A Rough Set Approach to Grouping Goods in Electronic Commerce

Qiguo Duan, Duoqian Miao
Department of Computer Science and Technology, Tongji University, Shanghai, 201804,China
The Key Laboratory of "Embedded System and Service Computing", Ministry of Education, Shanghai, 201804,China
dqgcn@126.com, miaoduoqian@163.com

***Abstract* — As Electronic Commerce (EC) become more and more prevalent to people, it is very critical to provide the right information to the right customers. The EC sites are generating large amount of data on customer purchases. Data mining techniques in EC domain is currently a hot research area. This paper proposes a rough set based approach to obtaining patterns of goods and classifying associated goods into groups, which can be used by the sites to group their goods according to the customers' preference and adopt appropriate selling policies.**

***Keywords* — rough set, electronic commerce, data mining.**

1. Introduction

Today, as the Electronic Commerce (EC) becomes more and more diverse, it is very critical to provide the right information to the right customers. In EC environment, how to find the association between goods is very important. If this kind of information is provided to the Web site manager, the performance of cross-selling should be improved by grouping associated goods in one Web page which helps customers purchase associated goods conveniently and quickly. In this paper, a novel approach to grouping associated goods based on an extended rough set, i.e., tolerance rough set, is proposed, in which the value of item co-occurrence is use to mine association of goods and then the tolerance rough class of item is generated to capture the relationship among different records or items in the transaction database.

2. Rough Set Theory

2.1. Rough Set

Rough set theory is a formal mathematical tool to deal with uncertain, vague and imprecise information introduced by Pawlak [1]. It has been successful in many applications [2] [5] [7]. In rough set theory, an information system, which is also called a decision table, is defined as $S = (U, A \bigcup D, V, f)$, where U is the finite set of objects, A a collection of condition attributes, D a collection of decision attributes, V a set of values of attributes in A and $f : A \to V$ a description function. For any $R \subseteq A$, there is an equivalence relation $I(R)$ as follows:

$$I(R) = \{(x, y) \in U^2 \mid \forall a \in R \ \ a(x) = a(y)\} \ . \tag{1}$$

If $(x, y) \in I(R)$, then x and y are indiscernible by attributes from R. The equivalence classes of the R-indiscernibility equivalence relation $I(R)$ are denoted by $[x]_R$. For any concept $X \subseteq U$ and attribute subset $R \subseteq A$, X could be approximated by the R-lower approximation and R-upper approximation as Figure 1.

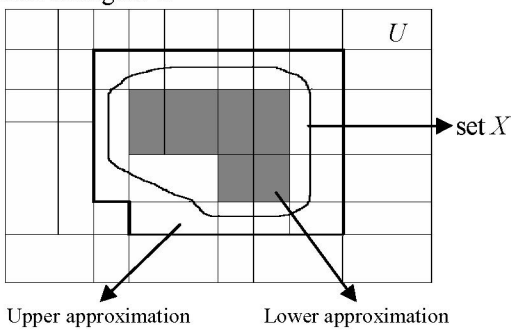

Figure 1. Lower and Upper approximations of a rough set

The R-lower approximation of X is the set of objects of U that are surely in X, defined as:

$$\underline{R}X = \{x \in U \mid [x]_R \subseteq X\} \ . \tag{2}$$

The R-upper approximation of X is the set of objects of U that are possibly in X, defined as:

$$\overline{R}X = \{x \in U \mid [x]_R \bigcap X \neq \varnothing\} \ . \tag{3}$$

The C-positive region of D is the set of all objects from the universe U which can be classified with certainty into classes of U/D employing attributes from C, that is:

$$POS_C(D) = \bigcup_{X \in U/D} \underline{C}X \ . \tag{4}$$

2.2. Tolerance Rough Set

The classical rough set theory is based on equivalence relation that divides the universe of objects into disjoint classes. However, the requirement for equivalent relation is not appropriate for some applications in practice. By relaxing

This work is supported by the National Natural Science Foundation of P.R. China (No. 60475019, 60175016).

the equivalence relation R (with reflexive, symmetric and transitive properties) to a tolerance relation (with reflexive and symmetric properties), where transitivity property is not required, a generalized tolerance space is conducted in [3]. Further, a tolerance rough set model (TRSM) which employs a tolerance relation instead of an equivalence relation in the original rough set model is introduced below [4] [6].

Let $I: U \rightarrow P(U)$ to denote a tolerance relation, if and only if $x \in I(x)$ for $x \in U$ and $y \in I(x) \Leftrightarrow x \in I(y)$ for any $x, y \in U$, where $P(U)$ are sets of all subsets of U. Thus, the relation $xIy \Leftrightarrow y \in I(x)$ is a tolerance relation and $I(x)$ is a tolerance class of x. Define the tolerance rough membership function $\mu_{I,V}$, as $x \in U, X \subseteq U$,

$$\mu_{I,V}(x, X) = v(I(x), X) = \frac{|I(x) \cap X|}{|I(x)|} . \quad (5)$$

The tolerance rough set for any $X \subseteq U$ are then defined as:

$$L_R(X) = \{x \in U \mid v(I(x), X) = 1\} . \quad (6)$$

$$U_R(X) = \{x \in U \mid v(I(x), X) > 0\} . \quad (7)$$

For the problem researched in this paper, how to make the most of subtle information among transaction data is helpful to classify goods into groups. With its ability to deal with vagueness, tolerance rough set is a promising tool to model relations between items. The usage of tolerance space and tolerance class to enrich items relation allows us to discover subtle associations. For instance, in transaction database, the items "beer" and "bread" co-occur frequently in some transaction records and the items "beer" and "beef" co-occur frequently in other transaction records, thus we can infer that the item "beer", "bread" and "beef" are latent associated items. Thus, the approach based on tolerance rough set was proposed as a way to enrich items associated relationship and group associated goods in tolerance class.

3. Application of Rough Set in Grouping Goods

3.1. Tolerance Space of Items in Transaction Database

Let $U = \{d_1, \ldots, d_M\}$ be a set of purchase records and $T = \{t_1, \ldots, t_N\}$ set of items for U. The tolerance space is defined over a universe of all items for U. The idea of items enrichment is to capture associated purchase items into classes. For this purpose, the tolerance relation I is determined as the co-occurrence of items in all purchase records from U. For example, item "beer" and "breed" co-occur frequently in some records, item "beer" and "beef" co-occur frequently in other records, then the item "beer", "beef" and "bread" can be contained in one tolerance class.

3.2. Tolerance Class of Item

Let $f_U(t_i, t_j)$ denotes the number of record in U in which both item t_i and t_j occurs. The uncertainty function I with respect to co-occurrence threshold θ defined as

$$I_\theta(t_i) = \{t_j \mid f_U(t_i, t_j) \geq \theta\} \cup \{t_i\} . \quad (8)$$

Obviously, the above function satisfies conditions of being reflexive: $t_i \in I(t_j)$ and symmetric: $t_j \in I(t_i) \Leftrightarrow t_i \in I(t_j)$ for any t_i, $t_j \in T$. Thus, $I(t_i)$ is the tolerance class of item t_i. A tolerance class represents a purchase demand that is characterized by items it contains. The membership function μ for $t_i \in T$, $X \subseteq T$ is then defined as:

$$\mu(t_i, X) = v(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|} . \quad (9)$$

Finally, the lower and upper approximations of any subset $X \subseteq T$ can be determined with the obtained tolerance relation respectively as:

$$L_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) = 1\} . \quad (10)$$

$$U_R(X) = \{t_i \in T \mid v(I_\theta(t_i), X) > 0\} . \quad (11)$$

4. System Architecture

This section gives an overview of system architecture for application of rough set to group goods in EC site. The system architecture is shown in Figure 2.

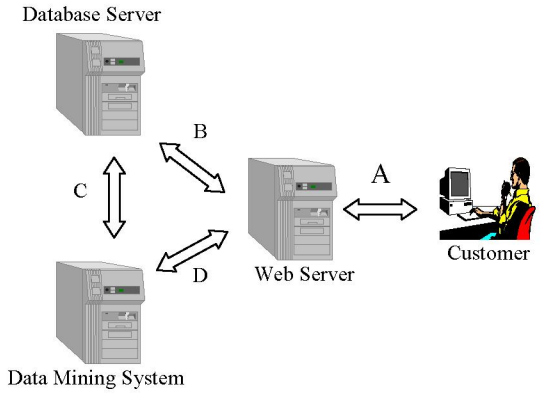

Figure 2. Schematic view of system architecture

The architecture is mainly composed of four components, i.e., Web server, database server, data mining system and customer computer. The procedure is described as follow:

A. Once customer visits the EC Web site, such as browser, buy goods through IE, the Web server accepts the requests and then fetches the information about goods in the group to IE.
B. Web server inserts, changes, deletes and gets transaction data from database server.
C. Data mining system analyses transaction data from database server and groups associated goods based on the proposed approach.

D. Web server calls the data mining system to group associated goods periodically.

5. Key Algorithm

We use TRS-Grouping to denote the proposed grouping techniques that generate tolerance class based on tolerance rough set to mine transaction database and put associated goods in groups. The TRS-Grouping algorithm is described as follow:

Input: *RMCM*, a matrix composed by record-goods pairs
θ, co-occurrence threshold.

Output: *MTCM*, goods tolerance classes matrix.

Step1. Construct a binary occurrence matrix *BOM* based on record-goods matrix *RMCM* as follows:

$BOM = [bom_{ij}]_{N \times M}$

where $bom_{ij} = 1$, if $tf_{ij} > 0$; otherwise $bom_{ij} = 0$. Each column in *BOM* is a bit vector which denotes goods occurrence in a record.

Step2. Construct goods co-occurrence matrix as follows:

$COM = [com_{x,y}]_{M \times M}$

where $com_{x,y} = card(bom[x]\ AND\ bom[y])$; $BOM[x]$, $BOM[y]$ are pair of goods x, y bit vectors in the *BOM* matrix; *AND* is a Boolean AND operation between bit vectors and *card* return cardinality of the vector. The $com_{x,y}$ is the co-occurrence frequency of goods x and y.

Step3. Set a co-occurrence threshold θ, and then calculate a goods tolerance binary matrix *MTCM* as follows:

$MTCM = [mtcm_{x,y}]_{M \times M}$

where $mtcm_{x,y} = 1$, if $com_{x,y} \geq \theta$; otherwise $mtcm_{x,y} = 0$. Thus, the value of $mtcm_{x,y}$ denotes whether goods x and y are in tolerance relation or not.

According to the TRS-Grouping algorithm, a tolerance class for given goods can be easily gotten by scanning the resulting matrix.

6. Conclusion

In this paper, a novel approach to grouping goods base on tolerance rough set was proposed, in which the value of item co-occurrence was used for mining association of goods and then the tolerance rough class of item was used for capturing the relationship among different records or items in the transaction database. It is feasible and useful to discovery the knowledge in data obtained from EC transactions by applying the rough set. In the future, more applications of data mining in EC will be developed, and EC sites that incorporate data mining results with its strategy is sure to be succeeded

REFERENCES

[1] Z. Pawlak, Rough sets: Theoretical aspects of reasoning about data, Kluwer Dordrecht, 1991.

[2] D.Q. Miao, L.S. Hou, "A comparison of rough set methods and representative inductive learning algorithms", Fundamenta Informaticae, v 59, n 2-3, February, 2004, pp. 203-219.

[3] A. Skowron, J. Stepaniuk, "Tolerance approximation spaces", Fundamenta Informaticae, 1996, pp.245–253.

[4] Tu Bao Ho, Ngoc Binh Nguyen, "Nonhierarchical document clustering based on a tolerance rough set model", International Journal of Intelligent Systems, Volume 17, 2002, pp.199-212.

[5] Y.Y. Yao, "Information granulation and rough set approximation", International Journal of Intelligent Systems 16, 2001, pp.87-104.

[6] Chi Lang Ngo, Hung Son Nguyen, "A tolerance rough set approach to clustering web search results", J.-F. Boulicaut et al. (Eds.): PKDD 2004, LNAI 3202, 2004, pp.515–517.

[7] Y.Y. Yao, J.T. Yao, "Granular computing as a basis for consistent classification problems", Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining, 2002, pp.101-106.