Customer Personalized System of eCRM Using Web Log Mining and Rough Set

Jae-Hoon Lee, Il-Yong Chung, Sung-Joo Lee

Dept. Computer Engineering, Chosun Univ.

nuridepo, iyc, sjlee@chosun.ac.kr

Abstract

In this paper, we propose a customer personalized system that presents the web pages to users which are customized to their individuality. It analyzes the action of users who visit the shopping mall, and preferentially supplies the necessary information to them. When they actually buy some items, it forecasts the users' access pattern to web site and their following purchasable items and improves their web page on the bases of their individuality. It reasons the relation among the web documents and among the items by using the log data of web server and the purchase information of DB. For reasoning, it employs Rough Set, which is a method that searches the association rule and offers most suitable cases by reduces cases.

It reasons the web pages by considering the users' access pattern and time by using the web log and reasons the users' purchase pattern by using the purchase information of DB. On the basis of the relation among them, it appends the related web pages to link of users' web pages and displays the inferred goods on users' web pages.

Key Words : Personalized system, Web log, Reasoning, Rough Set, eCRM

1. Introduction

It needs information about users' preference and access pattern in order to make a marketing strategy and to supply the user-oriented information in Internet. We can supply the dynamic web pages or the link information, which are customized to users' individuality, by using this information.

There have been many researches about this technology. This technology is based on user's convenience and usefulness, and is coming into the spotlight[3].

In this paper, we propose and implement a customer personalized system using Web Log Mining and Rough Set. It reasons the relation among the web documents and among the items by using the log data of web server and the purchase information of DB. For reasoning, it employs Rough Set, which is a method that searches the association rule and offers most suitable cases by reduces cases.

It reasons the users' access pattern and time by using the web log and reasons the users' purchase pattern by using the purchase information of DB.

On the basis of the relation among them, it appends the related web pages to link of users' web pages and displays the inferred goods on users' web pages.

2. Related Work

2.1 Web Log Mining

In web, basic information about users' activity is automatically collected by web server. Web server stores and manages them in form of 'Web Log'. Web Log data includes a visit information to web page that can recognize user who connect to web server and has characteristics that are different from general statistic data[5].

Basically, users sequentially visit to web page according to time, and log data is used in order to find users' visit pattern to web page mainly.

Web Mining analyzes such data as log file from web site and establishes strategy. It means adding other data to web log file and analyzing them. Here, other data are customer data, account data, electronic commerce data etc.[5, 6].

Usually, Web Mining is fallen into web structure mining, web contents mining and web usage mining according to web data that become target[7, 8].

2.2 Association Rule Inquiry Algorithm

Association rule inquiry is to find type of the event that happens frequently in unit transaction in high-capacity database[2].

Apriori algorithm is a representative

algorithm of association rule inquiry. Association rule reasoning process that uses this algorithm is consisted of two steps in Table 1.

Table 1. Association rule reasoning process

***Step 1* : Determining "large itemsets"**
- Find all combinations of items that have transaction support above minimum support
- Search a set of frequently occurred items which have a transaction support more than the predetermined smallest support.

***Step 2* : Generating rules**

for each large itemset L **do**
  for each subset c of L **do**
    if(support(L)/support($L-c$) $\geq$ minimum confidence) **then**
      output the rule($L-c$) $\rightarrow c$,
        with confidence =
          support(L)/support($L-c$)
          and support = support(L);

Fig 1. shows the search process of set of frequently occurred items in Apriori algorithm.

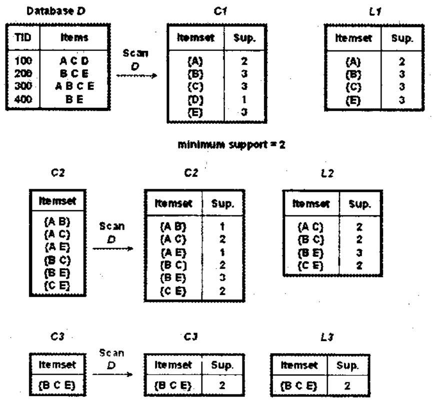

Fig. 2. Creation of candidate itemset and frequent occurrence item set

2.3 Approximations of Set

An equivalence relation induces a partitioning of the universe. These partitions can be used to build new subsets of the universe. Subsets that are most often of interest have the same value of the outcome attribute. It may happen, however, that a concept cannot be defined in a crisp manner.

Let $A=(U,A)$ be an information system and let $B\subseteq A$ and $X\subseteq U$. We can approximate X using only the information contained in B by constructing the *$B-$lower* and *$B-$upper approximations of* X, denoted $\underline{B}X$ and $\overline{B}X$ respectively, where $\underline{B}X=\{x|[x]_B\subseteq X\}$, $\overline{B}X=\{x|[x]_B\cap X\neq\varnothing\}$.

The objects in $\underline{B}X$ can be with certainty classified as member of X on the basis of knowledge in B, while the objects in $\overline{B}X$ can be only classified as possible members of X on the basis of knowledge in B.

The set $BN_B(X)=\overline{B}X-\underline{B}X$ is called the *$B-$boundary region of* X, and thus consists of those objects that we cannot decisively classify into Xon the basis of knowledge in B. The set $U-\overline{B}X$ is called the *$B-$outside region of* X and consists of those objects which can be with certainly classified as do not belong to X. A set is said to be *rough* if the boundary region is non-empty[9].

3. Personalization System

Reasoning engine is fallen into part that uses log file and that uses item set abstracted from customer and purchase information.

In part of used log file of web server collects web logs and preprocesses them for web log analysis. The preprocessing part finds out running route by user and finds out MFR(Maximal Forward Reference) because it must run Apriori algorithm. It recommends Documents using Apriori algorithm and the preprocessed data.

In part of used customer and purchase records of database, it composes customer's favorites and purchase items that connection deduction is possible. It reasons the recommended items by using rough set theory.

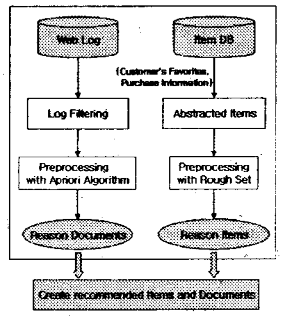

Fig 4. Reasoning Engine

4. Experiment and Estimation

4.1 Reasoning Web Page using Web Log

Reasoning using log data of Web Server collects and filters Web Log from Web server like the general method of web log analysis.

(a) User document running path(P)

The moving path using Fig 5. is as follow.

$$P=\{A,B,H,B,I,B,A,C,J,M,J,C,K,A,D,L\}$$

(b) MFR(Maximal Forward Reference)

It find out web documents that user passes through in case of proceeding except the case of backing. For example, MFR is as follow in Fig 8.

$MFR = \{\{ABH\},\{ABI\},\{ACJM\},\{ACK\},\{ADL\}\}$

If two preprocessing is finished, this system reasons by applying Apriori algorithm to the preprocessed data. It adds time-weight to the general Apriori algorithm. Adding time-weight increase the confidence of reasoning.

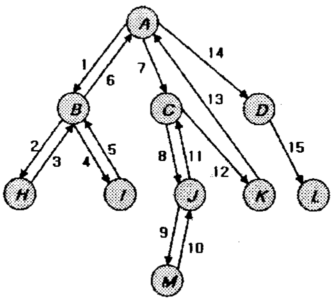

Fig 5. Running path of web document

Table 2. Web Log Data

User ID	(Web Page, Unit Time)
100	$(H,1),(A,1),(C,2)$
200	$(H,1),(A,1),(H,0),(B,2),(H,1),(A,1),(D,2)$
300	$(H,0),(A,2),(D,1),(E,2),(B,0)$
400	$(H,1),(B,1),(H,0),(E,1),(A,1),(D,3)$
500	$(H,0),(B,1),(E,1),(H,0),(A,1),(D,3)$

Table 2. is Web log data of some user who visits Web site. First data in parenthesis is Web page that he visits. Second data is a unit time. $(A,2)$ means that he visits Web page A during 2 unit time. $(H,0)$ means that he visits Web page H during less time than minimum unit time and that he pass through latter rather than visits Web site.

If pages having the visit time less than 30 seconds is ignored in Table 3. in order to apply Apriori algorithm considering time-weight, it become Table 4.

Table 3. Time Weight per Unit

Unit	Time	Weight
0	0~30sec.	0
1	30sec~5min	0.2
2	5min~30min	0.3
3	more than 30min	0.5

Table 4. Fine Web Log Data

User ID	(Web Page, Unit Time)
100	$(H,1),(A,1),(C,2)$
200	$(H,1),(A,1),(B,2),(H,1),(A,1),(D,2)$
300	$(A,2),(D,1),(E,2)$
400	$(H,1),(B,1),(E,1),(A,1),(D,3)$
500	$(B,1),(E,1),(A,1),(D,3)$

If it appoints the approval rating as 0.4 (minsup=0.4) and reasons FOIS(Frequent Occurrence Item Sets) by applying Apriori algorithm, result is as follow.

$\{H\}=(0.2+0.2+0.2+0.2)\times(3/5)=0.48$
$\{A\}=(0.2+0.2+0.2+0.3+0.2+0.2)\times(5/5)=1.3$
$\{B\}=(0.3+0.2+0.2)\times(3/5)=0.42$
$\{D\}=(0.3+0.2+0.5+0.5)\times(4/5)=1.2$
$\{C\}=(0.3)\times(1/5)=0.06$
$\{E\}=(0.3+0.2+0.2)\times(3/5)=0.42$

$\therefore\ 1-FOIS=\{H,A,B,D,E\}$

$2-FOIS=$
$\{\{H,A\},\{H,D\},\{A,D\},\{B,A\},\{B,D\},\{E,D\}\}$

$3-FOIS=\{\{H,A,D\},\{B,A,D\}\}$

4.2 Item Association using Rough Set

[*Step 1*] : Table 5. shows transaction of the purchase DB. It has 10 transaction, 7 items and $wminsup=1.24$.

Table 5. Transaction DB and Weights of Items

no	Items
1	A B C F G
2	A D
3	A B C F
4	A B C D E F
5	A B C E F G
6	A D
7	A B C E F G
8	B
9	A B C F
10	A B C E

Item set	total profit	weights	
{A}	900	0.9	$wminsup$ =1.24
{B}	300	0.3	
{C}	900	0.9	
{D}	100	0.1	
{E}	800	0.8	
{F}	100	0.1	
{G}	500	0.5	

Table 6. shows that apply (1) as follow and save item bound.

$$B(Y,k)=\frac{wminsup\times T}{W(Y,k)} \qquad \cdots(1)$$

Table 6. Bound of items

Itemset	B(Item,2)	B(item,3)	···	B(item,6)
{A}	7	5	···	4
{B}	11	6	···	4
{C}	7	5	···	4
{D}	13	7	···	4
{E}	8	5	···	4
{F}	13	7	···	4
{G}	9	6	···	4

[*Step 2*] : And include item that value is more than B(item,6) to C_1. In this case, C_1 ={{A}, {B}, {C}, {F}} and run algorithm of Table 1 using C_1. Fig 6. shows this process and the result FOIS is L_2={{AC}}, L_3 ={{ABC}}, L_4={{ABCF}}.

[*Step 3*] : Calculate item weight-minsup to (2) as follow and Table 7. shows new item bound.

$$NWminsup(j) = \frac{nw(j) \times (Wminsup \times n)}{\sum_{i=1}^{n} nw(i)}$$

$$nw(i) = \left[\sum_{k=1}^{n} \frac{w(k)}{sp(k)} \right] \times \frac{sp(i)}{w(i)} \quad \cdots (2)$$

Itemset	sp	Wsp	3.sp_bound	Large?	C_2?
{AB}	7	0.84	6	N	Y
{AC}	7	1.26	5	Y	Y
{AF}	6	0.6	7	N	N
{BC}	7	0.84	6	N	Y
{BF}	6	0.24	10	N	N
{CF}	6	0.6	7	N	N

⇓

Itemset	sup	Wsup	4.sp_bound	Large?	C_3?
{ACF}	6	1.14	5	N	Y
{ABC}	7	1.47	5	Y	Y
{ABF}	6	0.78	6	N	Y
{BCF}	6	0.78	6	N	Y

⇓

Itemset	sup	Wsup	4.sp_bound	Large?	C_3?
{ABCF}	6	1.32	5	Y	Y

Fig 6. Association Item of weight value

Table 7. new bound of items

Itemsets	NB(Item, 2)	NB(item, 3)	......	NB(item, 6)
{A}	4	4	......	2
{B}	15	15	......	5
{C}	3	3	......	2
{D}	19	19	......	6
{E}	2	2	......	1
{F}	37	37	......	11
{G}	3	3	......	2

In this paper, when we use the algorithm that proposed, we get lastly FOIS as L_2 ={{AC}}, L_3={{ABC}}, L_4 ={{ABCF}} and NL_2={{AE}, {AG}, {CE}}, NL_3={{AEG}, {ACE}, {ACG}, {CEG}}, NL_4={{ACEG}}.

5. Conclusion

By enlargement of business in Internet, the relation with customer become more and more important. For more smooth relation with customer, business constructs not Web-site presented one-side to customer but being able to communicate with customer.

In this paper, we proposed personalized system for this need. The proposed system presented one method of personalization to construct web-site according to individual's favorite.

This system analyzed user's pattern by using web log data and rough set, and reasoned goods that user buys actually and web pages that user visits frequently. It becomes more believable personalized system by this reasoning.

We increased confidence of personalization by using web log information and rough set at the same time and by using Apriori algorithm and approximation of rough setas the reasoning method.

In part that uses log data of web server, the proposed system analyzed user's access pattern to web page by considering time that user stays at web page as well as web pages that user frequently approaches, and could recommend the high believable web pages to user. In part that recommends goods by using rough set theory, this system recommended goods by applying Apriori algorithm that uses in traditional connection rule inquiry.

References

[1] Ramakrishnan Srikant, Yinghui Yang, "Mining Web Logs to Improve Website Organization", International World Wide Web Conferences(WWW 2001 in Hong Kong, China), pp. 430-437, 2001

[2] Hyuncheol Kang, Byoungcheol Jung, "A Study of Web Usage Mining for eCRM", The Korean Communications in Statistics, Vol.8, No.3, pp. 831-840, 2001.

[3] Jong-Su Park, Yeong-Kung Yu, "Inquiry and Application of Association Rule", SIGDB Spring Tutorial in The Korea Information Science Society, 1998.

[5] Jeong Yong Ahn, "A Study on the Mining Access Patterns from Web Log Data", IEICE Transactions on Information and System, Vol. E85-D, No. 4, pp. 782-785, 2002.

[6] Jin-Sung Kim, "Membership Functions and AHP-Based Negotiation Support in Electronic Commerce", Journal of Fuzzy Logic and Intelligent Systems Vol.12, No.4, pp.347-352, 1225-1127, 2002.

[7] Sug-Ki Kim, Hyun-Jung Koh, "Customer Support System of Web Based intelligent", Processing of KFIS Fall Conference, pp.265-268, 2003.

[8] http://www.webpro.co.kr

[9] Zdzislaw Pawlak, "Rough sets: theoretical aspects of reasoning about data", Kluwer Print Demand, 1992.