A Modeling Method of Interest using Rough Set Theory and Its Application

Akihiro OGINO, Toshikazu KATO

Chuo University, Science and Engeneering, 1-13-27 Kasuga Bunkyo-ku Tokyo 112-8551 JAPAN,
{ogino,kato}@chuo-u.ac.jp

Abstract: This study proposes a modeling method of each customer's interest in commodities using Rough Set theory. In order to detect which attributes of commodities are focused on by a customer, this study calculates attribute sets of a commodity as "Interest Set", which characterize customer's interest and utilize an "Interest Map", which displays the interest tendencies of customers using the interest set of each customer, visually. This paper also describes an application system of this method, which indicates interest clothes to each customer using the Interest Map.

***Key words*:** *Rough Set, Interest Modeling, Personalized Recommendation*

1. Introduction

Recently, customers have been exposed to a diverse range of commodities through the Internet and the general globalization of the market place. However, it can be difficult for the customers to find commodities that they want from the vast range of commodities available.

This paper proposes a method for aiding customers in finding commodities from a vast rage. In order to detect a customer's interest, we model the relation between an individual's interest in a commodity and the attributes of the commodity. This is done by analyzing a customer-answered questionnaire using Rough set theory[1]. Our method also utilized an "Interest Map", which displays the interest tendencies of customers, visually.

The paper is organized as follows: Section 2 describes the basic idea for detecting each customer's interest in commodities. Section 3 gives a system, which supports the making of an individual model about suitable commodities and informs them of their interest in commodities.

2. Method

In this section, we present an algorithm using Rough set theory for modeling the relation between an individual's interest in a commodity and the attributes of the commodity. The algorithm is organized as follows:

- Obtain personal data about interesting commodities through questionnaires
- Detect the important attributes of commodities selected or rejected by each customer's decision by analyzing the personal data using Rough set
- Make a personal model, which shows the individuals tendencies of interest in commodities, using data on important attributes.
- Create "Interest Map" using the personal model and show tendencies of interest in commodities to the customers.

2.1 Obtain personal data about interest

In order to model the individual's relation between his/her interest in a commodity and the attributes of the commodity, we obtain data on attributes of commodities in which each customer has an interest/uninterest, from each customer.

We ask questions concerning each customer's interest in sample commodities to each customer through a Web site, as in Figure 1, to get the individual's data.

And then, we accumulate matrix data, as in Table.1, of each customer. Table.1 shows the relation between customer X's interest/uninterest in commodities (I_1, I_2, I_3, I_4, I_5) and the attributes (Att_A, Att_B, Att_C, Att_D) of the commodities. In Rough set theory, the table is called "Decision Table".

2.2 Detect Important Attributes of Commodities

We aim to find the minimal set of consistent attributes of a commodity that characterize customer's interest using Rough Set theory[2]. This study calls the minimal set "Interest Set" and shows an interest set Y of customer X as S_{XY}. The method for extracting the interest set is to form a decision matrix, such as Table 2, corresponding to each individual value d of decision attribute Q.

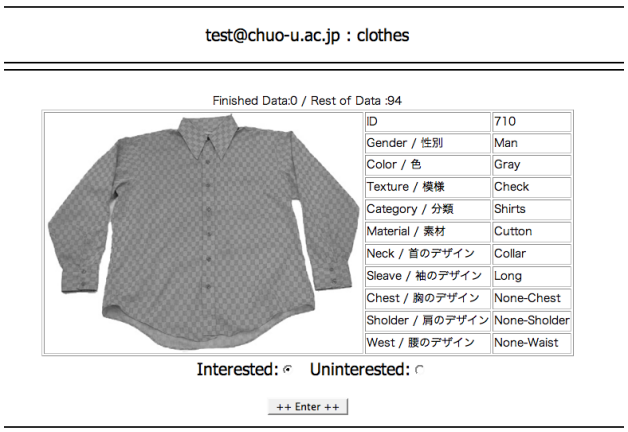

Fig.1 An example of the Web site for asking interest/uninterest in commodities to each customer

This study lists all attribute and its value that differ between commodities having $Q = d$ and $Q \neq d$, as the decision matrix for value d of decision attribute Q.

For example, Table 2 shows the decision matrix corresponding to Q (Evaluation) = d (Interest) in the customer X. We list attribute and its value, which characterize Q (Evaluation) = d (Interest) by comparing the customer's decision about whether a commodity is interested or not, in Table 2. And then, we translate the Table 2 from tabular form to logical form to extract the interest set by simply expressing the set of commodities as the following disjunction.

$$
\begin{aligned}
&(a_1 \vee b_1 \vee c_1) \wedge b_1 \wedge (a_1 \vee c_1 \vee d_2) \wedge (a_1 \vee b_1 \vee c_1 \vee d_2) \\
&= b_1 \wedge (a_1 \vee c_1 \vee d_2) \\
&= b_1 \wedge a_1 \vee b_1 \wedge c_1 \vee b_1 \wedge d_2 \\
&\therefore S_{X1}(b_1,a_1), S_{X2}(b_1,c_1), S_{X3}(b_1,d_2)
\end{aligned}
$$

Therefore, we can detect three interest sets as an interest tendency of customer X.

2.3 Model Personal Interest In Commodities

Our aim is to predict new commodities, which a customer doesn't know yet, but may be interested in. Therefore, we utilize the interest set to make a personal model of interest and use the model to create recommendations. For example, we calculate interest sets to evaluate interest in sample clothes. And then, if we find many similar interest sets that have the same attribute, such as Blue, we predict that the degree of "Blue" in interest sets is a priority for the customer.

We define a degree of an attribute in interest sets as " Interest Attribute Degree " and we detect which attributes are focused on by a customer.

Table 1. Decision Table: An example of a customer's relation between the customer's interest/uninterest in a commodity and the attributes in the commodity

Commodity	Att_A	Att_B	Att_C	Att_D	Evaluation
I_1	a_1	b_1	c_1	d_1	Interest
I_2	a_2	b_1	c_2	d_1	Interest
I_3	a_2	b_2	c_2	d_1	Uninterest
I_4	a_1	b_2	c_1	d_2	Interest
I_5	a_1	b_1	c_1	d_2	Interest

Table 2. The Decision Matrix Corresponding to *Q* (Evaluation) = *d* (Interest)

	I_3
I_1	a_1, b_1, c_1
I_2	b_1
I_4	a_1, c_1, d_2
I_5	a_1, b_1, c_1, d_2

And if we can know that the customer selected similar clothes, which all have the attributes of S_{XY}(Blue, Shirts), from the sample clothes, we can also detect the type of clothes for the customer is usually interested in. We define an existing degree of interest sets in all data as "Interest Set Degree".

In this study, we calculate the "Interest Attribute degree" and the "Interest Set Degree" from existing commodities. We model the relation between individual interest in commodities and attributes of commodities using the degrees.

2.3.1 Interest Attribute Degree

This study defines Interest Attribute Degree: IAD_X of customer X as follows:

$$
IAD_X(V_a) = \frac{m_k \times N_X}{n_{V_a} \times \sum_{i=1}^{i<r} (C_{V_a})_i \times r_{V_a}}
$$

V_a : Value: V of Attribute a in Commodities
r : Number of Interest Set
N_X : Number of Existing Commodities in Customer X
m_K : Number of Value included in Attribute K
n_{v_K} : Number of V_a in All Commodities
C_{v_k} : Number of Interest Set included in V_a
r_{v_k} : Number of V_a included in r

For example, we can calculate $IAD_X(b_B)$, which is Interest Attribute Degree of Value b of Attribute B in commodities as follows.

In this regard, each data are N_X=5, m_K = 3, n_{v_K}=5, C_{v_k}=2 and r_{v_k}=3.

$$IAD_X(b_B) = \frac{m_k \times N_X}{n_{V_a} \times \sum_{i=1}^{i<r} (s_{V_a})_i \times r_{V_a}} = \frac{3 \times 5}{5 \times 6 \times 3} = 0.166$$

We detect which attribute of commodities more important to the customer by searching IAD_X of all value in all attributes. For example, if the $IAD_X(Blue)$, which shows important attribute degree of blue, is high than other IAD_X, we can assume the customer priorities blue above other attributes.

2.3.2 Interest Set Degree

This study defines Interest Set Degree: ISD_X of a customer X as follows:

$$ISD_X(S_{Xy}) = \frac{n_{S_{Xy}}}{N_X}$$

S_{Xy}	:	Interest Set y of a Customer X
N_X	:	Number of Existing Commodities in Customer X
$n_{S_{Xy}}$	:	Number of commodities including S_{Xy} in Existing Commodities

For example, we can calculate $ISD_X(S_{X1})$ and $ISD_X(S_{X2})$, which is Interest Set Degree of a customer X about sample commodities as follows:

$$ISD_X(S_{X1}) = n_{S_{Xy}} / N_X = 1/5 = 0.2$$
$$ISD_X(S_{x2}) = n_{S_{Xy}} / N_X = 2/5 = 0.4$$

The above result indicates that the $ISD_X(S_{X2})$ is more popular than the $ISD_X(S_{X1})$ for the customer X and the customer will select commodities, which have the interest set S_{X2}, more than commodities, which have interest set S_{X1}.

2.4 Interest Map

We make the "Interest Map" by calculating "Reflected Degree" of commodities using IAD_X and ISD_X of a customer X. The reflected degree indicates how much a commodity includes the IAD_X and ISD_X of the customer X. Therefore, if a commodity includes much of the IAD_X and ISD_X of the customer, we predict that the commodity is suitable for the customer X.

This study defines the Reflected Interest Attribute Degree: $RIAD_X$ and Reflected Interest Set Degree: $RISD_X$ of a customer X as follows:

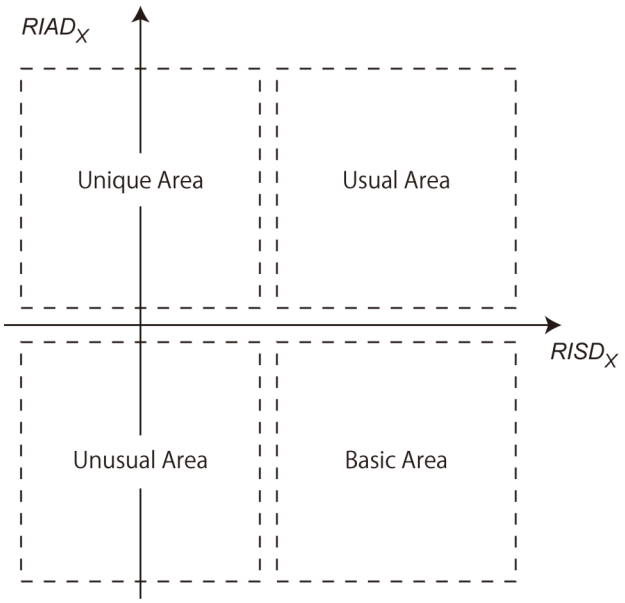

Fig.2 Interest Map

$$RISD_X(Co_i) = \frac{\sum_{y=1}^{y<m} ISD_X(S_{Xy})}{n_{S_{Xy}}}$$

$$RIAD_X(Co_i) = \sum IAD_X(V)$$

Co_i	:	A commodity i is one of All commodities
S_{Xy}	:	Interest Set y of a Customer X
$n_{S_{Xy}}$	:	Number of commodities including S_{Xy} in Existing Commodities
$ISD_X(S_{Xy})$	:	Interest Set Degree of a customer X about S_{Xy}
m	:	Number of S_{Xy} included in Co_i
V	:	Class about all values of attributes AT $V = \bigcup_{a \in AT} V_a$
V_a	:	Value: V of Attribute a in Commodities
AT	:	Class of Attribute about Commodities
$IAD_X(V_y)$	:	Interest Attribute Degree of a customer X about V_y

This study makes "Interest Map" by setting the $RISD_X$ on the axis of ordinate and $RIAD_X$ on the axis of abscissas. This study does a visualization of each customer's interest tendency using the Interest Map. The Fig.2 is the Interest Map.

We analyzed positions of commodities in this Interest Map and then This Interest Map is composed of 4 parts as follows:

- Usual Area

  Commodities included in this area are high degree of the $RISD_X$ and $RIAD_X$. Therefore, we think that these commodities are high frequently chosen by each customer in their shopping.

- Unique Area

  Commodities included in this area are low degree of the $RISD_X$ and high degree of $RIAD_X$. Therefore, we think that these commodities are a few but are most important for a customer. For example, they are commodities for the customer's new interest or for a special.

- Basic Area

  Commodities included in this area are high degree of the $RISD_X$ and low degree of $RIAD_X$. Therefore, we think that these commodities are many but aren't important for a customer. For example, they are probably a basic commodity for work or life.

- Unusual Area

  Commodities included in this area are low degree of the $RISD_X$ and low degree of $RIAD_X$. Therefore, we think that these commodities are a few and aren't important for a customer. For example, they are probably a present from someone or they were in fashion at one time.

By this classification of the interest map, we can utilize this interest map to make a recommendation of new commodities that are not known by a customer but are probably liked. For example, if we want to make a recommendation for usual commodities known and liked by a customer, we select a commodity mapped on a normal area in the interest map. If we want to make a recommendation for new commodities, which are not known but are probably liked by a customer, we select a commodity mapped on a unique area in the interest map.

3. Experiment

In this study, we had an experiment to make a model of seven customers about clothes. This study explained clothes by 10 attributes. An experiment system is built by Java 5.0, PostgreSQL and Jakarta Tomcat 5.5 on Mac OS X.

We showed 94 clothes, which are randomly selected 949 clothes in each customer, to each customer through a web site. We obtained evaluation data on interest/uninterest in these sample clothes. We define the 94 clothes as existing clothes of each customer in this study.

And then, we calculate the ISD_X, IAD_X, $RISD_X$ and $RIAD_X$ from obtained individual data and maps commodities based on $RISD_X$ and $RIAD_X$ of each commodity using the experiment system.

```
1.    gender=Man, neck=Henley
2.    color=Deep-Blue, neck=Henley
3.    sleave=Half, neck=Henley
4.    sleave=Half, color=Deep-Blue
5.    color=Deep-Green, category=Moss_Stitch_Polo-Shirts
6.    color=Black, category=Swetershirts_with_Hood
7.    color=Black, neck=Hood
8.    color=Black, waist=Pocket-Waist
9.    color=Deep-Blue, texture=Point
10.   color=Dark-Brown, neck=Henley
11.   sleave=Half, color=Dark-Brown
12.   color=Deep-Blue, category=Shirts
13.   color=Deep-Blue, neck=Collar
…
```

Fig.3 the abstract about S_{Xy} of Customer X

```
ISD_X31 = 0.010526315789473684
ISD_X30 = 0.010526315789473684
ISD_X21 = 0.021052631578947368
ISD_X7 = 0.021052631578947368
ISD_X6 = 0.021052631578947368
…
```

Fig.4 the abstract about ISD_X of Customer X

```
1.    None-Waist=0.02842064
2.    Cut_and_Sewn=0.02533348
3.    Crew=0.0233719
4.    Uniformity=0.020333764
5.    Cutton=0.017107809
6.    Collar=0.012774351
7.    Black=0.012547591
8.    Long=0.012475539
9.    None-Chest=0.011599176
10.   Sweter=0.011579025
11.   Shirts=0.011296534
12.   Dark-Brown=0.011208922
13.   V=0.010783055
14.   Deep-Blue=0.009484382
15.   Deep-Green=0.007527483
16.   Hood=0.007370452
17.   Man=0.007106672
18.   Point=0.006765144
19.   Swetershirts_with_Hood=0.006756248
20.   Blue=0.005132596
21.   Henley=0.00473684
22.   Pocket-Waist=0.004621309
23.   Wool=0.004311121
24.   Pink=0.004277205
25.   Moss_Stitch_Polo-Shirts=0.0042362
26.   Mark=0.003382542
27.   Half=0.001598202
```

Fig.5 the abstract about IAD_X of Customer X

Fig.3 is an abstract about S_{Xy} of customer X. In a case of Customer X, S_{Xy} about color and slave are extracted from his evaluation about interest/uninterest in sample clothes.Fig.4 is an abstract about ISD_X of customer X. We recognize that the ISD_{X7}, which is the interest set of S_{X7}, is more important than ISD_{X21} for customer X.

Fig.5 is an abstract about IAD_X of customer X. Customer X probably likes clothes which do not have an accent in a breast or which are a type of a cut and sewn because IAD_X about attributes of None-Waist and about Cut-and-Sewn is higher than others. Conversely, Customer X probably doesn't like clothes which have a mark in a breast or which are a half type.

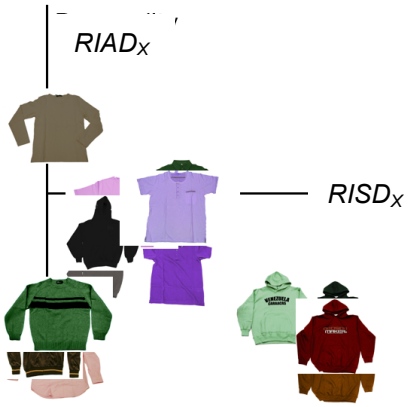

Fig.6 An Example of Interest Map of Customer *X* about Sample Clothes

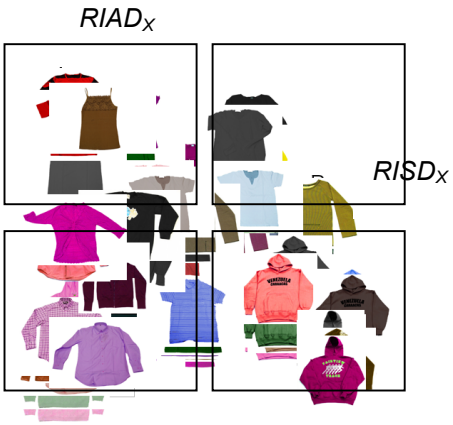

Fig.6 An Example of Interest Map of Customer *X* about Unknown Clothes

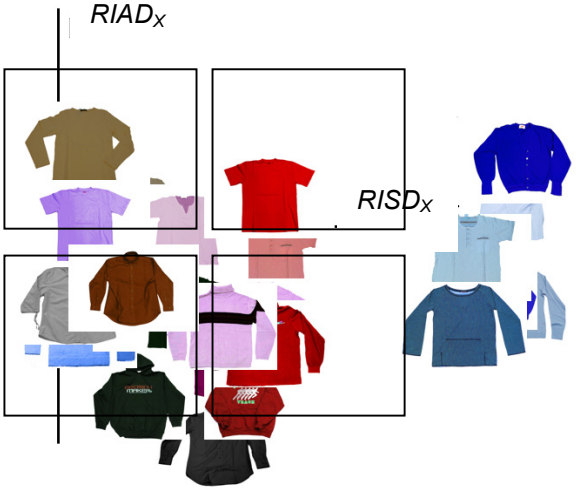

Fig.6 An Example of Interest Map of Customer *Y* about Sample Clothes

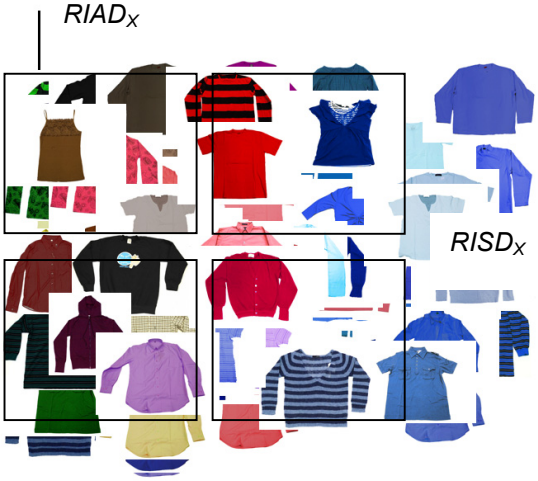

Fig.6 An Example of Interest Map of Customer *Y* about Unknown Clothes

Fig.5 and Fig.6 show an Interest Map of customer *X* and customer *Y* about sample clothes, respectably.

The Interest Map of customer *X* has many cut and sewn in normal area, but the Interest Map customer *Y* has many clothes of blue in normal area. As the result, we can detect that each customer has different important attribute by using the Interest Map.

Fig.7 and Fig.8 show an Interest Map of customer *X* and customer *Y* about unknown clothes, respectably. Unknown clothes are mapped at a same place as clothes with the same attribute and its value. Therefore, this system can predict customers' interest in unknown commodities by calculating the ISD_X, IAD_X, $RISD_X$ and $RIAD_X$.

4. Conclusions and Future works

This study proposed the ISD_X, IAD_X, $RISD_X$ and $RIAD_X$ as a modeling method for detecting an individual relation between customer's interests in commodities and attributes of commodities. This study explained the system to recommend unknown commodities for a customer using "Interest Map".

In future work, we try to improve ISD_X, IAD_X, $RISD_X$ and $RIAD_X$ and a recommendation method using "Interest Map".

Reference

[1] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning About Data",Kluwer Academic Publishers, (1991)

[2] N. Shan and W. Ziarko, "Data-based acquisition and incremental modification of classification rules", Computational Intelligence, Vol.11, pp.357-370, (1995)