TELKOMNIKA, Vol.15, No.3, September 2017, pp. 1344~1353
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i3.6382

A Soft Set-based Co-occurrence for Clustering Web User Transactions

Edi Sutoyo*[1], Iwan Tri Riyadi Yanto[2], Rd Rohmat Saedudin[3], Tutut Herawan[4]
[1,3]Department of Information Systems, Telkom University, Bandung, West Java, Indonesia, 40257
[2]Department of Information Systems, Ahmad Dahlan University, Yogyakarta, Indonesia, 55161
[4]Department of Information Systems, University of Malaya, Kuala Lumpur, Malaysia, 50603
*Corresponding author, e-mail: edisutoyo@telkomuniversity.ac.id, yanto.itr@is.uad.ac.id, rdrohmat@telkomuniversity.ac.id, tutut@um.edu.my

Abstract

Grouping web transactions into some clusters are essential to gain a better understanding the behavior of the users, which e-commerce companies widely use this grouping process. Therefore, clustering web transaction is important even though it is challenging data mining issue. The problems arise because there is uncertainty when forming clusters. Clustering web user transaction has used the rough set theory for managing uncertainty in the clustering process. However, it suffers from high computational complexity and low cluster purity. In this study, we propose a soft set-based co-occurrence for clustering web user transactions. Unlike rough set approach that uses similarity approach, the novelty of this approach uses a co-occurrence approach of soft set theory. We compare the proposed approach and rough set approaches regarding computational complexity and cluster purity. The result demonstrates better performance and is more effective so that lower computational complexity is achieved with the improvement more than 100% and cluster purity is higher as compared to two previous rough set-based approaches.

Keywords*: clustering, web user transactions, web mining, rough set theory, soft set theory*

1. Introduction

Data mining is the procedure of extracting important information from large databanks. It is a process of performing extraction automatically from a large data bank into knowledge or useful information. And also, it can be regarded as an algorithmic process that takes a sample as input and yield patterns such as classification, association rules, or clustering as an output [1]. One of the utmost advantageous role in the process of data mining is clustering in order to define the groups and to classify the distribution and patterns in the data. This is a process for grouping data into multiple clusters or groups so that data in one cluster has a maximum level of similarity and data among clusters has a minimum similarity [2]. The process that occurs is partitioning a collection of data objects into several subsets that have homogeneous characteristics called clusters. Objects within the cluster have similar characteristics between each other and are different from other clusters. Partitions are not done manually but with a clustering algorithm. Therefore, clustering is very useful and can find unknown groups or groups in the data. There are many approaches used for data clustering, some are able to handle uncertainty, and other improve computational complexity. Many practical and complex clustering problems in the area of engineering, health science, environmental science, and economics are often involving uncertain and vague data.

There are several well-known approaches to handling uncertainty during the clustering process. There are such as a fuzzy theory [3], a rough set theory [4], vague sets [5], and interval mathematics. But all of these theories have their inherent difficulties, as pointed out in [6]. As a result, Molodstov proposed a novel theory to deal with the uncertainty that is called a soft set theory [6]. It uses parameterization concept as its main vehicle, hence it offers wider applications in real problems. At present, many researchers and scholars have contributed to the development of soft set in theory as well as practice, including work of [7]–[16].

Clustering is the process of grouping objects based on information obtained from data explaining the relationships between objects with principles to make the most of the similarity

Received April 23, 2017; Revised July 9, 2017; Accepted July 25, 2017

between members of a group and minimize the similarity between groups. The goal is to find a quality cluster in a decent time. The similarity of objects is usually derived from the proximity of attribute values that describe data objects, whereas data objects are usually represented as a point in a multidimensional space. Clustering can be used to identify areas that are dense, group objects into groups that have homogeneous characteristics or variations of objects, distinguish between one cluster and an another, and find the distribution and interesting patterns between data attributes. In data mining, research focuses on the discovery methods for clusters in large-scale databases effectively and efficiently [17]. The many clustering approaches make it difficult to define universal quality measures. The many clustering approaches make it difficult to determine global quality measures. However, some things to note are the input parameters that do not complicate the user, the cluster of results that can be analyzed, and scalability to the addition of dimension size and record dataset. Clustering is another big issue of the soft set-based data analysis, particularly when it contains large and imprecise data, e.g., clustering of web user transactions. Web user transaction data can generally be obtained from the server log file. The server log is a log file created and managed automatically by a server that serves to record all information from requests generated by users with the right time along with the results of processing. A typical example is the web server log that stores the page request history.

Web clustering is related to the extraction of knowledge from a web page, user session or weblog data into some object groups [18]. The purpose of grouping web transactions into some clusters is to gain a better understanding the behavior of the users or to segment web visitors. One of the guidelines for optimizing the structure and navigation of a website can be seen from the trend of visitor access patterns. To obtain visitor behavior information, analyzing can be conducted on click pattern in clickstream website visitors. This can be obtained by extracting information from the weblog data. Recently, De and Khrisna presented a rough set-based algorithm for clustering web user transactions utilizing the similarity of upper approximations [19]. However, high complexity is still an outstanding issue for finding the similarity of upper approximations used to merger two or more clusters which have the same similarity. To handle this issue, Yanto, *et al.* proposed a soft set-based framework for clustering web user transactions [20]. Further, they proposed the RoCeT method for clustering web transaction by using the notion of similarity class showing how to transactions are allocated in the same cluster [21]. However, their technique still suffers from high computational complexity. Therefore, in this work, we propose a soft set-based approach for clustering web user transactions. To sum up, the main contribution of this work is listed as follows:

a. We propose a soft set approach for clustering web user transactions, which capable of achieving lower computational complexity and also higher clustering purity.
b. Unlike rough set approach that uses similarity approach, the novelty of this approach uses a co-occurrence approach of the soft set.
c. Also, this study presents comparison on theoretical analysis and performance results of the proposed approach, and two rough set-based approaches are presented.

The rest of this paper is systematized as follows. Section 2 recalls the rudimentary notion of rough set and soft set theory. Section 3 describes the proposed soft set-based technique. Section 4 elaborates the experimental and comparison results. As a final point, Section 5 concludes our works.

2. Theoretical Background

In this section, we discuss some rudimentary concepts of rough set theory and also soft set theory.

2.1. Rough Set Theory

A 4-tuple (quadruple) $S=(U,A,V,f)$ is definition of information system, where $U=\{u_1,u_2,u_3,\Lambda,u_{|U|}\}$ is a non-empty finite set of objects, $A=\{a_1,a_2,a_3,\Lambda,a_{|A|}\}$ is a non-empty finite set of attributes, $V=\bigcup_{a\in A}V_a$, V_a is the domain (value set) of attribute a, $f:U\times A\to V$ is an information function such that $f(u,a)\in V_a$, for every $(u,a)\in U\times A$, called information (knowledge) function. The rough set theory was initially developed by Pawlak

[4] for modeling imprecision and granularity in an information systems. Rough set is established on the assumption that with every object of the universe linked to some information (data, knowledge). Objects are marked by information that cannot be distinguished (indiscernible) by means of the information existing on the object. The indiscernibility relation produced in this way is the basic mathematical theory of the rough set theory. The set of all indiscernible objects are called the elementary set, and forms the basic granules (atom) of the knowledge of the universe. If in the information system, there are at least two objects that have the same feature then called indiscernible (indistinguishable).

Definition 2.1. *Two objects* $x, y \in U$ *are considered to be B-indiscernible (*$B \subseteq A$ *in S) if and only if* $f(x,a) = f(y,a)$*, for every* $a \in B$*.*

From Definition 2.1, it is clear that any subset of *A* generates its respective indiscernibility relation. Furthermore, the relation is an equivalence relation, i.e. reflexive, symmetric and transitive. Hence, the relation generates unique partition, let say U/B and the class containing an element *x* of *U* is represented by $[x]_B$. From such partition, we present the ideas of lower and upper approximations of a subset, which are defined as follows.

Definition 2.2. *(See* [4]*.) The B-lower approximation of X symbolized by* $\underline{B}(X)$ *and B-upper approximations symbolized by* $\overline{B}(X)$ *of X, are defined by:*

$$\underline{B}(X) = \{x \in U | [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U | [x]_B \text{ I } X\}.$$

2.2. Soft Set Theory

Let U be a non-empty universe of objects, E is a set of parameters in relation to objects of U, $P(U)$ is the power set of U. The following is the definition of soft set theory.

Definition 3.1. *(See* [6]*.) A pair* (F,E) *is called a soft set over U, where F is a mapping given by* $F: E \to P(U)$

Meaning is to say, a soft set is a parameterized family of subsets of the universe U. For $\varepsilon \in E$, $F(\varepsilon)$ can be considered as the set of ε-elements of the soft set (F,E) or as the set of ε-approximate elements of the soft set.

Example 2.1. Suppose we have a soft set (F,E) describes the "cars being considered" that Mr. Y is about to purchase. Assume that $U = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ and $E = \{e_1, e_2, e_3, e_4, e_5\}$, for instance there are six cars in the universe U and E is car feature (a set of parameters), e_i for $i = 1,2,3,4$ and 5, mean the parameters "expensive", "sport car", "family car", "cheap", and "saving fuel", respectively. Consider the mapping $F: E \to P(U)$ given by "car condition (.)", where (.) is to be filled in by one of the parameters $e \in E$. Assume that $F(e_1) = \{c_3, c_4\}$, $F(e_2) = \{c_1, c_3\}$, $F(e_3) = \{c_2\}$, $F(e_4) = \{c_1, c_2, c_5\}$, and $F(e_5) = \{c_1, c_4, c_5\}$. As from the example above, $F(e_2)$ means a car with "sport car" feature, whose functional value is the set $\{c_1, c_3\}$. Therefore, the soft set (F,E) can be presented as group of approximation as:

$$(F,E) = \left\{ \begin{matrix} e_1 = \{c_3, c_4\}, e_2 = \{c_1, c_3\}, \\ e_3 = \{c_2\}, e_4 = \{c_1, c_2, c_5\}, e_5 = \{c_1, c_4, c_5\} \end{matrix} \right\}.$$

In the following section, an alternative technique for clustering web user transactions employing soft set theory is proposed. This alternative technique is expected to perform better than previous rough set-based techniques in terms of two main concern, i.e. processing time and also cluster purity, respectively. This method will also be proved mathematically to ensure its performance.

3. Proposed Soft Set-based Technique

This section, the soft set-based clustering for web user transactions is proposed. It is based on the fact that user transactions can be represented as a soft set.

3.1. The Proposed Technique

Firstly, the relation of soft set theory and Boolean-valued information system is described as follows.

Proposition 3.1. *If* (F,E) *is a soft set over the universe U, then* (F,E)*is a binary-valued information system* $S=(U,A,V_{\{0,1\}},f)$

Proof. Let (F,E) is soft set over the universe *U*, we define a mapping $F=\{f_1,f_1,\mathrm{K},f_n\}$ where

$$f_1:U\to V_i \text{ and } f_i(x)=\begin{cases}1,x\in F(e_1)\\0,x\notin F(e)\end{cases}, \text{ for } 1\le i\le|A|$$

For that reason, if $A=E,V=\bigcup_{e_1\in A}V_E$, where $V_{e_1}=\{0,1\}$ then a soft set can be seen as a Boolean-valued information system $S=(U,A,V_{\{0,1\}},f)$.

From Proposition 3.1, a binary-valued information system can be easily represented as a soft set. As a result, a one-to-one correspondence between (F,E) over U and $S=(U,A,V_{\{0,1\}},f)$ can be made. To illustrate Proposition 3.1, let we consider Example 2.1. From the soft set in Example 2.1, Boolean-valued information system can be represented.

The representation of web user transactions using soft set theory is illustrated in the Example 3.1 as follows.

Example 3.1. The data of web user transactions are taken from [19] given in Table 1. This table contains four users $(|U|=4)$ and five hyperlinks $(|E|=5)$.

Table 1. Representation Example 2.1 into soft set (F,E)

(U,E)	e_1	e_2	e_3	e_4	e_5
c_1	0	1	0	1	1
c_2	0	0	1	1	0
c_3	1	1	0	0	0
c_4	1	0	0	0	1
c_5	0	0	0	1	1

Table 2. Sample of Web User 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

Table 2 above represents whether the user clicks the hyperlink hl_1or the user clicks on the hyperlink hl_2, and etc. The soft set representation of Table 2 is as follows.

$$(F,E)=\left\{\begin{matrix}hl_1=\{u_1,u_3\},hl_2=\{u_1,u_2,u_4\},\\hl_3=\{u_2,u_3,u_4\},hl_4=\{u_2\},\\hl_5=\{u_3,u_4\}\end{matrix}\right\}$$

From Proposition 3.1, the idea of similarity between two parameters (representing two transactions) *t* and *u* in *U* is proposed in the following definition. Firstly, the idea of occurrence of parameters in soft set theory is defined as follows.

Definition 3.1. *Let (F,E) be a soft set over the universe U representing data of web user transactions and a web user transaction $u \in U$. A parameter co-occurrence set of an object u can be defined as:*

$$cod(u)=\{e \in E: f(u,e)=1\}.$$

Clearly, $cod(u)=\{e \in E: f(e)=1\}$. Definition 3.1 can be illustrated in the Example 3.2 as follows.

Example 3.2. *From soft set (F,E) in Table 2, the parameter occurrence is defined as follows:*

$$cod(u_1)=\{hl_1,hl_2\}$$
$$cod(u_2)=\{hl_2,hl_3,hl_4\}$$
$$cod(u_3)=\{hl_1,hl_3,hl_5\}$$
$$cod(u_4)=\{hl_2,hl_3,hl_5\}.$$

The following definition is from Definition 3.1.

Definition 3.2. *Let (F,E) be a soft set over the universe U representing data of web user transactions and $t,u \in U$ are two user transactions. The similarity between t and u denoted by $sim(t,u)$ is defined as follows:*

$$sim(t,u)=\frac{cod(t) \text{ I } cod(u)}{cod(t) \text{ Y } cod(u)}.$$

From the similarity definition above, it can be concluded that $sim(t,u) \in [0,1]$. The $sim(t,u)=1$, when two transactions *t* and *s* are precisely the same. Meanwhile, $sim(t,u)=0$, if two transactions *t* and *s* does not the same items in common.

The representation of the Definition 3.2 is in the next example.

Example 3.3. *From soft set (F,E) in Table 2 and Example 3.2, the similarity between two user transactions is given as follows:*

$$sim(u_1,u_2)=\frac{cod(u_1) \text{ I } cod(u_2)}{cod(u_1) \text{ Y } cod(u_2)}$$
$$=\frac{\{hl_1,hl_2\} \text{ I } \{hl_2,hl_3,hl_4\}}{\{hl_1,hl_2\} \text{ Y } \{hl_2,hl_3,hl_4\}}$$
$$=\frac{\{hl_2\}}{\{hl_1,hl_2,hl_3,hl_4\}}=0.25$$
$$sim(u_1,u_3)=\frac{cod(u_1) \text{ I } cod(u_3)}{cod(u_1) \text{ Y } cod(u_3)}$$
$$=\frac{\{hl_1,hl_2\} \text{ I } \{hl_1,hl_3,hl_5\}}{\{hl_1,hl_2\} \text{ Y } \{hl_1,hl_3,hl_5\}}$$
$$=\frac{\{hl_1\}}{\{hl_1,hl_2,hl_3,hl_5\}}=0.25$$

and etc.

From Definition 2.1, soft set theory can be referred as a binary relation. Furthermore, from Definition 3.2, we present the notion a binary relation with respect to the similarity between two user transactions.

Definition 3.3. *Let* (F,E) *be a soft set over the universe U representing data of web user transactions and* $t,u \in U$ *are two user transactions. A binary relation R between t and u denoted by* tRu *is defined as follows*

$sim(t,u) \geq \text{th}$

where $\text{th} \in [0,1]$ *is a user pre-defined threshold value.*

This relation *R* in Definition 3.3 is both reflexive and symmetric, but may not a transitive. The representation of the Definition 3.3 is in the next example.

Example 3.4. *From soft set* (F,E) *in Table 2 and Example 3.3, the following binary relations are formed with a given threshold 0.4*

$$u_2 R u_4 \text{ since } sim(u_2,u_4) = \frac{\{hl_2, hl_4\}}{\{hl_2, hl_3, hl_4, hl_5\}} = 0.5 \geq 0.4$$

and

$$u_3 R u_4 \text{ since } sim(u_3,u_4) = \frac{\{hl_3, hl_5\}}{\{hl_1, hl_2, hl_3, hl_5\}} = 0.5 \geq 0.4.$$

From Definition 3.3, a similarity class can be defined as follows.

Definition 3.4. *Let* (F,E) *be a soft set over the universe U representing data of web user transactions and a web user transaction* $u \in U$ *.The similarity class of t, denoted by* $SC(u)$*, is defined as a set of transactions which are similar to t i.e.*

$$SC(t) = \{u \in T : tRu\}.$$

From Definition 3.4, for the given any threshold values, we can also get any similarity classes. An expert can choose a threshold based on their preferable value to get expected similarity classes. The representation of the Definition 3.4 is in the following example.

Example 3.5. *From soft set* (F,E) *in Table 2, we the following similarity classes of each transaction with a given threshold 0.4*

$$SC(u_1) = \{u_1\},\ SC(u_2) = \{u_2, u_4\},\ SC(u_3) = \{u_3, u_4\}, \text{ and } SC(u_4) = \{u_2, u_3, u_4\}.$$

3.2. Correctness Proof

The following definition states that two web user clusters in U to be similar if their union are equal.

Definition 3.5. *Let* (F,E) *be a soft set over the universe U representing data of web user transactions. Two web user clusters* C_i *and* C_j *in* U *, for* $i \neq j$ *are said to be the same if* $C_i = \bigcup SC(u_i)$*, for* $i = 1,2,\Lambda\ ,|U|$.

From similarity classes in Definitions 3.4 and 3.5, we can form a cluster of web user transactions as shown in Proposition 3.2.

Proposition 3.2. *Let* (F,E) *be a soft set over the universe U representing data of web user transactions and* $SC(u_i)$ *be a similarity class of transaction u, for* $i = 1,2,\Lambda\ ,|U|$*. If* $\bigcap SC(u_i) \neq \phi$*, then* $\bigcup SC(u_i) = C_i$.

Proof. We suppose that $\bigcup SC(u_i) \neq C_i$, then from Definition 3.5, we have $C_i \neq C_j$ and further $C_i \cap C_j = \phi$. The consequence, we get the following:

$$\left(\Upsilon C(u_i)\right) \cap \left(\Upsilon C(u_j)\right) = \phi$$
$$\cap\ C(u_i) = \phi$$

This is a contradiction from the hypothesis.

3.3. Algorithm and Its Complexity

The algorithm of the proposed technique is presented in Figure 1.

Algorithm: Soft set technique
Input: Web user transactions data set Output: Web user transactions clusters *Begin* *Step 1. Calculate the degree of the similarity between two object transactions.* *Step 2. Get the similarity class with the given threshold value.* *Step 3. Combine the transaction if there is two of similarities have a non-empty intersection.* *End*

Figure 1. The pseudo-code of the soft set-based web user transactions

From Step 3 in Figure 1, the clusters formed are based on a non-void intersection of the two similarity classes, i.e.

$$C_i = \Upsilon\left\{SC(u_i) \mid \cap\ SC(u_i) \neq \phi\right\}.$$

From the proposed technique, suppose that there are *n* objects in a soft set (F, E) over the universe U representing data of web user transactions. As a result, there are at most *n* similar classes. For getting this done, the technique needed is $\frac{n^2}{2}$ computation for determining the similarity matrix. Since the computation of union of relation similarity matrix to obtain the cluster is $\frac{n^2}{2}$. Accordingly, the computing complexity as a whole is polynomial $O(n^2)$.

Table 3. Complexity Comparison

	De & Krishna [27]	Yanto *et al.* [28], [29]	Proposed Technique
Complexity	$O(2n^2 + 2n)$	$O(2n^2)$	$O(n^2)$

Based on Table 3, the proposed approach achieves lower complexity as compared to others.

4. Results and Discussion

The experimental results of the clustering web user transactions technique are elaborated in this section. To compare the proposed technique and techniques proposed by [19]–[21] are carried out on a 1.86 GHz Intel Core i3-3217U machine with 8 GB memory using Windows 10 operating system. The techniques for clustering web transactions are implemented in MATLAB 8.4 (R2014b). For experiment test, the two UCI Repository of Machine Learning Databases benchmark datasets are used. Those datasets are obtained from [22]. This data only contains the server-side logs because the data from client-side is not recorded. This

experiments use 2000 data transactions, and then those data are split into five different sample size groups i.e.100, 200, 500, 1000, and 2000 data transactions, respectively.

4.1. Cluster Purity

In general terms, web user transactions clustering algorithms are based on criteria to assess the quality of a given class. Especially, they take some parameters as an input, e.g. number of classes. Since the clustering algorithms are an apriori technique, then they need to be assessed in term of the purity of classes [44]. Formally, the purity of classes (clusters) is defined as follows:

$$\text{Purity}(i) = \frac{\sum t_{it_h}}{\sum t_n},$$

where $\sum t_{it_h}$ is the number of data occurring in both the *i-th* cluster under the given threshold and $\sum t_n$ is the number of data in the dataset. Meanwhile, the overall purity of clusters is defined as follows:

$$\text{OverallPurity}(i) = \frac{\sum_{i=1}^{\#\text{of cluster}} \text{Purity}(i)}{\#\text{of cluster}}$$

In accordance with the above equations, the highest value of overall class purity reflects the best clustering result, where perfect clustering results are closing to have a value of 100 %. In the next sub-section, we elaborate the experimental results of our proposed soft set-based technique employing the benchmark datasets.

4.2. Msnbc.com Dataset

This msnbc.com dataset describes the pages of websites visited by visitors. The data is recorded at the level URLs category in chronological order taken from the Internet Information Server (IIS) server log file from the msnbc.com website. Each line of the dataset corresponds to a request from a user for a web page.

Table 4 shows the comparative results in terms of execution time (in seconds) and cluster purity on each number of different data samples.

Table 4. Performance Comparison of Msnbc.com data set

Number of Transactions	Executing Time				Clusters Purity			
	Technique [27]	Technique [28], [29]	Proposed	Improvement	Techniqe [27]	Techniqe [28], [29]	Proposed	Improvement
100	15.6	9.8	5.9	115.3%	93.0%	100%	100%	3.5%
200	171.2	121.2	65.9	121.9%	96.0%	100%	100%	2.0%
500	819.6	493.7	259.1	153.4%	95.5%	100%	100%	2.3%
1000	3723.6	2092.9	1283.9	126.5%	95.5%	100%	100%	2.3%
2000	5371.2	4352.8	2901.8	67.6%	96.9%	100%	100%	1.6%
Average improvement				116.9%	Average improvement			2.3%

Based on the results shown in Table 4, the rough set-based approaches proposed by [19]–[21] took longer processing time, because to find the similarity of upper approximations requires many iterations. Meanwhile, the soft set-based approach shows significant improvement of more than 116 % for user transactions starting from 100 until 2000 data sample. For cluster purity, the proposed approach improves up to 2.3% as compared to two rough set-based approaches.

4.3. Microsoft.com data set

This dataset describes the pages visited by users who visit www.microsoft.com. For each user, the data lists all areas of the website (Vroot) that a user visits within a week. The log files do not contain personally identifiable information, and the title and also URL identifies the Vroots.

Table 5. Performance comparison of Microsoft.com data set

Number of Transactions	Executing Time				Clusters Purity			
	Technique [27]	Technique [28], [29]	Proposed	Improvement	Technique [27]	Technique [28], [29]	Proposed	Improvement
100	12.1	6.4	4.1	125.6%	76.0%	100%	100%	12.0%
200	51.1	32.1	19.3	115.5%	79.5%	100%	100%	10.3%
500	431.7	215.9	173.9	86.2%	87.4%	100%	100%	6.3%
1000	2375.8	1598.4	1079.1	84.2%	91.5%	100%	100%	4.3%
2000	4751.6	2375.8	1663.1	114.3%	93.2%	100%	100%	3.4%
		Average improvement		105.2%		Average improvement		7.2%

Table 5 shows the comparison of the processing time (in second) and clusters purity. As we can see that, the proposed approach has shown improvement of more than 105% in terms of execution time for user transactions starting from 100 until 2000. For cluster purity, the proposed approach improves up to 7.2% from the two rough set-based approaches.

5. Conclusion

In this paper, we have examined the clustering web user transactions which focusing on reducing computational complexity and increasing cluster purity. The proposed method is the first study that has proposed for clustering web user transactions by using soft set theory. We proposed an algorithm with varying lower computational complexity and higher cluster purity. Although several existed baseline techniques address the issues concerning web user transactions clustering, none of these techniques provide low computational complexity and high cluster purity. We have carried out a comparative analysis of the proposed technique concerning final computation complexity and cluster purity. The results showed that the proposed technique outperforms two rough set-based techniques on computation complexity and cluster purity.

References

[1] J Han, M Kamber, *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.
[2] M Vazirgiannis, M Halkidi, D Gunopulos, Uncertainty handling and quality assessment in data mining. *Springer*, 2003.
[3] LA Zadeh. Fuzzy sets. *Inf. Control*, 1965; 8(3): 338-353.
[4] Z Pawlak, J Grzymala-Busse, R Slowinski, W Ziarko. Rough sets. *Commun. ACM*, 1995; 38(11): 88-95.
[5] W-L Gau, DJ Buehrer. Vague sets. *IEEE Trans. Syst. Man. Cybern.*, 1993; 23(2): 610-614.
[6] D Molodtsov. Soft set theory-first results. *Comput. Math. with Appl.*, 1999; 37(4-5): 19-31.
[7] MI Ali, M Shabir, M Naz. Algebraic structures of soft sets associated with new operations. *Comput. Math. with Appl.*, 2011; 61(9): 2647-2654.
[8] F Feng, Y Li, N Ça\ugman. Generalized< i> uni--int</i> decision making schemes based on choice value soft sets. *Eur. J. Oper. Res.*, 2012; 220(1): 162-170.
[9] PK Maji, R Biswas, AR Roy. Soft set theory. *Comput. Math. with Appl.*, 2003; 45(4): 555-562.
[10] J Mao, D Yao, C Wang. Group decision making methods based on intuitionistic fuzzy soft matrices. *Appl. Math. Model.*, 2013; 37(9): 6425-6436.
[11] M Shabir, M Irfan Ali, T Shaheen. Another approach to soft rough sets. *Knowledge-Based Syst.*, 2013; 40: 72-80.
[12] W Xu, Z Xiao, X Dang, D Yang, X Yang. Financial ratio selection for business failure prediction using soft set theory. *Knowledge-Based Syst.*, 2014; 63: 59-67.
[13] E Sutoyo, M Mungad, S Hamid, T Herawan. An efficient soft set-based approach for conflict analysis. *PLoS One*, 2016; 11(2).

[14] ITR Yanto, I Azhari. Alternative Technique Reducing Complexity of Maximum Attribute Relation. *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2015; 13(4): 1361-1367.
[15] MAT Mohammed, WMW Mohd, RA Arshah, M Mungad, E Sutoyo, H Chiroma. Analysis of Parameterization Value Reduction of Soft Sets And Its Algorithm. *Int. J. Softw. Eng. Comput. Syst.*, 2016; 2(1): 51-57.
[16] RR Saedudin, E Sutoyo, S Kasim, H Mahdin, ITR Yanto. *Attribute selection on student performance dataset using maximum dependency attribute*. in Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on, 2017: 176-179.
[17] AK Jain, RC Dubes. Algorithms for clustering data. 1988.
[18] G Xu, Y Zhang, L Li. Web mining and social networking: techniques and applications. 2010; 6. *Springer*.
[19] SK De, PR Krishna. Clustering web transactions using rough approximation. *Fuzzy Sets Syst.*, 2004; 148(1): 131-138.
[20] ITR Yanto, T Herawan, MM Deris. A Framework of Rough Clustering for Web Transactions. in *Advances in Intelligent Information and Database Systems*, Springer, 2010: 265-277.
[21] ITR Yanto, T Herawan, MM Deris. RoCeT: Rough set approach for clustering web transactions. *Int. J. Biomed. Hum. Sci.*, 2010; 16(2): 135-145.
[22] M Lichman, {UCI} Machine Learning Repository. 2013.