Knowledge-Based Systems 84 (2015) 224–238

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Pattern set mining with schema-based constraint

Luca Cagliero [a], Silvia Chiusano [a], Paolo Garza [a,*], Giulia Bruno [b]

[a] Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[b] Dipartimento di Ingegneria Gestionale e della Produzione, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form 5 March 2015
Accepted 21 April 2015
Available online 1 May 2015

Keywords:
Pattern set mining
Itemset mining
Data mining

ABSTRACT

Pattern set mining entails discovering groups of frequent itemsets that represent potentially relevant knowledge. Global constraints are commonly enforced to focus the analysis on most interesting pattern sets. However, these constraints evaluate and select each pattern set individually based on its itemset characteristics.

This paper extends traditional global constraints by proposing a novel constraint, called schema-based constraint, tailored to relational data. When coping with relational data itemsets consist of sets of items belonging to distinct data attributes, which constitute the itemset schema. The schema-based constraint allows us to effectively combine all the itemsets that are semantically correlated with each other into a unique pattern set, while filtering out those pattern sets covering a mixture of different data facets or giving a partial view of a single facet. Specifically, it selects all the pattern sets that are (i) composed only of frequent itemsets with the same schema and (ii) characterized by maximal size among those corresponding to that schema. Since existing approaches are unable to select one representative pattern set per schema in a single extraction, we propose a new Apriori-based algorithm to efficiently mine pattern sets satisfying the schema-based constraint. The experimental results achieved on both real and synthetic datasets demonstrate the efficiency and effectiveness of our approach.

1. Introduction

Frequent itemsets represent recurrent correlations among data items [1], which are usually selected by considering their local interestingness in the analyzed data [2,3]. However, since itemset mining from real-life data commonly entails discovering a large number of itemsets that are fairly correlated with each other, the manual inspection of the mining result could be a challenging task. To overcome this issue, pattern set mining with global constraints aims at discovering worthwhile groups of itemsets [4]. Instead of evaluating and selecting itemsets individually, pattern sets (i.e., sets of itemsets) are generated and evaluated as a whole to analyze the correlations among data from a high-level viewpoint.

Relational data is characterized by a fixed schema, which consists of a set of attributes representing peculiar data features. Itemsets mined from relational data are sets of items belonging to distinct data attributes. Hence, they are characterized by a schema too. Frequent itemsets with the same schema are, to a certain extent, semantically correlated with each other because they are recurrent instances of the same data facet. Hence, the itemset schema can be considered to be particularly suitable for clustering recurrent co-occurrences among data items related to the same facet into pattern sets. Furthermore, instead of generating all the pattern sets complying with a given schema, for each schema only the largest pattern set should be considered, because all the others are partial representations of the same data facet. However, to evaluate pattern set interestingness existing algorithms just evaluate one pattern set at a time. Therefore, they cannot extract for each schema only the best representative pattern set unless generating all the pattern sets first and then postprune the uninteresting ones.

This paper addresses the problem of pattern set mining with global constraints from relational data. To generate only the groups of itemsets containing all the pertinent information related to a given facet, we propose a new global constraint, namely the *schema-based constraint*, tailored to relational data. The schema-based constraint selects all the pattern sets that are (i) composed only of frequent itemsets with the same schema and (ii) characterized by maximal size among those corresponding to that schema. To provide a condensed and potentially useful representation of different data facets we select *at most* one pattern set per schema, i.e., the pattern set that consists of *all and only* the frequent itemsets with that schema.

To improve the manageability of the mined pattern sets two parallel strategies are commonly adopted [4]: (i) enforcing a

* Corresponding author. Tel.: +39 011 090 7084; fax: +39 011 090 7099.
E-mail addresses: luca.cagliero@polito.it (L. Cagliero), silvia.chiusano@polito.it (S. Chiusano), paolo.garza@polito.it (P. Garza), giulia.bruno@polito.it (G. Bruno).

http://dx.doi.org/10.1016/j.knosys.2015.04.023
0950-7051/

maximum number of itemsets per pattern set, or (ii) enforcing a minimum percentage of data that must be covered by each mined pattern set. The former constraint, called cardinality constraint, can be exploited to discard very large and thus unmanageable pattern sets. The latter constraint, named coverage constraint, prevents the extraction of pattern sets representing a small and thus not significant portion of data. Note that our goal is to characterize data using recurrent patterns, rather than pinpointing abnormal (rare) patterns. To efficiently perform pattern set mining with schema-based constraint, we present a new Apriori-based algorithm [5], namely *CO*strained *PA*ttern *S*et mining algorithm (COPAS), which adopts a level-wise approach to discovering itemsets and pattern sets at the same time. The COPAS algorithm pushes the newly proposed schema-based constraint, in conjunction with one of the two traditional constraints (cardinality or coverage, based on users needs), deep into the mining process. In such a way, the pattern sets of interest can be extracted in a single extraction without the need for postprocessing. The result can be directly explored by domain experts for advanced analyses or further processed by using ad hoc strategies.

The paper is organized as follows. Section 2 presents a motivating example. Section 3 compares our work with previous approaches. Section 4 states the mining problem addressed by the paper. Section 5 presents the COPAS algorithm, while Section 6 describes the experiments performed. Finally, Section 7 draws conclusions and discusses future work.

2. Motivating example

A company would like to plan advertising campaigns targeted to customers located in Italy according to their most peculiar features. To personalize advertisements the company clusters customers into segments, which consist of subsets of customers having similar features. However, deciding the features (or the feature combinations) according to which customers should be clustered is a non-trivial task in large databases.

Table 1 collects some relevant information about the customers under analysis. Each row corresponds to a different customer and it reports the values of a subset of attributes, in particular the city of provenance, gender, year of birth, and job. To achieve their goal, company analysts mine from the input data itemsets like {(City, Turin), (Gender, M)}, where each itemset is characterized by a given schema (e.g., {City, Gender}). To guarantee itemset relevance, the mined itemsets must hold for at least 30% of the customers, i.e., their frequency of occurrence (support) in the source dataset must be equal to or above a given threshold *minsup* = 30%. Then, itemsets with the same schema are analyzed together because they represent the same data facet. For the sake of simplicity, let us consider the itemsets related to pairs of attributes. Since analysts do not know a priori what are the most significant schemata to consider, they have to (i) generate all the itemsets satisfying *minsup*, (ii) cluster the mined itemsets into pattern sets according to their schema, and (iii) rank the pattern sets by decreasing coverage (i.e., the percentage of customers in the dataset for which any itemset in the pattern set holds) and discard those not satisfying a minimum coverage threshold (e.g., *mincov* = 60%). At Step (ii) the aforesaid procedure generates 24 pattern sets, because all the possible combinations of the four data attributes are considered. However, only half of them satisfy the coverage constraint and thus they are considered for planning advertising campaigns.

Table 1
Example relational dataset.

Rid	*City*	*Gender*	*Year*	*Job*
1	Turin	F	1980	Teacher
2	Turin	M	1945	Lawyer
3	Turin	M	1945	Lawyer
4	Milan	F	1957	Teacher
5	Rome	M	1976	Clerk
6	Milan	F	1978	Teacher

Table 2
Pattern sets satisfying the schema-based and the minimum coverage constraints mined from the dataset in Table 1 (*minsup* = 30%, *mincov* = 60%).

Pattern set	Itemsets (support)	Coverage (%)
P_{City}	{(City, Turin)} (50%) {(City, Milan)} (33.3%)	83.3
P_{Gender}	{(Gender, M)} (50%) {(Gender, F)} (50%)	100
P_{Job}	{(Job, Teacher)} (50%) {(Job, Lawyer)} (33.3%)	83.3
$P_{City,Gender}$	{(City, Turin), (Gender, M)} (33.3%) {(City, Milan), (Gender, F)} (33.3%)	66.6
$P_{City,Job}$	{(City, Turin), (Job, Lawyer)} (33.3%) {(City, Milan), (Job, Teacher)} (33.3%)	66.6
$P_{Gender,Job}$	{(Gender, F), (Job, Teacher)} (50%) {(Gender, M), (Job, Lawyer)} (33.3%)	83.3

Our approach allows analysts to efficiently extract the subset of pattern sets of interest without generating all the possible itemsets and itemset combinations. Table 2 reports the subset of mined pattern sets. Among the pattern sets related to pairs of attributes, the pattern set with highest coverage is {*Gender*, *Job*} (83.3%). Each itemset in the pattern represents a combination of customer gender and job, which targets a specific subset of customers. For example, according to customer gender and job, analysts could figure out different advertising policies for female teachers and male lawyers. Together, the previously mentioned segments cover 83% of the customers thus represent potential targets of advertising campaigns.

3. Related works

Pattern set mining entails discovering groups of itemsets that satisfy a set of global constraints. Instead of selecting patterns based upon their individual merits, global constraints evaluate each pattern set as a whole [6]. Pattern set mining approaches focus on (i) selecting the pattern set that maximizes a certain global quality measure [6–14] or (ii) discovering all the pattern sets that satisfy a given constraint [4,15,16]. Examples of problems related to Task (i) are (a) database tiling [8], which concerns the extraction of the pattern set that covers all the dataset transactions, (b) data compression based on the Minimum Description Length (MDL) principle [12], and (c) pattern set selection by means of constraint programming techniques [9]. Unlike [6–14], this work addresses the more general Task (ii), i.e., it selects not only the best pattern set but a set of potentially interesting pattern sets.

In [4] the authors formally introduce many different global constraints. Rather than performing pattern set mining as a postprocessing step that follows the traditional itemset mining task [1], in [15,16] the authors formulate the global constraints directly on the entire itemset space and then accomplish the pattern set mining task using constraint programming techniques. An overview of the constraints used in pattern set mining is given in [4]. For all the previously proposed constraints the selection of a pattern set depends only on the characteristics of its itemsets. Hence, a pattern set cannot be selected based upon the comparison with other candidate pattern sets. Unlike [4,15,16], this paper proposes a new constraint whereby pattern sets are selected not only based upon their own characteristics but also based upon those of other pattern sets. Specifically, the newly proposed schema-based