On Computing Minimal Generators in Multi-Relational Data Mining with respect to θ-Subsumption

Noriaki Nishio, Atsuko Mutoh, and Nobuhiro Inuzuka

Nagoya Institute of Technology,
Gokiso-cho, Showa, Nagoya 466-8555, Japan
nishio@nous.nitech.ac.jp, mutoh@nitech.ac.jp, inuzuka@nitech.ac.jp

Abstract. We study the minimal generators (mingens) in multi-relational data mining. The mingens in formal concept analysis are the minimal subsets of attributes that induce the formal concepts. An intent for a formal concept is called a closed pattern. In contrast to the wide attention to closed patterns, the mingens have been paid little attention in Multi-Relational Data Mining (MRDM) field. We introduce an idea of non redundant mingens in MRDM. The notion of mingens in MRDM is led by θ-subsumption relation among patterns, and is useful to grasp the structure and information in the concepts.

1 Introduction

Formal Concept Analysis (FCA) [1] is an important tool for data analysis and knowledge discovery [2]. A formal concept is determined by its extent and its intent. The intent of a formal concept is the closure of the attributes, itemsets, or patterns that form a maximum characterization of the formal concept. Mining the closed patterns [3, 4] has attracted a lot of attentions because it reduces the number of patterns by selecting only representative patterns of their equivalent patterns in the sense that they produce the same extent.

While a closure is the maximal pattern presenting a concept, a minimal generator (mingen) [5] is a minimal pattern. The mingens play an important role in many contexts, e.g., database design (as key sets), graph theory (as minimal transversals), and data mining (as minimal premises of association rules). Dong et al. study the mingens and define Succinct System of Mingens (SSMG) [6] which removes redundant mingens. In this paper, we state that SSMGs of a formal context for relational patterns have further redundancy and propose a novel concept of non redundant mingens based on θ-subsumption of Multi-Relational Data Mining (MRDM) [7].

Sections 2 and 3 introduce FCA and MRDM. Section 4 describes about mingens. Section 5 provides a definition of minimal generators consisted of relational patterns. Then section 6 reports experimental results on compactness.

	a	b	c	d	e	g	h	i
t_1	×	×	×	×	×	×	×	×
t_2	×		×	×		×		
t_3		×	×	×		×	×	×
t_4	×	×		×			×	×
t_5		×	×		×	×	×	×

Fig. 1. A context that involves relations between objects and attributes.

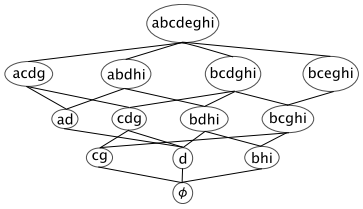

Fig. 2. A concept lattice for the context of Fig. 1.

2 Formal Concept Analysis

We review the basis of Formal Concept Analysis. Start with an arbitrary relation, $I \subseteq G \times M$, between G, a set of objects, and M, a set of attributes, and define

$$A \mapsto A^I = \{m \in M \mid (g, m) \in I \text{ for all } g \in A\} \text{ for } A \subseteq G,$$
$$B \mapsto B^I = \{g \in G \mid (g, m) \in I \text{ for all } m \in B\} \text{ for } B \subseteq M.$$

A triple $\mathbb{K} = (G, M, I)$ is called a *formal context.*

Definition 1 (formal concept). *A pair* (X, Y) *is called a* formal concept *of a* formal context $\mathbb{K} = (G, M, I)$*, if it satisfies*

$$X \subseteq G, Y \subseteq M, X^I = Y, X = Y^I. \qquad \square$$

When we define an order by $(X_1, Y_1) \leq (X_2, Y_2) \Longleftrightarrow X_1 \subseteq X_2$ $(\Leftrightarrow Y_2 \subseteq Y_1)$, among formal concepts of a formal context $\mathbb{K}$, it forms a complete lattice. We call it the *concept lattice* of $\mathbb{K}$.

Example 1. A Fig. 1 shows a formal context where each object has a set of attributes. A pair $(t_1t_2t_3, cdg)$ (set brackets are omitted) is a formal concept, where $t_1t_2t_3^I = cdg$ and $cdg^I = t_1t_2t_3$. Fig. 2 shows the concept lattice. Each concept is labeled by its intent. $\square$

3 Multi-Relational Data Mining

While propositional data mining algorithms look for patterns in a single data table, MRDM algorithms look for relational patterns, represented by logical formulae, that involve multiple tables (relations).

Example 2. A Database $R_{\mathbf{fam}}$ in Fig. 3 includes four relations on families, where $\mathsf{grandfather}(x)$ meaning x is someone's grandfather, $\mathsf{parent}(x, y)$ meaning x is a parent of y, $\mathsf{male}(x)$ for male x, and $\mathsf{female}(x)$ for female x. Then a pattern, such as $\mathsf{grandfather}(X) \leftarrow \mathsf{parent}(X, Y), \mathsf{parent}(Y, Z), \mathsf{female}(Z)$, can be found. $\square$

grandfather
person01
person07
person12
person19
person20

parent	
person01	person02
person02	person03
person02	person04
person03	person05
...	...

male
person01
person05
person07
person10
...

female
person02
person03
person04
person06
...

Fig. 3. The family DB $R_{\mathbf{fam}}$, including grandfather, parent, male and female.

4 Succinct System of Mingens

The mingens are defined bellow.

Definition 2 (mingens). *A set $P \subseteq M$ is called a* mingen *for a formal concept (X, Y) of a formal context $\mathbb{K} = (G, M, I)$ if $P^I = X$ but for every proper subset $P' \subset P$, $P'^I \neq X$.* □

Example 3. In Fig. 1, *bc*, *bg*, *ch*, *ci*, *gh*, and *gi* are mingens for a formal concept $(t_1t_3t_5, bcghi)$. □

In the above example, the closed itemset *bcghi* has six mingens, where *b*, *h* and *i* always appear together in each object and thus can be exchanged each other, and similarly for *c* and *g*. An SSMG is a representative of each equivalence class, which is defined bellow. A criterion which selects a representative is left to users, because dependence between items is not defined.

Definition 3 (C-equivalence). *X, $Y \subseteq M$ are* C-equivalent *for a formal concept C of a formal context $\mathbb{K} = (G, M, I)$, denoted $X \approx_C Y$, if they satisfy either following condition.*

1. *There is a concept $C' \geq C$ such that both X and Y are mingens of C'.*
2. *There are subsets $Z, Z', M \subseteq M$ such that $X = W \cup Z$, $Y = W \cup Z'$, and $Z \approx_C Z'$.* □

Definition 4 (SSMG). *Given an order $\sqsubseteq$ on M, a* succinct system of mingens *(SSMG) by the order $\sqsubseteq$ for a formal concept C of a formal context $\mathbb{K}$ consists of elements satisfied either following condition.*

1. *If C is a maximal formal concept in the sense of the concept lattice except $(G, \emptyset)$, an SSMG for C holds the following conditions.*
   - *It is a mingen for C.*
   - *It is minimal in the sense of $\sqsubseteq$ among all mingens in a C-equivalence class for C.*
2. *Otherwise, a mingen in a C-equivalence class is an SSMG for C if it does not include any mingen which is not an SSMG for $C' \geq C$.* □

The definition of SSMGs in [6] uses the alphabetic lexicographic order for $\sqsubseteq$ above. Since the alphabetic order is linear, there is a unique SSMG for an equivalence class. Our extended definition allows a partial order and then there are more than one SSMGs.

Table 1. Attributes by expressed formulae.

$a = \mathsf{gf}(A) \leftarrow \mathsf{m}(A).$
$b = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{f}(B).$
$c = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{p}(B,C), \mathsf{f}(C).$
$d = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{p}(B,C), \mathsf{p}(C,D), \mathsf{f}(D).$
$e = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{p}(B,C), \mathsf{p}(C,D), \mathsf{m}(D).$
$f = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{f}(B), \mathsf{p}(B,C), \mathsf{f}(C)$
$g = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{f}(B), \mathsf{p}(B,C), \mathsf{f}(C), \mathsf{p}(C,D), \mathsf{m}(D).$
$h = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{p}(B,C), \mathsf{f}(C), \mathsf{p}(C,D), \mathsf{p}(D,E), \mathsf{m}(E).$
$i = \mathsf{gf}(A) \leftarrow \mathsf{p}(A,B), \mathsf{p}(B,C), \mathsf{p}(C,D), \mathsf{f}(D), \mathsf{p}(B,E), \mathsf{p}(E,F), \mathsf{m}(F).$

	a	b	c	d	e	f	g	h	i
gf(01)	×	×	×	×	×	×	×	×	×
gf(07)	×	×	×	×	×	×	×	×	×
gf(12)	×	×	×		×	×	×	×	
gf(19)	×		×	×	×			×	×
gf(20)	×	×	×			×			

Fig. 4. $\mathbb{K}'_{\mathbf{fam}} = (G, M, I)$ w.r.t $R_{\mathbf{fam}}$

Table 2. Formal concepts of $\mathbb{K}'_{\mathbf{fam}}$

(G, ac)
$(\mathsf{gf}(01)\mathsf{gf}(07)\mathsf{gf}(12)\mathsf{gf}(19), aceh)$
$(\mathsf{gf}(01)\mathsf{gf}(07)\mathsf{gf}(12)\mathsf{gf}(20), abcf)$
$(\mathsf{gf}(01)\mathsf{gf}(07)\mathsf{gf}(19), acdehi)$
$(\mathsf{gf}(01)\mathsf{gf}(07)\mathsf{gf}(12), abcefgh)$
$(\mathsf{gf}(01)\mathsf{gf}(07), M)$

5 Mingens of MRDM

Though SSMGs remove redundant patterns, a simple application of SSMGs to relational patterns does not remove all of redundancy. Mapix [8, 9], a miner in MRDM, enumerates patterns consisted of *property items* [8] which are restricted sets of literals. Though a search space of MRDM has no limit as long as adding literals, that of Mapix restricts into a meaningful form by modes of predicates.

We construct a formal context $\mathbb{K}' = (G, M, I)$ of property items produced by Mapix, which we call a *formal relational context*, where G is a set of instances of a target (key) relation (e.g., grandfather relation in Fig. 3), M is a set of property items (e.g., in Table 1), and I is relation among G and M, which indicates whether an instance satisfies a property item (e.g., $\mathbb{K}'_{\mathbf{fam}}$ in Fig. 4). The notion of the formal relational context was discussed in [10]. Then we can also compute formal concepts and SSMGs of the formal relational context $\mathbb{K}'$.

Logical Mingens We reduce further redundancy of SSMGs in relational patterns by θ-subsumption relation, i.e. C θ-subsumes D, denoted by $C \succeq D$, if $C\theta \subseteq D$, for a substitution θ, where C and D are clauses.

Definition 5 (LMG). *A mingen for a formal concept C of a formal relational context $\mathbb{K}'$ is called a* logical mingen *(LMG) for C of $\mathbb{K}'$, if it satisfies the following conditions.*

- *It is an SSMG by θ-subsumption order ($\preceq$).*
- *It is a minimal in the sense of $\preceq$ in among all SSMG in C-equivalence class.*

□

Table 3. LMG_Miner : the algorithm for enumerating LMG

LMG_Miner($\mathbb{K}'$, $\sup_{\min}$):

input $\mathbb{K}'$: A formal relational context;
$\sup_{\min}$: A support threshold;

output LMG : logical mingens;

1. **let** $LMG := \emptyset$;
2. **let** $I :=$ {all attributes associated with property items};
3. **let** $LC :=$ {items occurring in all transactions};
4. **call** $DFS(H := \emptyset, T := I - LC, LC)$;
5. **return** LMG;

$DFS(H, T, LC)$:

1. **if** $\sup(H) < \sup_{\min}$ *return*;
2. **for each** $x \in T$
3. **if** $\sup(H \cup \{x\}) = \sup(H)$ **let** $T := T - \{x\}$, $LC := LC \cup \{x\}$;
4. **if** $(H : LC,\ \sup(H))$ construct a new concept with $\sup(H)$
5. **for each** $p \in LC$ **do if** $p \succeq H$ **then** p is removed from LC;
6. **add** $(H : LC,\ \sup(H))$ to LMG;
7. **else** remove clutter; // see [6] for details
8. **for each** $x \in T$
9. **let** $H_x := H \cup \{x\}$ *and* $T_x := \{y \in T \mid y > x\}$;
10. **call** $DFS(H_x, T_x, LC)$;

Note that the definition above uses the θ-subsumption twice, for the selection of mingens and for the selection of SSMG.

Example 4. Table 2 shows formal concepts in a formal context $\mathbb{K}'_{\mathbf{fam}}$. SSMG for $D = (\mathsf{gf}(01)\mathsf{gf}(07)\mathsf{gf}(12), abcefgh)$ is g, fe, fh, and LMG for D is only fe. □

The Mining Algorithm The algorithm in Table 3 follows the depth-first search framework using a set-enumeration tree (SE-tree) [11]. A node v, including a head H and a tail T, has a search space for all itemsets $Z = H \cup T'$, where T' is a nonempty subset of T. For the node labelled by ab in the SE-tree for $\{a, b, c, d\}$, we have $H = ab$ and $T = cd$, and its search space consists of abc, abd and $abcd$. A local closure of H, $LC(H) = \{x \in H \cup T \mid H^I = Hx^I\}$, is a closure w.r.t ancestor nodes. For all ancestor nodes v' of v with head H' and tail T', $LC(H')$ is a proper subset of $LC(H)$. Hence H is considered as the local mingen for $LC(H)$.

6 Experimental Results and Conclusion

We have done experiments on two data sets and compared between the number of patterns, the first one was with $R_{\mathbf{fam}}$ in Fig. 3 and the latter was with

Table 4. Patterns on $R_{\mathbf{fam}}$.

$\sup_{\min}(\%)$	80	60	40	20
MAPIX	17	109	1063	4601
SSMG	12	21	39	-
LMG	6	13	22	-

Table 5. Patterns on Mutagenesis-Bonds.

$\sup_{\min}(\%)$	90	80	70	60	50	40	30	20	10
MAPIX	336	360	614	721	721	925	1467	2948	6630
SSMG	9	9	13	16	16	19	31	67	149
LMG	6	6	9	12	12	14	25	58	137

Mutagenesis-Bonds. Tables 4 and 5 show the number of patterns generated by MAPIX, SSMG, and LMG. In both data sets, though SSMG and LMG had large reduction of patterns compared with MAPIX, LMG reduces patterns even more than SSMG. Because of a complex structure of $R_{\mathbf{fam}}$, SSMG and LMG fault the computation with $\sup_{\min} = 20\%$.

We still need revise of the algorithm for scalability. We also need examine the efficacy in the intuitive sense, such as readability.

References

1. Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1998.
2. Jonas Poelmans, Paul Elzinga, Stijn Viaene, and Guido Dedene. Formal concept analysis in knowledge discovery: A survey. In *ICCS*, pp. 139–153, 2010.
3. Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT'1999*, Vol. 1540 of *LNCS*, pp. 398–416. Springer, 1999.
4. Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *Discovery Science'2004*, Vol. 3245 of *LNCS*, pp. 16–31. Springer, 2004.
5. Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lotfi Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Computational Logic'2000*, Vol. 1861 of *LNCS*, pp. 972–986. Springer, 2000.
6. Guozhu Dong, Chunyu Jiang, Jian Pei, Jinyan Li, and Limsoon Wong. Mining succinct systems of minimal generators of formal concepts. In *DASFAA'2005*, Vol. 3453 of *LNCS*, pp. 175–187. Springer, 2005.
7. Saso Dzeroski. Multi-relational data mining: an introduction. *SIGKDD Explorations*, Vol. 5, No. 1, pp. 1–16, 2003.
8. Jun-ichi Motoyama, Shinpei Urazawa, Tomofumi Nakano, and Nobuhiro Inuzuka. A mining algorithm using property items extracted from sampled examples. In *ILP'2006*, Vol. 4455 of *LNCS*, pp. 335–350. Springer, 2007.
9. Yusuke Nakano and Nobuhiro Inuzuka. Multi-relational pattern mining based-on combination of properties with preserving their structure in examples. In *ILP'2010*, Vol. 6489 of *LNCS*, pp. 181–189. Springer, 2011.
10. Gerd Stumme. Iceberg query lattices for datalog. In *ICCS*, pp. 109–125, 2004.
11. Ron Rymon. Search through systematic set enumeration. In *KR'1992*, KR Proceedings, pp. 539–550. Morgan Kaufmann, 1992.