the above-mentioned rules, the calculation of confidence is performed via the following formula (1).

$$\frac{\text{Support \% of Antecedent and Consequent}}{\text{Support \% of the Antecedent}} = \text{Confidence\%} \tag{1}$$

Using Rule 1 as an example, in formula (1), the Support % of the Antecedent is 10.5%. There are 82 instances in which customers whose favorite activity is sports and who have rented crime movies and science fiction videos (82/778 = 10.5%). Support % of Antecedent and Consequent means there are 74 instances of those customers whose favorite activity is sports and who rented crime films, science fiction movies and action videos (9.47% of 778). Thus, Support % of Antecedent divided by Support % of Antecedent and Consequent gives 90.2% (74/82 = 90.2%).

Rule	Antecedent	Consequent	Instances of Antecedent	Support % of Antecedent	Instances of Antecedent and Consequent	Support % of Antecedent and Consequent	Confidence %
1	crime movies and physical activities and science fiction movies	action movies	82	10.5%	74	9.47%	90.2%
2	crime movies and physical activities and playing on-line games	action movies	82	10.5%	72	9.22%	87.8%
3	crime movies and science fiction movies and enjoying culture	action movies	96	12.3%	84	10.76%	87.5%
4	shopping and enjoying culture and action movies	comedy movies	78	10.0%	68	8.72%	87.2%
5	physical activities and travel and enjoying culture	comedy movies	82	10.5%	70	8.97%	85.4%

Table 1. Rules Produced by Apriori Association Rule Algorithm

4. Result analysis

This study applied the decision tree algorithm to construct a detailed exploration and analysis of the video rental database of a video rental store in central Taiwan. Three interesting results were found. First, the profession of customers who most often rented videos was students. Those customers whose favorite activity was playing on-line games and who liked action movies the best were generally male. Second, the number two customer group renting videos consisted of four professions: engineering and commerce, finance and insurance, freelancers, and government officials and teachers. Most of these customers preferred to rent crime movies and most of them are male. Romance movies are their second choice. The rate of male customers renting romance videos is the same as that of females. Third, the least common occupation for video customers is the manufacturing industry. Most of these customers prefer to rent romance movies and most of them are female. In this group, action movies are the second most popular preference. Most of the customers who rent action videos are male.

Further, after performing a rough exploration and analysis of the video rental database using the Apriori association rule algorithm, two possible marketing strategies are suggested. First, no matter what the customer's favorite activity is, clerks should recommend action movies to the customers who come to rent videos. Second, clerks should recommend comedy videos to customers who rented action movies before or whose favorite leisure activities are not dining or playing on-line games. The study suggests that the customers will rent more videos with only a simple reminder.

5. Conclusion

Performing data mining using decision tree algorithm and Apriori association rule algorithm, we suggested a marketing approach for a single store of a local chain of video rental stores. The store should recommend videos to customers based on their professions and leisure activities, increasing store performance. Using information techniques and data mining of customer data and rental records, video rental marketing strategies may be designed. By expanding the size and time period of the data on customer rentals, the precision of the research may be improved. This method also shows promise for application to similar retail situations.

6. References

Agrawal, R. & Srikant, R. (1994). Fast algorithm for mining association rules, In Proceedings of the 20th International Conference on Very Large Data Bases, pp.487-499, Santiago, Chile

Berry, M. J. A. & Linoff G. (1997). Data mining technique for marketing, sale, and customer support, Wiley Computer and Sons, Inc.

Bazerman, M. H. (2001). Consumer research for consumers, Journal of Consumer Research, Vol. 27, No.4, pp.499-504.

Berson, A.; Smith, S. & Thearling, K. (2001). Building data mining application for CRM, McGraw-Hill Inc, New York.

Curt, H. (1995). The Devile's in the detail: techniques, tool, and applications for data mining and knowledge discovery-part 1, Intelligent Software Strategies, Vol. 6, No. 9, pp.1-15.

Engel, J. F.; Roger, D. B. and David, T. K. (1982). Consumer behavior, 4th ed., Hwa-Tai Co, Taipei.

Frawley, W. J.; Paitetsky-Shapiro, G. & Matheus, C. J. (1991). Knowledge discovery in database: an overview, Knowledge Discovery in Database, pp.1-27, AAAI/MIT Press, California.

Fayyad, U. M.; Piatesky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: an overview, Advances in knowledge Discovery and Data Mining, pp.1-34, MIT Press, California.

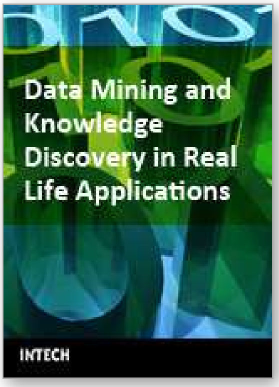
Hand, D. J.; Blunt, G.; Kelly, M. G. & Adams, N. M. (2000). Data mining for fun and profit, Statistical Science, Vol.15, No. 2, pp.111-131.

Peacock, P. R. (1998). Data mining in marketing: part 1, Marketing Management, Vol. 6, No.4, pp.8-18.

Schiffman, L. G., & Kanuk, L. Lazar. (2004) Consumer behavior, Prentice-Hall Inc Englewood Cliffs, New Jersey.

Wilkie William L. (1994). Consumer behavior, 3th ed., John Wiley & Sons, New York.

UDNJOB.Com CO. LTD. (2006). http://udnjob.com/fe/jobseeker/index.shtml.



Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821