Association-rule-based User Segmentation: An Empirical Study

Ming Ren[1, 2] Qiang Wei[3] Wei Xu[1, 4]

[1]Key Lab of Data Engineering & Knowledge Engineering of the Ministry of Education, Renmin University of China, China
[2]School of Information Resource Management, Renmin University of China, China
[3]School of Economics & Management, Tsinghua University, China
[4]School of Information, Renmin University of China, China
Email: renm@ruc.edu.cn

Abstract

Segmentation is becoming crucial than ever for proper exploration of information and delivery of services to the users in a personalized manner. This paper proposes a user segmentation method based on association rules discovered in large databases, and represents the hierarchical segmentation by the Entity-Relationship model, which is easy to understand. To evaluate the proposed model, the understandability of the model is studied from the perspective of a modeler. The experiment result shows that the models were understandable and richer in semantics, and that the level of segmentation hierarchy might affect the degree of understandability.

Keywords: segmentation, association rule, the Entity-Relationship (ER) model, understandability, empirical study

1. Introduction

With the ever-increasing amount of information resources and essential difference of web users' demands, segmentation is becoming crucial for proper exploration of information and delivery of services to the users in a personalized manner. By using characteristics of interest, users can be partitioned into user segmentations, or customer segmentations in commerce field, with similar needs or characteristics. Then the segmentation can be used for the decision makers to determine particular competitive strategies (e.g. differentiation, low cost, or focus strategy) [1].

Typically, the statistics-based approaches compute user segmentation based on demographic variables and lifestyles (e.g., income, address, and education) or transaction data (e.g., recency, frequency, and monetary values). Then users can be partitioned by applying clustering algorithms in the space of the statistics above. The selected variables are crucial to successful segmentation in that irrelevant variables will distort the clustering structure and make the results useless [10]. Further, the inherent assumption that users with similar demographics and lifestyles will exhibit similar behavior is questionable [5]. Nowadays users can easily obtain abundant information resources and pursues personalized products and services even within groups with similar demographics and lifestyles. General variables such as demographics may not work well in segmentation.

It is quite natural to explore the daily transaction records in the database to understand users' behaviors and interests. Recent years witness an increasing interest in using user-item data to predict us-

ers' interests, recommend items, and promote cross-selling [7,8,13].

Association rule (AR) mining, as one of the most popular and well studied data mining methods, discovers hidden associations among data items [2]. For example, the association rule *Ticket_rock=>young* discovered from purchase records tells that tickets for rock are bought by young guys at a certain percentage, which is usually considerable. Generalized association rules reflect associations between items at any level of the taxonomy on items [16]. Constrained association rules enable users to clearly specify what associations to be mined, by defining constraints to be satisfied by the antecedent or the consequent of an association rule [12].

Since association rules directly reflect users' interests and preferences, this paper proposes a user segmentation approach based on association rules discovered in transaction data. For example, based on the above rule, young users who have ever bought tickets for rock could be segmented, in which both demographic data and transactional records are considered. Of course, not all association rules could be used for segmentation, but those with their measures (e.g. support, confidence) satisfying the pre-defined thresholds.

In this paper, the segmentation is represented by the Entity-Relationship (ER) model [4], which is one of the most widely accepted and used tools in conceptual modeling of relational databases. It is natural to use the ER model here since the data used for segmentation is usually stored in databases, and ER diagrams are simple and easy to understand. Typically an important purpose of segmentation is to predict user preferences and behaviors, and so a segmentation method is often evaluated by testing the performance of the prediction. From a different perspective of a modeler, this paper evaluates the ER model representing the segmentation in terms of understandability, since any desired change to a model (e.g. further segmentation), must be preceded by a valid understanding of the model. It is considered important whether a model can be easily comprehensible [15,17]. An experiment conducted with MIS (i.e., management information systems) undergraduates shows that the segmentation models were understandable and richer in semantics, and that the level of segmentation hierarchy might affect the degree of understandability.

2. An AR-based user segmentation

As discussed above, the user segmentation will be represented by the ER model. In our previous study [3], association rules are used to define sub-classes (i.e., specialization) in ER models. Here specialization refers to the attribute-defined specialization, which defines subclasses based on an attribute-and-value tuple <A, *v*>, where A is an attribute and *v* is supposed to be a subset of the domain of A, and instances in a subclass take the values from *v* on attribute A. For example, subclasses of user can be defined based on <age, "<35">, <age, "(35, 55)"> and <age, ">55">, resulting in three subclasses, young, mid and aged. An association rule "*Ticket_rock=>age<35*" will lead to a specialization that young users order tickets for rock, and the AR "*Ticket_drama=>age∈(35,55)*" and "*Ticket_concert=>age>55*" will lead to specializations that mid-aged order tickets for drama and aged order tickets for concert (see Figure 1).

Before an association rule is used for segmentation, it must satisfy the thresholds of measures. Take *Ticket_rock=>age<35* as an example. The measure *support* is satisfied if the transactions young guys buying rock ticket have reached certain scale among all the

transactions. The measure *confidence* is satisfied if more than a certain proportion of the tickets for rock are bought by young guys. These measures will help assure that the pattern is representational to be used for segmentation. Once association rules are satisfactory with respect to the threshold, those most meaningful to decision makers will be selected for segmentation.

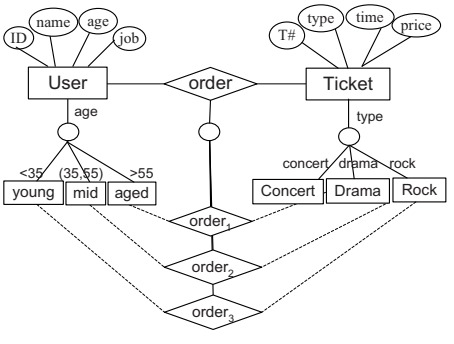

Fig. 1: An example of user segmentation.

Several benefits can be obtained from the user segmentation model. First, decision makers could target a smaller market with greater precision, and thus forges a good relationship with the users. For example, aged people will receive information about concert performances and favorable conditions. Second, it enables easy and quick access to the segmented users and related information. For example, a quick response to online queries like "what rock tickets are bought by young people?" In addition, the result of segmentation can be used as the basis of further analysis of user data, as discussed below.

As an effort to advance the understanding of users, the transactions that young people ordered rock tickets can be analyzed in a similar way to find association rules like:

<job, college_student>⇒<type, punk>

which reflects a representative pattern in the young rock fans group. Accordingly, this piece of knowledge motivates hierarchical segmentation, and a more precise targeting.

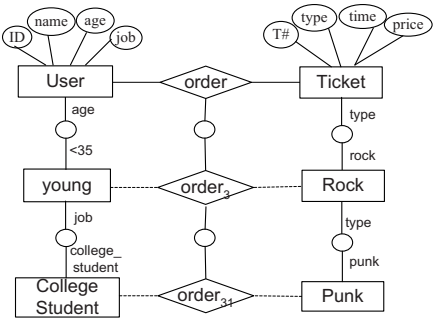

Fig. 2: A hierarchical user segmentation.

During the incremental process of hierarchical segmentation, the hierarchy problem may be of concern due to potential dependencies in knowledge. Inference rules are provided to construct and maintain the hierarchies, which will be studied in a separate work.

3. Understandability of the model

As an important purpose of segmentation is to predict user preferences and behaviors, a segmentation method is often evaluated by testing the performance of the prediction. From a different perspective of a modeler, this paper evaluates the segmentation model in terms of understandability. This measure is selected since it is considered as a key factor of a model, and any change to the model (e.g. further segmentation) must be preceded by a valid understanding of the model. Our purpose is to test whether the understandability of the segmentation model varies with the evolution.

3.1. Model understandability

Model understandability, also called comprehensibility, is defined as the ease with which the data model can be understood [11]. Understandability can only be evaluated with close participation of users and/or developers [11]. It is common to use laboratory experiment investigating model understandability from the human factors perspective [6,14,17]. In these studies, subjects are usually MIS students with knowledge in conceptual modeling, who are given conceptual models (typically ER diagrams), and required to finish comprehension tasks or problem solving tasks.

Understandability is generally measured using performance and/or attitude of the subjects. The performance-based measures include the number of correct answers (or the degree a solution corresponds to a pre-defined one), the time to finish the task, and the number of correct answers divided by the time. The attitude-based measures include preference to use a certain model, perceived value of the modeling formalism, and perceived ease-of-use, etc.

3.2. Experiment

This evaluation study attempts to test if there is any variance in understandability during the hierarchical segmentation. As the conceptual model is concerned, a traditional ER model first evolves into the one with one-layer segmentation, and then into the one with two-layer segmentation, which three are selected respectively as M1: the ER model, M2: the model with one-layer segmentation, and M3: the model with two-layer segmentation. Although the three models are different in richness in semantics expressed, it is regarded interesting to see if the levels of understandability for the three models differ. Adjacent two models will be tested if there is a difference in understandability between them (i.e., M1 and M2, M2 and M3).

In our empirical test, measures of concern include: a). understanding effect (UE): the number of correct answers, and b). perceived ease of understandability (PEU). Hence, we have the following hypotheses:

H1: UE of M2 is the same as M1.
H2: PEU of M2 is the same as M1.
H3: UE of M3 is the same as M2.
H4: PEU of M3 is the same as M2.

A laboratory experiment was designed to test these hypotheses. The questionnaire contains three diagrams of various domains: a diagram in the university domain (M1), a diagram in the supermarket domain (M2), and a diagram in the online ticket-booking service domain (M3). For each diagram there are five TRUE/FALSE statements to check if the information conveyed by the diagram is correctly understood. After that, it will be asked how easy/difficult subjects think to understand this diagram, rated on a 1-7 point scale. Two versions of questionnaires were prepared with diagrams arranged in different orders.

The experiment was conducted during the course of *Database: principle and application* for second-year undergraduates in the information system department at Tsinghua University. Thirty MIS students participated in this experiment, and were randomly divided into two groups. The two versions of questionnaires were assigned to the two groups respectively.

3.3. Results

We computed the number of correct answers per task (i.e. diagram) by each subject, and then applied *t*-statistic to test the significance of differences between the mean grades. Also, the average perceived ease of understandability was computed and compared. The results are shown in Table 1.

As indicated, there is no significant difference in either understanding effect (UE) or perceived ease of understandability (PEU) between M1 and M2. That is, in an early stage of evolution when there is only one layer of segmentation, modelers familiar with ER could understand the segmentation model without difficulty, although it is conveying more information.

When comparing M2 with M3, though both models were all considered understandable with the average UE at over 64% (=4.467/7) and 55% (=3.833/7) respectively, there was a significant difference in both understanding effect and perceived ease of understandability for the two models. That is, in a further segmentation when two-layer segmentation is introduced, modelers might start to be different in understanding the models.

It seems that the difference in understandability, apart from the problem domains, may first of all be due to the fact that model M3 is an extension of M2 and therefore more complex inherently (but for richer semantics). In addition, it may partly result from the structural complexity of the segmentation hierarchies. This also conforms to some other findings, at a similar spirit, that the more attributes and relationships an ER diagram contains, the less understandable it is [6].

Overall, the experiment results revealed that, although segmentation models represent advanced concepts and constructs to reflect focused aspects of the domain in light of richer semantics, which is often inevitable as the business evolves, the level of segmentation hierarchy is a factor and needs to be considered in modeling.

	Model	Mean	Observations	t Stat	P(T≤t) one-tail
H1: UE	M1	4.233	30	-0.909	**0.186**
	M2	4.467	30		
H2: PEU	M1	4.300	30	-0.619	**0.270**
	M2	4.400	30		
H3: UE	M2	4.467	30	2.993	**0.003**
	M3	3.833	30		
H4: PEU	M2	4.400	30	3.319	**0.001**
	M3	3.833	30		

Table 1: Results (α=0.05).

4. Conclusions

This paper has proposed to use association rules for user segmentation and have discussed the hierarchical segmentation with the context of the ER model. The segmentation model has been evaluated in terms of understandability. The experiment has revealed that the models were understandable and richer in semantics, and that the level of specialization hierarchy might affect the degree of understandability.

There are still some open questions to be explored. In particular, there is a potential challenge that such a model tends to be overloaded with many layers of segmentation. A segmentation with more layers could enable more personalized strategy, but will turn to cost more. As a matter of fact, there is always a trade-off between benefit and cost in user segmentation. A balance between is desired,

which requires a set of suitability measures (e.g., importance, interestingness, and stability) to inspect association rules before they are used for segmentation.
Future work includes empirical evaluations from other possible perspectives (e.g., the performance of predicting).

Acknowledgements

The work was partly supported by the National Natural Science Foundation of China (70890083), Key Lab of Data Engineering & Knowledge Engineering of the Ministry of Education (Renmin University of China), and Fundamental Research Funds for the Central Universities and Research Funds of Renmin University of China (10XNB057) and Beijing Natural Science Foundation (9092006).

References

[1] D. A. Aaker, *Strategic market management*. New York: John Wiley and Son, 2001.

[2] R. Agrawal, T. Imielinski, A. Swarmi, "Mining association rules between sets of items in large databases", in *Proceedings of the ACM-SIGMOD 1993 International Conference on Management of Data*, pp. 207-216, 1993.

[3] G.Q. Chen, M. Ren, P. Yan, and X.H. Guo, "Enriching the ER schema based on discovered association rules from large databases", *Information Sciences*, 177(7), 1558-1566, 2007.

[4] P.P. Chen, "The Entity-Relationship model - toward a unified view of data". *ACM Transactions on Database Systems*, 1(1), 9-36, 1976.

[5] R.G. Drozdenko, P.D. Drake, Optimal database marketing: Strategy, development, and data mining. London: Sage, 2002.

[6] M. Genero, G. Poels, & M. Piattini, "Defining and validating metrics for assessing the understandability of entity-relationship diagrams", *Data & Knowledge Engineering,* 64, 534-557, 2008.

[7] Z. Huang, D. Zeng, H. Chen, "Analyzing consumer-product graphs: Empirical findings and applications in recommender systems". *Management Science*, 53(7): 1146-1164, 2007.

[8] T.Y. Jiang, A. Tuzhilin, "Dynamic micro-targeting: fitness-based approach to predicting individual preferences", *Knowlege Information System*, 19:337-360, 2009.

[9] J.F. Li, K.L. Wang, L. Xu, "Chameleon based on clustering feature tree and its application in customer segmentation", *Annual Operation Research*, 168: 225-245, 2009.

[10] H. H. Liu, C. S. Ong, "Variable selection in clustering for marketing segmentation using genetic algorithms", *Expert Systems with Applications*, 34: 502-510, 2008.

[11] D.L. Moody, "Metrics for evaluating the quality of entity relationship models". *Proc. Of nternational conference on Conceptual Modelling* (ER'98), pp. 213-225, 1998.

[12] R. T. Ng, V.S. Lakshmanan, A. Pang, J.W. Han, "Exploratory mining and pruning optimizations of constrained associations rules", *ACM SIGMOD Record*, 27(2), 13-24, 1998.

[13] M. Ren, Q. Wei, F. Li, G.Q. Chen, "Personalized Recommendation Using Association Rule Mining in Neighborhood", *Proc. Of the Asia Pacific Conference on Information Management (APCIM'09)*, Beijing, China, 2009.

[14] P. Shoval, R. Danoch, & M. Balabam, "Hierarchical entity-relationship diagrams: the model, method of creation and experimental evaluation", *Requirements Engineering*, 9, 217-228, 2009.

[15] P. Shoval, & I. Frumermann, "OO and EER conceptual schemas: a comparison of user comprehension". *Journal of Database Management*, 5(4), 28-38, 1994.

[16] R. Srikant, & R. Agrawal, "Mining generalized association rules". In *Procs of the 21st VLDB Conference*, pp. 407~419, 1995.

[17] H. Topi, & V. Ramesh, "Human factors research on data modeling: a review of prior research, an extended framework and future research directions". *Journal of Database Management*, 13(2), 3-19, 2002.

[18] T. Zhou, J. Ren, M. Medo, "Bipartite network projection and personal recommendation". *Phys Rev E*, 76: 046115, 2007.