Beyond Hypergraph Dualization

Lhouari Nourine, Jean-Marc Petit

Title:	Beyond Hypergraph Dualization
Name:	Lhouari Nourine[1], Jean Marc Petit[2]
Affil./Addr. 1:	Clermont-Université, CNRS, Université Blaise Pascal, LIMOS, France.
Affil./Addr. 2:	Université de Lyon, CNRS, INSA Lyon, LIRIS, France
Keywords:	Dualization; Hypergraph Transversal; Partially ordered set; Enumeration
SumOriWork:	2012; Nourine, Petit

Beyond Hypergraph Dualization

Lhouari Nourine[1], Jean Marc Petit[2]

[1] Clermont-Université, CNRS, Université Blaise Pascal, LIMOS, France.
[2] Université de Lyon, CNRS, INSA Lyon, LIRIS, France

Years aud Authors of Summarized Original Work

2012; Nourine, Petit

Keywords

Dualization; Hypergraph Transversal; Partially ordered set; Enumeration

Problem Definition

This problem concerns hypergraph dualization and generalization to poset dualization.

A *hypergraph* $\mathcal{H} = (V, \mathcal{E})$ consists of a finite collection $\mathcal{E}$ of sets over a finite set V, i.e. $\mathcal{E} \subseteq \mathcal{P}(V)$ (the powerset of V). The elements of $\mathcal{E}$ are called *hyperedges*, or simply *edges*. A hypergraph is said simple if none of its edges is contained within another. A *transversal* (or *hitting set*) of $\mathcal{H}$ is a set $T \subseteq V$ that intersects every edge of $\mathcal{E}$. A transversal is *minimal* if it does not contain any other transversal as a subset. The set of all minimal transversal of $\mathcal{H}$ is denoted by $Tr(\mathcal{H})$. The hypergraph $(V, Tr(\mathcal{H}))$ is called the *transversal hypergraph* of $\mathcal{H}$. Given a simple hypergraph $\mathcal{H}$, the hypergraph dualization problem (Trans-Enum for short) concerns the enumeration without repetitions of $Tr(\mathcal{H})$.

The Trans-Enum problem can also be formulated as a dualization problem in posets. Let $(P, \leq)$ be a poset (i.e. $\leq$ is a reflexive, antisymmetric, and transitive relation on the set P). For $A \subseteq P$, $\downarrow A$ (resp. $\uparrow A$) is the downward (resp. upward) closure of A under the relation $\leq$ (i.e. $\downarrow A$ is an ideal and $\uparrow A$ a filter of $(P, \leq)$). Two antichains $(\mathcal{B}^+, \mathcal{B}^-)$ of P are said to be *dual* if $\downarrow \mathcal{B}^+ \cup \uparrow \mathcal{B}^- = P$ and $\downarrow \mathcal{B}^+ \cap \uparrow \mathcal{B}^- = \emptyset$. Given an implicit description of a poset P and an antichain $\mathcal{B}^+$ (resp. $\mathcal{B}^-$) of P, the poset dualization problem (Dual-Enum for short) enumerates the set $\mathcal{B}^-$ (resp. $\mathcal{B}^+$), denoted by $Dual(\mathcal{B}^+) = \mathcal{B}^-$ (resp. $Dual(\mathcal{B}^-) = \mathcal{B}^+$). Notice that the function dual is self-dual or idempotent, i.e. $Dual(Dual(\mathcal{B})) = \mathcal{B}$.

Trans-Enum is a particular case of Dual-Enum. Indeed, consider P be the poset $(\mathcal{P}(V), \subseteq)$ for some set V. Then for every dual set $(\mathcal{B}^+, \mathcal{B}^-)$ of P, we have $\mathcal{B}^- = Tr(\overline{\mathcal{B}^+}) = Dual(\mathcal{B}^+)$, or equivalently $\mathcal{B}^+ = \overline{Tr(\mathcal{B}^-)} = Dual(\mathcal{B}^-)$ with $\overline{\mathcal{E}} = \{V \setminus E \mid E \in \mathcal{E}\}$ where $\mathcal{E} \subseteq \mathcal{P}(V)$.

Now we ask the following question: For which posets Dual-Enum can be reduced to Trans-Enum? To do so, we introduce the notions of duality gap, convex embedding and poset reflexion.

Let $(P, \leq P)$ and $(Q, \leq_Q)$ be two posets and $f : P \to Q$ an *injective reflection*, i.e. for all $x, y \in P, f(x) \leq_Q f(y)$ implies $x \leq_P y$. Notice that the reflection f preserves incomparability i.e. if x and y are incomparable in P then $f(x)$ and $f(y)$ are incomparable in Q. Therefore, for every dual set $(\mathcal{B}^+, \mathcal{B}^-)$ of P, $Dual(f(\mathcal{B}^+))$ contains $f(\mathcal{B}^-)$. The difference between the size of $Dual(f(\mathcal{B}^+))$ and the size of $f(\mathcal{B}^-)$ is a positive integer, called the *duality gap*. We speak about *weak duality* when the gap is strictly positive, *strong duality* otherwise.

Duality gaps are important in enumeration problems because they provide an upper bound on the difference between the number of enumerated solutions and the number of solutions of the original problem.

Key Results

Trans-Enum has been intensively studied in the last two decades, and several results show that it is equivalent to many problems in computer science area (see the paper by Eiter and Gottlob [3]). The question whether Trans-Enum admits an output-polynomial time algorithm is still open. In fact, despite the number of papers on Trans-Enum, the best known algorithm is the one by Fredman and Khachiyan [8] which runs in time $O(n^{log(n)})$ where n is the size of the hypergraph plus the number of minimal transversals. Others results on complexity can be found in [5; 14; 6; 12; 11]. For general posets, it is shown in [7] that the dualization over the products of some posets can be done with the same complexity as Trans-Enum. Recently, Nourine and Petit [16] have investigated dualization problems in general posets for which the duality gap is bounded by a polynomial.

Strong Duality

The following characterization theorem of the zero gap, is a reformulation of a known result in [15; 10], where the poset Q is the powerset for some set.

Theorem 1. *Let $(P, \leq_P)$ and $(Q, \leq_Q)$ be two posets. Then the duality gap is zero iff there exists a map $f : P \to Q$ such that f is a bijective embedding, i.e. for all $x, y \in P$ $f(x) \leq_Q f(y)$ iff $x \leq_P y$.*

Many instances of problems have such a property, for example frequent itemsets, monotone boolean functions, minimal keys, inclusion dependencies or minimal dominating sets [15; 10; 13]. Nevertheless, the bijective embedding between two posets does not always exist. In the following we give a relaxation of the bijection embedding in order to capture some polynomial reductions between enumeration problems.

Weak Duality

Let $(P, \leq_P)$ and $(Q, \leq_Q)$ be posets. A function $f : P \to Q$ is a *convex embedding* if for all $x, y \in P$ and $z \in Q$, $x \leq_P y$ iff $f(x) \leq_Q f(y)$ and $f(x) \leq_Q z \leq_Q f(y)$ implies there exists $t \in P$ such that $f(t) = z$.

The following result can be seen as a relaxation of the bijective embedding given in Theorem 1.

Proposition 1. *Let $(P, \leq_P)$ and $(Q, \leq_Q)$ be two posets and $f : P \to Q$ a convex embedding. Then there exist two antichains $\mathcal{B}_0^+$, $\mathcal{B}_0^-$ of Q such that $P \setminus \{\bot_P\}$ is isomorphic to $Q \setminus (\downarrow \mathcal{B}_0^+ \cup \uparrow \mathcal{B}_0^-)$, where $\bot_P$ is the bottom of P if it exists. Furthermore, the duality gap is bounded by $\mid \mathcal{B}_0^+ \mid + \mid \mathcal{B}_0^- \mid$.*

Complexity

For strong duality, [15; 10] points out how the result of Fredman and Khachiyan [8] can be re-used to devise an incremental quasi-polynomial time algorithm, called `Dualize and Advance`, for some pattern mining problems. For weak duality, whenever the duality gap remains polynomial in the size of the problem and $(Q, \leq_Q)$ isomorphic to $(\mathcal{P}(E), \subseteq)$ for some set E, the `Dualize and Advance` algorithm can be re-used with the same complexity if the following assumptions hold:

1. The reflexion f of $(P, \leq)$ to $(\mathcal{P}(E), \subseteq)$ and its inverse, is computable in polynomial time.
2. Given two elements $x, y \in P$, checking $x \leq y$ is polynomial time.

Applications

The hypergraph dualization is a crucial step in many applications in logics, databases, artificial intelligence and pattern mining [11; 3; 8; 4; 15], especially for hypergraphs, i.e. boolean lattices. The main application domain concerns pattern mining problems, i.e. the identification of maximal interesting patterns in database by asking membership queries (predicate) to a database. In the rest of this section, we give two examples of pattern mining problems related to Dual-Enum and weak duality.

Frequent conjunctive queries

We consider the problem statement defined in [9]. Let $\mathbf{R} = \{R_1, \ldots, R_n\}$ be a database schema, $\mathcal{D}$ the domain of $\mathbf{R}$ and $sch(\mathbf{R}) = \{R_i.A | R_i \in \mathbf{R}, A \in R_i\}$. A (simple) conjunctive queries Q over $\mathbf{R}$ is of the form $\pi_X(\sigma_F(R_1 \times \ldots \times R_n))$ ($\pi_X(\sigma_F)$ for short) where $X \subseteq sch(\mathbf{R})$ and F a conjunction of equalities of the form $R_i.A = R_j.B$ or $R_i.A = c$ with $R_i.A, R_j.B \in sch(\mathbf{R})$ and $c \in \mathcal{D}$. Let $\mathcal{Q}_r$ be the set of all possible conjunctive queries over $\mathbf{R}$. For a given database d over $\mathbf{R}$, we note $Adom(d) \subseteq \mathcal{D}$ the active domain of d and $Q(d)$ the result of the evaluation of Q against d. We note $\mathcal{F}$ the finite set of all possible selection formula over $\mathbf{R}$ and $Adom(d)$, i.e. $\mathcal{F} = \{\{A, B\} \mid A \neq B, A \in \mathbf{R}, B \in \mathbf{R} \cup Adom(d)\}$.

Let Q_1, Q_2 be two conjunctive queries over $\mathbf{R}$. Q_1 is *contained* in Q_2, denoted $Q_1 \subseteq Q_2$, if for every database d over $\mathbf{R}$, $Q_1(d) \subseteq Q_2(d)$. Q_1 is *diagonally contained* in Q_2, denoted $Q_1 \subseteq^{\Delta} Q_2$, if Q_1 is contained in a projection of Q_2, i.e. for instance $Q_1 \subseteq \pi_X(Q_2)$ The frequency of $\pi_X(\sigma_F)$ in d is defined by $|\pi_X(\sigma_F)(d)|$. A query $\pi_X(\sigma_F)$ is *frequent* in d with respect to a given threshold ϵ if $|\pi_X(\sigma_F)(d)| \geq \epsilon$. The frequency is anti-monotonic with respect to $\subseteq^{\Delta}$ [9].

Proposition 2. *Let $Q_1 = \pi_{X_1}(\sigma_{F_1})$ and $Q_2 = \pi_{X_2}(\sigma_{F_2})$ be two queries of $\mathcal{Q}_r$. Then $Q_1 \subseteq^{\Delta} Q_2$ iff $X_1 \subseteq X_2$ and $F_2 \subseteq F_1$. Equivalently, $Q_1 \subseteq^{\Delta} Q_2$ iff $X_1 \cup (\mathcal{F} \setminus F_1) \subseteq X_2 \cup (\mathcal{F} \setminus F_2)$.*

From Proposition 2, $f : \mathcal{Q}_r \to \mathcal{P}(\mathbf{R} \cup \mathcal{F})$ with $f(\pi_X(\sigma_F)) = X \cup (\mathcal{F} \setminus F)$ is a bijective embedding. Thus $\mathcal{Q}_r$ ordered under $\subseteq^{\Delta}$ is a boolean lattice and Theorem 1 can be applied. It is interesting to consider the subclass of $\mathcal{Q}_r$ restricted to *consistent queries*, i.e. queries for which there exists at least one database such that their evaluations return values different from zero. For instance, $\sigma(B = 1 \wedge B = 2)$ or $\sigma(A = B \wedge A = 1 \wedge B = 2)$ are not consistent. Let us consider the set $\mathcal{Q}_C \subset \mathcal{Q}_r$ of all consistent queries.

Lemma 1. *Let $Q_1 = \pi_{X_1}(\sigma_{F_1})$ and $Q_2 = \pi_{X_2}(\sigma_{F_2})$ be two queries of $\mathcal{Q}_r$ such that $Q_1 \subseteq^{\Delta} Q_2$. Then Q_2 is consistent implies Q_1 is consistent.*

Notice that the restriction of f to $\mathcal{Q}_C$ is still a convex embedding, but no longer bijective. More interestingly, the associated duality gap is not polynomial. Indeed, $B_0^+ = \emptyset$ but B_0^- has a size exponential in the size of $\mathbf{R} \cup Adom(d)$ since the number of selections of the form $\sigma(A_1 = A_2 \wedge \ldots \wedge A_{n-1} = A_n \wedge A_1 = v \wedge A_n = v')$ is exponential in the number of attributes.

Rigid sequences

Let us consider sequences with or without wildcard (denoted $\star$), see e.g. [1]. Let Σ be an alphabet and $\star \notin \Sigma$. A rigid sequence $s[n]$ is a word of size n of $(\Sigma \cup \{\star\})^*$ such that $s[1] \neq \star$ and $s[n] \neq \star$. The set of all rigid sequences of size at most n are denoted by Σ_R^n and the empty sequence by ϵ. Let $s[l], t[k] \in \Sigma_R^n$. We consider the following classical (prefix and factor) partial orders on rigid sequences.

- $s \sqsubseteq_f t$, if there exists $j \in [1..k]$ such that for every $i \in [1..l]$, either $s[i] = t[j + i - 1]$ or $s[i] = \star$ (factor).
- $s \sqsubseteq_p t$, if for every $i \in [1..l]$, either $s[i] = t[i]$ or $s[i] = \star$ (prefix).

The following theorem shows that the duality gap between the dualization in prefix posets of rigid sequences and Trans-Enum is bounded by a polynomial in n and $\mid \Sigma \mid$.

Theorem 2. *[16] Let $f : (\Sigma_R^n \setminus \{\epsilon\}, \sqsubseteq_p) \to (\mathcal{P}(\{1, \ldots, n\} \times \Sigma), \subseteq)$ be a function defined by $f(s) = \{(i, s[i]) \mid s[i] \neq \star, i \leq n\}$. Then f is a convex embedding with $\mathcal{B}_0^+ = \{\{(i, x) \mid x \in \Sigma, i \in [2..n]\}\}$ and $\mathcal{B}_0^- = \{\{(1, x), (1, y)\} \mid x, y \in \Sigma, x \neq y\} \cup \{\{(1, x), (i, y), (i, z)\} \mid x, y, z \in \Sigma, y \neq z, i \in [2..n]\}$.*

Proposition 3. *[16] There is a poset reflection $f : (\Sigma_R^n, \sqsubseteq_f) \to (\Sigma_R^n, \sqsubseteq_p)$ with a duality gap bounded by a polynomial in n.*

Using Theorem 2 and Proposition 3 we conclude that the duality gap between the dualization in factor posets of rigid sequences and Trans-Enum is bounded by a polynomial the size of Σ and n [16].

Open Problems

1. The challenging question is to find an output-polynomial time algorithm for Trans-Enum.
2. Lattices are a particular class of posets. For example, the dualization over product of chains can be done with the same complexity as Trans-Enum which is equivalent to dualization in boolean lattices. For distributive lattices class which contains boolean lattice and the product of chains, the dualization is open.

3. Many connections have to be done between Trans-Enum and graph theory problems, such as minimal dominating sets [13].
4. Many problems in data mining can be formulated as dualization in posets, e.g. frequent subgraphs or frequent subtrees. An interesting direction is to identify posets for which the dualization is equivalent to Trans-Enum.

URLs to Code and Data Sets

Program Codes and Instances for Hypergraph Dualization can be found on the Takeaki Uno's webpage at `http://research.nii.ac.jp/~uno/dualization.html`. Some pattern mining problems, reducible to Trans-Enum with strong duality, can be found on the iZi webpage at `http://liris.cnrs.fr/izi/`.

Cross-References

Minimal Dominating Set Enumeration
Polynomial Time Algorithms for Dualization; Solvable Cases
Frequent Itemset Mining

Recommended Reading

1. H. Arimura, T. Uno Polynomial-delay and polynomial-space algorithms for mining closed sequences, graphs, and pictures in accessible set systems. In: SDM. (2009) 1087–1098
2. E. Boros and K. Makino. A fast and simple parallel algorithm for the monotone duality problem. In S. Albers, A. Marchetti-Spaccamela, Y. Matias, S. E. Nikoletseas, and W. Thomas, editors, ICALP (Part I), LNCS 5555, pp. 183-194. Springer, 2009.
3. T. Eiter and G. Gottlob. Identifying the minimal transversals of a hypergraph and related problems. *SIAM J. Comput.*, 24(6):1278–1304, 1995.
4. T Eiter, G. Gottlob, and K Makino. New results on monotone dualization and generating hypergraph transversals. *SIAM J. Comput.*, 32:514–537, February 2003.
5. T. Eiter, G. Gottlob, and K. Makino. New results on monotone dualization and generating hypergraph transversals. SIAM J. Comput., 32(2):514-537, 2003.
6. K.M. Elbassioni On the complexity of monotone dualization and generating minimal hypergraph transversals. Discrete Applied Mathematics, 156(11):2109-2123, 2008.
7. K.M. Elbassioni Algorithms for dualization over products of partially ordered sets. SIAM J. Discrete Math. **23**(1) (2009) 487–510
8. M. L. Fredman and L. Khachiyan. On the complexity of dualization of monotone disjunctive normal forms. *J. Algorithms*, 21(3):618–628, 1996.
9. Bart Goethals and Jan Van den Bussche. Relational association rules: Getting warmer. In *Pattern Detection and Discovery*, pages 125–139, 2002.
10. D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, R.S. Sharm Discovering all most specific sentences. ACM Trans. Database Syst. **28**(2) (2003) 140–174
11. G. Gottlob Deciding monotone duality and identifying frequent itemsets in quadratic logspace. In Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2013, New York, NY, USA - June 22 - 27, 2013.
12. M. Hagen Algorithmic and Computational Complexity Issues of MONET. Cuvillier Verlag, 2008.
13. M. Moustapha Kanté, V. Limouzy, A. Mary, and L. Nourine On the Enumeration of Minimal Dominating Sets and Related Notions Arxiv 2014.
14. D. J. Kavvadias and E. C. Stavropoulos Monotone Boolean Dualization is in co-NP[log2n]. Information Processing Letters, 85:1-6, 2003.
15. H. Mannila and H. Toivonen Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery **1**(3) (1997) 241–258
16. L. Nourine and J-M. Petit Extending Set-Based Dualization: Application to Pattern Mining. In Press, I., ed.: ECAI 2012. (August 2012)