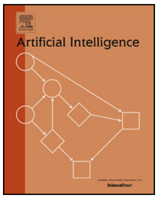
Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems

Willy Ugarte [a], Patrice Boizumault [a], Bruno Crémilleux [a,*], Alban Lepailleur [b], Samir Loudni [a], Marc Plantevit [c], Chedy Raïssi [d], Arnaud Soulet [e]

[a] GREYC (CNRS UMR 6072), University of Caen, F-14032 Caen, France
[b] CERMN (UPRES EA 4258 – FR CNRS 3038 INC3M), University of Caen, F-14032 Caen, France
[c] Université de Lyon, CNRS, Université Lyon 1, LIRIS (UMR5205), F-69622, France
[d] INRIA Nancy Grand-Est, France
[e] LI (EA 2101), Université François Rabelais de Tours, F-41029 Blois, France

ARTICLE INFO

Article history:
Received in revised form 7 April 2015
Accepted 14 April 2015
Available online xxxx

Keywords:
Skypatterns
Pattern mining
Constraint programming
Dynamic CSP
User preferences

ABSTRACT

Data mining is the study of how to extract information from data and express it as useful knowledge. One of its most important subfields, pattern mining, involves searching and enumerating interesting patterns in data. Various aspects of pattern mining are studied in the theory of computation and statistics. In the last decade, the pattern mining community has witnessed a sharp shift from efficiency-based approaches to methods which can extract more meaningful patterns. Recently, new methods adapting results from studies of economic efficiency and multi-criteria decision analyses such as Pareto efficiency, or skylines, have been studied. Within pattern mining, this novel line of research allows the easy expression of preferences according to a dominance relation. This approach is useful from a user-preference point of view and tends to promote the use of pattern mining algorithms for non-experts. We present a significant extension of our previous work [1,2] on the discovery of skyline patterns (or *"skypatterns"*) based on the theoretical relationships with condensed representations of patterns. We show how these relationships facilitate the computation of skypatterns and we exploit them to propose a flexible and efficient approach to mine skypatterns using a dynamic constraint satisfaction problems (CSP) framework.

We present a unified methodology of our different approaches towards the same goal. This work is supported by an extensive experimental study allowing us to illustrate the strengths and weaknesses of each approach.

1. Introduction

The process of extracting useful patterns from data, called *pattern mining*, is an important subfield of data mining, and has been used in a wide range of applications and domains such as bioinformatics [3], chemoinformatics [4], social network analysis [5], web mining [6] and network intrusion detection [7]. Since the key papers of Agrawal et al. [8], Mannila et al. [9], a considerable number of patterns, such as itemsets, strings, sequences, trees and graphs, have been studied and

[*] Corresponding author.
E-mail address: bruno.cremilleux@unicaen.fr (B. Crémilleux).

http://dx.doi.org/10.1016/j.artint.2015.04.003

used in real-world applications. Nowadays, many pattern extraction problems like subgroup discovery [10], discriminative pattern mining [11], and tiling [12] are understood from both theoretical and computational perspectives.

Most existing pattern mining approaches enumerate patterns with respect to a given set of constraints that range from simple to complex. For instance, given a transaction database, a well-known pattern mining task is to enumerate all itemsets (i.e. sets of items) that appear in at least s transactions. However, the output of pattern mining operations can be extremely large even for moderately sized datasets. For instance, in the worst case, the number of frequent itemsets is exponential in the number of the items in the dataset.

So far, the community has expended much effort on developing sophisticated algorithms which push the constraints deep into the mining process [13]. But also in on compression (i.e. reduction) techniques to limit the number of output patterns depending on the application contexts [14–16]. The pattern mining community, however, has paid less attention to combining mining constraints. In practice, many constraints entail choosing threshold values such as the well-used minimal frequency. This notion of "*thresholding*" has serious drawbacks. Unless specific domain knowledge is available, the choice is often arbitrary and may lead to a very large number of extracted patterns which can reduce the success of any subsequent data analysis. This drawback is even more pronounced when several thresholds have to be combined. A second drawback is the *stringent enumeration aspect*: a pattern is either above or below a threshold. But what about patterns that respect only some thresholds? Should they be discarded? It is often very difficult to apply *subtle selection* mechanisms. There are very few works such as [17,18] which propose to introduce a softness criterion into the mining process. Other studies attempt to integrate user preferences into the mining task in order to limit the number of extracted patterns such as the *top-k* pattern mining approaches [19,20]. By associating each pattern with a *rank score*, this approach returns an ordered list of the k patterns with the highest score to the user. However, combining several measures in a single scoring function is difficult and the performance of top-k approaches is often sensitive to the size of the datasets and to the threshold value, k.

We present a unified methodology of two approaches that aim to make the results of pattern mining *useful from a user-preference point of view*. To this end, we integrate into the pattern discovery process the idea of skyline queries [21] in order to mine *skyline patterns* in a threshold-free manner. Such queries have attracted considerable attention due to their importance in multi-criteria decision making and economics where they are usually called "*Pareto efficiency or optimality queries*". Briefly, in a multidimensional space where a preference is defined for each dimension, a point a dominates another point b if a is better (i.e. more preferred) than b in at least one dimension, and a is not worse than b on every other dimension. For example, a user selecting a set of patterns may prefer a pattern with a high frequency, a large length and a high confidence. In this case, we say that pattern a *dominates* another pattern b if $a.frequency \geq b.frequency$, $a.length \geq b.length$, $a.confidence \geq b.confidence$, where at least one strict inequality holds. Given a set of patterns, the skyline set contains the patterns that are not dominated by any other pattern.

Skyline pattern mining is interesting for several reasons. First, skyline processing does not require any threshold selection. In addition, for many pattern mining applications it is often difficult (or impossible) to find a reasonable global ranking function. Thus the idea of finding all optimal solutions in the pattern space with respect to multiple preferences is appealing. Second, the formal property of dominance satisfied by the skyline pattern defines a global interestingness measure with semantics easily understood by the user. These semantics are discussed at length in the economics literature, where the Pareto efficiency is applied to the selection of alternatives in resource distributions. However, while this notion of skylines has been extensively developed in engineering and database applications, it has remained unused for data mining purposes until recently [1]. Thirdly, skyline pattern mining is appealing from an efficiency and usability point of view. The authors of [22] established a loose upper-bound on the average number of skyline tuples $O((\ln n)^{d-1})$ (with n tuples and d dimensions) which contrasts with the usual worst-case number of possible itemsets $O(2^{|\mathcal{I}|})$ (where $|\mathcal{I}|$ represents the cardinality of the set of items).

Contributions and roadmap We present significant extensions of our recent papers [1,2] on the discovery of skyline patterns, or "*skypatterns*". First, we detail a *static* method (called Aetheris) based on the theoretical relationships with condensed representations of patterns (representations which return a subset of the patterns having the same expressiveness as the whole set of patterns [23]). Second, we describe a *dynamic* method (called CP+Sky) which involves a continuous refinement of the skyline constraints based on the extracted patterns. This is achieved through a dynamic CSP (Constraint Satisfaction Problems) framework (denoted by DynCSP). Third, the key notion of "*skylineability*" which constitutes the cornerstone of our two methods is explained in more detail. Finally, we present an extensive empirical study which includes a wide range of datasets and comparisons of our techniques. This enables us to draw some lessons about the strengths and weaknesses of each method and to better understand the advantages/weaknesses of the CSP machinery (see Sections 7.1.2 and 7.1.3).

The rest of this paper is organized as follows. Section 2 surveys the works related to skyline pattern analysis. Section 3 introduces some basic definitions, the formal problem statement and an overview of our work. The key notion of skylineability is then studied in Section 4. Section 5 discusses the computation of condensed representation of patterns for skypattern queries. Section 6 discusses skylineability but within a DynCSP framework. We report an empirical study on several datasets and a case study from the chemoinformatics domain in Section 7. Finally, Section 8 discusses the learnt lessons.

Please cite this article in press as: W. Ugarte et al., Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems, Artificial Intelligence (2015), http://dx.doi.org/10.1016/j.artint.2015.04.003