Journal of King Saud University – Computer and Information Sciences (2014) **26**, 99–109

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

Unexpected rules using a conceptual distance based on fuzzy ontology

Mohamed Said Hamani [a], **Ramdane Maamri** [b,*], **Yacine Kissoum** [c], **Maamar Sedrati** [d]

[a] *Lire Laboratory, University of M'sila, Algeria*
[b] *Lire Laboratory, University of Constantine 2, Algeria*
[c] *Lire Laboratory, University of Skikda, Algeria*
[d] *University of Hadj Lakhdar Batna, Algeria*

Received 31 July 2012; revised 20 April 2013; accepted 3 June 2013
Available online 18 June 2013

KEYWORDS

Fuzzy ontology;
Unexpectedness;
Association rule;
Domain knowledge;
Interestingness;
Conceptual distance

Abstract One of the major drawbacks of data mining methods is that they generate a notably large number of rules that are often obvious or useless or, occasionally, out of the user's interest. To address such drawbacks, we propose in this paper an approach that detects a set of unexpected rules in a discovered association rule set. Generally speaking, the proposed approach investigates the discovered association rules using the user's domain knowledge, which is represented by a fuzzy domain ontology. Next, we rank the discovered rules according to the conceptual distances of the rules.

1. Introduction

Knowledge discovery in data mining has been defined in Fayyad et al. (1996) as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data. Association rule algorithms (Agrawal et al., 1993) are rule-discovery methods that discover patterns in the form of IF-THEN rules. It has been noticed that most of the algorithms that perform data mining generate a large number of rules that are valid but obvious or not interesting to the user (Liu and Hsu, 1996; Piatetsky-Shapiro, 1996; Piatetsky-Shapiro and Matheus, 1991; Silberschatz and Tuzhilin, 1996). To address this issue, most of the approaches to knowledge discovery use objective measures of interestingness for the evaluation of the discovered rules, such as confidence and support measures (Agrawal et al., 1993). These approaches capture the statistical strength of a pattern. The interestingness of a rule is essentially subjective (Liu and Hsu, 1996; Piatetsky-Shapiro and Matheus, 1991; Silberschatz and Tuzhilin, 1996; Klemettinen et al., 1994). Subjective measures of interestingness, such as unexpectedness (Zimmermann, 2001; Asa and Mangano, 1995; Uthurusamy et al., 1991), assume that the interestingness of a pattern depends on the decision-maker and does not solely depend on the statistical strength of the pattern. Although objective measures are

* Corresponding author. Tel.: +213 661693966.
E-mail addresses: saidhamani@hotmail.com (M.S. Hamani), rmaamri@ yahoo.fr, rmaamri@umc.edu.dz (R. Maamri), kissoumyacine @yahoo.fr (Y. Kissoum), msedrati2006@yahoo.fr (M. Sedrati).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

1319-1578 © 2013 Production and hosting by Elsevier B.V. on behalf of King Saud University.
http://dx.doi.org/10.1016/j.jksuci.2013.06.001

useful, they are insufficient in the determination of the interestingness of the rules. One way to address this problem is by focusing on discovering unexpected patterns (Liu and Hsu, 1996; Silberschatz and Tuzhilin, 1996; Liu et al., 1997; Padmanabhan and Tuzhilin, 1998; Padmanabhan and Tuzhilin, 1999; Silberschatz and Tuzhilin, 1995), where the unexpectedness of the discovered patterns is usually defined relative to a system of prior expectations.

Moreover, ontology represents knowledge. Ontology is organized as a DAG (Directed Acyclic Graph) hierarchy. Ontologies allow domain knowledge to be represented explicitly and formally in such a way that it can be shared among human and computer systems. Unfortunately, knowledge about a system can contain ambiguity and vagueness. For this reason, fuzzy ontologies have been used to address such fuzzy knowledge (xxx, 1291), where the concepts are related to each other in the ontology with a degree of membership μ $(0 \leqslant \mu \leqslant 1)$. In this paper, we propose a new approach that adds intelligence and autonomy for ranking rules according to their conceptual distance (the distance between the antecedent and the consequent of the rule) relative to the hierarchy. In other words, highly related concepts are grouped together in the hierarchy. The more distant the concepts are, the less they are related to each other. For concepts that are part of the definition of a rule, the less the concepts are related to each other, the more the rule is surprising and therefore interesting. With such a ranking method, a user can check fewer rules on the top of the list to extract the most pertinent ones.

1.1. Association Rules

Association rule mining finds interesting associations and/or correlation relationships among a large set of data items. Association rules show attribute value conditions that occur frequently together in a given dataset. A typical and widely used example of association rule mining is Market Basket Analysis (http://www.resample.com/xl). In market basket analysis, customers' buying habits are analyzed to find associations between different items that customers place in their shopping cart. Two different items, 'a' and 'b', in an item set are assumed to have a relation if they are purchased together in the same transaction. The more those two items are purchased together in the same transaction, the more they have a stronger relation. The discovery of such associations can help retailers develop marketing strategies by gaining an insight into which items are frequently purchased together by customers. Association rules provide information of this type in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, the association rules are probabilistic in nature (http://www.resample.com/xl). Objective measures such as support and confidence are often used to evaluate the interestingness of the association rules.

The support is simply the number of transactions that include all of the items in the antecedent and consequent parts of the rule. The support is sometimes expressed as a percentage of the total number of records in the database.

The confidence is the ratio of the number of transactions that include all of the items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all of the items in the antecedent.

One way to think of support is that it is the probability that a randomly selected transaction from the database will contain all of the items in the antecedent and consequent, whereas the confidence is the conditional probability that a randomly selected transaction will include all of the items in the consequent given that the transaction includes all of the items in the antecedent (http://www.resample.com/xl).

Interestingness measures are called fitness functions in Ykhlef (2011). The study in Ykhlef (2011) divides fitness functions into two types, basic and complex. Support and confidence are considered to be basic measures, whereas certain other fitness functions are derived from information theory and are considered to be complex fitness functions.

Many algorithms can be used to discover association rules from data to extract useful patterns. The Apriori algorithm is one of the most widely used and famous techniques for finding association rules (Agrawal et al., 1993; Agrawal, 1994). The Apriori algorithm requires two thresholds of minconfidence (representing minimum confidence) and minsupport (representing minimum support). These two thresholds determine the degree of association that must hold before a rule will be mined. The algorithm operates in two phases. In the first phase, all of the item sets with minimum support (frequent item sets) are generated. The second phase of the algorithm generates rules from the set of all frequent item sets.

1.2. Rule interestingness measures

Past research in data mining has shown that the interestingness of a rule can be measured using objective measures and subjective measures. Objective measures involve analyzing the rule's structure, predictive performance, and statistical significance. In association to rule mining, such measures include support and confidence (Liu et al., 2000). However, it is noted in Piatetsky-Shapiro and Matheus (1991) that such objective measures are insufficient for determining the interestingness of a discovered rule. Indeed, subjective measures are needed.

There are two main subjective interestingness measures, namely unexpectedness (Liu and Hsu, 1996; Silberschatz and Tuzhilin, 1996) and actionability (Piatetsky-Shapiro and Matheus, 1991; Silberschatz and Tuzhilin, 1996).

- Unexpectedness: Rules are interesting if they are unknown to the user or contradict the user's existing knowledge (or expectations).
- Actionability: Rules are interesting if the user can do something with them to his/her advantage.

In this research, we focus only on unexpectedness.

1.3. Ontology

The term ontology has been widely used in recent years in the field of Artificial Intelligence and computer and information science, especially in domains such as cooperative information systems, intelligent information integration, information retrieval and extraction, knowledge representation, and database management systems (Guarino, 1998). Although there is no universal consensus on the definition of an ontology, it is generally accepted that ontology is a specification of conceptualization (Leacock and Chodorow, 1998). The prior knowledge