Journal of Southeast University (English Edition) Mar. 2002 Vol.18 No.1 ISSN 1003—7985

Web Mining Model Based on Rough Set Theory

Wu Bing* Zhao Lindu

(School of Economics and Management, Southeast University, Nanjing 210096, China)

Abstract: Due to a great deal of valuable information contained in the Web log file, the result of Web mining can be used to enhance the decision-making for electronic commerce (EC) operation and management. Because of ambiguous and abundance of the Web log file, the least decision-making model based on rough set theory was presented for Web mining. And an example was given to explain the model. The model can predigest the decision-making table, so that the least solution of the table can be acquired. According to the least solution, the corresponding decision for individual service can be made in sequence. Web mining based on rough set theory is also currently the original and particular method.

Key words: Web mining, rough sets, electronic commerce, knowledge reasoning, Web log

Electronic commerce (EC) is a kind of service activity. By using EC, all kinds of commercial activity, business activity, financial activity and so on can be put up and completed[1]. Although abundant data in point to customers can be recorded in the database of EC web station, the deviation of data and knowledge came into being as a result of the absence of the approach to mine the connotative knowledge and to predicate the tendency. Newer and more effective methods are needed for data mining so that the potential data can be deplored. Just under such requirement, data mining was brought out and it developed rapidly.

Data mining is a crossed subject, touching upon such subjects as machine studying, model recognizing, statistics, database, on-line analyzing, fuzzy logic, artificial neural network, uncertainty reasoning and data visualization, etc[2]. The core conception of data mining is rote learning in the field of artificial intelligence.

Professor Z.Pawlak in Poland brought forward rough set theory in 1982. The theory is a kind of intelligent decision-making tool. So the imprecise, different and incomplete information can be analyzed by rough set theory[3]. And through data analyzing and reasoning, rough set theory can discover connotative knowledge and even prompt potential regulations. Therefore rough set theory has become the practicable technology for Web mining[4,5].

In general, Web mining tasks can be classified into three categories (as shown in Fig.1). Useful information from relevant files on the artificial structured Web can be obtained by content mining. And useful knowledge from artificial links can be got from structural mining. While usage mining can discover user access patterns of Web pages[6].

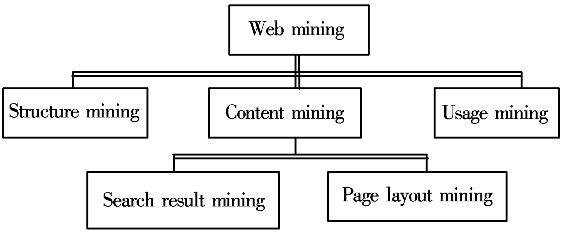

Fig.1 Varieties of mining

Based on the basic concepts of rough set theory, the least decision-making model will be prompted in section 2.4, so that profitable information in EC can be acquired from the Web log and linked structure. Thus individual service can be provided for customers.

1 Basic Concepts of Rough Set Theory

Rough set theory is based on the ability to classify the observed and measured data. The core of the theory is upper and lower approximations.

1.1 Knowledge denotation system and decision-making system

Faced net condition, the base of decision-making and the source of knowledge are the knowledge denotation system and decision-making system. Based on Refs.[7,8], the following definitions can be given.

Definition 1 $S = (U, A, \{Va\}, a)$ is the knowledge denotation system, where U is a no empty finite set, named discussed field. A is a no empty finite set

Received 2001-08-24.

* Born in 1974, female, graduate.

too, named property set. Va is the value field of the property of $a \subseteq A$. $U \rightarrow Va$ is a mapping relation, which makes any element of discussed field U have the exclusive value when getting property a from Va. If A is composed by condition attribute set C and conclusion attribute set D, meanwhile C and D satisfy $C \cup D = A, C \cap D = \varnothing$, then S is a decision-making system.

To show simply, $(U, C \cup \{d\})$ can be used to express decision-making system.

Definition 2 For knowledge denotation system $S = (U, A, \{Va\}, a)$, suppose $R \in A, X \in U, \mathrm{POS}_R(X) = R_{-} X, \mathrm{NEG}_R(X) = U - R_{-} X$ and $\mathrm{BN}_R(X) = R_{-} X - R^{-} X$ are respectively called positive fields, negative fields and border of R below X.

1.2 Core and reduced knowledge

Due to the increasing of the capacity of information, knowledge rooted in information increased accordingly, the acquired knowledge needs to be refined. The aim of refining knowledge is to construct the core of knowledge according to the degree of knowledge dependence and to wipe out redundant knowledge under the condition of insuring the core of knowledge unchanged.

Definition 3 $U_B(x, X) = \mathrm{card}([x]_B \cap X)/\mathrm{card}([x]_B)$ is the reliance degree of element x to set X. where, card denotes the base of gather.

Definition 4 Given R is an equivalent relation family and $r \in R$, when $\mathrm{ind}(R) = \mathrm{ind}(R - \{r\})$, r is omissible for R, or else r is not omissible. All sets of no omissible relation in R are called the core of R, noted as $\mathrm{CORE}(R)$. Thereinto, $\mathrm{CORE}(R) = \cap \mathrm{RED}(P)$, $\mathrm{RED}(P)$ is all the reduced family of P.

2 Analysis of Web Mining Model

The potential of the technique for Web mining lies in the newest arithmetic applied in data mining, so that the Web log, external data of customers, sales and products on the WWW servers can be analyzed[9,10].

2.1 Functions of Web mining

The change of economic model not only adds the mode of virtual trade on the net based on the traditional entity shop, but also changes the mutual act model between the customer and the trader. With the increasing number of the customer on the net, the focus that people pay attention to is changing all the time. The challenge, which the trader on the net has to affront to, has ranged from how to introduce product effectively to how to find out the taste of customers. Faced EC, Web mining technology has three benefits.

1) Comprehend customers' behavior

- Optimize the EC model of fare in virtue of dynamic behavior of visitors;
- Capture the fancy model of customers;
- Mine the model of purchasing of customers and the model of browsing of visitors.

2) Judge the efficiency of Web station

- Modify, design the structure and the appearance of the Web station no longer depending on the qualitative direction, but on the information of visitors;
- Supply individual service aimed at different customers.

3) Evaluate the success or the loss of the mode of EC

- Evaluate the investment return rate of advertising;
- Get the reliant feedback information of the marketing.

To carry out the idea that customers are the center, the application of Web mining technology is useful, so that the enterprise can hold out the ability of strong competition in the age of EC.

2.2 System frame of Web mining

Implementing of Web mining technology needs a basic frame; the frame should not only describe the data flow of the system but also describe the basic structure of the system (as shown in Fig.2).

2.3 Formation of Web log

For Web mining, the formation of Web log must be known. The formation of Web log is shown in Tab.1.

Tab.1 Formation of Web log

Contents	Descriptions
Date	Date, time, and time zone of requesting for page layout
Client IP	IP or DNS entrance of long-distance host computer
User name	User name of telnet
Bytes	Byte of transfer (send or receive)
Servers	Server, address of IP and port
Request	URL query
Status	State marks returned to http
Service name	Requested service name
Time of cost	Time of finishing browsing
Edition of agreement	Edition of agreement for transferring
The agent of users	Suppliers of service
Cookie	ID of Cookie
Consult page	Previous page

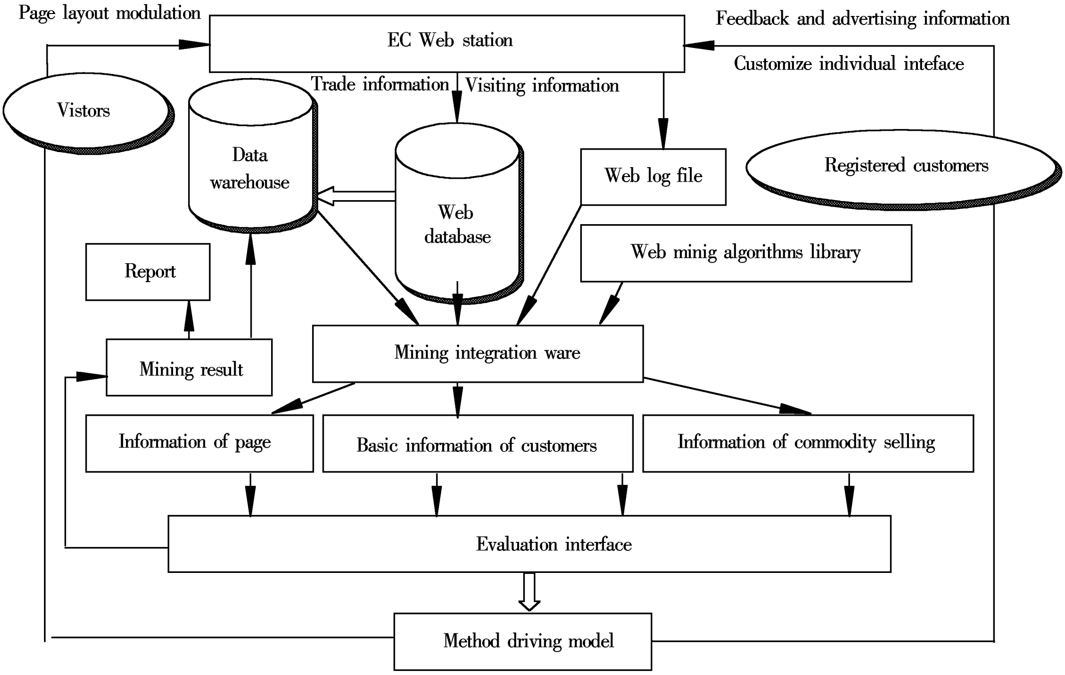

Fig.2 System frame of Web mining

2.4 The least decision-making model for Web mining

Generally, the decision-making table acquired from the data of Web log is complex. To draw required information directly from the table is difficult. The least decision-making model based on rough set theory for Web mining will be provided for simplifying the decision-making table.

The steps of the least decision-making model based on rough set theory are given below:

1) To decide the set of the attribute, and to choose the value range for the attribute respectively. So $(U, C \cup \{d\})$ is obtained.

2) If the result of the table is incompatible, the table should be decomposed to two sub tables.

3) To compute the core value, the duplicate rows and columns should be eliminated.

4) The answer to the question can be optimized in terms of different requirements.

The steps above mainly aimed at mining compatible decision-making table. When mining incompatible decision-making table, the steps are not suitable.

3 An Example

3.1 Description

To simply analyze the whole decision-making table for the frequency of visiting some pages, so that reasonable decision for page layer out can be made[10]. The example was based on the Web log of an EC web station. One row of the Web log is shown as follows:

10.1.34.164 -- [14/Jan/2001:19:12:16 + 0800] “GET/pls/admin-/gatewaty.htm HTTP/1.0” 302 290

3.2 Solution

The decision-making model above will be used to simplify the regulation.

In EC environment, if EC traders want to enhance the decision-making for EC operation and management, they should analyze the browse trace of visitors. In order to analyze easily, visitors’ Client IP set {‘61.133.169.48’, ‘10.1.34.164’, ‘127.0.0.1’, ‘131.188.23.138’, ‘202.111.46.135’, ‘203.93.36.5’, and ‘210.12.79.30’} is delegated by U set {1,2,3,4,5,6,7}. Page set {‘teshucaigouguanli1.htm’, ‘synopsis.htm’, ‘gatewaty.htm’, ‘shihua-top.htm’} is represented by attribute set $\{a, b, c, d\}$. The frequency of visiting page layout $\{a, b, c, d\}$ is described as 0,1,2. where 0 represents the frequency between 0 and 100, 1 represents the frequency between 100 and 200, and 2 represents the frequency between 200 and ∞. Based on the information offered by the log file, according to their Client IP, the sort of decisions for visitors is described as e set {0,1,2} (as shown in Tab.2). 7 visitors are classified into 3 groups, namely group 1: {1,2}, group 2: {3,4} and group 3: {5,6,7}.

Tab.2 One decision table

U	a	b	c	d	e
1	1	0	0	1	1
2	1	0	0	0	1
3	0	0	0	0	0
4	1	1	0	1	0
5	1	1	0	2	2
6	2	1	0	2	2
7	2	2	2	2	2

Firstly, let $C = \{a, b, c, d\}$ and $D = \{e\}$, then $\mathrm{CF} = \mathrm{card}(C \cap D)/\mathrm{card}(C) = 1$, namely, the result of the table is consistent.

Secondly, after simplified the attribute, condition attribute c is omitted.

For decision-making regulation 1:

$F = \{[1]_a, [1]_b, [1]_d\} = \{\{1,2,4,5\}, \{1,2,3\}, \{1,4\}\}, a(1) = 1, b(1) = 0, d(1) = 1, [1](a, b, d) = [1]_a \cap [1]_b \cap [1]_d = \{1,2,4,5\} \cap \{1,2,3\} \cap \{1,4\} = \{1\}$

To work out the omitted category, one category should be eliminated at a time, then it should be checked whether it meets of other categories still containing in the decision-making category $[1]_e = \{1, 2\}$, namely

$$[1]_b \cap [1]_d = \{1,2,3\} \cap \{1,4\} = \{1\} \tag{1}$$

$$[1]_a \cap [1]_d = \{1,2,4,5\} \cap \{1,4\} = \{1,4\} \tag{2}$$

$$[1]_a \cap [1]_b = \{1,2,4,5\} \cap \{1,2,3\} = \{1,2\} \tag{3}$$

The core value of decision-making regulation 1 is $b(1) = 0$. The cores of other regulations are shown in Tab.3.

Tab.3 Cores of other regulations

U	a	b	d	e
1	—	0	—	1
2	1	—	—	1
3	0	—	—	0
4	—	1	1	0
5	—	—	2	2
6	—	—	—	2
7	—	—	—	2

To simplify the count family F, all filial family $\cap\ \theta \in F$, and $\cap\ \theta \in [1]_e = \{1,2\}$ must be figured out. Three filial family of F are described as formulae (1) - (3), thereinto, formulae (1) and (3) have been predigested for F. Thus, $b(1) = 0, d(1) = 1$ and $b(1) = 0, a(1) = 1$. The values of attribute a and attribute b, or the values of attribute b and attribute d are characters of decision-making regulation 1. The predigested attribute value form of every decision-making regulation in Tab.2 is shown in Tab.4.

Tab.4 Predigested attribute value of each decision-making regulation

U	a	b	d	e
1	1	0	×	1
$1'$	×	0	1	1
2	1	0	×	1
$2'$	1	×	0	1
3	0	×	×	0
4	×	1	1	0
5	×	×	2	2
6	×	×	2	2
$6'$	2	×	×	2
7	×	×	2	2
$7'$	×	2	×	2
$7''$	2	×	×	2

Obviously, decision-making regulations 1,2,6 have two predigested forms, while decision-making regulations 3,4,5 have only one predigested form, and decision-making regulation 7 has three predigested forms. In this way, the problem has 24 solutions, two of which are respectively shown in Tab.5 and Tab.6.

Tab.5 Predigested table of one regulation

U	a	b	d	e
1	1	0	×	1
2	1	×	0	1
3	0	×	×	0
4	×	1	1	0
5	×	×	2	2
6	×	2	×	2
7	2	×	×	2

Tab.6 Predigested table of another decision-making regulation

U	a	b	d	e
1	1	0	×	1
2	1	0	×	1
3	0	×	×	0
4	×	1	1	0
5	×	×	2	2
6	×	×	2	2
7	×	×	2	2

The least answer of the decision-making table for knowledge denotation system is not exclusive. So decision-makers should make choices by their experience and tropism. For example, if this EC trader thinks customers of the group 1 are important, he should make necessary decisions based on Tab.7. Namely, for customer 1, page layout a will be put on the first page, then page layout b and page layout d, while page layout c will be hidden.

Tab.7 A least solution of decision-making regulation

U	a	b	d	e
1	1	0	×	1
2	1	0	×	1
3	0	×	×	0
4	×	1	1	0

4 Conclusion

To support EC management and operation, the least decision-making model for Web mining based on rough set theory is constructed in this paper. By this model, the knowledge on formed decision-making table is predigested to educe the predigested decision-making table, finally the least answer is found. According to different individual standpoint, the answer is not sole.

References

[1] Zhao Lindu. *Theory and practice of electronic commerce*[M]. Beijing: People's Post & Telecommunications Publishing House,2001.20 - 24.(in Chinese)

[2] Chan Chien-Chung. A rough set approach to attribute generalization in data mining[J]. *Journal of Information Sciences*, 1998(107):169 - 176.

[3] Griffin G, Chen Z. Rough set extension of Tcl for data mining [J]. *Knowledge-Based System*,1998(11):249 - 253.

[4] Wong S K M. Rough sets: probabilistic versus deterministic approach[J]. *Int J Man-Machine studies*,1998(29):81 - 95.

[5] Li Yongmin, Zhu Shanjun, Cheng Xianghui, et al. Data mining model based on rough set theory[J]. *Journal of Tsinghua University*,1999,**39**(1):110 - 113.(in Chinese)

[6] Han Jiawei, Kamber Micheline. *Data Mining Concepts and Techniques*[M]. Kaufmann Publishers,2001.435 - 443.

[7] Walczak B, Massart D L. Rough set theory[J]. *Chemo metrics and Intelligent Laboratory Systems*,1999(47):1 - 16.

[8] Zeng Huanglin. Rough set theory and application[J]. *Journal of Sichuan institute of light industry and chemical technology*, 1996,**9**(2):1 - 7.(in Chinese)

[9] Xie Danxia, Li Xiaodong. Data mining technology applying on Web and tool designing. *Computer Application*,2001,**21**(2): 42 - 44.(in Chinese).

[10] Zhou Xianchun, Xie Zhong, Zhou Yanhui. EC and Web data mining[J]. *Computer Application*,2001,**21**(5):21 - 23.(in Chinese)

基于粗糙集理论的 Web 挖掘模型

吴 冰　赵林度

(东南大学经济管理学院,南京 210096)

摘　要　在电子商务网站的 Web 日志中,蕴含着大量有价值的信息.利用 Web 挖掘技术能够有效获取这些信息,这将有助于提高电子商务运营管理的经营决策.在 Web 挖掘研究过程中,结合 Web 日志具有的数据量大、不确定等特点,提出了一种基于粗集理论的最小化决策模型.运用这一模型,通过对决策表进行知识简化,可以导出简化决策表,最后获得最小解.电子商务系统的决策人员就可以依据得到的最小解,为提供个性化服务进行决策.应用基于粗集理论的数据挖掘方法,对 Web 日志进行挖掘,已经成为当前研究的热点问题.

关键词　Web 挖掘,粗糙集,知识推理,电子商务,Web 日志

中图分类号　TP311.131