INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN
ICED 05 MELBOURNE, AUGUST 15-18, 2005

CUSTOMER DEMAND DISCOVERY FOR NEW PRODUCT DESIGN

Xiang Li, Junhong Zhou, Junxin Ren, Qizhen Yang and Wen Feng Lu*

Keywords: Customer demand, Rule mining, New product design

1. Introduction

Collecting and investigating customer demand is a critical success factor for new product design and development. This paper presents a set of customer requirement discovery methodologies to achieve wide and complex market investigations. Our approach supports flexible questionnaire model and uses data mining technologies. The discovery rule can be flexible defined. By using our rule mining methodology, the complete customer multi-preference patterns are discovered and the statistic analysis results are calculated for new product design. A software system that allows for on-line customer feedback collection, digitisation of the language feedbacks and numerical descriptions of customer preferences of a product is developed for this approach. Our approach also links the customer demand analysis data with new product design procedure. The system could significantly shorten the survey and analysis time and is thus expected to help companies to reduce design cycle time for new products.

2. Background

To successfully develop a new product, discovering customer demand or preference is always important yet difficult [1]. In most cases, there are usually many ideas and options for selection during conceptual design. One of the key factors for designers to consider during this selection is the preference of customers. To investigate this customer preference, one approach is to have a quantitative description of the preference analysis for different group of people. The most common practice for establishing such understanding is through survey with predefined questionnaire. New products or new features are listed and options are designed for customers to choose. The feedbacks are tabulated and analyzed for the desired information. Product designers will be able to use this information in their product design.

Although a good survey firstly depends on the design of questionnaire, the analysis methodology is also critical, as it will decide how much information can be extract from a collection of feedbacks. Currently, it is easy to get a preference percentage of the survey group for one particular feature. However, to get a statistical analysis on the preference of a combination of multiple features would be relatively difficult, even with the help of commonly available software tools such as MS Excel. It is even impossible to manually get the preference statistical analysis for multiple feature combinations when the number of features in consideration is large and/or huge amount of feedbacks are received. It is, therefore, desirable to have a proper information handling methodology for such survey

* Singapore-MIT Alliance Fellow

collection and analysis.

Since 1995, there have been efforts to use data mining technology to extract implicit, previously unknown and potentially useful knowledge from data [?]. The knowledge here means relationships and patterns between data elements. Further, there have been quite a number of researches on data mining technology to solve problems in marketing, planning, optimization, prediction [2]. There are also new survey tools in the market like Host Survey by Hostedware Corporation [3], XPO Online Survey System by XPO Show Services Inc. [4], EZSurvey by David Hull & Associates Ltd [5], and Survey Pro Software by Survey Professional [6], etc. Most of them have the capability of Web enabled, flexible documentation and powerful statistical analysis. However, these tools have their limitation to perform survey analysis as described above and we have not come across any other software tool that can satisfy this requirement.

In this paper, an associate rule mining methodology is presented for on-line customer feedback collection, digitisation of the language feedbacks, and numerical descriptions of customer preferences for either single or multi-features of a new product. A software prototype is developed based on this associate rule mining to test the feasibility of the proposed methodology. A scheme is also developed to convert linguistic comments or feedbacks to digital values so that statistical analysis can be performed. This will provide freedom for both survey designers in designing survey questions and customers in answering these survey questions. The system could significantly shorten the survey and analysis time and is thus expected to help companies to reduce design cycle time for new product development. In addition, the algorithm in the system is generic and can be adapted to other scenarios that require customer preference and motivation analysis.

3. CUSTOMER MULTI-PREFERENCE PATTERNS DISCOVERY (CMPD) WITH ASSOCIAITON RULE MINING

3.1 ARM Definition

Association rule mining (ARM) is one important method for knowledge discovery in databases. An ARM process searches for interesting relationships among items in a given data set [7] by finding a collection of item subsets (called itemsets) that frequently occur in database records or transactions, and then extracting the rules representing how one subset of items influences the presence of another subset [8]. For a given pair of confidence and support thresholds, the problem of mining association rules is to discover all rules that have confidence and support greater than the corresponding thresholds. For example, for a sale in a computer hardware shop, the association rule of “CD Writer $\Rightarrow$ Lens Cleaner” can mean that whenever customers buy CD writers, they also buy lens cleaners c% of the time and this trend occurs s% of the time. Therefore, an ARM operation performs two tasks [9]: discovery of frequent itemsets, and generation of association rules from frequent itemsets.

Many academic researchers have tackled the first sub-problem that is more computationally extensive and less straightforward [10]. These works designed many algorithms for efficient discovery of frequent itemsets. In the present work, we lever the ability of available ARM algorithms to rapidly discover frequent itemsets of customer preference patterns in new product design.

3.2 Approach for CMPD

There could be many ways to define questionnaire and handle customer feedbacks. In the present work, we adopted the most common practice that the survey questions are set with prescribed options. For example, if a company has a few new design options (A, B, C) for a mini-HIFI system, the survey questionnaire could be appeared as follows.

1. Which product do your perforate among A, B, C;
2. Where do you prefer to put the product; either the personal room or family room; and table top or other location.

To discover customer preference patterns from this type of survey feedbacks is to find the frequent itemsets of association rules with customer preference patterns and its support percentage. In addition, designers might have one or few important particular products or product features in mind (feature of interest, or FOI) and would like to know customer preference patterns associated with the FOI.

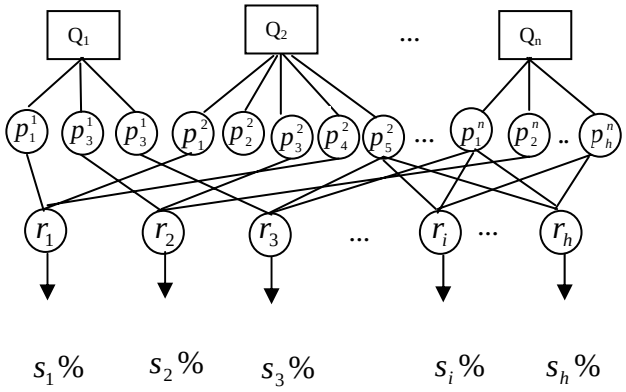

Figure 1. CMPD Network Structure

Based on this perception on the CMPD tasks, we adopted a network approach to handle the feedback collection and analysis. A three-layered network structure is designed as shown in Fig. 1

The nodes at the first layer are the input questions that are linguistic statement q_ks, such as ‘my favorite product is’, or ‘I like to set the product in’. The task of layer one is to translate these linguistic statements to numeric numbers. The u_is are their input numeric values. We set $u_i = i$, e.x. Q_1=1; Q_2=2.

The nodes at the second layer are the question options, such as ‘product A’, ‘product B’, ‘product C’, which are associated with certain questions. The function of this layer is to transfer the question numbers from layer one and translate the natural answers to binary format (explained in next section).

Layer 3 contains the rule nodes. The links here define the preconditions of the rule nodes, while the node outputs define their consequences. To determine the number of nodes in this layer, one should select one main node (main question) in the first layer relevant to the FOI. For each rule node in this layer there must have at least two links from the relevant nodes in the second layer and one of the links must go to an option node of the main node in the first layer. The operation of the network will find out the association rules of the options of other questions to the main question.

The nodes at this layer perform the heuristic AND operations, as well sometimes they also represent 'IF-AND-THEN' rules. The outputs of this layer are the percentage of supports s% to the rule. For example, an association rule can be as follows:

Age (X, "20…29)∧Gender(X, Male)

⇒ Favorite(X, "Product B") ∧Location(X, "Own room")∧Function(X, "Connect with VCD/DVD") [support = 16.7%]

where X is a variable representing the choice of a customer.

3.3 Digitalization of Survey Feedback

The function of the second layer is to digitalize the linguistic customer survey feedback to numerical data representation. Fig. 2 illustrates the linguistic-to-numerical transformation method and process. The linguistic answer to a question is represented by a special formatted numbers, consisting of a decimal portion and a binary portion. The question number in the survey questionnaire is represented by using a decimal number in the decimal portion. The options to the question are represented by binary digits in their sequence: 1 if the option is chosen and 0 if not chosen. For example, the output of layer two for question #1 will be '1|010' corresponding to a customer answer of 'my favorite product is B'. Similarly, the output of second layer for question two will be '2|1010' if the customer's answer is option one and option three. To save storage space in database and further calculations, the binary digits can be further transferred into decimal numbers when store the customer data to the database.

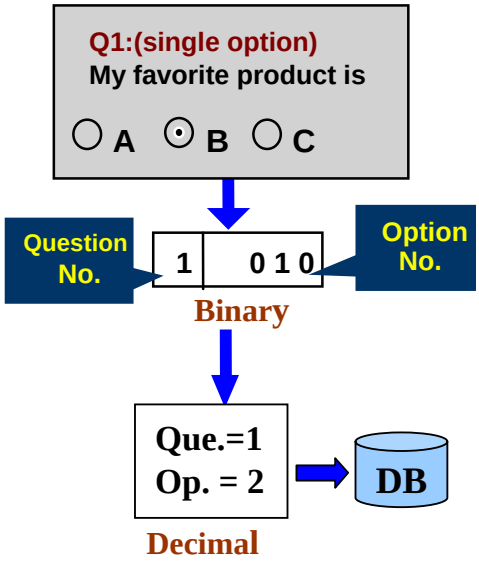

Figure 2. Illustration of the process for survey feedback digitalization

3.4 Rule Mining Algorithm

The algorithm is designed to find the support percentage of all the customer preference patterns based on CMPD network structure digitization mechanism described above and frequent itemsets principle.

Input: Database, D, of transactions; customer survey data include total number of questions *n*; survey question q_k (*k* = 1,2,…, n), sub-total number of option-itemset T(k); survey options p_j^k (*j* = 1,2,…,*m*) of *k*th question; minimum support threshold $s_{\min}\%$.

Output: Rule support percentage s%, that is the popularity of customer preference patterns percentage.

Method:

1. Convert customer questions q_k (k=1,2,…,n) and options p_j^k (*j*=1,2,…,*m*) to binary data format, and save them to the database *D* as decimal data format;
2. Transmit question values q_k to the layer one of CMPD architecture, forming binary data value of options p_j^k to the relevant nodes in layer two.
3. Select main question as q_1.
4. Work out the rule pattern collection and do statistic analysis for each rule. There are two ways to work out the rule pattern collection and do the statistic analysis. The first methodology is to count all of the question-option combination patterns first. Determine the total rule numbers to identify total scenarios of customer preference patterns assumed that all questions are defined with single chose. Then, for each rule pattern, the customer survey records will be gone throng. The statistic result will be calculated.

$$R_{total} = \sum_{i=1}^{n-1}[\sum_{j_1=2}^{n-i+1}\sum_{j_2=j_1+1}^{n-i+2}\ldots\sum_{j_i=j_1+i-1}^{n}T(1)\cdot T(j_1)\cdot T(j_2)\cdot\ldots\cdot T(j_i)] \quad (1)$$

(*i* = 1, 2, …, *n*-1 ; j_i = 1,2, …, *i* +1)

where

n — the total number of questions;

$T(1)$ — the numbers of option-itemset of main question;

$T(j_i)$ — the numbers of option-itemset of j_i th question;

The second methodology is to generate all the question combination patters by using the following recursive algorithm:

For each question combination pattern, customer survey records will be gone throng. One customer selection pattern is a rule pattern. If the rule pattern already exists, increase the frequent. Otherwise, add the rule pattern to rule pattern collection.

For the second methodology, the maxim rule number and loop times are equal:

$$R_{\max} = CU \sum_{i=1}^{n-1} C_{n-1}^{i}$$

where

n — the total number of questions;

CU — the total numbers of customer

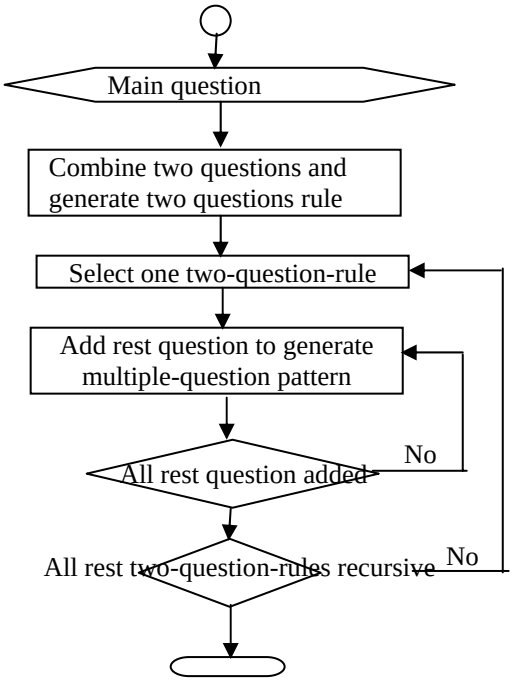

Figure 3. Question combination pattern generation flow chart

As the option will join the recursive in the first method, usually, the second way is more effective. But when the number of customers is a lot and the questionnaire is simple, the first method may take less time.

5. Filter out frequent rules R_u (u = 1,2,…,h) at the layer 3. R_{total} is a superset of R_u, that is, its members may or may not be frequent rules, but all of the frequent rules R_u are included in R_{total}. R_{total} can be huge. To reduce the size of rule base and heavy computation, a scan of the database to determine the count of each candidate in R_{total} is carried out to result in the determination of R_u. That is, all candidates having a rule count value s% no less than the minimum support count s_{min}% are frequent by definition, and therefore belong to R_u.

6. Result in the rule count value s% as the output of the node of rule layer.

7. Save the frequent rules R_us and the rule support percentage s%, (the popularity of customer preference patterns percentage) to the rule base and the database respectively.

8. Convert the discovered frequent rules R_us back to their original natural questions and options.

4. The prototype System

The CPPD architecture described in the previous section has been implemented in a prototype software system with Java and VB languages. The system is designed to be web enabled for capturing customer's survey data and analyzing customer requirement with a standalone system. Figure 3 shows the prototype system architecture. The core of the system is the data mining engine consisting of three components: statistical analyzer, multi-preference analyzer, and reasoning analyzer. The CMPD algorithm is embedded in the component of multi-

preference analyzer. The data mining engine is integrated with five supporting modules, namely User Account Manager, Data Manager, GUI controller, Database and Rule Base. In operation, the data manager receives customer data through Web-based survey form by GUI controller, and then converts the natural answers to binary formats and decimal formats to the database. The engine extracts the data and information from the database directly when performing a dictated task. Finally, the outcome, analysis results are sent to the designer

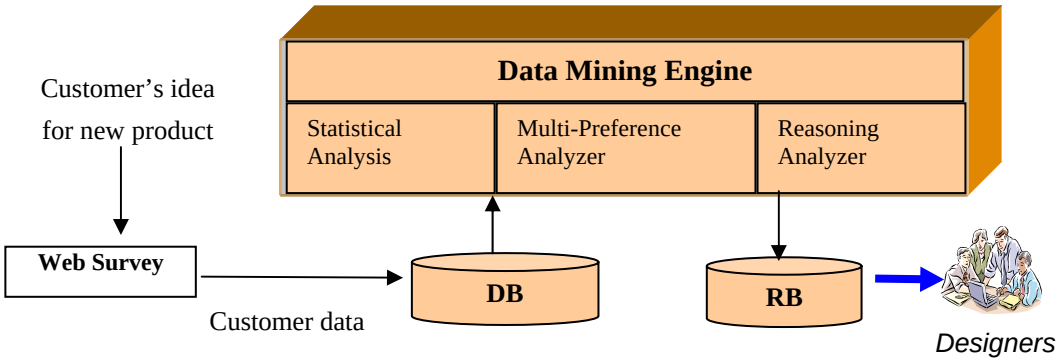

Figure 4. Architecture of the prototype

5. An Illustrative Case Study

The prototype system has been tested in several practical applications. One such case is presented here to illustrate the quantification process and the working of the prototype system.

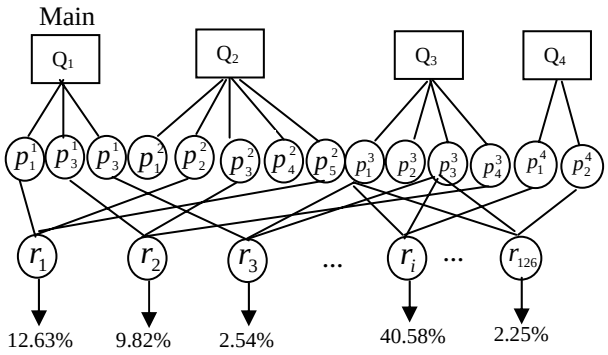

Figure 5. The case study

The case is associated with an analysis of the customer preference patterns for a new HIFI system design in an electronic company. The designer in the company would like to have quantitative understanding on which product option was preferred and how their customers would use their products.

Table 1 shows a few representative customer survey questions and their corresponding options. A total of 499 survey feedbacks are collected and their digitized results are represented in Table 2 (binary data format) and Table 3 (decimal data format). Based on the customer answer data, the CMPD architecture is built as shown in Fig. 5.

Table 1. Questions and options of customer survey for NPD

Question No.	Questions	Options (p_i^k)
1	My favorite product is	a. Audio system A b. Audio system B c. Audio system C
2	I plan to put it in/on	a. my room b. family room c. bookshelf d. table top e. other location
3	I plan to use the audio system	a. stand alone b. connected to PC c. connected to game machine d. connected to VCD/DVD
4	I have at home a	a. VCD b. DVD

Assign the question values q_k (k = 1,2,3,4) to the nodes in Layer 1 and the binary data value of options p_j^k to the relevant nodes in Layer 2. Taking the first question q_1 as the main question, the total number of rules is calculated as shown in eq. (2):

$$
\begin{aligned}
R_{total} &= \sum_{i=1}^{n-1}[\sum_{j_1=2}^{n-i+1}\sum_{j_2=j_1+1}^{n-i+2}\ldots\sum_{j_i=j_1+i-1}^{n}T(1)\cdot\cdot T(j_1)\cdot T(j_2)\cdot\ldots\cdot T(j_i)] \\
&= T(1)\cdot[\sum_{j_1=2}^{4}T(j_1)+\sum_{j_1=2}^{3}\sum_{j_2=3}^{4}T(j_1)\cdot T(j_2)+\sum_{j_1=2}^{2}\sum_{j_2=3}^{3}\sum_{j_3=4}^{4}T(j_1)\cdot T(j_2)\cdot T(j_3)] \\
&= T(1)\cdot\{[T(2)+T(3)+T(4)]+[T(2)\cdot T(3)+T(2)\cdot T(4)+T(3)\cdot T(4)] \\
&+[T(2)\cdot T(3)\cdot T(4)]\} \\
&= 3\times[5+4+2+4\times5+5\times2+4\times2+4\times5\times2] \\
&= 267
\end{aligned}
\qquad (2)
$$

Table 2. Binary customer data format

ID		1	2	3	4	…..	499
Q1	p_1^1	0	0	1	0	…..	0
	p_2^1	1	1	0	0	…..	1
	p_3^1	0	0	0	1	…..	0
Q2	p_1^2	1	1	0	0	…..	1

	p_2^2	0	0	1	1	…..	0
	p_3^2	1	0	1	0	…..	0
	p_4^2	0	1	0	1	…..	0
	p_5^2	0	0	0	0	…..	0
Q3	p_1^3	1	1	1	0	…..	1
	p_2^3	0	1	0	0	…..	1
	p_3^3	1	0	0	1	…..	0
	p_4^3	0	0	1	0	…..	0
Q4	p_1^4	1	1	0	0	…..	0
	p_2^4	0	0	1	1	…..	1

Table 3. Decimal customer data format

ID	1	2	3	4	…..	499
Q1	2	2	1	3	…..	2
Q2	20	18	12	10	…..	16
Q3	10	12	9	2	…..	12
Q4	2	2	1	1	…..	1

It can seen that for such a relatively small number of questions and simple option combinations, the R_{total} is as many as 267. The R_{total} can be huge if more questions and options are involved. To reduce the size of rule base as well as the computation loading, a rule filtering is carried out to pick out the frequent rules R_u (u = 1,2,…,h) at the layer 3 based on a assigned minimum rule support, such as 5% of the total feedbacks. Those rules with support of less than 1% are ignored in further calculation. In the illustrative case, the calculation results of the frequent rule R_u and their support s% are shown in Table 4. It is not difficult to translate the rules back to natural language. For example, the rule number #39 can be explained as:

“The percentage of customers, who prefer product B, like to put it in their own rooms, wants to use it as a stand alone unit and have a separate VCD at home, is 17%”.

Figure 6 shows the prototype user interface of customer multi-preference analysis.

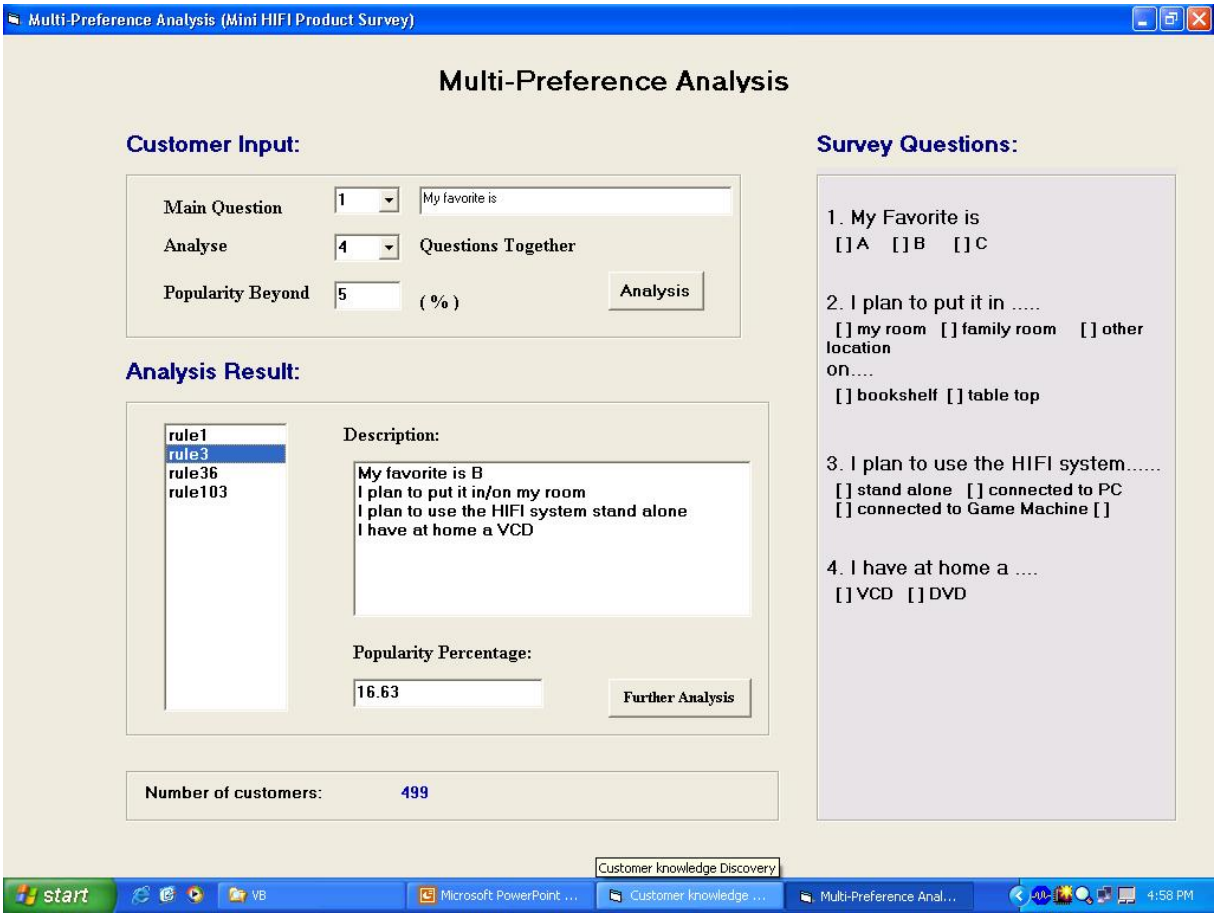

Figure 6. The Prototype User Interface of Customer Multi-preference Analysis

Table 4. Frequent rule results

Frequent Rule	**IF**	**THEN**	**Rule Support**
(R_u)	Option of main question	Option of other questions' combination	s (%)
#1	$p_1^1 \Rightarrow$	p_1^2	12.63
#7	$p_1^1 \Rightarrow$	p_1^3	9.82
#8	$p_1^1 \Rightarrow$	$p_1^2 \cap p_4^2$	2.43
#19	$p_2^1 \Rightarrow$	p_1^2	40.58
#20	$p_2^1 \Rightarrow$	p_1^3	35.68

#29	$p_2^1 \Rightarrow$	$p_1^2 \cap p_4^2$	3.81
#39	$p_2^1 \Rightarrow$	$p_1^2 \cap p_1^3 \cap p_2^4$	17.24
#42	$p_3^1 \Rightarrow$	$p_1^2 \cap p_4^2$	2.25
#50	$p_2^1 \Rightarrow$	$p_1^2 \cap p_1^3 \cap p_1^4$	8.93
#66	$p_2^1 \Rightarrow$	$p_1^2 \cap p_4^2 \cap p_2^4$	3.81
#67	$p_2^1 \Rightarrow$	$p_1^3 \cap p_1^4$	10.52
#73	$p_2^1 \Rightarrow$	$p_1^2 \cap p_2^4$	11.46
#104	$p_1^1 \Rightarrow$	$p_1^3 \cap p_1^4$	2.54
#111	$p_3^1 \Rightarrow$	$p_1^2 \cap p_1^3 \cap p_2^4$	4.73
#135	$p_3^1 \Rightarrow$	$p_1^2 \cap p_4^3 \cap p_1^4$	1.85
……	……	……	……
Total Frequent Rules	126	S_{min}	1.00

6. Customer Demand Discovery Utilization in New Product Design

As a key factor to consider in new product design, providing customer demand information will greatly help the designer in new product design. In our approach, the customer demand survey results are stored in PDM system. During conceptual design, the designer can view the customer preference of different design patterns. They can also set the features related with new product or component. When the designer wants to look at the related customer demand survey results, our service will search the records and pop-up the analysis information to designer.

As shown in Figure 7, for the up-cover of the new Hi-Fi system, the designer will consider the color, the location of the Hi-Fi system in customer's home, whether it will be connected with VCD and whether the customer will use it to play games. So the designer set these four factors as the features of up-cover. When the designer wants to see how about the customer demand, our program will go through the rule patterns. If the rule description includes any word of "Color", "Location", "Connect to VCD", and "Play games", the statistic analysis result of this rule will be collected. Finally, all the related statistic results of the related rule patterns will be pop-up. So that the designer can know what kinds of feature can meet the customer requirement. Especially, when the features have relationships or conflictions, the combination rule patterns will give a very helpful guide for designer.

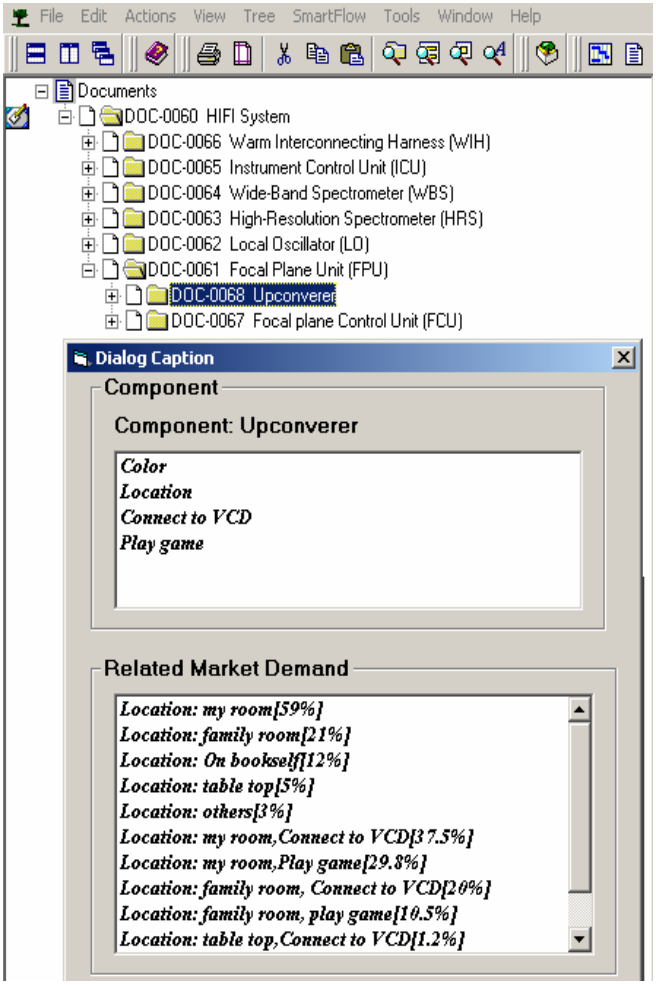

Figure 7. Show related customer demand information with component.

7. Conclusion

A new approach to the discovery of customer multi-preference patterns in new product design has been presented. An association rule mining algorithm for the research has been developed and embedded into a prototype system. The algorithm is based on a mechanism of digitization of the linguistic feedback survey and an association rule mining method. The algorithm and the prototype system allow on-line customer feedback collection through Internet and extracts customer's multi-preferences from predefined data format. The on-line feedback collection feature and the new analysis algorithm embedded in the system could significantly shorten the survey and analysis time from months to a few days and is thus expected to reduce design cycle time of new products. Although the discussion and illustrations are with product design applications, the algorithm in the system is generic and can be adopted to other scenarios that require customer preference and motivation analysis.

REFERENCES

[1] Crow, K. *Voice of the Customer*, *Product Development Forum*, DRM Associates, 2002.

[2] Adriaans, P. and Zantinge, D. *Data Mining*, Addison-Wesley Longman 1996.

[3] Hostedware Corporation, *Hosted Survey*, USA, 2004. http://www.hostedsurvey.com

[4] XPO Show Services Inc., *XPO Online Survey*, USA, 2004. http://xposhow.com

[5] David Hull & Asociates Ltd., *EZSurvey*, Canada, 2002. http://www.survey-software.com.

[6] The Survey Professional Ltd, *Survey Pro Software*, USA, 2004. http://www.hostedsurvey.com

[7] Han, J. and Kamber, M., *Data Mining: Concepts and Technique*, Morgan Kaufmann Publishers, San Francisco, 2001.

[8] Agrawal, R., Imielinski, T. and Swami, A.N., *Mining association rules between sets of items in large databases, Proceddings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C. 1993.

[9] Agrawal, R. and Srikant, R., *Fast Algorithms for Mining Association Rules*. In: Proc. 20th Int. Conf. on Very Large Databases. Santiago, Chile, 1994.

[10] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann Publishers, 2001.

Xiang Li

Singapore Institute of Manufacturing Technology

71 Nanyang Drive

Singapore 638075

Singapore

65-67938429

65-67916377

xli@SIMTech.a-star.edu.sg