



**HAL**  
open science

# Qualitative Reasoning and Data Mining

Yakoub Salhi

► **To cite this version:**

Yakoub Salhi. Qualitative Reasoning and Data Mining. International Symposium on Temporal Representation and Reasoning (TIME), 2019, Malaga, Spain. 10.4230/LIPIcs.TIME.2019.9 . hal-03301176

**HAL Id: hal-03301176**


**<https://univ-artois.hal.science/hal-03301176>**

Submitted on 28 Mar 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Qualitative Reasoning and Data Mining

Yakoub Salhi 

CRIL - CNRS & Université d'Artois, Lens, France.

salhi@cril.fr

---

## Abstract

In this paper, we introduce a new data mining framework that is based on qualitative reasoning. We consider databases where the item domains are of different types, such as numerical values, time intervals and spatial regions. Then, for the considered tasks, we associate to each item a constraint network in a qualitative formalism representing the relations between all the pairs of objects of the database w.r.t. this item. In this context, the introduced data mining problems consist in discovering qualitative covariations between items. In a sense, our framework can be seen as a generalization of gradual itemset mining. In order to solve the introduced problem, we use a declarative approach based on the satisfiability problem in classical propositional logic (SAT). Indeed, we define SAT encodings where the models represent the desired patterns.

**2012 ACM Subject Classification** Information systems → Data mining; Information systems → Association rules; Theory of computation → Constraint and logic programming; Computing methodologies → Knowledge representation and reasoning

**Keywords and phrases** Qualitative Database, Qualitative Pattern Mining, Declarative Approach, SAT Modeling

**Digital Object Identifier** 10.4230/LIPIcs...

## 1 Introduction

Data mining techniques are applied on different data types, such as transactions, sequences, graphs, texts, etc. In order to consider complex aspects of the real world, it is interesting to extend these techniques for knowledge discovery to new complex data, such as spatio-temporal pieces of information. However, it is important in this context to take into account the simplicity of the pattern structure. Thus, the challenge in this work is to propose a framework that allows us to deal with different complex data types and discovering patterns having a simple structure.

Qualitative reasoning is concerned with facilitating reasoning about complex entities and pieces of information through symbolic representation formalisms. In particular, this kind of reasoning is strongly related to human one and, for instance, it can be used for dealing with pieces of information that come from natural language. In the literature, the qualitative formalisms are widely used for reasoning about two physical entities of the world that are time and space (e.g. see [21]). Indeed, qualitative spatial and temporal reasoning is an important research field in Artificial Intelligence in general, and knowledge representation in particular. The spatial and temporal representation formalisms allow reasoning about configurations by abstracting numerical quantities of space and time thanks to qualitative relations, such as *inside*, *before*, *after*, etc. One of the best known qualitative representation formalisms is the Point Algebra [31], which allows representing and reasoning about the possible relative positions between two points on the timeline. The Interval Algebra [2, 3], for its part, is used for reasoning about the possible positions between two intervals. Furthermore, regarding qualitative spatial reasoning, the Region Connection Calculus RCC8 [25] is one of the most studied formalisms in qualitative reasoning, which concerns topological relations between two spatial regions.



© Yakoub Salhi;

licensed under Creative Commons License CC-BY

Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

47 In this work, we propose a framework for data mining using qualitative reasoning, which  
 48 allows considering different data types, such as numerical values, time intervals and spatial  
 49 regions. To this end, we first introduce the notion of qualitative database, which is defined  
 50 by associating to each item a constraint network in a qualitative formalism representing the  
 51 relations between the pairs of objects of the database w.r.t. this item. Then, we describe  
 52 data mining tasks for discovering qualitative covariations, called qualitative itemsets, in  
 53 the previous kind of databases. For instance, the desired patterns can capture pieces of  
 54 information of the form "a variation of an item  $a$  w.r.t. the qualitative relation  $r_1$  is associated  
 55 with a variation of  $b$  w.r.t. the qualitative relation  $r_2$ ". In a sense, the proposed tasks can be  
 56 seen as a generalization of those related to gradual itemsets where the qualitative relations that  
 57 are considered in the extracted patterns are only  $\leq$  and  $\geq$  on numerical values [7, 10, 11, 20].  
 58 We express the interestingness predicate on the qualitative itemsets in a database through  
 59 two different definitions of support. The first definition takes into consideration a local view  
 60 by reasoning about the pairs of objects that satisfy the partial order induced by the itemset,  
 61 while the second is obtained by reasoning about the sequences respecting the previous partial  
 62 order. These two definitions allow extracting interesting recurrent pieces of information.  
 63 Finally, we use a declarative and flexible solution for solving the introduced data mining  
 64 tasks based on the use of the satisfiability problem in classical propositional logic (SAT).  
 65 Indeed, we define for each task a SAT encoding whose models allow us to obtain all the  
 66 desired patterns. Thus, we follow in our solution the constraint programming based approach  
 67 for data mining initiated in [24, 13], which offers a declarative and flexible representation  
 68 model.

69 The rest of the paper is organized as follows. After describing related works in Section 2,  
 70 we introduce in Section 3 the notion of qualitative database. In Section 4, we present the  
 71 data mining tasks proposed in this work. In Section 5, we describe our SAT-based encodings  
 72 for solving these tasks, while Section 6 concludes the paper.

## 73 **2 Related Works**

74 The most related data mining tasks to our framework are those concerned with extracting  
 75 gradual itemsets [7, 10, 11, 20]. A gradual itemset is a pattern expressing covariations of  
 76 items having as domains sets of numerical values. For instance, the gradual itemset containing  
 77 three gradual items  $\{sport^{\geq}, weight^{\geq}, diseases^{\leq}\}$  can be used to express the fact "the higher  
 78 the time of physical activity, the higher the weight loss, and the fewer the number of diseases".  
 79 The gradual itemset structure allows analyzing numerical data in a simple and intuitive way,  
 80 since it avoids the quantitative aspects of the considered data.

81 The data mining framework introduced in this work can be seen as a generalization of  
 82 that of mining gradual itemsets in the case of numerical data. Indeed, instead of using only  
 83 the inequality relations  $\leq$  and  $\geq$ , many binary qualitative relations on different data types  
 84 can be used in our framework, in particular qualitative relations on time intervals and spatial  
 85 regions.

86 It is worth noting that we use in our framework measures for determining the quality of  
 87 a qualitative itemset similar to those proposed in the case of gradual itemsets. In fact, in  
 88 the same way as in gradual itemset mining, we consider two distinct definitions of support:  
 89 the first definition considers the pairs of objects that respect the itemset, while the second  
 90 definition is obtained by reasoning about the length of the sequences that respect the pattern.  
 91 More precisely, the first definition of support corresponds to the numbers of pairs of objects  
 92 that satisfy the partial order associated to the pattern, and the second definition corresponds

93 the length of the longest sequences of objects that are ordered using the partial order induced  
94 by the pattern.

95 The use of a declarative approach for data mining was originally proposed in [24] for  
96 performing different tasks. Specifically, the authors have demonstrated that constraint  
97 programming is an appropriate tool in many respects in itemset mining. One of the main  
98 motivations lies in the fact that this framework offers a flexible and generic representation  
99 model. Indeed, new constraints often require new implementations for specialized data mining  
100 approaches, which can often be integrated in a fairly simple way into declarative frameworks,  
101 since it is not needed to change the solving tools. In addition, the continual evolution in the  
102 efficiency of tools dedicated to problems that can be used for data mining modeling, like ASP  
103 (*Answer Set Programming*), CSP (*Constraint Satisfaction Problem*) and SAT, is a strong  
104 argument in favor of using approaches based on these problems. Thus, from this first work,  
105 a new line of research has emerged within the data mining community. Indeed, in recent  
106 years, many works using CSP and SAT for different data mining tasks have been proposed in  
107 the literature (e.g. [13, 16, 12, 30, 19, 9]). In particular, in [8], the authors show that their  
108 SAT-based approach achieves better performance than state-of-the-art specialized techniques.  
109 In this work, we use a SAT-based approach for solving all the considered data mining tasks.  
110 Let us note that a SAT-based approach was recently used for extracting gradual patterns  
111 in [22].

### 112 3 Qualitative Database

113 In this section, we introduce the notion of qualitative database. The main idea consists in  
114 associating to each item a constraint network in a qualitative formalism representing the  
115 relations between the pairs of objects of the database w.r.t. this item. To illustrate our  
116 proposal, we consider three distinct qualitative formalisms for reasoning about time and  
117 space, namely Point Algebra [31], Interval Algebra [2, 3] and Region Connection Calculus  
118 RCC8 [25].

119 Given a finite set  $S$ , we use  $\mathcal{P}(S)$  and  $|S|$  to denote respectively the powerset and the  
120 cardinality of  $S$ . Given a finite set of items  $\mathcal{I}$ ,  $V_a$  is used to denote the domain of the  
121 item  $a \in \mathcal{I}$ . The domain of an item can be a numerical value, a temporal interval, a  
122 spatial region, etc. Further, we associate to each item  $a$  a finite set of qualitative base  
123 relations  $B_a$ , which consists of *jointly exhaustive* and *pairwise disjoint* relations, i.e., for  
124 each  $(v, v') \in V_a \times V_a$ , there exists exactly one  $b \in B_a$  such that  $(v, v') \in b$ . Further, we  
125 only consider the set of qualitative base relations  $B_a$  that contains the identity relation  
126  $id = \{(v, v') \in V_a \times V_a \mid v = v'\}$ , and is closed under the inverse operation  $(\cdot)^{-1}$ , namely  
127 whenever  $b$  is in  $B_a$ , the inverse  $(b)^{-1}$  is also in  $B_a$ . A qualitative relation is said to be  
128 *universal* if it contains all the base relations.

129 The *weak composition* of two base relations  $b$  and  $b'$  in  $B_a$ , denoted  $b \diamond b'$ , is defined as  
130 the set of base relations  $\{b'' \in B_a \mid \exists (v, v') \in b \ \& \ (v', v'') \in b' \ \& \ (v, v'') \in b''\}$ . The weak  
131 composition operation is extended to the relations in  $\mathcal{P}(B_a)$  as follows:  $r \diamond r' = \bigcup_{b \in r, b' \in r'} b \diamond b'$ .  
132 In this context, it is worth mentioning that the *composition*  $\circ$  of two relations is defined as  
133 follows:  $r \circ r' = \{(v, v') \mid \exists v'', (v, v'') \in r \ \& \ (v'', v') \in r'\}$ . In other words,  $r \diamond r'$  is the largest  
134 set of base relations where each one shares at least one value with  $r \circ r'$ .

135 For example, consider the point algebra (PA) qualitative formalism described in Figure 1,  
136 which has been mainly used for temporal reasoning. Indeed, PA can be used to encode  
137 temporal relations between two points in the timeline. We also describe in Figure 2 the base  
138 relations of two other qualitative formalisms: interval algebra (IA) and region connection

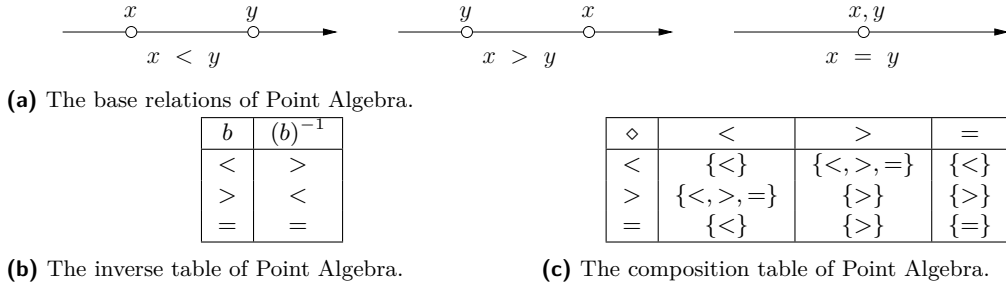


Figure 1 Point Algebra

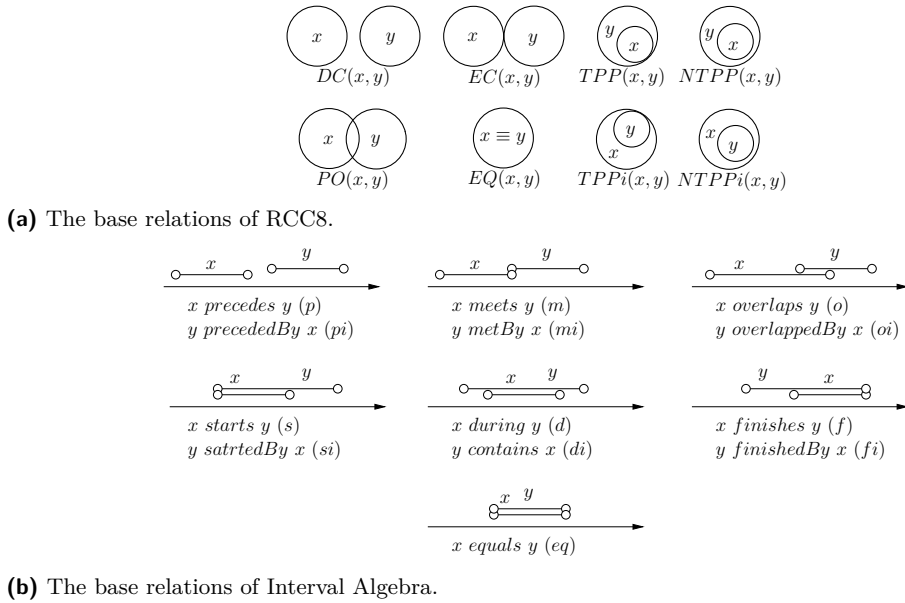


Figure 2 The qualitative formalisms RCC8 and Interval Algebra.

139 calculus RCC8. The formalism IA allows encoding relative relations between intervals, while  
 140 RCC8 allows encoding topological relations between two regions. For instance, the expression  
 141  $DC(Region1, Region2)$  represents the fact that the two spatial regions  $Region1$  and  $Region2$   
 142 are disconnected.

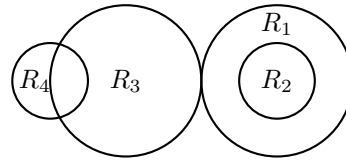
143 ► **Definition 1** (Qualitative Column). A q-column is a structure of the form  $c = (a, \mathcal{O}, R)$ ,  
 144 where  $a$  is an item, denoted  $item(c)$ ,  $\mathcal{O}$  is a finite non empty set of objects, denoted  $obj(c)$ ,  
 145 and  $R$  is a mapping from  $\mathcal{O} \times \mathcal{O}$  to  $B_a$ , denoted  $rel(c)$ .

146 Let us now introduce the notion of qualitative database, which is defined by associating to  
 147 each item a constraint network in a qualitative formalism representing the relations between  
 148 the pairs of objects of the database w.r.t. this item.

149 ► **Definition 2** (Qualitative Database). A qualitative database is a structure of the form  
 150  $(\mathcal{O}, \mathcal{I}, \mathcal{C})$ , where  $\mathcal{O}$  is a finite non empty set of objects,  $\mathcal{I}$  is a finite non empty set of items  
 151 and  $\mathcal{C}$  is a set of q-columns s.t. (i)  $|\mathcal{C}| = |\mathcal{I}|$ , (ii)  $\forall c \in \mathcal{C}, obj(c) = \mathcal{O}$ , and (iii)  $\forall a \in \mathcal{I}$ , there  
 152 exists exactly one  $c \in \mathcal{C}$  s.t.  $item(c) = a$ .

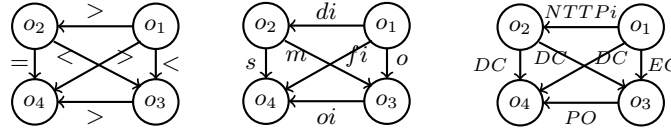
153 In the sequel, we sometimes use  $R_a$  to denote  $rel(c)$  where  $c$  is the qualitative column  
 154 associated to the item  $a$ .

objects	a	b	c
$o_1$	2	[1,4]	$R_1$
$o_2$	1	[2,3]	$R_2$
$o_3$	4	[3,6]	$R_3$
$o_4$	1	[2,4]	$R_4$



(b) A representation of the real situation the regions  $R_1, R_2, R_3$  and  $R_4$ .

(a) A database using values in item domains.



(c) The qualitative database corresponding to the database in (a).

■ **Figure 3** A Qualitative Database

155 For example, we describe in Figure 3 a qualitative database: we provide in (a) a database  
 156 using values in item domains, in (b) the concret situation of the considered spatial regions,  
 157 and in (c) the qualitative database. For instance, the edge between  $o_1$  and  $o_2$  in the left-hand  
 158 graph represents the qualitative base relation in  $PA > (o_1, o_2)$ , usually denoted  $o_1 > o_2$ .

#### 4 Mining Qualitative Itemsets

160 In this section, we introduce data mining tasks for discovering qualitative covariations in  
 161 qualitative databases. For instance, the patterns in this context can be used to capture pieces  
 162 of information of the form "a variation of  $a$  w.r.t. the qualitative relation  $r_1$  is associated  
 163 with a variation of  $b$  w.r.t. the qualitative relation  $r_2$ ".

164 ► **Definition 3** (Qualitative Itemset). *A qualitative itemset is a finite non empty set of*  
 165 *qualitative items  $I$ , where a qualitative item is a structure of the form  $a^r$  where  $a$  is an item*  
 166 *and  $r \subseteq B_a$ .*

167 Let us now describe the partial order on the objects of a database that is induced by a  
 168 qualitative itemset, and also the notion of ordered sequence that is used for defining the  
 169 support of a qualitative itemset.

170 ► **Definition 4** (Induced Partial Order). *Let  $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{C})$  be a qualitative database,  $o, o' \in \mathcal{O}$*   
 171 *and  $I = \{a_1^{r_1}, \dots, a_k^{r_k}\}$  a qualitative itemset. Then, we say that  $o$  precedes  $o'$  w.r.t.  $I$ , written*  
 172  *$o \preceq_I o'$ , if for all  $i \in 1..k$ ,  $R_a(o, o') \in r_i$  holds.*

173 ► **Definition 5** (Ordered sequence of objects). *Let  $\mathcal{D}$  be a qualitative database,  $L = \langle o_1, \dots, o_k \rangle$*   
 174 *a sequence of distinct objects in  $\mathcal{D}$  and  $I$  a qualitative itemset. We say that  $L$  respects  $I$  if it*  
 175 *is ordered with respect to  $\preceq_I$ , i.e.,  $o_i \preceq_I o_{i+1}$  for every  $i \in 1..k - 1$ .*

176 We here use  $\mathcal{L}(\mathcal{D}, I)$  to denote all the sequences of objects occurring in  $\mathcal{D}$  that respect  
 177 the qualitative itemset  $I$ .

178 In the same way as in gradual itemset mining, we express the quality of an itemset in a  
 179 database through two different definitions of *support*. The first definition captures a local  
 180 view by taking into consideration the number of pairs that satisfy the partial order induced  
 181 by the qualitative itemset ( $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{C})$ ):

$$supp_1(I, \mathcal{D}) = \frac{|\{\{o, o'\} \subseteq \mathcal{O} \mid o \neq o', o \preceq_I o'\}|}{|\mathcal{O}| \cdot (|\mathcal{O}| - 1)/2}.$$

The second definition is obtained by reasoning about the sequences that respect the qualitative itemset. Indeed, it corresponds to the length of the longest sequences that respect the considered itemset:

$$supp_2(I, \mathcal{D}) = \frac{\max\{|L| \mid L \in \mathcal{L}(\mathcal{D}, I)\}}{|\mathcal{O}|}.$$

182 Furthermore, we consider that it is more appropriate to allow the user to select the  
 183 relations that can be associated to every item in a pattern. For example, it is not interesting  
 184 to consider the universal or empty relations because they do not describe any variation.

185 Thus, we define two problems of enumerating qualitative itemsets as follows: given a  
 186 function  $f$  that maps each item  $a$  to a subset of relations  $f(a) \subseteq \mathcal{P}(B_a)$  which is closed  
 187 under the inverse operation and the inclusion, and a minimum support threshold  $v$ , the  
 188 problems QIE1 and QIE2 consist in computing respectively the sets of qualitative itemsets  
 189  $QIE1(\mathcal{D}, f, v) = \{I \mid supp_1(I, \mathcal{D}) \geq v \ \& \ \forall a^r \in I, r \in f(a)\}$  and  $QIE2(\mathcal{D}, f, v) = \{I \mid$   
 190  $supp_2(I, \mathcal{D}) \geq v \ \& \ \forall a^r \in I, r \in f(a)\}$ .

191 Let us consider now two condensed representations, which are similar to those that are  
 192 widely considered in itemset mining. Before that, we need the following partial order relation.  
 193 Given two qualitative itemsets  $I$  and  $J$ , we have  $I \sqsubseteq J$  if,  $\forall a^r \in I, \exists a^{r'} \in J$  s.t.  $r' \subseteq r$ .  
 194 Moreover, we have  $I \sqsubset J$  if  $I \sqsubseteq J$  and  $I \neq J$ .

195 ► **Definition 6** (Closedness). *Let  $\mathcal{D}$  be a database and  $I$  a qualitative itemset. Then,  $I$  is*  
 196 *said to be a closed qualitative itemset in  $\mathcal{D}$  w.r.t.  $supp_1$  (resp.  $supp_2$ ) if, for all qualitative*  
 197 *itemset  $J$  with  $I \sqsubset J$ ,  $supp_1(I, \mathcal{D}) > supp_1(J, \mathcal{D})$  (resp.  $supp_2(I, \mathcal{D}) > supp_2(J, \mathcal{D})$ ) holds.*

198 In other words, a qualitative itemset is closed if there is no more informative qualitative  
 199 itemset that has the same support.

200 ► **Definition 7** (Maximality). *Let  $\mathcal{D}$  be a database,  $v$  a minimum support threshold and  $I$  a*  
 201 *qualitative itemset. Then,  $I$  is said to be a maximal qualitative itemset w.r.t.  $supp_1$  (resp.*  
 202  *$supp_2$ ) and the threshold  $v$  if, for all qualitative itemset  $J$  with  $I \sqsubset J$ ,  $supp_1(J, \mathcal{D}) < v$  (resp.*  
 203  *$supp_2(J, \mathcal{D}) < v$ ) holds.*

204 A qualitative itemset is maximal if there is no more informative qualitative itemset that  
 205 has a support greater than or equal to the minimum support threshold.

206 In the context of the condensed representations, one can easily see that we have the  
 207 following property.

208 ► **Proposition 8** (Anti-Monotonicity). *Let  $\mathcal{D}$  be a qualitative database and  $I$  and  $J$  two*  
 209 *qualitative itemsets in  $\mathcal{D}$ . If  $I \sqsubseteq J$  then  $supp_1(I, \mathcal{D}) \geq supp_1(J, \mathcal{D})$  and  $supp_2(I, \mathcal{D}) \geq$   
 210  $supp_2(J, \mathcal{D})$ .*

211 Therefore, using the anti-monotonicity property, computing either the closed itemsets or the  
 212 maximal itemsets in  $QIE1(\mathcal{D}, f, v)$  and  $QIE2(\mathcal{D}, f, v)$  allows getting all the elements of these  
 213 two sets. Furthermore, the anti-monotonicity property can be used for defining Apriori-like  
 214 algorithms for solving the problems QIE1 and QIE2 in a fairly simple way. Let us recall that  
 215 Apriori algorithm was originally proposed in [1] for mining frequent itemsets.

216 It is worth mentioning that the qualitative relations are not necessarily transitive. For  
 217 example, we have  $1\{<, >\}2\{<, >\}1$  in PA ( $x\{<, >\}y$  means that  $x$  is different from  $y$ ) without  
 218 having  $1\{<, >\}1$ . This has as a consequence the fact that a sequence respects a qualitative

219 itemset does not implies that its sub-sequences (by avoiding intermediate objects) respect  
 220 also this pattern. Thus, in order to have transitivity, a solution can consist in restricting  
 221 our mining task to the relations that satisfy  $\diamond$ -idempotence: a qualitative relation  $r$  is said  
 222 to be  $\diamond$ -idempotent if  $r \diamond r = r$ . For example, in PA the  $\diamond$ -idempotent relations are  $\{=\}$ ,  
 223  $\{<\}$ ,  $\{<,=\}$ ,  $\{>\}$ ,  $\{>,=\}$  and  $\{<,=,>\}$ , i.e., all the relations except  $\{\}$  and  $\{<,>\}$ . That  
 224 being said, we provide in this work general methods for solving QIE1 and QIE2 without  
 225 considering transitivity.

226 In order to illustrate the mining tasks described previously, we provide now a simple  
 227 example. Consider the database described in Figure 4. It represents pieces of information  
 228 related to a set of workers about time at work, productivity and satisfaction degree. For the  
 229 corresponding qualitative database, we consider interval algebra for time at work, and point  
 230 algebra for both productivity and satisfaction degree. Moreover, we only consider QIE2 with  
 231 a support threshold equal to 3 without any restriction on the considered qualitative relations  
 232 in the patterns on **time**, but we only consider  $\{<,\leq,>,\geq\}$  on both **productivity** and  
 233 **satisfaction**. A first interesting qualitative pattern is  $I = \{\mathbf{time}^{\{p,o,m\}}, \mathbf{productivity}^{\leq}\}$ ,  
 234 which has a support equal to 4 since it is satisfied by the sequence  $\langle w_1, w_2, w_3, w_4 \rangle$ . In a  
 235 sense, it expresses that starting work earlier increase productivity. The pattern  $I$  is not  
 236 closed since it has the same supports as  $J = \{\mathbf{time}^{\{p,o,m\}}, \mathbf{productivity}^{<}\}$ . Moreover,  $J$   
 237 is closed since  $J \cup \{\mathbf{satisfaction}^{\leq}\}$  and  $J \cup \{\mathbf{satisfaction}^{\geq}\}$  are respectively 2 and 3.  
 238 Moreover,  $J \cup \{\mathbf{satisfaction}^{>}\}$  is a maximal patterns since its support is equal to the fixed  
 239 threshold and it is not included in any other pattern.

worker	time	productivity	satisfaction
$w_1$	5am to 9am	100	1
$w_2$	8am to 12am	80	4
$w_3$	12am to 4pm	60	5
$w_4$	5pm to 9pm	50	3

■ **Figure 4** A description of a database.

## 240 5 SAT-based Approach for Enumerating Qualitative Itemsets

241 In this section, we introduce a SAT-based approach for solving the problems QIE1 and  
 242 QIE2. We first describe the satisfiability problem in classical propositional logic. We then  
 243 introduce our SAT encodings for QIE1 and QIE2: the computation of the models of each  
 244 encoding corresponds to the computation of the desired qualitative itemsets. We here follow  
 245 the constraint programming based approach for data mining initiated in [24, 13].

### 246 5.1 Classical Propositional Logic

We here describe the syntax and the semantics of classical propositional logic. We use  $\mathbf{Prop}$  to denote the set of propositional variables. The propositional formulas of classical propositional logic ( $CPL$ ) are built using  $\mathbf{Prop}$ , the constants  $\top$ , denoting *true*, and  $\perp$ , denoting *false*, the unary logical connective  $\neg$  and the usual binary connectives  $\wedge$ ,  $\vee$ ,  $\rightarrow$  and  $\leftrightarrow$ . The grammar is defined as follows:

$$\phi ::= p \mid \top \mid \perp \mid \phi \wedge \phi \mid \phi \vee \phi \mid \phi \rightarrow \phi \mid \phi \leftrightarrow \phi \mid \neg \phi$$

247 with  $p \in \mathbf{Prop}$ . The set of propositional formulas is denoted  $\mathbf{Form}$ . We use the letters  
 248  $p, q, r, s$  to denote the propositional variables, and the Greek letters  $\phi, \psi$  and  $\chi$  to denote



249 the propositional formulas. Moreover, given a syntactic object  $o$ , we use  $Var(o)$  to denote  
250 the set of propositional variables occurring in  $o$ .

251 A *Boolean interpretation*  $\mathcal{B}$  of a formula  $\phi$  is defined as a function from the set of  
252 variables  $Var(\phi)$  to  $\{0, 1\}$  (0 stands for *false* and 1 for *true*). It is inductively extended to  
253 propositional formulas as usual:

$$\begin{aligned}
& \mathcal{B}(\top) = 1 & \mathcal{B}(\perp) = 0 \\
& \mathcal{B}(\neg\phi) = 1 - \mathcal{B}(\phi) & \mathcal{B}(\phi \rightarrow \psi) = \max(1 - \mathcal{B}(\phi), \mathcal{B}(\psi)) \\
& \mathcal{B}(\phi \wedge \psi) = \min(\mathcal{B}(\phi), \mathcal{B}(\psi)) & \mathcal{B}(\phi \vee \psi) = \max(\mathcal{B}(\phi), \mathcal{B}(\psi)) \\
& \mathcal{B}(\phi \leftrightarrow \psi) = 0 \text{ if } \mathcal{B}(\phi) \neq \mathcal{B}(\psi), \mathcal{B}(\phi \leftrightarrow \psi) = 1 \text{ otherwise}
\end{aligned}$$

256 A formula  $\phi$  is satisfiable if there exists a Boolean interpretation  $\mathcal{B}$  of  $\phi$  such that  $\mathcal{B}(\phi) = 1$ ,  
257 and  $\mathcal{B}$  is called a *model* of  $\phi$  in this case. We use  $Mod(\phi)$  to denote the set of all the models  
258 of  $\phi$ .

259 Consider for instance the formula  $(p \wedge q) \leftrightarrow p$ , which has exactly three models:  $\mathcal{B}_1$  with  
260  $\mathcal{B}_1(p) = \mathcal{B}_1(q) = 0$ ;  $\mathcal{B}_2$  with  $\mathcal{B}_1(p) = \mathcal{B}_1(q) = 1$ ; and  $\mathcal{B}_3$  with  $\mathcal{B}_3(p) = 0$  and  $\mathcal{B}_1(q) = 1$ .

262 A propositional formula in *Conjunctive Normal Form* (CNF) is a conjunction of clauses,  
263 where a *clause* is a disjunction of literals. It is well-known that every propositional formula  
264 can be translated to CNF w.r.t. the satisfiability problem using Tseitin's linear encoding [29].  
265 The problem of determining whether there exists a model that satisfies a given CNF formula,  
266 abbreviated as SAT, is one of the most studied NP-complete problems.

268 A *cardinality constraint* is an inequality of the form  $\sum_{i=1}^n p_i \geq m$ . Several polynomial  
269 encodings of this kind of constraints into propositional formulas have been proposed in  
270 the literature (e.g. [4, 26, 5]). An *AtMostOne constraint* is a particular case of the form  
271  $\sum_{i=1}^n p_i \leq 1$ , which can be linearly encoded in SAT. For instance, the encoding using  
272 sequential counter [26, 23] is defined as follows:

$$\begin{aligned}
& (\neg p_1 \vee q_1) \wedge (\neg p_n \vee q_{n-1}) \\
& \bigwedge_{1 < i < n} ((\neg p_i \vee q_i) \wedge (\neg q_{i-1} \vee q_i) \wedge (\neg p_i \vee \neg q_{i-1}))
\end{aligned}$$

274 where  $q_i$  is a fresh propositional variable for  $i = 1, \dots, n - 1$ .

## 275 5.2 A SAT Encoding for QIE1

276 In this section, we propose a SAT encoding for the problem of enumerating qualitative  
277 itemsets QIE1. More precisely, we associate to every instance of QIE1 a propositional  
278 formula so that its models allow us to obtain all the corresponding qualitative itemsets.

280 Let  $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{C})$  be a qualitative database,  $f$  a function that maps each  $a \in \mathcal{I}$  to a  
281 subset of  $\mathcal{P}(B_a)$  closed under the inverse operation and the inclusion, and  $v$  a minimum  
282 support threshold. We here use the integer  $\alpha$  defined as the value  $v \cdot (|\mathcal{D}| \cdot (|\mathcal{D}| - 1)/2)$ .

283 In order to define our encoding, we associate to each pair of an item  $a$  and a relation  
284  $r \in f(a)$  a distinct propositional variable denoted  $p_{ar}$ . The variable  $p_{ar}$  is used to express  
285 the qualitative itemset in the sense that it is true if and only if  $a^r$  belongs to the current  
286 qualitative itemset. Furthermore, we associate to each ordered pair of different objects  
287  $(o, o')$  in  $\mathcal{D}$  a distinct propositional variable denoted  $q_{(o, o')}$ . In the proposed encoding, a  
288 variable  $q_{(o, o')}$  is true if and only if  $o$  precedes  $o'$  with respect to the current qualitative

289 itemset. In order not to take into account both symmetric couples of objects in support com-  
 290 putation, we also associate a variable denote  $s_{\{o,o'\}}$  to each pair of distinct objects  $\{o,o'\}$  in  $\mathcal{D}$ .

291

292 The first propositional formula of our encoding for QIE1 allows avoiding the empty  
 293 itemset by requiring at least one item:

$$294 \quad \bigvee_{a \in \mathcal{I}} \bigvee_{r \in f(a)} p_{ar}. \quad (1)$$

295 Indeed, this formula corresponds to a single clause that expresses that there is at least one  
 296 variable of the form  $p_{ar}$  assigned to true.

297 The following conjunction of AtMostOne constraints allows avoiding the association of  
 298 multiple variations to an item in the same pattern:

$$299 \quad \bigwedge_{a \in \mathcal{I}} \sum_{r \in f(a)} p_{ar} \leq 1. \quad (2)$$

300 More precisely, each AtMostOne constraint is associated to a distinct item and means that  
 301 there is at most one qualitative relation associated to this item in the pattern.

302 The following formula allows establishing that each variable  $q_{(o,o')}$  is true if and only  $o$   
 303 precedes  $o'$  w.r.t. the qualitative itemset:

$$304 \quad \bigwedge_{o,o' \in \mathcal{O}, o \neq o'} \neg q_{(o,o')} \leftrightarrow \bigvee (\{p_{ar} \mid a \in \mathcal{I}, r \in (f(a) \setminus \{r' \in f(a) \mid R_a(o,o') \in r'\})\}). \quad (3)$$

305 We exactly express in the previous formula that  $q_{(o,o')}$  is false if and only if there is a  
 306 qualitative item  $a^r$  such that  $r(o,o')$  does not hold.

307 We now introduce the formula that is used for symmetry breaking by considering in the  
 308 support computation at most one of the couples  $(o,o')$  and  $(o',o)$ :

$$309 \quad \bigwedge_{o,o' \in \mathcal{O}, o \neq o'} s_{\{o,o'\}} \leftrightarrow (q_{(o,o')} \vee q_{(o',o)}). \quad (4)$$

310 Finally, the following cardinality constraint expresses that support of every qualitative  
 311 itemset in  $\mathcal{D}$  has to be greater than or equal to  $v$ :

$$312 \quad \sum_{o,o' \in \mathcal{O}, o \neq o'} s_{\{o,o'\}} \geq \alpha. \quad (5)$$

313 Let us note that the use of  $\alpha$  in the previous constraint is clearly equivalent to the use of  $v$   
 314 as a minimum support threshold.

315

316 We use  $\mathcal{ENC}(\mathcal{D}, f, v)$  to denote the conjunction of the previous formulas:  $(1) \wedge (2) \wedge (3) \wedge$   
 317  $(4) \wedge (5)$ .

318 There are three important properties related to our encoding  $\mathcal{ENC}(\mathcal{D}, f, v)$ . First, the  
 319 soundness property means that every model encodes a frequent qualitative itemset. Second,  
 320 the completeness property expresses that every frequent qualitative itemset is encoded in  
 321 a model of the encoding. Third, the non-redundancy property is used to capture the fact  
 322 that there is a bijective mapping between the set of the models and the set of the frequent  
 323 qualitative itemsets.

324 ► **Proposition 9 (Soundness).** *Given an instance  $(\mathcal{D}, f, v)$  of QIE1, if  $\mathcal{B}$  is a model of*  
 325  *$\mathcal{ENC}(\mathcal{D}, f, v)$  then  $I_{\mathcal{B}} = \{a^r \mid \mathcal{B}(p_{ar}) = 1\} \in \mathcal{QIE1}(\mathcal{D}, f, v)$ .*

## XX:10 Qualitative Reasoning and Data Mining

326 **Proof.** First, using the formula (1), we clearly have  $|I_{\mathcal{B}}| \geq 1$ . Then, using (2), we know that  
 327 an item occurs at most once in every pattern. Moreover, using (3)  $\wedge$  (4), we obtain  $\{s_{\{o,o'\}} \mid$   
 328  $\mathcal{B}(s_{\{o,o'\}}) = 1\} = \{\{o,o'\} \subseteq \mathcal{O} \mid o \neq o', o \preceq_{I_{\mathcal{B}}} o'\}$ . Thus, using the cardinality constraint (5),  
 329 we obtain  $|\{s_{\{o,o'\}} \mid \mathcal{B}(s_{\{o,o'\}}) = 1\}| \geq \alpha$  and we have thereby  $\text{supp}_1(I_{\mathcal{B}}, \mathcal{D}) \geq v$ . Therefore,  
 330  $I_{\mathcal{B}}$  belongs to  $\mathcal{QIE1}(\mathcal{D}, f, v)$ .  $\blacktriangleleft$

331 **► Proposition 10 (Completeness).** *Given an instance  $(\mathcal{D}, f, v)$  of QIE1, if  $I \in \mathcal{QIE1}(\mathcal{D}, f, v)$*   
 332 *then there exists a Boolean interpretation  $\mathcal{B}_I$  that satisfies the encoding  $\mathcal{ENC}(\mathcal{D}, f, v)$ , where*  
 333  $I = \{a^r \mid \mathcal{B}_I(p_{a^r}) = 1\}$ .

334 **Proof.** Let us define  $\mathcal{B}_I$  as follows:

- 335 1. for every pair of an item  $a$  and a relation  $r \in f(a)$ ,  $\mathcal{B}_I(p_{a^r}) = 1$  iff  $a^r \in I$ ;
- 336 2. for every ordered pair of distinct objects  $(o, o')$ ,  $\mathcal{B}_I(q_{(o,o')}) = 1$  iff  $o \preceq_I o'$ ;
- 337 3. for every pair of distinct objects  $\{o, o'\}$ ,  $\mathcal{B}_I(s_{\{o,o'\}}) = 1$  iff  $o \preceq_I o'$  or  $o' \preceq_I o$ .

338 Using the fact that  $|I| \geq 1$ ,  $\mathcal{B}_I$  satisfies (1). Then, using the fact that an item cannot occur  
 339 more than once in  $I$ ,  $\mathcal{B}_I$  satisfies (2). Further, using the properties 1 and 2 in the definition  
 340 of  $\mathcal{B}_I$ , we obtain that  $\mathcal{B}_I$  satisfies (3). Using the fact that  $\mathcal{B}_I$  satisfies (3) and the property 3  
 341 in the definition of  $\mathcal{B}_I$ , we also obtain that (4) is also satisfied by  $\mathcal{B}_I$ . Moreover, the formula  
 342 (5) is satisfied since  $\text{supp}_1(I, \mathcal{D}) \geq v$ .  $\blacktriangleleft$

343 **► Proposition 11 (Non-Redundancy).** *Given an instance  $(\mathcal{D}, f, v)$  of QIE1, for all two distinct*  
 344 *models  $\mathcal{B}$  and  $\mathcal{B}'$  of  $\mathcal{ENC}(\mathcal{D}, f, v)$ ,  $\{a^r \mid \mathcal{B}(p_{a^r}) = 1\} \neq \{a^r \mid \mathcal{B}'(p_{a^r}) = 1\}$  holds.*

345 **Proof.** This property is a direct consequence of the fact that we use the equivalence logical  
 346 connective in the formulas (3) and (4). Indeed, the support is encoded using the variables of  
 347 the form  $q_{(o,o')}$  and  $s_{\{o,o'\}}$ , and a qualitative itemset cannot have two distinct values for the  
 348 support.  $\blacktriangleleft$

349 It is worth noting that having a bijective mapping between the set of the models and the  
 350 set of the frequent qualitative itemsets allows us to adapt in a fairly simple way our encoding  
 351 for many variants of QIE1, such as counting the number of patterns.

352

353 Let us now introduce the notion of complementary qualitative itemset, which is mainly  
 354 used for reducing the search space.

355 **► Definition 12 (Complementary Qualitative Itemset).** *Let  $I = \{a_1^{r_1}, \dots, a_k^{r_k}\}$  be a qualitative*  
 356 *itemset. The complementary of  $I$ , denoted  $I^c$ , is the qualitative itemset  $\{a_1^{(r_1)^{-1}}, \dots, a_k^{(r_k)^{-1}}\}$ .*

357 We clearly have the following proposition.

358 **► Proposition 13.** *The following two properties are satisfied, for all qualitative database  $\mathcal{D}$*   
 359 *and for all qualitative itemset  $I$ :*

- 360  $\blacksquare$   $\text{supp}_1(I, \mathcal{D}) = \text{supp}_1(I^c, \mathcal{D})$
- 361  $\blacksquare$   $\text{supp}_2(I, \mathcal{D}) = \text{supp}_2(I^c, \mathcal{D})$ .

362 Proposition 13 can be used to avoid unnecessary computations. Indeed, at each found  
 363 model, we can avoid in the next step both the corresponding qualitative itemset and its  
 364 complementary itemset. It is worth noting that a similar property is used in the case of  
 365 gradual patterns [7, 10, 11, 20].

366

367 Let us now consider the condensed representations corresponding to the closed and the  
 368 maximal qualitative itemsets. In order to obtain the closed qualitative itemsets, we first need  
 369 to conjunctively add to the encoding  $\mathcal{ENC}(\mathcal{D}, f, v)$  the following formula:

$$370 \quad \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a)} \left( \left( \bigwedge_{o, o' \in \mathcal{O}, o \neq o'} (q_{(o, o')} \rightarrow R_a(o, o') \in r) \right) \rightarrow \bigvee_{r' \subseteq r} p_{a^{r'}} \right). \quad (6)$$

371 Indeed, this propositional formula means that, for all qualitative item  $a^r$ , if we have  
 372  $\text{supp}_1(I, \mathcal{D}) = \text{supp}_1(I \cup \{a^r\}, \mathcal{D})$ , then there exists  $r' \subseteq r$  such that  $a^{r'}$  belongs to  $I$ ,  
 373 where  $I$  is the qualitative itemset associated to the current model. In other words, it allows  
 374 making the current qualitative itemset more informative without changing the support.

375 Then, we add the following formula to express that it is not possible to reduce the size of  
 376 any relation in the pattern without changing the support:

$$377 \quad \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a), |r| > 1} (p_{a^r} \rightarrow \bigwedge_{r' \subset r} \left( \bigvee_{o, o' \in \mathcal{O}, o \neq o'} q_{(o, o')} \wedge R_a(o, o') \notin r' \right)). \quad (7)$$

378 We use  $\mathcal{ENC} - \mathcal{C}(\mathcal{D}, f, v)$  to denote the SAT encoding for the problem of enumerating  
 379 the closed qualitative itemsets:  $\mathcal{ENC}(\mathcal{D}, f, v) \wedge (6) \wedge (7)$ .

380 Similarly, to compute the maximal qualitative itemsets, we only need to conjunctively  
 381 add to  $\mathcal{ENC}(\mathcal{D}, f, v)$  the following two formulas:

$$382 \quad \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a)} \left( \sum_{o, o' \in \mathcal{O}, o \neq o'} (q_{(o, o')} \wedge R_a(o, o') \in r) \geq \alpha \rightarrow \bigvee_{r' \subseteq r} p_{a^{r'}} \right) \quad (8)$$

383

$$384 \quad \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a), |r| > 1} (p_{a^r} \rightarrow \bigwedge_{r' \subset r} \sum_{o, o' \in \mathcal{O}, o \neq o'} (q_{(o, o')} \wedge R_a(o, o') \notin r') < \alpha). \quad (9)$$

385 The formula (8) allows maximizing the size of the current qualitative itemset while keeping the  
 386 support greater than or equal to  $v$ , (9) states that it is not possible to reduce the size of any  
 387 relation without reducing the support to a value smaller than  $v$ . We use  $\mathcal{ENC} - \mathcal{M}(\mathcal{D}, f, v)$   
 388 to denote the SAT encoding  $\mathcal{ENC}(\mathcal{D}, f, v) \wedge (8) \wedge (9)$ .

### 389 5.3 A SAT Encoding for QIE2

390 We here propose a SAT encoding for the problem QIE2, which combines formulas defined for  
 391 QIE1 and new ones that are described in this section.

392

393 Let  $\mathcal{D} = (\mathcal{O}, \mathcal{I}, \mathcal{C})$  be a database,  $f$  a function that maps each  $a \in \mathcal{I}$  to a subset of  $\mathcal{P}(B_a)$   
 394 closed under the inverse operation and the inclusion, and  $v$  a minimum support threshold.  
 395 We here use the integer  $\beta$  defined as the value  $v \cdot |\mathcal{D}|$ . We now describe an encoding that  
 396 allows one to obtain all the elements of  $\text{QIE2}(\mathcal{D}, v)$ .

397 In the same way as the previous encoding, we also use in the same way the propositional  
 398 variables of the forms  $p_{a^r}$  and  $q_{(o, o')}$ : the variables of the form  $p_{a^r}$  are used to encode the  
 399 qualitative itemset, and those of the form  $q_{(o, o')}$  to encode its support. Moreover, we associate  
 400 to each integer  $i \in 1..\beta$  and object  $o$  in  $\mathcal{D}$  a fresh propositional variable  $t_o^i$ , which is used to  
 401 express that the object  $o$  is used at the location  $i$  in a sequence in  $\mathcal{L}(\mathcal{D}, I)$ , where  $I$  is the  
 402 current qualitative itemset.

403

404 The first formula in our encoding is the conjunction of (1)  $\wedge$  (2)  $\wedge$  (3) of the previous  
 405 encoding  $\mathcal{ENC}(\mathcal{D}, f, v)$ . Indeed, (1) is used to express that every qualitative itemset contains

## XX:12 Qualitative Reasoning and Data Mining

406 at least one qualitative item, (2) is used to avoid multiple occurrences of an item in the same  
 407 itemset, and (3) says that  $q_{(o,o')}$  is false if and only if there is a qualitative item  $a^r$  such  
 408 that  $R_a(o, o') \in r$  does not hold. As a consequence, every model of the previous conjunction  
 409 encodes a qualitative itemset, where the variables of the form  $q_{(o,o')}$  encode the pairs of  
 410 objects that satisfy the partial order induced by this itemset.

411 Using the fact that the propositional variables of the form  $t_o^i$  are used to build an ordered  
 412 sequence of objects, the following formula means that an object cannot be used more than  
 413 once in a sequence:

$$414 \quad \bigwedge_{o \in \mathcal{O}} \sum_{i=1}^{\beta} t_o^i \leq 1. \quad (10)$$

415 The following formula says that there is exactly one object at each location:

$$416 \quad \bigwedge_{i=1}^{\beta} \sum_{o \in \mathcal{O}} t_o^i = 1. \quad (11)$$

417 Clearly, the previous formula allows us to only consider the qualitative itemsets that have  
 418 supports greater than or equal to  $v$  w.r.t.  $\text{supp}_2$ .

419 In order to require the ordering induced by the qualitative itemset, the following formula  
 420 is used to capture the fact that if two objects  $o$  and  $o'$  occur in successive locations, then the  
 421 couple  $(o, o')$  respects the qualitative itemset, which is expressed by the truth of the variable  
 422  $q_{(o,o')}$ :

$$423 \quad \bigwedge_{o, o' \in \mathcal{O}, o \neq o'} \bigwedge_{i=1}^{\beta-1} ((t_o^i \wedge t_{o'}^{i+1}) \rightarrow q_{(o,o')}). \quad (12)$$

424 We use  $\mathcal{ENC2}(\mathcal{D}, f, v)$  to denote the encoding that corresponds to the following conjunc-  
 425 tion:  $(1) \wedge (2) \wedge (3) \wedge (10) \wedge (11) \wedge (12)$ .

426 ► **Proposition 14 (Soundness).** *Given an instance  $(\mathcal{D}, f, v)$  of QIE2, if  $\mathcal{B}$  is a model of*  
 427  *$\mathcal{ENC2}(\mathcal{D}, f, v)$  then  $I_{\mathcal{B}} = \{a^r \mid \mathcal{B}(p_{a^r}) = 1\} \in \mathcal{QIE2}(\mathcal{D}, f, v)$ .*

428 **Proof.** The soundness can be shown in the same way as in the case of QIE1. Using (1),  
 429 we know that  $I_{\mathcal{B}}$  contains at least one qualitative item. Then, using (2), each item occurs  
 430 at most once in every qualitative itemset. Further, using (3), we obtain  $a^r \in I_{\mathcal{B}}$  iff, for  
 431 all  $o, o' \in \mathcal{O}$ ,  $\mathcal{B}(q_{(o,o')}) = 1$  iff  $R_a(o, o') \in r$ . Thus, using (10)  $\wedge$  (11)  $\wedge$  (12), we know that  
 432 there exists a sequence  $\langle o_1, \dots, o_{\beta} \rangle$  which respects  $I_{\mathcal{B}}$ , where  $\mathcal{B}(t_{o_i}^i) = 1$  for  $i \in 1..\beta$ . As a  
 433 consequence,  $\text{supp}_2(I_{\mathcal{B}}, \mathcal{D}) \geq v$  and  $I_{\mathcal{B}}$  belongs to  $\mathcal{QIE2}(\mathcal{D}, f, v)$  ◀

434 ► **Proposition 15 (Completeness).** *Given an instance  $(\mathcal{D}, f, v)$  of QIE2, if  $I \in \mathcal{QIE2}(\mathcal{D}, f, v)$*   
 435 *then there exists a Boolean interpretation  $\mathcal{B}_I$  that satisfies the encoding  $\mathcal{ENC2}(\mathcal{D}, f, v)$  where*  
 436  *$I = \{a^r \mid \mathcal{B}_I(p_{a^r}) = 1\}$ .*

437 **Proof.** First, given a sequence  $s = \langle o_1, \dots, o_{\beta} \rangle$  respecting  $I$ , we define  $\mathcal{B}_I$  as follows:

- 438 ■ for every pair of an item  $a$  and a relation  $r \in f(a)$ ,  $\mathcal{B}_I(p_{a^r}) = 1$  iff  $a^r \in I$ ;
- 439 ■ for every couple of distinct objects  $(o, o')$ ,  $\mathcal{B}_I(q_{(o,o')}) = 1$  iff  $o \preceq_I o'$ ;
- 440 ■ for every object  $o$  and location  $i \in 1..\beta$ ,  $\mathcal{B}_I(t_o^i) = 1$  iff  $o = o_i$ .

441 For the same reasons described in the proof of Proposition 10,  $\mathcal{B}_I$  satisfies (1)  $\wedge$  (2)  $\wedge$  (3).  
 442 Then, using the fact that the length of  $s$  is  $\beta$  and the objects in this sequence are pairwise  
 443 distinct,  $\mathcal{B}_I$  satisfies also (10)  $\wedge$  (11). Finally, using the fact that  $s$  respects the partial order  
 444 induced by  $I$ ,  $\mathcal{B}_I$  satisfies (12). ◀

445 Contrary to our previous encoding,  $\mathcal{ENC2}(\mathcal{D}, f, v)$  does not satisfy the non-redundancy  
 446 property, since the same qualitative itemset may be associated to distinct sequences. How-  
 447 ever, this is not a problem for enumerating the qualitative itemsets without redundancy,  
 448 because we only need to conjunctively add the negation of the found qualitative itemset at  
 449 each step instead of the negation of the found model. More precisely, if we found a model  
 450 representing the qualitative itemset  $I = \{a_1^{r_1}, \dots, a_k^{r_k}\}$ , then we conjunctively add the clause  
 451  $\neg p_{a_1^{r_1}} \vee \dots \vee \neg p_{a_k^{r_k}} \vee \bigvee_{a^r \notin I} p_a^r$  to avoid this itemset in the next steps.

452  
 453 In  $\mathcal{ENC2}(\mathcal{D}, f, v)$ , we use propositional variables that are associated to only  $\beta$  locations,  
 454 since we aim at computing the qualitative itemsets having supports at least equal to  $v$ .  
 455 However, for computing the closed qualitative itemsets, we need to have the exact value of  
 456 the support, which means that we have to encode one of the longest sequences in each model  
 457 of the SAT encoding. In order to avoid this problem, we propose an intermediate solution by  
 458 restricting  $\mathcal{ENC2}(\mathcal{D}, f, v)$  to the closed qualitative itemsets w.r.t. QIE1. In this context, we  
 459 clearly have the following property.

460 ► **Proposition 16.** *Let  $\mathcal{D}$  be a qualitative database and  $I$  a qualitative itemset. If  $I$  is closed*  
 461 *in  $\mathcal{D}$  w.r.t.  $\text{supp}_2$ , then it is also closed in  $\mathcal{D}$  w.r.t.  $\text{supp}_1$ .*

462 **Proof.** This property is a direct consequence of the fact that if  $\text{supp}_2(I, \mathcal{D}) > \text{supp}_2(J, \mathcal{D})$ ,  
 463 then  $\text{supp}_1(I, \mathcal{D}) > \text{supp}_1(J, \mathcal{D})$  holds for every qualitative itemsets  $I$  and  $J$  with  $I \sqsubset J$ . ◀

464 Thus, the set of closed qualitative itemsets w.r.t. QIE2 is included in that of the qualitative  
 465 itemsets obtained from the encoding  $\mathcal{ENC2}(\mathcal{D}, f, v) \wedge (6) \wedge (7)$ . As a consequence, the previous  
 466 SAT encoding can be used for enumerating all the closed qualitative itemsets w.r.t. QIE2.  
 467 Indeed, we only need in this context to select the largest patterns w.r.t.  $\sqsubseteq$  for every value  
 468 for the support.

469 Let us now consider the problem of enumerating the maximal qualitative itemsets. In  
 470 this context, consider the following formulas:

$$471 \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a)} \bigwedge_{o, o' \in \mathcal{O}, o \neq o'} \bigwedge_{i=1}^{\beta-1} ((t_o^i \wedge t_{o'}^{i+1} \wedge R_a(o, o') \in r) \rightarrow \bigvee_{r' \subseteq r} p_{a^{r'}}), \quad (13)$$

$$473 \bigwedge_{a \in \mathcal{I}} \bigwedge_{r \in f(a), |r| > 1} (p_{a^r} \rightarrow \bigwedge_{r' \subset r} \bigvee_{o, o' \in \mathcal{O}, o \neq o'} \bigwedge_{i=1}^{\beta-1} (t_o^i \wedge t_{o'}^{i+1} \wedge R_a(o, o') \notin r')). \quad (14)$$

474 These two formulas express that, for a sequence of length equal to  $\beta$ , the associated qualitative  
 475 itemset has to be the largest w.r.t.  $\sqsubseteq$ . Therefore, in the same way as our encoding for  
 476 enumerating the closed qualitative itemsets, the encoding  $\mathcal{ENC2}(\mathcal{D}, f, v) \wedge (13) \wedge (14)$  allows  
 477 one to compute a set of patterns that contains all the maximal qualitative itemsets.

478 It is worth noting that the strategies proposed in [15, 18] for adapting Conflict-Driven  
 479 Clause-Learning (CDCL) based SAT-solvers to the task of model enumeration can be directly  
 480 used in the case of our encoding. Furthermore, it is also possible to directly use the  
 481 decomposition method introduced in [17] for improving the SAT-based approach in solving  
 482 data mining problems.

## 483 6 Conclusion and Perspectives

484 The first main contribution of this article is a definition of a framework for data mining  
 485 using qualitative reasoning. This framework allows considering different data types, such

486 as numerical values, time intervals and spatial regions. Moreover, the data mining tasks  
 487 introduced in this work can be seen as a natural generalization of those related to gradual  
 488 itemsets. The second main contribution is our declarative and flexible solution for solving  
 489 the proposed data mining tasks based on the satisfiability problem in classical propositional  
 490 logic (SAT): each task is modeled as a propositional formula whose models correspond to  
 491 the desired patterns.

492 In our future work, we intend to further study qualitative reasoning in data mining  
 493 following three main directions: (1) the use of disjunctions of base relations between objects,  
 494 which allows, for instance, modeling vagueness; (2) considering qualitative formalisms that  
 495 are not closed under the inverse operation, such as cardinal direction calculus [27, 28];  
 496 (3) considering some qualitative relations with arities greater than two in the case of some  
 497 particular data types (e.g. [14, 6]). Furthermore, we plan to implement the proposed  
 498 SAT-based methods to provide an experimental study on the use of our framework.

---

#### 499 — References —

- 500 **1** Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining Association Rules between  
 501 Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International  
 502 Conference on Management of Data, Washington, DC, USA*, pages 207–216. ACM Press, 1993.
- 503 **2** James F. Allen. An Interval-Based Representation of Temporal Knowledge. In *Proceedings of  
 504 the 7th International Joint Conference on Artificial Intelligence, IJCAI '81, Vancouver, BC,  
 505 Canada*, pages 221–226. William Kaufmann, 1981.
- 506 **3** James F. Allen. Maintaining Knowledge about Temporal Intervals. *Communications of the  
 507 ACM*, 26(11):832–843, 1983.
- 508 **4** Olivier Bailleux and Yacine Boufkhad. Efficient CNF Encoding of Boolean Cardinality  
 509 Constraints. In *Principles and Practice of Constraint Programming - CP 2003, 9th International  
 510 Conference, CP 2003, Kinsale, Ireland*, pages 108–122, 2003.
- 511 **5** Olivier Bailleux, Yacine Boufkhad, and Olivier Roussel. A Translation of Pseudo Boolean  
 512 Constraints to SAT. *JSAT*, 2(1-4):191–200, 2006.
- 513 **6** Philippe Balbiani, Jean-François Condotta, and Gérard Ligozat. Reasoning about Cyclic  
 514 Space: Axiomatic and Computational Aspects. In *Spatial Cognition III, Routes and Navigation,  
 515 Human Memory and Learning, Spatial Representation and Spatial Learning*, pages 348–371.  
 516 Springer, 2003.
- 517 **7** Fernando Berzal, Juan C. Cubero, Daniel Sánchez, María Amparo Vila Miranda, and José-  
 518 María Serrano. An Alternative Approach to Discover Gradual Dependencies. *International  
 519 Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(5):559–570, 2007.
- 520 **8** Abdelhamid Boudane, Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. A SAT-Based Approach  
 521 for Mining Association Rules. In *Proceedings of the Twenty-Fifth International Joint Conference  
 522 on Artificial Intelligence, IJCAI 2016, New York, NY, USA*, pages 2472–2478. IJCAI/AAAI  
 523 Press, 2016.
- 524 **9** Abdelhamid Boudane, Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. SAT-Based Data  
 525 Mining. *International Journal on Artificial Intelligence Tools*, 27(1):1–24, 2018.
- 526 **10** Lisa Di-Jorio, Anne Laurent, and Maguelonne Teisseire. Fast extraction of gradual association  
 527 rules: a heuristic based method. In *CSTST 2008: Proceedings of the 5th International  
 528 Conference on Soft Computing as Transdisciplinary Science and Technology, Cergy-Pontoise,  
 529 France*, pages 205–210. ACM, 2008.
- 530 **11** Lisa Di-Jorio, Anne Laurent, and Maguelonne Teisseire. Mining Frequent Gradual Itemsets  
 531 from Large Databases. In *Advances in Intelligent Data Analysis VIII, 8th International  
 532 Symposium on Intelligent Data Analysis, IDA 2009, Lyon, France*, pages 297–308. Springer,  
 533 2009.
- 534 **12** Tias Guns, Anton Dries, Siegfried Nijssen, Guido Tack, and Luc De Raedt. Miningzinc: A  
 535 declarative framework for constraint-based mining. *Artificial Intelligence*, 244:6–29, 2017.

- 536 13 Tias Guns, Siegfried Nijssen, and Luc De Raedt. Itemset mining: A constraint programming  
537 perspective. *Artificial Intelligence*, 175(12-13):1951–1983, 2011.
- 538 14 Amar Isli and Anthony G. Cohn. A new approach to cyclic ordering of 2d orientations using  
539 ternary relation algebras. *Artificial Intelligence*, 122(1-2):137–187, 2000.
- 540 15 Saïd Jabbour, Jerry Lonlac, Lakhdar Sais, and Yakoub Salhi. Extending modern SAT solvers for  
541 models enumeration. In *Proceedings of the 15th IEEE International Conference on Information  
542 Reuse and Integration, IRI 2014, Redwood City, CA, USA*, pages 803–810. IEEE Computer  
543 Society, 2014.
- 544 16 Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. Boolean satisfiability for sequence mining. In  
545 *22nd ACM International Conference on Information and Knowledge Management, CIKM'13,  
546 San Francisco, CA, USA*, pages 649–658. ACM, 2013.
- 547 17 Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. Decomposition Based SAT Encodings for  
548 Itemset Mining Problems. In *Advances in Knowledge Discovery and Data Mining - 19th  
549 Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam*, pages 662–674. Springer,  
550 2015.
- 551 18 Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. On SAT models enumeration in itemset  
552 mining. *CoRR*, abs/1506.02561, 2015.
- 553 19 Saïd Jabbour, Lakhdar Sais, and Yakoub Salhi. Mining Top-k motifs with a SAT-based  
554 framework. *Artificial Intelligence*, 244:30–47, 2017.
- 555 20 Anne Laurent, Marie-Jeanne Lesot, and Maria Rifqi. GRAANK: Exploiting Rank Correlations  
556 for Extracting Gradual Itemsets. In *Flexible Query Answering Systems, 8th International  
557 Conference, FQAS 2009, Roskilde, Denmark*, pages 382–393. Springer, 2009.
- 558 21 Gérard Ligozat. *Qualitative Spatial and Temporal Reasoning*. ISTE. Wiley, 2013.
- 559 22 Jerry Lonlac, Saïd Jabbour, Engelbert Mephu Nguifo, Lakhdar Sais, and Badran Raddaoui.  
560 Extracting Frequent Gradual Patterns Using Constraints Modeling. *CoRR*, abs/1903.08452,  
561 2019.
- 562 23 João P. Marques-Silva and Inês Lynce. Towards Robust CNF Encodings of Cardinality Con-  
563 straints. In *Principles and Practice of Constraint Programming - CP 2007, 13th International  
564 Conference, CP 2007, Providence, RI, USA*, pages 483–497, 2007.
- 565 24 Luc De Raedt, Tias Guns, and Siegfried Nijssen. Constraint Programming for Itemset Mining.  
566 In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery  
567 and Data Mining, Las Vegas, Nevada, USA*, pages 204–212, 2008.
- 568 25 David A. Randell, Zhan Cui, and Anthony Cohn. A Spatial Logic Based on Regions and  
569 Connection. In *Proceedings of the 3rd International Conference on Principles of Knowledge  
570 Representation and Reasoning (KR'92). Cambridge, MA, USA*, 1992.
- 571 26 Carsten Sinz. Towards an Optimal CNF Encoding of Boolean Cardinality Constraints. In  
572 *Principles and Practice of Constraint Programming - CP 2005, 11th International Conference,  
573 CP 2005, Sitges, Spain*, pages 827–831, 2005.
- 574 27 Spiros Skiadopoulos and Manolis Koubarakis. Composing cardinal direction relations. *Artificial  
575 Intelligence*, 152(2):143–171, 2004.
- 576 28 Spiros Skiadopoulos and Manolis Koubarakis. On the consistency of cardinal direction  
577 constraints. *Artificial Intelligence*, 163(1):91–135, 2005.
- 578 29 Gregory S. Tseitin. On the complexity of derivations in the propositional calculus. In *Structures  
579 in Constructives Mathematics and Mathematical Logic, Part II*, pages 115–125, 1968.
- 580 30 Willy Ugarte, Patrice Boizumault, Bruno Crémilleux, Alban Lepailleur, Samir Loudni, Marc  
581 Plantevit, Chedy Raïssi, and Arnaud Soulet. Skypattern mining: From pattern condensed  
582 representations to dynamic constraint satisfaction problems. *Artificial Intelligence*, 244:48–69,  
583 2017.
- 584 31 Marc B. Vilain and Henry A. Kautz. Constraint Propagation Algorithms for Temporal  
585 Reasoning. In *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia,  
586 PA, USA. Volume 1: Science*, pages 377–382. Morgan Kaufmann, 1986.