Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Visualizing Method based on Item Sales Records and its Experimentation

Yoshihiro Hayashi, and Hiroshi Tsuji
Graduate School of Engineering
Osaka Prefecture University
Sakai, Japan
hayashi@mis.cs.osakafu-u.ac.jp, tsuji@cs.osakafu-u.ac.jp

Ryosuke Saga
Department of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Japan
saga@ic.kanagawa-it.ac.jp

***Abstract*— A method for improving a visualized preference transition network by screening nodes in the network, where a node represents a product item, is described. The original preference transition network was developed not only for visualizing customer movements/trends in selecting items but also for finding the features of items. However, understanding such movements/trends and features is difficult when the network has many nodes and links. To solve this problem, the proposed method is a sensitivity analysis for identifying redundant nodes and links with adjustment of the threshold expressed by the Simpson coefficient. The effectiveness of this method was shown through a numerical experiment for 172 kinds of products, 2,227 customers, and their 90,000 sales records.**

***Keywords*—Data Mining, Data Visualization, Preference Analysis, Information Filtering**

I. INTRODUCTION

Data mining [1] is a major topic for knowledge engineering. In particular, there are many works on mining large amounts of sales records for useful knowledge, such as the similarities among customers [2], customer preferences [3], and the preferred products that are selected [4]. Technique for visualizing knowledge including data is also an interesting topic [5]. Techniques include using Key-graph [6] and FACT-Graph [7]. While Key-graph has focused on co-occurrence among topics, FACT-Graph has focused on recency as well as frequency for expressing trends. Such visualizing techniques are possible due to drawing tools such as Graphviz [8].

Previously, for commodity items such as beer and soap, the authors proposed a data mining method based on preference transition [4]. A preference transition indicates customer movements/trends among items. The method represents the transition by a statistical technique and is used for recommending more items that may be preferred. By visualizing a network called a "preference transition network", which consists of product items (called nodes) and transitions (called links), the method also helps analysts understand all the relations among product items. This method is also used for visualizing product competition states [9]. However, sometimes there are too many nodes and links while sometimes there are too few nodes and links for analysis. This makes the visualization method ineffective and is due to the proper threshold for selecting items being unclear.

To alleviate this problem, irrelevant nodes for the visualization-based analysis must be identified, and these nodes and their related links must be filtered out of the preference transition network. However, the criterion for identifying relevance or irrelevance is unclear.

The prospects of sensitivity analysis through changing the threshold, which is expressed as a Simpson coefficient, are described. We aim to identify the proper threshold and effectiveness of our method. The paper is organized as follows. Chapter II gives an overview of the previously proposed visualizing method [4]. Chapter III identifies the problem with the previous method and describes an improved process for visualization. Then, Chapter IV shows the findings from a numerical experimentation and presents speculation from the results. Finally, Chapter V summarizes and concludes this paper.

II. VISUALIZING PREFERENCE TRANSISTION NETWORK

This chapter introduces our previously proposed visualizing method and describes the features of the product items at which our method was aimed [4].

A. Products Characteristics

Because products have many kinds of properties, developing a general method for identifying the relations of any product is not realistic. Therefore, first we describe the targeted class of products and available sales records.

- There are many product instances (items) in a specific class. 'Many kinds' is assumed to be more than about 100.
- Customers buy the items frequently. 'Frequently' means several times a year at least.
- Customer preferences may change from one item to another. However, even after the preference changes, the customers may buy previously preferred items again.

Examples are commodity items such as beer, beverages, soap, and toothbrushes. Customers buy such items on many occasions. Reservation of business hotels is also included. Such sales records should include at least these three attributes: who, which item, and when..

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

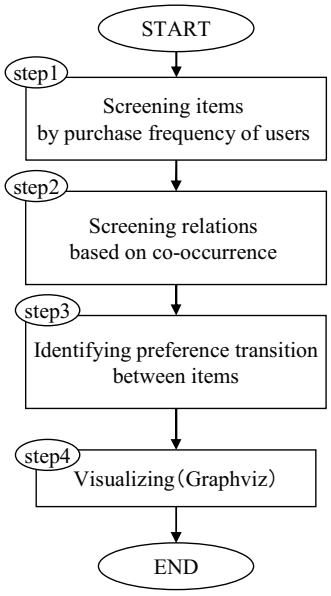

Figure 1. Outline of visualizing method

B. *Visualizing preference transition network*

The authors previously proposed a visualization method for analyzing the above items on the basis of preference transitions [4]. The method can be used not only for recommending items [4] but also for determining competitive status [9].

The method generates a preference transition network by following the four steps shown in Fig. 1: 1) extracting the items that are recorded to a certain number of sales logs, 2) screening relations on the basis of co-occurrence strength, 3) identifying preference transitions between items, and 4) visualizing. The network is generated from sales records as a directed graph. In this network, a node shows a product item (instance), a directed link from node *A* to node *B* shows customer preference changing from item *A* to item *B*. Note that a bidirectional link between items *A* and *B* shows that the items are equally preferred by customers.

1) *Expression of the relations between items*

First, the items are screened according to the purchase frequency of the customers. Next, the relations between these items are generated from sales records. The network generated from sales records is called an *item network*. This process to create the network consists of three sub-steps as shown in Fig. 2.

a) Creating binary graph between customers and items: From the filtered sales records, the relations between customers and items are generated as a binary graph.

b) Connecting item nodes: Two items that are bought by the same customer are linked. For example, in Fig. 2, if user u_1 bought both items i_1 and i_2, then item i_1 and item i_2 are regarded as relatable or alike, and the system connects the two items with a link.

2) *Screening relations based on co- occurrence*

To limit the number of relations to analyze, the relations of the item network are filtered based on co-occurrence strength. By setting a threshold, we regard the links with a lesser degree of co-occurrence than the threshold as nonqualified and remove those links. For example, in Fig. 3, if the threshold is 0.3, then the link between i_2 and i_6 is removed because the co-occurrence of the link is 0.1 (<0.3). This method uses the Simpson coefficient as the co-occurrence strength, defined as formula (1):

$$Simpson(X,Y)=\frac{count(X\cap Y)}{\min(count(X),count(Y))} \quad (1)$$

Here, *Simpson*(*X*, *Y*) indicates the strength of the co-occurrence between items *X* and *Y*, count(*X*) and count(*Y*) are the number of customers who bought items *X* and *Y*, respectively, and count (*X*∩*Y*) is the number of customers who bought both items *X* and *Y*.

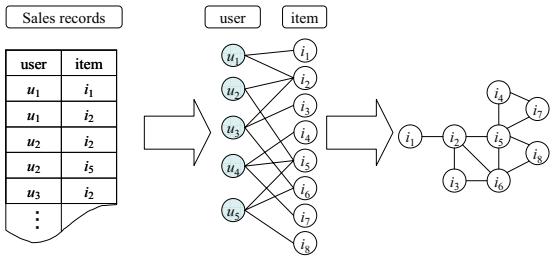

Figure 2. Expression of the relations between items

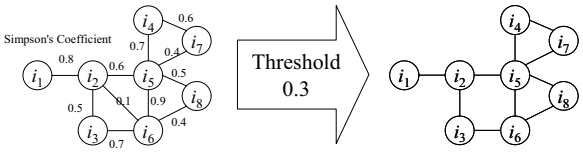

Figure 3. Filtering relations based on co-occurrence

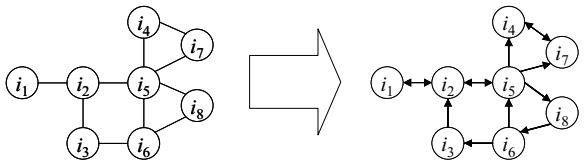

Figure 4. Identification of preference transition

TABLE I. PATTERN OF TRANSITION

Pattern of transition		
mean	no difference between *i* and *j*	*i* is better than *j*

TABLE II. DECISION OF TRANSITION EXAMPLE

Time of sales	1/2004	12/2003	10/2003	5/2003	4/2003	1/2003	10/2002	7/2002
Rank	1	2	3	4	5	6	7	8
Item	A	A	B	A	B	B	B	B

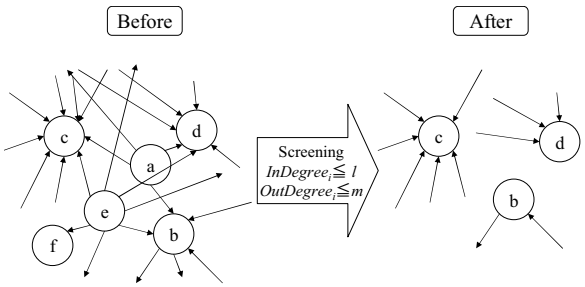

Figure 5. Screening nodes and links by $InDegree_i$ and $OutDegree_i$

3) Identification of preference transition

Customer preference seems apparent for an item recently purchased. However, the current item chosen might have been an aberration. Therefore, the preference transition between two items is identified statistically from sales records.

To identify the preference transition, we used the Mann-Whitney *U* test [10]. This is a rank test for assessing whether or not two populations are different. If the null hypothesis of the test is rejected, the populations differ from each other, and if not, the populations seem to not be different. In our current work, we used recent *k* records from sales records of items *i* and *j* to conduct the Mann-Whitney *U* test.

If the null hypothesis of the test for items *A* and *B* is not rejected, we consider that the preferences for the two items do not differ. That is, the preference for item *A* may be equal to that for item *B*. Thus, as a matter of convenience, the result is shown in the item network as a bidirectional link. However, if the null hypothesis of the test for items *A* and *B* is rejected, we consider that the preferences seem to be different, i.e., one item is more preferable than the other. The result is shown in the item network as a directional link that leads from the node with the higher average rank to the node with the lower average rank (Table I).

For example, let us decide a direction among product items. Here, let us assume that *k* (purchase count considered) is 8. Table II shows the sales records for two items (*A* and *B*) ranked in order of the most recent purchase. In this case, the statistics value of the Mann-Whitney *U* test is 0.036. If the level of significance is 5%, the null hypothesis is rejected and the system concludes that the preferences for items *A* and *B* are different. Here, the average rank of *A* is (1+2+4)/3=2.3 and that of *B* is (3+5+6+7+8)/5=5.8. Consequently, the system judges that the user's preference seems to move from *B* to *A*, so it draws a directional link from item *B* to item *A*. Note that this can show customer movements/trends among items.

Let us show other examples. If the sales records include "*AABBBBAB*", the null hypothesis is not rejected. Thus, the preferences for *A* and *B* are not different, and a bidirectional link is drawn between items *A* and *B*. If the sales records are "*ABABABAB*", the null hypothesis is also not rejected, and a bidirectional link is also drawn.

In this way, preference transition is added to the network in Fig. 3 as shown in Fig. 4, and a preference transition network is created.

III. SCREENING NODE BASED ON LINKS

The previously proposed preference transition network was used for a recommender system [4]. It was also used for analyzing and understanding an overview of the competitive relationships among all the items in a network [9]. However, the visualized network sometimes had a problem in our experimentation because of the product characteristics shown in Chapter II A. It was difficult to find notable items in the case of a preference transition network having many links and nodes.

The reason is derived from the thresholds used for filtering nodes and links in steps 1 and 2 (Fig. 1). If the thresholds are low for observing the relationships among items, then the numbers of nodes and links in the network increase. Having many nodes and links means having large amounts of information, but excess information prevents us from pinpointing highly important information. To alleviate the problem, a proper threshold must be configured and nodes and links in the network must be screened.

To filter out irrelevant nodes for analysis in a preference transition network, $InDegree_i$ and $OutDegree_i$ are defined. $InDegree_i$ is the number of links to item *i*, and $OutDegree_i$ is the number of links from item *i*. $InDegree_i$ indicates the degree of strength of item *i* and shows how many preferences transfer to item *i*. $OutDegree_i$ indicates the degree of weakness of item *i* and shows how many preferences are transferred from item *i*.

Using $InDegree_i$ and $OutDegree_i$, we can identify whether or not a node is necessary for our analysis and can then delete irrelevant nodes. Here, let us define a node as irrelevant when $InDegree_i$ is less than threshold l and $OutDegree_i$ is less than threshold m (Fig. 5). Such nodes do not have much information. In fact, when a node does not have many

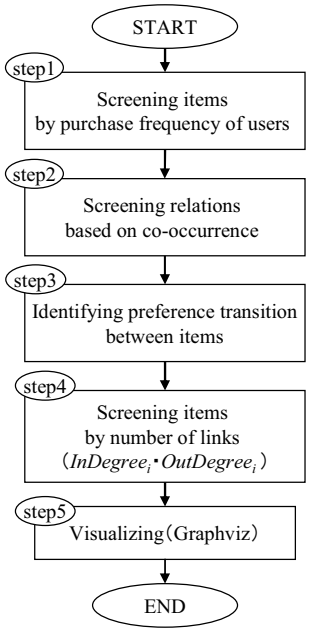

Figure 6. Outline of proposed visualizing method

preference transitions, the node has little information and little value for the analysis. Through deletion of nodes and related links in phases, the notable nodes remain in the network and analysts can discover customer movements/trends.

In the current [7] and the previously proposed methods [4][9], the co-occurrence coefficient is mainly used in screening. However, to screen a directed graph in the current method, $InDegree_i$ and $OutDegree_i$, i.e., values based on transitions, are newly used. Screening according to the concentrated degree of a link becomes possible.

This process is embedded between steps 3 and 4 in Fig. 1 and is shown in Fig. 6. However, the proper values of $InDegree_i$ and $OutDegree_i$ are not clear. Thus, we tried to find the proper values through experimentation.

IV. EXPERIMENTATION

This chapter describes calculated outcomes of sensitivity analysis using a threshold, expressed as a Simpson coefficient, for data on business hotels in Tokyo.

A. Experimentation Environment

We conducted an experiment for forecasting the proper $InDegree_i$ and $OutDegree_i$. We used sales records for 2,227 customers who have stayed in hotels in Tokyo more than 20 times for this experiment. The Simpson coefficient was assumed to range from 0.1 to 1 in intervals of 0.1. In this experiment, nodes are deleted under two cases: (1) $InDegree_i$ =0 and $OutDegree_i$ =1 (i.e., threshold l =0 and m =1) and (2) $InDegree_i$ is 0 (i.e., threshold l =0). For each Simpson coefficient and each case, we deleted the nodes in phases and measured the number of deleted nodes and the number of deletion candidates that were accompanied by deleted nodes.

B. Result

The results of the experiment for condition (1) with the numbers of deleting phases, deleted nodes, and deletion candidates for each Simpson coefficient are shown in Table III. We can see that there are few deleted nodes. The result when there are three deleting phases is shown in Table IV. The number of deleted nodes is more than that in Table III in the case of condition (1).

An example of a preference transition network before screening the hotels by the number of links (with Simpson coefficient of 0.4) is shown in Fig. 7; a partial close-up of Fig. 7 is in Fig. 8. A part of a preference transition network after screening the hotels by the number of links three times (with Simpson coefficient of 0.4) is shown in Fig. 9.

C. Speculation

In the case of condition (1), only the nodes (hotels) at the edges of the graph are deleted because the deleting phase only occurs once (Table III). In the case of condition (2), screening is done more than in condition (1) because all the nodes whose preference transitions only move to other nodes are deleted (Table IV).

In Fig. 8, there are many nodes and links, and it is difficult to see on which nodes the preference transitions concentrate.

TABLE III. SCREENING ITEM BY NUMBER OF LINKS (1)

Simpson coefficient	count	a number of deleted nodes	a number of nodes deleted together	rest	Total
0.1	1	0	0	262	262
0.2	1	1	0	261	262
0.3	1	11	0	238	249
0.4	1	22	0	185	207
0.5	1	16	0	156	172
0.6	1	9	2	96	107
0.7	1	6	1	77	84
0.8	1	4	0	77	81
0.9	1	4	0	77	81
1	1	4	0	77	81

TABLE IV. SCREENING ITEM BY NUMBER OF LINKS (2)

Simpson coefficient	count	a number of deleted nodes	a number of nodes deleted together	rest	Total
0.1	1	8	0	254	262
0.2	1	17	0	245	262
	2	3	0	242	245
	3	1	0	241	242
0.3	1	38	9	202	249
	2	17	1	184	202
	3	1	0	183	184
0.4	1	44	34	129	207
	2	15	7	107	129
	3	2	1	104	107
0.5	1	36	63	73	172
	2	5	12	56	73
0.6	1	19	63	25	107
	2	2	1	22	25
0.7	1	12	62	10	84
	2	1	1	8	10
0.8	1	10	63	8	81
0.9	1	10	63	8	81
1	1	10	63	8	81

However, in Fig. 9, we can quickly see that preference transitions move to hotels 1 and 6 from the other hotels because the nodes whose preference transitions only move to other nodes were deleted along with their related links. In this way, finding the hotels that should be analyzed becomes easier.

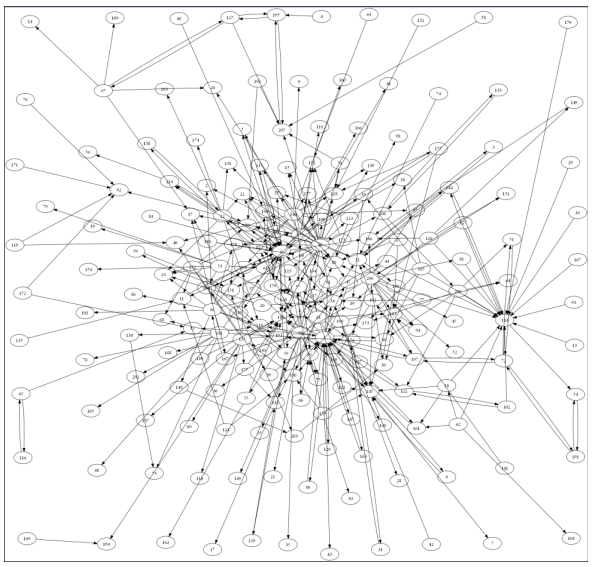

Figure 7. Preference transition network example
(Simpson coefficient: 0.4)

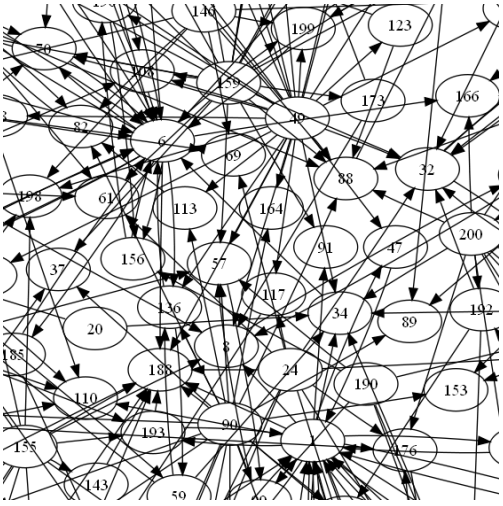

Figure 8. Before screening item by number of link
(Simpson coefficient: 0.4)

V. CONCLUSION

A method for improving the visualized preference transition network used for visualizing customer movements/trends in selecting items as well as for finding the features of items was described. The difficulty in understanding such movements/trends and features in the case of a network having many nodes and links was demonstrated, and a sensitivity analysis for identifying redundant nodes and links to overcome this difficulty was presented. The optimal choice of threshold for the Simpson coefficient for 172 kinds of products, 2,227 customers, and their 90,000 sales records was also shown.

In future work, we will develop the use of the depicted graph and apply our method to a variety of domains.

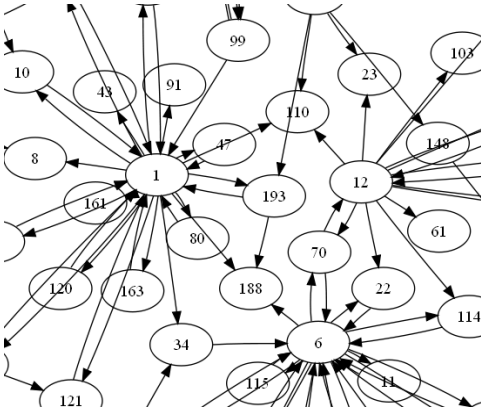

Figure 9. After screening item by number of link
(Simpson coefficient: 0.4)

REFERENCES

[1] P. Adriaans, D. Zantinge, “Data Mining Addison-Wesley,” 1996.

[2] P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, ans J. Riedl, “Grouplens: An open architecture for collaborative filtering of netnews,” in Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work. Chapel Hill, North Carolina: ACM, pp. 175-186, 1994.

[3] R. Saga, H. Tsuji, and J. Onoda, “Agent System for Notifying Hotel Room Reservation Alternatives,” 11th International Conference on Human Computer Interaction (HCII2005) , Volume 5 - Emergent Application Domains in HCI, pp.1-10 (In CD-Rom), 2005.

[4] R. Saga, Y. Hayashi, and H. Tsuji, “Hotel Recommender System based on User's Preference Transition,” IEEE International Conference on Systems, Man & Cybernetics (IEEE/SMC 2008), pp.2437-2442, 2008.

[5] T. Yoshikawa, “Visualization Techniques of Multi-Dimensional Data,” Systems, Control and Information, Vol. 52, No.7, pp.232-238

[6] Y. Ohsawa, N. E. Benson, M. Yachida, ”KeyGraph : Automatic Indexing by Segmenting and Unifing Co-occurrence Graphs”, *IEICE D-I*, Vol. J82-D-Ⅰ, No. 2, pp. 391-400, 1999

[7] M. Terachi, R. Saga, Z. Sheng, and H. Tsuji, “Visualized Technique for Trend Analysis of News Articles,” in Lecture Notes on Artificial Intelligence (LNAI 5027: Ed by M. Ali & N.T. Nguyen), Springer-Verlag Berlin Heidelberg, pp.659-668, 2008

[8] Graphviz, http://www.graphviz.org/

[9] Y. Hayashi, R. Saga, and H. Tsuji, “Competition State Visualization for Sales Record Mining,” Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE 2009), 2009.

[10] E. L. Lehmann, Nonparametric Statistical Methods Based on Ranks. New York, McGraw-Hill, 1975.