Soft Maximal Association Rule for Web User Mining

Iwan Tri Riyadi Yanto, Arif Rahman
Department of Information system
Ahmad Dahlan University
Yogyakarta, Indonesia
{yanto.itr,arif.rahman}@is.uad.ac.id

Youes Saaadi
Departement of Information System
University of Malaya
Kualalunpur, Malaysia
younessaadi@gmail.com

***Abstract*— Association rule mining of the web user transaction is one of important techniques for extracting information from web data, including its content, link, and user information using data mining tool. This technique finds a pattern and causal relation between items on given databases. The Maximal Association Rule is a data mining tools to determine the association rule the rough set theory based. Accordingly, the rough set can be defined in a form of soft set. This paper presents an implementation of Soft Maximal Association Rule which is the soft set theory based for web mining. The experiment shows that the computation of the proposed technique outperforms comparing to the baseline technique.**

Keywords—web user transaction; association rule; soft maximal association rule; soft set.

I. INTRODUCTION

Web Information Gathering (WIG) is a mechanism consists of finding relevant information on the Web to meet information needs of their user. The "overload" problem is considered as a popular issue related to the effectiveness of WIG [1]. This problem is produced when the number of useful web information is large and included in the users requests [2].

Web mining is a technique in data mining to knowing and extracting knowledge on the website automatically [3]. Cooley, Mobasher, and Srivastava are more highlighting for the importance of web users behavior [4]. Web mining is divided in some steps, those are finding information resource, selecting information, finding knowledge or pattern, and analyzing the pattern founded [5].

Web mining technique can be generally defined as a method for extracting the information from web data consisting of content, link, and user information with data mining tool. Between web data, user clicking sequence such as data usage can describe user navigation pattern and identify usage load. After the data has been characterized effectively, it can be giving benefit the next web application such as facilitating and increasing the quality of web service for both web supplier and user [6]–[8].

Association rule mining is one of some important techniques in data mining. This technique is used for knowing pattern, causal relation between the items of a set from given database. The maximal association rule is introduced by Feldman et all [9], [10] considered as an association rule mining modification. Guan, Bell, Liu, [11] and Anderson, McClean, S.[12] introduced maximal association rules method via the rough set theory [13]–[15]. It is known that a rough set can be represented in a soft set [16]. Therefore, we propose further recitation about how to apply soft set to solve the proble of association rule in web mining.

II. SOFT SET THEORY

Lets U refers to the universe that is not empty, E is a parameter set which is clearing the object in U, the power set of U is (P, U) and $A \subseteq E$.

Definition 1. (look at [16]). The pair (F, A) is called a soft set in U, where F is the mapping that is presented in (1):

$$F: A \rightarrow P(U) \tag{1}$$

From another point of view, a soft set (F, A) in U is parameter subset of U. $\alpha \in A, f(\alpha)$ can be reputed as a set of element α of soft set $F(A)$ or approximation element α of soft set(F, A)[16].

Example 2.1. : as in [16],Given universe $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ is a set of house candidates we want to buy and parameter set $E = \{e_1, e_2, e_3, e_4, e_5\}$ is a set of house conditions contain of "beautiful," "modern," "cheap," "many reparations," and "little reparations". Suppose G is the mapping of parameter E in a group of subsets which belongs to the universe U. It can be denoted as below:

$G(e_1) = \{u_1, u_2, u_3, u_4, u_5, u_6\}$,

$G(e_2) = \{u_1, u_2, u_6\}$,

$G(e_3) = \{u_1, u_5\}$,

$G(e_4) = \{u_1, u_4, u_5\}$,

$G(e_5) = \{u_3, u_6\}$.

Soft set (G, E), describes the condition of the house we want to buy. According to the data, soft set (F, E) can be written as:

$$(G, E) = \begin{cases} beautiful = \{u_1, u_2, u_3, u_4, u_5, u_6\} \\ modern = \{u_1, u_2, u_6\} \\ cheap = \{u_1, u_5\} \\ many\ reparations = \{u_1, u_4, u_5\} \\ little\ reparations = \{u_3, u_6\} \end{cases}$$

III. REPRESENTING WEB USER TRANSACTION IN SOFT SET

The connection inter a given soft set and a information system having binary values can be represented in propostion 1.

Proposition 1. If (G,E) denotes a soft set defined in universe U. (G,E) constitutes a binary-valued information system denoted by $S = \left(U, A, V_{\{0,1\}}, g\right)$

Proof. Let (G,E) represents a soft set in the universe U. It is possible to have the mapping $G = \{g_1, g_2, \dots, g_n\}$ where:

$$g_1 : U \to V_i \tag{2}$$

And

$$g_i(x) = \begin{cases} 1, x \in G(e_1) \\ 0, x \notin G(e) \end{cases}, 1 \le i \le |A| \tag{3}$$

Hence, if $A = E, V = \bigcup_{e_1 \in A} V_E$, where $V_{e_1} = \{0,1\}$. The pair of (G,E) can be presented by $S = (U, A, V_{\{0,1\}}, g)$ a Boolean-valued information system.

Proposition 1 illustrate, a particular binary-valued information system can be denoted as a soft set. It is possible to establish a direct connection inter the pair (G,E) over U and $S = \left(U, A, V_{\{0,1\}}, g\right)$.

Example 3.1. To illustrate Proposition 1, let we consider from the example 2.1, we are able to representing the soft set as table or information system having binary values as in Table II.

Table 1.A Boolean-tabular example 2.1.

U/E	e_1	e_2	e_3	e_4	e_5
u_1	0	1	0	1	1
u_2	1	0	0	0	0
u_3	0	1	0	1	0
u_4	1	0	0	0	0
u_5	0	0	0	1	0
u_6	0	0	0	0	0

Web access transaction is defined as a pair of finite groups: the transaction affected by the user and the web links [17]. The transactions affected by the user is denoted by the set *U,* which is represented by a sequence of elements belong to *m* users. In contrast *A* is represented by a group of different *n* clicks affected by a specific user. Hence $U = \{t_1, t_2, \dots, t_m\}$ and $A = \{hl_1, hl_2, \dots, hl_n\}$, where for every $t_i \in T \subseteq U$ is a subset of U where t_i is non empty. The provisional arrangement of the clicks for a specific transaction has been taken into consideration. The transaction is denoted by a vector $t = [u_1^t, u_2^t, \dots, u_n^t]$, where $u_i^t = 1$ if $hl_i \in t$, and $u_i^t = 0$ if otherwise.

The following example illustrates how to represent web user transactions using soft set theory.

Example 3.2. Given data of user transactions visiting web there are four ($|U| = 4$)) user visit five ($|E| = 5$) hyperlinks on the web.

$t_1 = \{hl_1, hl_2\}$,

$t_2 = \{hl_2, hl_3, hl_4\}$,

$t_3 = (hl_1, hl_3, hl_5)$,

$t_4 = \{hl_2, hl_3, hl_5\}$.

The web transaction of users can be described as in Table II.

Table 2. Data transactions

U/A	hl_1	hl_2	hl_3	hl_4	hl_5
u_1	1	1	0	0	0
u_2	0	1	1	1	0
u_3	1	0	1	0	1
u_4	0	1	1	0	1

From Table 2, we can easily to understand that is user clicks the hyperlinkhl_1, then the user will click on hyperlinkhl_2, and etc. The Table 2 above can be represented into a soft set as follow

$$(F,E) = \begin{cases} hl_1 = \{u_1, u_3\} \\ hl_2 = \{u_1, u_2, u_4\} \\ hl_3 = \{u_2, u_3, u_4\} \\ hl_4 = \{u_3, u_4\} \end{cases}$$

IV. SOFT MAXIMAL ASSOCIATION RULE FOR WEB MINING

Consider to the web user transaction, it can be described as a soft set theory. Therefore, the soft maximal association can be used to find the rule of web user transaction.

Definition 4.1. Suppose soft set (G,E) over U, $u \in U$, where E represents a collection of hyperlinks (hl) visited by user. The set of co-occurrence in the transaction u are denoted as (4) :

$$Co(u) = \{hl \in E : g(u, hl) = 1\} \tag{4}$$

Definition 4.2. Soft Maximal association supports a parameter D to be noted as $SMsup(D)$, defined as U transaction number that maximal support D, as in (5):

$$SMsup(U) = \{u : D = Co(u) \cap E\} \tag{5}$$

Definition 4.3. Soft Maximal association rule C and D, where two maximal hyperlink sets $C, D \subseteq E$ and $C \cap D = \emptyset$ are the implication of $C \underset{\Rightarrow}{max} D$. Item set C is called maximal antecedents and D is the consequent. Maximal association support from maximal association rule $C \Rightarrow D$, notated as $Msup\left(X \underset{\Rightarrow}{max} Y\right)$, defined as (6) :

$$Msup\left(C \underset{\Rightarrow}{max} D\right) = Msup(C \cup D) = \{a : C \cup D = Co(u) \cap E\} \tag{6}$$

Definition 4.4. Soft Maximal confidence from maximal association rule $\underset{\Rightarrow}{CmaxD}$ that is notated as $Mconf\left(\underset{\Rightarrow}{CmaxD}\right)$, is defined as (7)

$$\begin{gathered}\{u: C = Co(u) \cap E\} \\ \{u: C \cup D = Co(u) \cap E\} \\ Mconf\left(\underset{\Rightarrow}{CmaxD}\right) = \frac{Msup(C \cup D)}{Msup(C)}\end{gathered} \tag{7}$$

V. Computation

In concern to analyze proposed method performance, the experiment is conducted on two UCI benchmark datasets i.e: Microsoft and MNBC web user transaction. Then results in form of the support, confident and Response time are compared with baseline method. The algorithms are implemented PhP software with the Windows 10 operating system. The data sets are executed sequentially either Intel processor GHz i5. The amount of main memory is 4G.

A. Microsoft data set

The Microsoft log data is recorded from anonymous user www.microsoft.com which is selected 38000 randomly. The data constitutes all the areas of the website (Vroots) visited by users in February 1998. The data comes from one week containing 294 Vroots. Identification of user is only by a sequential number. The log file contains no personally identifiable information.

Combination on-air-misc(47) --> Sup(misc) => 88	Confidence 72.81%
Combination frontpage health(49) --> Sup(health) => 88	Confidence 57.35%
Combination frontpage living(36) --> Sup(living) => 71	Confidence 50.7%
Combination news-opinion(21) --> Sup(opinion) => 42	Confidence 50%
Combination news-travel(9) --> Sup(travel) => 14	Confidence 64.29%
Combination frontpage-travel(8) --> Sup(travel) => 14	Confidence 57.14%
Combination living-travel(7) --> Sup(travel) => 14	Confidence 50%
Combination news-bbs(4) --> Sup(bbs) => 6	Confidence 66.67%
Combination frontpage bbs(4) --> Sup(bbs) => 6	Confidence 66.67%
Combination summary bbs(3) --> Sup(bbs) => 6	Confidence 50%

Fig. 1 The Microsoft web user rule using Maximal Association Rule

With news-msn-news(94) <--> M Sup(news,) => 126	Confidence 74.6%
With frontpage-news(72) <--> M Sup(news,) => 120	Confidence 77.14%
With frontpage sports(54) <--> M Sup(frontpage,news,) => 56	Confidence 96.43%
With news-on-air(58) <--> M Sup(on-air,) => 92	Confidence 57.61%
With on-air-misc(47) <--> M Sup(on-air,) => 92	Confidence 51.09%
With on-air-misc(47) <--> M Sup(misc,) => 61	Confidence 77.05%
With frontpage-on-air(47) <--> M Sup(frontpage,news,) => 56	Confidence 83.93%
With frontpage-on-air(47) <--> M Sup(on-air,) => 92	Confidence 51.09%
With news-local(44) <--> M Sup(local,) => 73	Confidence 60.27%
With news-misc(43) <--> M Sup(misc,) => 61	Confidence 70.49%

Fig. 2 The Microsoft web user rule using Soft Maximal Association Rule

B. MSNBC data set

The data set is recorded from mscnbc.com website internet server (IIS) logs and its news related. The data is obtained in 28th September 1999 (Pacific Standard Time), where the datasets sequence represented by a page views of a particular user in one day period. An the order of the events indicates the total requests of a page by a particular user. All requests are registered at the page category level, which is determined by the administrator. The categories defined as follows: "msn-news," "opinion," "news," "local," "on-air," "weather," "FrontPage," "health," "living," "business," "tech," "sports," "summary," "music," "bbs (bulletin board service)," "travel," and "msn-sports". The data do not contain web page demands delivered with a caching mechanism, because were not saved in the server logs.

.

Combination ssafe-ssafesupport(263) --> Sup(support) => 524	Confidence 58.19%
Combination support-ssafesupport(263) --> Sup(support) => 524	Confidence 58.19%
Combination support ssafe(263) --> Sup(support) => 524	Confidence 58.19%
Combination word support(213) --> Sup(word) => 339	Confidence 62.86%
Combination vba vbasic(163) --> Sup(vba) => 165	Confidence 98.79%
Combination powerpoint-support(156) --> Sup(powerpoint) => 260	Confidence 53.79%
Combination msp-mspress(155) --> Sup(msp) => 196	Confidence 79.08%
Combination support-msp(111) --> Sup(msp) => 196	Confidence 56.63%
Combination exceldev-excel(97) --> Sup(excel) => 106	Confidence 91.51%
Combination support-vbe(96) --> Sup(vba) => 165	Confidence 58.18%

Fig. 3 The MSNBC web user rule using Maximal Association Rule

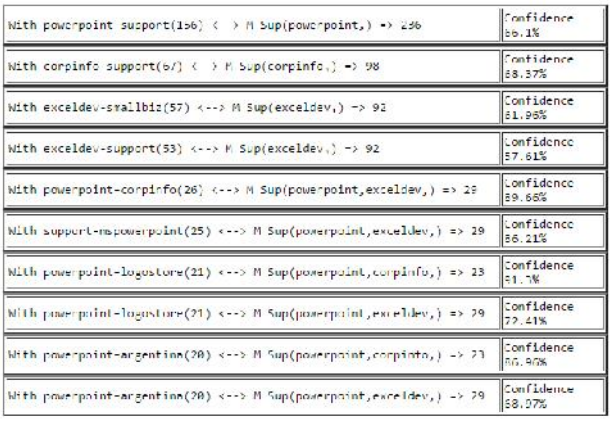

Fig. 4 MSNBC web user rule using Soft Maximal Association Rule

Fig. 1 and 2 show the ten highest support and confidences of the Microsoft web user transaction rule using maximal association rule and soft maximal association rule, respectively. The highest confidence of Microsoft web user transaction rule using maximal association rule is reaching up to 60%, and run into more 90% when using soft maximal association rule. Fig. 3 and 4 show the ten highest support and confidences of the MSNBC web user transaction rule using maximal association rule and soft maximal association rule, respectively. Both of the methods achieve more than 90 % of confidence. The executing time of web association rule using maximal association rule and soft maximal association rule is shown in Fig 5.

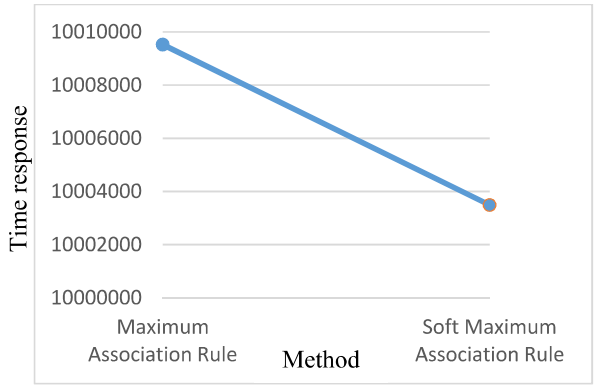

Fig. 5 The comparison of executing time.

VI. Conclusion

In this study, a novel technique for soft maximal association rule for web mining based on soft set theory is presented. Firstly, the representing web user transaction in soft set is explained. Thus, the soft maximal association rule can be applied to find the pattern of web visitor or user. We elaborate the technique trough the benchmark datasets of Microsoft and MSNBC. The computational results show that the soft maximal association rule discover rule with high confidence and achieve faster executing time.

References

[1] Y. Li and N. Zhong, "Web mining model and its applications for information gathering," *Knowledge-Based Syst.*, vol. 17, no. 5–6, pp. 207–217, 2004.

[2] Y. Li and N. Zhong, "Rough Association Rule Mining in Text Documents for Acquiring Web User Information Needs."

[3] O. Etzioni, "The World-Wide Web: Quagmire or Gold Mine?," *Commun. ACM*, vol. 39, no. 11, pp. 65–68, Nov. 1996.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowl. Inf. Syst.*, vol. 1, no. 1, pp. 5–32, 1999.

[5] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[6] A. G. Büchner and M. D. Mulvenna, "Discovering Internet Marketing Intelligence Through Online Analytical Web Usage Mining," *SIGMOD Rec.*, vol. 27, no. 4, pp. 54–61, Dec. 1998.

[7] E. Cohen, B. Krishnamurthy, and J. Rexford, "Improving End-to-end Performance of the Web Using Server Volumes and Proxy Filters," *SIGCOMM Comput. Commun. Rev.*, vol. 28, no. 4, pp. 241–253, Oct. 1998.

[8] T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A Tour Guide for the World Wide Web," in *IN PROCEEDINGS OF THE FIFTEENTH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 1997, pp. 770–775.

[9] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir, "Text mining at the term level," in *Principles of Data Mining and Knowledge Discovery: Second European Symposium, PKDD '98 Nantes, France, September 23--26, 1998 Proceedings*, J. M. Żytkow and M. Quafafou, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 65–73.

[10] A. Amir, Y. Aumann, R. Feldman, and M. Fresko, "Maximal Association Rules: A Tool for Mining Associations in Text," *J. Intell. Inf. Syst.*, vol. 25, no. 3, pp. 333–345, 2005.

[11] J. W. Guan, D. A. Bell, and D. Y. Liu, "The rough set approach to association rule mining," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 529–532.

[12] Y. Bi, T. Anderson, and S. McClean, "A rough set model with ontologies for discovering maximal association rules in document collections," *Knowledge-Based Syst.*, vol. 16, no. 5–6, pp. 243–251, 2003.

[13] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.

[14] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers Norwell, MA, USA, 1992.

[15] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Inf. Sci. (Ny).*, vol. 177, no. 1, pp. 3–27, Jan. 2007.

[16] D. Molodtsov, "Soft set theory—First results," *Comput. Math. with Appl.*,

vol. 37, no. 4–5, pp. 19–31, Feb. 1999.

[17] S. K. De and P. R. Krishna, "Clustering web transactions using rough approximation," *Fuzzy Sets Syst.*, vol. 148, no. 1, pp. 131–138, 2004.