A Framework for Synthesizing Arbitrary Boolean Queries Induced by Frequent Itemsets

Animesh Adhikari, Department of Computer Science, S. P. Chowgule College, Margao, Goa, India

ABSTRACT

Frequent itemsets determine the major characteristics of a transactional database. It is important to mine arbitrary Boolean queries induced by frequent itemsets. In this paper, the author proposes a simple and elegant framework for synthesizing arbitrary Boolean queries using conditional patterns in a database. Both real and synthetic databases were used to evaluate the experimental results. The author presents an algorithm for mining a set of specific itemsets in a database and a model of synthesizing a query in a database. Finally, the author discusses an application of the proposed framework for reducing query processing time.

Keywords: Boolean Query, Conditional Pattern, Data Mining, Generator of a Boolean Expression, Synthesis of Query

INTRODUCTION

An itemset could be thought as the basic type of pattern in a transactional database. Itemset patterns influence research in knowledge discovery in databases (KDD) in the following ways: Firstly, many interesting algorithms have been reported on mining itemset patterns in a database (Agrawal & Srikant, 1994; Han et al., 2000; Savasere et al., 1995). Secondly, many patterns are defined based on the itemset patterns in a database. They may be called as derived patterns. Some examples of derived patterns are positive association rules (Agrawal et al., 1993), negative association rules (Antonie & Zaïane, 2004), and conditional patterns (Adhikari & Rao, 2008) in a database. A good amount of work has been reported on mining / synthesizing such derived patterns. Thirdly, the solutions of many problems are based on the analysis of patterns in a database. Such applications (Adhikari & Rao, 2008; Wu et al., 2005) process patterns in a database for the purpose of making some decisions. Thus, the mining and analysis of itemset patterns in a database is an interesting as well as important issue. Also, mining Boolean expressions induced by frequent itemsets could lead to significant nuggets of knowledge with many potential applications in market basket data analysis, web usage mining, social network analysis and bioinformatics. There are two important goals of this paper. First, we design a framework for synthesizing an arbitrary Boolean expression

DOI: 10.4018/ijkbo.2013040104
induced by frequent itemsets. Afterwards, we present a technique for reducing query processing time using such synthesized knowledge.

The support (Agrawal et al., 1993) of an itemset \( X \) in database \( D \) could be defined as the fraction of transactions in \( D \) containing all the items of \( X \), denoted by \( S(X, D) \). The importance of an itemset could be judged by its support. An itemset \( X \) is frequent in \( D \) if \( S(X, D) \geq \text{minimum support} \) (\( a \)). Let \( SFIS(D) \) be the set of frequent itemsets in \( D \). Frequent itemsets determine the major characteristics of a database. Wu et al. (2005) have proposed a solution of inverse frequent itemset mining. Authors argued that one could efficiently generate a synthetic market basket dataset from the frequent itemsets and their supports. Let \( X \) and \( Y \) be two itemsets in \( D \). The characteristics of \( D \) are revealed more by the pair \( (X, S(X, D)) \) than that of \( (Y, S(Y, D)) \), if \( S(X, D) > S(Y, D) \). So it is important to study frequent itemsets more than infrequent itemsets. Hence, we propose a framework for synthesizing arbitrary Boolean queries induced by frequent itemsets in \( D \). Zhao et al. (2006) have proposed the BLOSOM framework for mining arbitrary Boolean expressions. The framework suffers from the following limitations:

- It does not handle NOT operator effectively.
- Let \( \{a, b, c\} \) be a frequent itemset of our interest. We wish to mine some functions induced by \( \{a, b, c\} \). It proposes a framework to mine minimal generators of (i) closed OR-clauses, (ii) closed AND-clauses, (iii) closed maximal min-DNF, and (iv) closed maximal min-CNF. It requires establishing a mapping from the space of minimal generators to the space of arbitrary Boolean expressions, so that we could study the desired Boolean expressions induced by \( \{a, b, c\} \). Thus, the BLOSOM framework might not provide the knowledge of Boolean expression that one wishes to study.
- A specific framework for a specific type of Boolean expressions is introduced.

Moreover, most of the existing works (Pei & Han, 2000; Bonchi & Lucchese, 2005) have attempted to answer queries during the mining process. First, we propose here a simple and elegant approach for synthesizing arbitrary Boolean queries induced by frequent itemsets. The proposed framework of synthesizing Boolean queries is based on conditional patterns in \( D \). Afterwards, we have presented an application of such synthesized knowledge by presenting a framework for answering arbitrary queries.

First we explain the concept of conditional pattern and then we present a framework for synthesizing Boolean queries induced by frequent itemsets in \( D \). The concept of conditional pattern is not new (Adhikari & Rao, 2008). For the purpose of completeness, we discuss the notion of conditional pattern in this section. Let \( X = \{a, b, c\} \) be an itemset in \( D \). The study of items in \( X \) might be incomplete if we have only the following information about \( X \): (i) the supports of \( X \) and its subsets, (ii) the association rules generated from \( X \). The answers to some queries on the items of \( X \) are not immediately available from (i) and (ii). A few examples of such queries are given below:

- Find the support that a transaction contains item \( a \), but not items \( b \) and \( c \) with respect to frequent itemset \( \{a, b, c\} \).
- Find the support that a transaction contains items \( a \) and \( b \) but not the item \( c \) with respect to frequent itemset \( \{a, b, c\} \).

The above queries correspond to a specific type of pattern in a database. Some of these patterns could have significant supports, since \( \{a, b, c\} \) is a frequent itemset. In general, let \( X \) be a frequent itemset in \( D \). If we wish to study the distribution of items of itemset \( Y \subseteq X \), but not the items of itemset \( X-Y \), then such analysis of items is not be immediately available. Analysis of such patterns is interesting, since their supports could be high. Therefore, one needs to identify this type of patterns. Let \( (Y, X) \) be a pattern that a transaction in a database contains all the items of itemset \( Y \), but not the items of itemset \( X-Y \), for a given itemset \( X \) in \( D \). Let \( S(Y, X, D) \) be the support that a transaction in \( D \) contains all the items of \( Y \), but not the items of \( X-Y \), for a given itemset \( X \) in \( D \). The pattern of type \( (Y, X) \) is called conditional pattern. A
Roles and Impacts of Automatic Item Generation on Assessment Research, Practice, and Policy
www.igi-global.com/chapter/roles-and-impacts-of-automatic-item-generation-on-assessment-research-practice-and-policy/205402?camid=4v1a

Knowledge Management and Systematic Innovation Capability
www.igi-global.com/article/knowledge-management-and-systematic-innovation-capability/170543?camid=4v1a