Available online at www.sciencedirect.com
ScienceDirect
Computers and Mathematics with Applications 54 (2007) 1303–1318
An International Journal
computers & mathematics
with applications
www.elsevier.com/locate/camwa

Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan

R.J. Kuo*, C.W. Shih

Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan, ROC

Received 17 March 2006; accepted 21 March 2006

Abstract

In the field of data mining, an important issue for association rules generation is frequent itemset discovery, which is the key factor in implementing association rule mining. Therefore, this study considers the user's assigned constraints in the mining process. Constraint-based mining enables users to concentrate on mining itemsets that are interesting to themselves, which improves the efficiency of mining tasks. In addition, in the real world, users may prefer recording more than one attribute and setting multi-dimensional constraints. Thus, this study intends to solve the multi-dimensional constraints problem for association rules generation.

The ant colony system (ACS) is one of the newest meta-heuristics for combinatorial optimization problems, and this study uses the ant colony system to mine a large database to find the association rules effectively. If this system can consider multi-dimensional constraints, the association rules will be generated more effectively. Therefore, this study proposes a novel approach of applying the ant colony system for extracting the association rules from the database. In addition, the multi-dimensional constraints are taken into account. The results using a real case, the National Health Insurance Research Database, show that the proposed method is able to provide more condensed rules than the Apriori method. The computational time is also reduced.

Keywords: Data mining; Multiple dimensional constraints; Ant colony system; Apriori

1. Introduction

Mining association rules from a large database of business data, such as transaction records, has been an important issue in the field of data mining. The problem of association rule mining can be divided into two sub-problems: (1) frequent itemset discovery and (2) association rules generation. It has also been shown that the overall performance of mining is seriously determined by the first sub-problem.

Frequent itemset mining algorithms often generate a very large number of frequent itemsets and rules, which reduce both the efficiency and also the effectiveness of the mining algorithms since only the subset of the complete frequent itemsets and association rules is of interest to users. In addition, the users need an additional post-processing step to filter the large number of mined rules to determine the useful ones. Recent work [1–4] has highlighted the importance

* Corresponding author. Tel.: +886 2 27712171x4541; fax: +886 2 27317168.
E-mail address: rjkuo@ntut.edu.tw (R.J. Kuo).

0898-1221/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2006.03.043

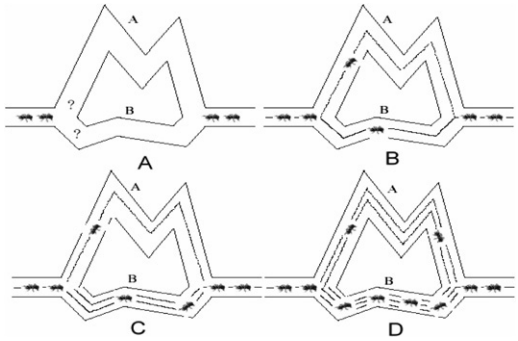

Fig. 1. The behavior of real ants.

of constraint-based mining. They exploit user-specific constraints in the mining process to improve performance, or efficiency. With multi-dimensional items, constraints can be imposed on multiple dimensional attributes. We classify multi-dimensional constraints into two cases according to the number of sub-constraints including: (1) single constraint against multiple dimensions, such as $\max(X, \text{cost}) \leq (X, \text{price})$, where X is an itemset and each item in X contains two attributes "cost" and "price", and (2) conjunction and/or disjunction of multiple sub-constraints, such as $(C_1 : X, \text{cost} \leq v_1) \wedge (C_2 : X, \text{price} \leq v_2)$, where v_1 and v_2 are constant values, respectively.

Therefore, this study intends to use the *ant colony system*, which has recently been shown to be very promising in the areas of the traveling salesman problem and scheduling [5,6], for multiple dimensional constraints mining association rules. Furthermore, since data mining has rarely been applied to solve questions in medical science, this study uses data from the National Health Insurance Research Database of Taiwan to find disease association rules. Here, an important issue is to find the potential disease and early prevention. The evaluation results show that the proposed method, using the ant colony system, really can provide more concise and accurate information than the conventional Apriori-based algorithm.

The rest of this paper is organized as follows. Section 2 summarizes some important background information, and the proposed method is described in Section 3. Section 4 presents the evaluation results and discussion. Finally, concluding remarks are made in Section 5.

2. Background

This section reviews three aspect of the literature. Two of these are the main component of the proposed method, namely Ant Algorithms, and Association Rule Mining, which is a technique for mining patterns. Finally, a related survey of Multi-Dimensional Constraints Mining is presented. Detailed information is presented below.

2.1. Ant algorithms

2.1.1. Concept of ant theorem

In the real world, ants communicate with other ants by a trail of chemicals called "*pheromones*" which are deposited by ants when they search for food. Then, the other ants encounter the previously laid pheromones and decide how many probabilities they will follow. As more and more ants pass by the same path, the pheromones on the shorter path would be increased, but the pheromone would evaporate on the other paths, as illustrated in Fig. 1.

2.1.2. The evolution and applications of ant algorithms

The first ant algorithm was introduced in Dorigo's dissertation, called *Ant System* [7], and was inspired by observation of real ant colonies. One of the first applications for the ant algorithm was for the traveling salesman problem (TSP). Recently, ACO has successfully been applied to several combinatorial optimization problems and yielded many promising developments. For instance, Gambardella and Dorigo [8] proposed *Ant-Q* station transition rules and used Q-learning to strengthen and renew pheromone trails. Bullnheimer et al. [9] proposed selecting a quantity of elite ants to renew the pheromone trail, this was called the $\mathbf{AS}_{\text{rank}}$ algorithm. Dorigo and Gambardella [10] presented a method to improve the ant system, called *Ant Colony System*. The current development of ant algorithms