Theoretical Computer Science 618 (2016) 107–121

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Extended dualization: Application to maximal pattern mining

Lhouari Nourine [a,*], Jean-Marc Petit [b]

[a] Université Blaise Pascal, CNRS, LIMOS, France
[b] Université de Lyon, CNRS, INSA-Lyon, LIRIS, France

ARTICLE INFO

Article history:
Received 13 March 2015
Received in revised form 17 November 2015
Accepted 16 January 2016
Available online 27 January 2016
Communicated by P. Widmayer

Keywords:
Hypergraph dualization
Enumeration algorithms
Patterns mining

ABSTRACT

The dualization in arbitrary posets is a well-studied problem in combinatorial enumeration and is a crucial step in many applications in logics, databases, artificial intelligence and pattern mining.

The objective of this paper is to study *reductions* of the dualization problem on arbitrary posets to the dualization problem on boolean lattices, for which output quasi-polynomial time algorithms exist. Quasi-polynomial time algorithms are algorithms which run in $n^{o(\log n)}$ where n is the size of the input and output. We introduce *convex embedding* and *poset reflection* as key notions to characterize such reductions. As a consequence, we identify posets, which are not boolean lattices, for which the dualization problem remains in quasi-polynomial time and propose a classification of posets with respect to dualization. From these results, we study how they can be applied to maximal pattern mining problems. We deduce a new classification of pattern mining problems and we point out how known problems involving sequences and conjunctive queries patterns, fit into this classification. Finally, we explain how to adapt the seminal DUALIZE & ADVANCE algorithm to deal with such patterns.

As far as we know, this is the first contribution to explicit non-trivial reductions for studying the hardness of maximal pattern mining problems and to extend the DUALIZE & ADVANCE algorithm for complex patterns.

1. Introduction

The dualization in arbitrary finite[1] partially ordered sets (poset for short) is well-studied in combinatorial enumeration such as minimal transversals of a hypergraph, the blocker of a clutter, minimal dominating sets and maximal cliques of a graph. The dualization problem is the following: Given a compact representation of a poset P and an antichain $\mathcal{B}^+$ of P, find another antichain $\mathcal{B}^-$ of P such that the union of the ideal induced by $\mathcal{B}^+$ and the filter induced by $\mathcal{B}^-$ is exactly P.[2] The dualization problem has been popularized largely through the work on artificial intelligence and pattern mining [9,12,21,10,18,19], where P is a boolean lattice and $\mathcal{B}^+$ is given by a monotonically decreasing predicate. This link has been done through the well known DUALIZE & ADVANCE algorithm [21,15,3,4,23,22]. Many authors have investigated the existence of an output-polynomial time algorithm for listing without duplications the antichain $\mathcal{B}^-$. An output-polynomial algorithm is an algorithm whose running time is bounded by a polynomial depending on the sum of the sizes of the input and output. The

* Corresponding author.
E-mail addresses: nourine@isima.fr (L. Nourine), jean-marc.petit@insa-lyon.fr (J-M. Petit).

[1] It also works for infinite partially ordered sets that are well ordered, i.e. all antichains are finite.

[2] $\mathcal{B}^+$ and $\mathcal{B}^-$ are dual sets, also known as blocker and anti-blocker or positive and negative borders.

http://dx.doi.org/10.1016/j.tcs.2016.01.020
0304-3975/

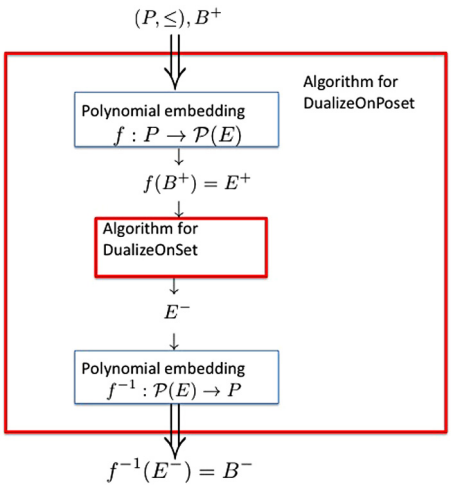

Fig. 1. Reduction from **DualizeOnPoset** to **DualizeOnSet**.

existence of an output-polynomial algorithm for the enumeration of minimal transversals of a hypergraph is a widely open question and is closely related to many data mining problems [11].

In this paper, we are interested in characterizing posets for which the dualization is equivalent to the enumeration of minimal transversals of a hypergraph. The strategy is based on *reductions* of the dualization problem on arbitrary posets to the dualization problem on boolean lattices. On posets, the dualization problem can be stated as follows:

DualizeOnPoset
Input: A representation of a poset $(P, \leq)$, B^+ an antichain of P.
Output: B^- such that (B^+, B^-) are dual sets.

On boolean lattices, it is stated as follows:

DualizeOnSet
Input: A finite set E, B^+ an antichain of $\mathcal{P}(E)$ (the powerset of E).
Output: B^- such that (B^+, B^-) are dual sets in $\mathcal{P}(E)$.

The complexity of **DualizeOnSet** is known to be quasi-polynomial time while the complexity of **DualizeOnPoset** is still open in most posets (for example, lattice) [11]. In this setting, we are interested in studying the *reduction* from **DualizeOnPoset** to **DualizeOnSet**, i.e. under which conditions **DualizeOnSet** is polynomially equivalent to **DualizeOnPoset**. Notice that reductions for the hardness of enumeration problems are not well established as for decision problems. In this paper, we consider only polynomial time reductions, inspired from classic polynomial reductions of decision problems (cf. Fig. 1).

Contribution on dualization We introduce *convex embedding* and *poset reflection* as key notions to characterize such reductions. As a consequence, we identify posets, which are not boolean lattices, for which the dualization problem remains quasi-polynomial time (cf. Fig. 1) and propose a classification of posets with respect to dualization.

From these results, we study how they can be applied to maximal (or more specific) pattern mining problems. Mining interesting patterns in databases has been extensively studied in the data mining community over the last twenty years, from association rules and frequent itemset mining to frequent graph mining or functional dependency inference to mention a few. For studying their complexity, the underlying dualization problem has been identified as the main bottleneck in [21, 15,23]. Roughly speaking, for a partial order $(P, \leq)$ (representing patterns) and some monotonically decreasing predicate Q over P, the dualization consists in identifying all maximal elements of P verifying Q from all minimal elements of P not verifying Q, and vice versa.

The seminal work of Mannila and Toivonen [21] proposes a general framework, especially they classify pattern mining problems that are (isomorphically) equivalent to frequent itemset mining (FIM). Nevertheless, the isomorphism requirement is too restrictive for many "complex" patterns such as sequences, episodes or graphs to mention a few. The ambition of this paper is to take into account such complex patterns and to propose a new framework for studying their complexity. From a practical point of view, the idea is to be able to re-use as much as possible the myriad of techniques and algorithms developed for FIM to such complex patterns.

Contribution on maximal pattern mining From the contributions on dualization, we deduce a new classification of pattern mining problems. We point out how known problems involving sequences and conjunctive queries patterns, fit into this