Knowledge-Based Systems 24 (2011) 186–195

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A soft set approach for association rules mining

Tutut Herawan [a], Mustafa Mat Deris [b,*]

[a] Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang, Gambang 26300, Pahang, Malaysia
[b] Faculty of Information Technology and Multimedia, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat 86400, Johor, Malaysia

ARTICLE INFO

Article history:
Received 7 June 2009
Received in revised form 23 August 2010
Accepted 24 August 2010
Available online 9 September 2010

Keywords:
Association rules mining
Maximal association rules mining
Boolean-valued information systems
Soft set theory
Items co-occurrence

ABSTRACT

In this paper, we present an alternative approach for mining regular association rules and maximal association rules from transactional datasets using soft set theory. This approach is started by a transformation of a transactional dataset into a Boolean-valued information system. Since the "standard" soft set deals with such information system, thus a transactional dataset can be represented as a soft set. Using the concept of parameters co-occurrence in a transaction, we define the notion of regular and maximal association rules between two sets of parameters, also their support, confidence and maximal support, maximal confidences, respectively properly using soft set theory. The results show that the soft regular and soft maximal association rules provide identical rules as compared to the regular and maximal association rules.

1. Introduction

Data mining is a generic term which covers research results, techniques and tools used to extract useful information from large databases. Association rule is one of the most popular data mining techniques and has received considerable attention, particularly since the publication of the AIS and Apriori algorithms [1,2]. They are particularly useful for discovering relationships among data in huge databases and applicable to many different domains including market basket and risk analysis in commercial environments, epidemiology, clinical medicine, fluid dynamics, astrophysics, and crime prevention. The association rules are considered interesting if it satisfies certain constraints, i.e. predefined minimum support (*minsupp*) and minimum confidence (*minconf*) thresholds. Many algorithms of association rules mining have been proposed, including the works of [3–5]. The association rules method was developed particularly for the analysis of transactional databases, whose attributes posses Boolean values and has been shown to be a very efficient data structure for association rules mining. In other words, the occurrence of an item can be viewed as a Boolean variable and its value is "1" if it appears in that particular transaction and "0" otherwise. This conforming that a transactional dataset can be converted into a Boolean-valued information system, $S = (U, A, V_{\{0,1\}}, f)$.

Soft set theory [6], proposed by Molodtsov in 1999, is a new general method for dealing with uncertain data. Soft sets are called (binary, basic, elementary) neighborhood systems. As for standard soft set, it may be redefined as the classification of objects in two distinct classes, thus confirming that soft set can deal with a Boolean-valued information system. Molodtsov [6] pointed out that one of the main advantages of soft set theory is that it is free from the inadequacy of the parameterization tools, unlike in the theories of fuzzy set [7], probability and interval mathematics. In recent years, research on soft set theory has been active, and great progress has been achieved, including the works of the using of fundamental soft set theory, soft set theory in abstract algebra and soft set theory for data analysis, particularly in decision making [8–12]. Since the "standard" soft set (F,E) over the universe U can be represented by a Boolean-valued information system, thus a soft set can be used for representing a transactional dataset. Therefore, one of the applications of soft set theory for data mining is for mining association rules. However, not many researches have been done on this application.

In this work, we propose an alternative approach for regular and maximal association rules mining using soft set theory. We use the Boolean-valued information system that has been shown to be a very efficient data structure for association rules mining. The Boolean-valued information system is a conversion from a transactional dataset. We define the regular support, regular confidence and maximal support and maximal confidence of the rules, respectively, based on the concept of co-occurrences and maximal co-occurrences of parameters in a transactional dataset under soft set theory. There are three main contributions of this work. First, we present that a soft set can be used to represent a transactional data via a Boolean-valued information system. Second, we present the applicability of the soft set theory for mining "regular" association rules and maximal association rules. Third, we show that by using soft set theory, association rules and maximal rules

* Corresponding author. Tel.: +60 7 4538001; fax: +60 7 4532199.
E-mail addresses: tutut@ump.edu.my (T. Herawan), mmustafa@uthm.edu.my (M.M. Deris).

0950-7051/$ - see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2010.08.005

discovered are identical to that rules in association rules and maximal association rules approaches.

The rest of this paper is organized as follows. Section 2 describes fundamental concept of "regular" association rules and maximal association rules mining. Section 3 describes the fundamental concept of information system and soft set theory. Section 4 describes a transformation of a transaction table into a soft set via a Boolean-valued information system. Section 5 describes soft set approach for association rules and maximal association rules mining. Section 6 describes the results of the proposed approaches. Finally, the conclusion of this work and future work are described in Section 7.

2. Preliminaries

2.1. Association rules

Let $I = \{i_1, i_2, \ldots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called *set of items* and the set $D = \{t_1, t_2, \ldots, t_{|U|}\}$, for $|U| > 0$ refers to the transactional dataset, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \ldots, i_{|M|}\}$, $1 \leqslant |M| \leqslant |A|$ and each transaction can be identified by a distinct identifier *TID*. Let, a set $X \subseteq t \subseteq I$ called an *itemset*. An itemset with k-items is called a *k-itemset*. The *support* of an itemset X, denoted sup (X) is defined as a number of transactions contain X. An *association rule* between sets X and Y is an implication of the form $X \Rightarrow Y$, where $X \cap Y = \phi$. The itemsets X and Y are called *antecedent* and *consequent*, respectively. The *support* of an association rule $X \Rightarrow Y$, denoted sup $(X \Rightarrow Y)$, is defined as a number of transactions in D contain $X \cup Y$. The *confidence* of an association rule $X \Rightarrow Y$, denoted cfi $(X \Rightarrow Y)$ is defined as a ratio of the numbers of transactions in D contain $X \cup Y$ to the number of transactions in D contain X. Thus, $\text{cfi}(X \Rightarrow Y) = \frac{\sup(X \Rightarrow Y)}{\sup(X)}$.

A huge number of association rules can be found from a transactional dataset. To find the interesting association rules in a transactional dataset, we must define a specified minimum support (called *minsup*) and specified minimum confidence (called *minconf*). The itemset $Y \subseteq I$ is called *frequent itemset* if sup $(X) \geqslant$ minsup. It is known that a subset of any frequent itemset is a frequent itemset, a superset of any infrequent itemset is not a frequent itemset. Finally, the association rule $X \Rightarrow Y$ holds if conf $(X \Rightarrow Y) \geqslant$ minconf.

The association rules are said to be strong if it meets the minimum confidence threshold. However, while association rules provide means to discover many interesting associations, they fail to discover other, no less interesting associations, which also hidden in the data. Maximal association rules introduced by Feldman et al. [13] is a variant of association rules which is designed to handle the above problem. It allows the discovery of associations pertaining to items that most often do not appear alone, but rather together with closely related items, and hence associations relevant only to these items tend to obtain low confidence. These rules are very important in discovering maximal association, particularly from documents text collection. The idea is inspired from the fact that many interesting rules in databases cannot captured by regular rules. Feldman et al. noted that maximal association rules are not designed to replace regular association rules, but rather to complement them. Every maximal association rule is also regular association, with perhaps different support and confidence [14]. While association rules are based on the notion of frequent itemsets which appears in many records, maximal association rules are based on frequent maximal itemsets which appears maximally in many records [15]. Using only maximal association rules, many interesting regular associations may and will be lost. The initial step to discover maximal rules is a partition on the set of items from a transactional dataset so-called a taxonomy and categorization of items.

2.2. Taxonomy and category

Let $I = \{i_1, i_2, \ldots, i_{|A|}\}$ be a set of items. A *taxonomy* T of I is a partition of I into disjoint subsets, i.e., $T = \{T_1, T_2, \ldots, T_n\}$. Each member of T is called a *category*. For an item i, we denote $T(i)$ the category that contain i. Similarly, if X is an itemset all of which are from a single category, then we denote this category by $T(X)$.

Example 1. There is a dataset consisting of the 10 transactions [13]; 2 articles referring to Countries "Canada, Iran, USA" and refers to Topics "crude, ship"; 1 article referring to "USA" and refers to "earn"' 2 articles referring to "USA" and refers to "jobs, cpi"; 1 article referring to "USA" and refers to "earn, cpi"; 1 article referring to "Canada" and refers to "sugar, tea"; 2 articles referring to "Canada, USA" and refers to "trade, acq" and 1 article referring to "Canada, USA" and refers to "earn". We can present such transactions in the following table.

We can create a taxonomy based on Table 1, which is contains two categories "countries" and "topics", i.e., $T = \{\text{countries, topics}\}$, where

$$\text{countries} = \{\text{Canada, Iran, USA}\}$$

and

$$\text{topics} = \{\text{crude, ship, earn, jobs, cpi, sugar, tea, trade}\}.$$

2.3. Maximal association rules

To illustrate the notion of maximal association rules, let we consider the idea which are quoted directly from [14]. In maximal association rule $X \overset{\max}{\Rightarrow} Y$, we are interested in capturing the notion that whenever X appears alone then Y also appears, with some confidence. For this, we must fist define the notion of alone. We do so with respect to the categories of T as follows.

For a transaction t, a category T_i and an itemset $X \subseteq T_i$, we say that X is *alone* in t if $t \cap T_i = X$. That is, X is alone in t if X is the largest subset of T_i which is in t. In this case we also say that X is *maximal* in t and that t M-supports X. For a database D, the M-support of X in D, denoted $S_D^{\max}(X)$ is the number of transaction $t \in D$ that M-support X.

A *maximal association rule* or M-association rule is a rule of the form $X \overset{\max}{\Rightarrow} Y$, where X and Y subsets distinct categories, $T(X)$ and $T(Y)$, respectively. The M-support of the M-association rule $X \overset{\max}{\Rightarrow} Y$, denoted by $S_D^{\max}(X \overset{\max}{\Rightarrow} Y)$ is defined as

$$S_D^{\max}(X \overset{\max}{\Rightarrow} Y) = |\{t : tM - \text{supports } X \text{ and } t \text{ supports } Y\}|.$$

That is, $S_D^{\max}(X \overset{\max}{\Rightarrow} Y)$ is the number of transactions in D that M-support X and also support Y in the regular sense. The intuitive meaning

Table 1
A data of transactions from [13].

TID	Items
1	Canada, Iran, USA, crude, ship
2	Canada, Iran, USA, crude, ship
3	USA, earn
4	USA, jobs, cpi
5	USA, jobs, cpi
6	USA, earn, cpi
7	Canada, sugar, tea
8	Canada, USA, trade, acq
9	Canada, USA, trade, acq
10	Canada, USA, earn