Efficient Closed Pattern Mining in Strongly Accessible Set Systems*
(Extended Abstract)

Mario Boley[1], Tamás Horváth[1], Axel Poigné[1], and Stefan Wrobel[1,2]

[1] Fraunhofer IAIS, Schloss Birlinghoven, Sankt Augustin, Germany
[2] Dept. of Computer Science, University of Bonn, Germany
{mario.boley,tamas.horvath,axel.poigne,stefan.wrobel}@iais.fraunhofer.de

Abstract. Many problems in data mining can be viewed as a special case of the problem of enumerating the closed elements of an *independence system* w.r.t. some *specific* closure operator. We consider a generalization of this problem to *strongly accessible set systems* and *arbitrary* closure operators. For this more general problem setting, the closed sets can be enumerated with polynomial delay if deciding membership in the set system and computing the closure operator can be solved in polynomial time. We discuss potential applications in graph mining.

1 Introduction

Over the past years, a large body of research has been devoted to finding efficient algorithms for the frequent item set enumeration problem, and it has turned out that by looking at *closed* frequent sets, important gains can be made in the design of efficient algorithms (see, e.g., [3]). A closed frequent set is a frequent set that cannot be further enlarged without changing its support in the database. Unfortunately, similar results do not yet exist for (closed) pattern enumeration tasks in many of the more complex representations that are becoming increasingly popular due to applications in highly structured domains. Consider for example the task of finding closed frequent *connected* subgraphs of movements of people or cars in a street network given a database of GPS-based recordings of spatiotemporal movements (so called *tracks*) [2]. In mining such tracks instead of sets, some important properties that are true for the frequent set mining problem no longer hold. In particular, it is not true that all subpatterns of a frequent connected pattern must necessarily also be frequent connected, since subpatterns need not be connected.

Technically, for problems like the track mining problem mentioned above, we note that unlike for the simple frequent set case, where the underlying set system is an *independence system*, here we are dealing with the weaker properties of a set system which is only *strongly accessible*. In this paper, we show that for this generalized problem, it is possible to design an algorithm that enumerates all

* A longer version of this extended abstract has been accepted at PKDD 2007.

closed frequent patterns, for arbitrary closure operators, with polynomial delay (provided deciding membership in the set system and computing the closure operator can be done in polynomial time, which usually is the case as we show in different settings in the long version of this paper). To our knowledge, this result gives the first efficient closed pattern enumeration algorithm for this generalized and practically important task.

Due to space limitations, we only outline our results (Sections 3 and 4) and omit the proofs and the description of our algorithm in this extended abstract.

2 Preliminaries

In this section we define some notions used in this paper. We assume familiarity with standard definitions from graph theory and frequent pattern mining. A *set system* is a pair $(E, \mathcal{F})$, where E is the ground set and $\mathcal{F} \subseteq 2^E$. A set system is called *finite* if its ground set is finite. A set system $(E, \mathcal{F})$ with $\emptyset \in \mathcal{F}$ is called

- *accessible* if for all $X \in \mathcal{F} \setminus \{\emptyset\}$ there is an $e \in X$ such that $X \setminus \{e\} \in \mathcal{F}$,
- *strongly accessible* if for every $X, Y \in \mathcal{F}$ satisfying $X \subsetneq Y$, there is an $e \in Y \setminus X$ such that $X \cup \{e\} \in \mathcal{F}$, and
- an *independence system* if $Y \in \mathcal{F}$ and $X \subseteq Y$ implies $X \in \mathcal{F}$.

The definitions imply that (i) every independence system is strongly accessible and (ii) every finite strongly accessible set system is accessible. However, the converse of (i) and (ii) does not hold.

Let $(E, \mathcal{F})$ be a set system. A function $\rho : \mathcal{F} \to \mathcal{F}$ is called a *closure operator* if (i) $X \subseteq \rho(X)$, (ii) $X \subseteq Y$ implies $\rho(X) \subseteq \rho(Y)$, and (iii) $\rho(X) = \rho(\rho(X))$ hold for all $X, Y \in \mathcal{F}$. A set $F \in \mathcal{F}$ is *ρ-closed* if $\rho(F) = F$. The *family of ρ-closed sets* of $(E, \mathcal{F})$ is denoted by $\rho(\mathcal{F})$. A set $F \in \mathcal{F}$ is a *generator* of a ρ-closed set C if $\rho(F) = C$.

3 The General Problem

Many problems in data mining (e.g., closed frequent itemset mining) can be considered as a special case of the following listing problem:

The Closed Set Mining (CSM) Problem: *Given* a finite set E, a membership oracle $M_{\mathcal{F}} : 2^E \to \{0, 1\}$ defining a family $\mathcal{F} \subseteq 2^E$ satisfying $\emptyset \in \mathcal{F}$, and a closure operator $\rho : \mathcal{F} \to \mathcal{F}$, *compute* $\rho(\mathcal{F})$.

Usually, $\mathcal{F}$ can be enumerated efficiently. Even then the naïve algorithm enumerating each set $S \in \mathcal{F}$ and testing whether S is ρ-closed is inefficient because $|\mathcal{F}|$ can be exponential in $|\rho(\mathcal{F})|$.

There are several results on efficient enumeration of ρ-closed sets for the case that the underlying set system is finite and it is an independence set system or at least closed under intersection (see, e.g., [1, 3]). In contrast to these results, we do *not* require the set system to be closed under intersection. Instead, we consider

finite set systems $(E, \mathcal{F})$ associated with closure operators $\rho : \mathcal{F} \to \mathcal{F}$ satisfying the following property: for any ρ-closed element of $\mathcal{F}$, there exists an *inductive generator*. An inductive generator of a ρ-closed element $C \in \rho(\mathcal{F})$ is an element $C' \cup \{e\} \in \mathcal{F}$ such that $C' \in \rho(\mathcal{F})$, $e \in E \setminus C'$, and $C = \rho(C' \cup \{e\})$. These generators can then be used to enumerate all closed sets with a DFS algorithm resulting in the following positive result:

Lemma 1 *The CSM problem can be solved with polynomial delay for instances satisfying (i) the membership oracle $M_{\mathcal{F}}$ and the closure operator ρ can be computed in polynomial time and (ii) for every ρ-closed set except $\rho(\emptyset)$, there exists an inductive generator.*

One can show that property (ii) holds for arbitrary closure operator if the underlying set system is finite and strongly accessible, implying the following result:

Theorem 2 *For any finite strongly accessible set system $(E, \mathcal{F})$ given by a polynomial membership oracle and for any polynomially computable closure operator $\rho : \mathcal{F} \to \mathcal{F}$, $\rho(\mathcal{F})$ can be enumerated with polynomial delay.*

We note that accessibility alone is not enough to guarantee the existence of inductive generators.

4 Applications

Let $G = (V, E)$ be an undirected graph and $\mathcal{D}$ be a database of subgraphs of G. We assume w.l.o.g. that the graphs in $\mathcal{D}$ are represented by subsets of E, i.e., $\mathcal{D}$ is considered as a transaction database over E. Transaction databases of this type occur e.g. in *track mining* applications (see, e.g., [2]) where we have a network represented by an undirected graph $G = (V, E)$ and points moving in the network within a time interval T. For each point i, let $E_i \subseteq E$ be the set of edges of G visited by point i in T and let G_i be the subgraph of G induced by E_i. The collection of the graphs G_i for every i forms the database $\mathcal{D}$. In contrast to other frequent subgraph mining problems defining the embedding usually by *subgraph isomorphism*, we define it by the *subset* relation. That is, for the subgraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ of G, G_1 can be *embedded* into G_2 iff $V_1 \subseteq V_2$ and $E_1 \subseteq E_2$.

Given a database $\mathcal{D}$ of the above form and frequency threshold $t > 0$, the closed frequent subgraph mining problem is equivalent to the closed frequent itemset mining problem. We face, however, a different problem if we require the closed frequent subgraphs to be *connected*. From an algorithmic point of view, we consider the following problem:

Closed Frequent Connected Subgraph Mining (CFCSM) Problem:
Given an undirected graph G, a transaction database $\mathcal{D}$ of G's subgraphs, and an integer $t > 0$, *list* all closed t-frequent *connected* subgraphs of $\mathcal{D}$.

As an application example, closed frequent connected subgraphs of a network can e.g. be considered as *homogeneous* connected subnetworks. For the above problem, the following result holds:

Theorem 3 *The CFCSM problem can be solved with polynomial delay.*

The proof is based on showing that there is a strongly accessible set system $(E, \mathcal{F})$ and a closure operator $\rho : \mathcal{F} \to \mathcal{F}$ such that $\rho(\mathcal{F})$ corresponds to the family of closed frequent connected subgraphs. In particular, $\mathcal{F}$ corresponds to the set of frequent connected subgraphs. We note that since the intersection of connected graphs need not be connected, $(E, \mathcal{F})$ is not closed under intersection and is thus not an independence system.

The setting defined in this paper actually goes beyond the standard definition of "closedness" usually employed in data mining, as it can be used to resolve an anomaly with this notion. In "well-behaved" cases (e.g., the above CFCSM problem), the design of closure operators is straightforward because there exists a closure operator inducing the set of closed frequent patterns. However, it does not hold in general that such a closure operator exists. As an example, one can consider the problem of mining closed frequent *paths* in the above defined transaction database $\mathcal{D}$. By the "canonical" definition of data mining, a path P is *closed frequent* if it is frequent and the support of P' is a *proper* subset of the support of P for every path P' containing P. It turns out, however, that there is no set system and closure operator inducing the set of closed frequent paths w.r.t. this definition. Using our approach, this anomaly can be resolved by considering another natural notion of "closedness" which is induced by a closure operator. We note that the underlying set system corresponding to the set of frequent paths is again strongly accessible, but not closed under intersection.

5 Conclusion

In this paper, we have presented a positive result on efficient enumeration of the family of closed sets of strongly accessible set systems w.r.t. arbitrary closure operators. The significance of our result in the context of data mining is that most of the closed frequent pattern mining algorithms are restricted to the case that the underlying set system corresponding to the set of frequent patterns is an independence system or at least closed under intersection. Strongly accessible set systems, however, are not necessarily independence systems or closed under intersection. We have presented graph mining applications motivated by track mining, where the underlying set systems are strongly accessible, but not closed under intersection.

References

1. B. Ganter and K. Reuter. Finding all closed sets: A general approach. *Order*, 8(3):283–280, 1991.
2. B. Kuijpers, M. Nanni, C. Körner, M. May, and D. Pedreschi. Spatio-temporal data mining. In F. Giannotti and D. Pedreschi, editors, *Geography, mobility, and privacy: a knowledge discovery vision*. (to appear)
3. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25–46, 1999.