Exploring sequential data with relational concept analysis

Cristina Nica

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE Mathématiques, Sciences de l'Information et de l'Ingénieur

Laboratoire ICube – UMR 7357

THÈSE présentée par :

Cristina NICA

soutenue le : **13 octobre 2017**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Informatique

Exploring Sequential Data with Relational Concept Analysis

THÈSE dirigée par :

Mme LE BER Florence	Ingénieure en chef des Ponts, des Eaux et des Forêts, HDR, ENGEES, Université de Strasbourg, ICube

RAPPORTEURS :

M. PONCELET Pascal	Professeur, Université Montpellier, LIRMM
M. SOLDANO Henry	Maître de Conférences, HDR, Université Paris-Nord, LIPN

AUTRES MEMBRES DU JURY :

M. FERRÉ Sébastien	Maître de Conférences, HDR, Université Rennes 1, IRISA
M. BEISEL Jean-Nicolas	Professeur, ENGEES, Université de Strasbourg
Mme BRAUD Agnès	Maître de Conférences, Université de Strasbourg, ICube

This thesis is dedicated to my late mother, Cecilia,
to whom I owe all that I am today.

Acknowledgements

First of all, I would like to express my gratitude to both my PhD advisers, Florence LE BER and Agnès BRAUD, who have been actively interested in my work. I would like to thank them as well for encouraging my research and for allowing me to grow as a research scientist.

Besides my advisers, I am very thankful to Pascal PONCELET and Henry SOLDANO for accepting to read and review my thesis manuscript. I gratefully acknowledge Sébastien FERRÉ and Jean-Nicolas BEISEL for accepting to be members of my thesis committee.

I would also like to thank my mid-thesis committee, Amedeo NAPOLI and Jens GUSTEDT, for their insightful comments and encouragement. In particular I would like to thank Marianne HUCHARD and Xavier DOLQUES for helping me in the preliminary steps of this thesis. In addition, I am very thankful to Corinne GRAC for her patience and her help in validating the experimental results presented in this thesis.

I will forever be thankful to Cornelia TUDORIE and Emilia PECHEANU, my former advisers at the "Dunărea de Jos" University of Galați (Romania), who were the reason why I decided to go to pursue a career in research.

I would like to thank the University of Strasbourg for funding my PhD studies. In addition, I would like to thank ICube and ENGEES administrative stuff.

Lastly, I would like to thank my family and my close friends for supporting me throughout these past 3 years of my PhD studies.

Contents

List of Figures

List of Tables

1 General Introduction

Contents

Nowadays, tremendous amounts of data are found in databases. These large collections of data bring up the basic question: *"Can valuable information be gleaned from these data?"*. Therefore, special methods have been devised to automatically discover in these data potentially useful and understandable regularities from which practical insights can be drawn. In this context, the process *Knowledge Discovery in Databases* (KDD, [Fayyad et al., 1996]) has emerged.

1.1 Knowledge Discovery and Data Mining Techniques

Fayyad et al. [1996] presented KDD as the process concerned with the development of methods and techniques for discovering knowledge units from data stored in databases. The KDD process is interactive and iterative and each of its subsequent step relies on the output of the previous step. In addition, Fayyad et al. define the KDD process as a sequence of five steps:

1. **data selection** that aims at gathering only the data relevant to the analysis task. This step is guided by domain experts;
2. **data preprocessing** that employs techniques for dealing with the outliers, the noise and/or the missing values from the selected data;

3. **data transformation** that converts the preprocessed data to an appropriate format for data mining;

4. **data mining** that explores the transformed data by means of specific techniques in order to extract regularities;

5. **evaluation** that represents the interpretation and the assessment of the discovered regularities with respect to the motivation behind the analysis task. The obtained knowledge units can then be used by decision-makers.

Data mining is the core step in the KDD process and it has two widely accepted main goals: prediction and description [Fayyad et al., 1996]. *Prediction* focuses on prognosticating the identity of one thing based on the descriptions of other things whereas *description*, the goal around which this thesis revolves, aims at revealing a simple and concise description of the analysed data by means of user-friendly structures to capture regularities. Han et al. [2011] outline several fundamental data mining tasks that can be used to achieve these goals: classification, regression, clustering and pattern mining. These tasks have been used in various applications, e.g. e-commerce [Ansari et al., 2001], social communities [Jay et al., 2008], education [Al-Twijri and Noaman, 2015], agriculture [Pitarch et al., 2015], anomaly detection [Agrawal and Agrawal, 2015], healthcare [Jothi et al., 2015] and accounting [Amani and Fadlalla, 2017].

The choice of the data mining technique depends on the type of the analysed data, e.g. sequences, graphs, intervals or streams. **In this thesis we deal with sequences of itemsets**. Indeed, the exploration of sequential data is a major challenge in current research due to the progress in storing information regarding, e.g. customer purchase behaviours, patient physical examinations, football player evolutions and web access history. Accordingly, to simplify and summarise a set of sequences in a manner that domain experts can understand we rely on two data mining techniques, namely *sequential pattern mining* and *Formal Concept Analysis*. Before stating the motivation of this thesis, we present a brief overview of the two aforementioned techniques.

1.1.1 Sequential Pattern Mining

Discovering sequential patterns [Agrawal and Srikant, 1995] is a well-known data mining task whose aim is to find relevant subsequences in a sequence database (set of sequences) with respect to a measure of interest. For example, the measure of interest can be a user-defined minimum support, i.e. the minimum number of sequences that have to contain a discovered subsequence.

The objective of the sequential pattern mining task is to enumerate all relevant subsequences from a sequence database. Naively, this task can be tackled by computing the support of all possible subsequences, and next by enumerating only those for which the support is greater than or equal to a user-defined minimum support. Consequently, researchers have been working on efficiency-based methods that can enumerate the sequential patterns discovered in a sequence database without checking over all possible subsequences.

Recently, Fournier-Viger et al. [2017] have surveyed the up-to-date studies on sequential pattern mining and its applications. Given a particular input, we obtain the same set of sequential patterns by applying any of the classical methods, e.g. UDDAG [Chen, 2010] or CM-SPADE [Fournier-Viger et al., 2014]. However, these methods employ different data structures and algorithmic paradigms. Let us note that these methods rely on *propositional algorithms*, i.e. algorithms that explore data from a single table. In addition, Fournier-Viger et al. have underlined the key drawbacks of these classical methods such as:

- the huge number of generated sequential patterns that overwhelm domain experts during the pattern evaluation step;
- the limited amount of data captured by sequential patterns from the analysed sequence databases.

Therefore, researchers have shifted their focus to discovering either concise representations (i.e. patterns that summarise sequential patterns) or more informative sequential patterns (i.e. patterns that capture additional data recorded in a sequence database). Firstly, *closed sequential patterns* [Yan et al., 2003] and *closed partially-ordered patterns* [Casas-Garriga, 2005] are two concise representations. In addition, the number of sequential patterns can be reduced by pushing constraints into the mining process, e.g. regular expressions [Garofalakis et al., 1999]. Secondly, to discover more informative sequential patterns, e.g. Chowdhury Farhan et al. [2010] proposed the *high-utility sequential patterns* that capture in the context of the market basket problem the quantities of purchased items.

Finally, sequential pattern mining is useful in many real-life applications, e.g. bioinformatics [Liao and Chen, 2013], e-learning [Ziebarth et al., 2015] and text analysis [Pokou et al., 2016].

1.1.2 Formal Concept Analysis and Its Extensions

Formal Concept Analysis (FCA) was devised by Wille [1982] as a mathematical theory based on both the lattice and set theories [Barbut and Monjardet, 1970, Birkhoff, 1967]. FCA is a well-founded mathematical framework [Ganter and Wille, 1999] that can be used for various purposes, e.g. data analysis [Poelmans et al., 2010b] and information retrieval [Codocedo and Napoli, 2015].

FCA is appropriate for the unsupervised machine learning task. Indeed, given a bunch of binary data, FCA clusters the objects that have common attributes. A cluster is called *formal concept* and represents a pair of two maximal sets: objects (*extent*) and attributes (*intent*). FCA reveals a conceptual hierarchy (*concept lattice* represented as a directed acyclic graph (DAG)) that helps to visualise the considered data and to exhibit its intrinsic structure. In addition, the concept lattice is built without loss of information. Hence, on the one hand, no relevant details are overlooked, but on the other hand, the computation of the concept lattice becomes a time-consuming task. Kuznetsov and Obiedkov [2002] compared the existing methods for computing concept lattices. Recently, Andrews [2017] has proposed a fast method for deriving formal concepts, precisely IN-Close4. Furthermore, the complexity of concept lattices can be diminished by using, e.g. *iceberg lattices* [Stumme, 2002], *alpha Galois lattices* [Ventos and Soldano, 2005] or *expandable concept trees* [Melo et al., 2011].

Usually, real-life data are more complex (e.g. sequences, graphs, logical formulas or intervals) than binary data on which revolve classical FCA-based approaches. Therefore, researches have been focusing on introducing various theoretical extensions, e.g. *Conceptual Scaling* [Ganter and Wille, 1989], *Triadic Concept Analysis* [Lehmann and Wille, 1995], *Pattern Structures* [Ganter and Kuznetsov, 2001], *Relational Concept Analysis* [Rouane-Hacene et al., 2013] and *Graph-FCA* [Ferré, 2015].

Lastly, FCA and its extensions are useful in many real-life applications, e.g. software engineering [Wermelinger et al., 2009], environment [Braud et al., 2011] and chemistry [Stumpfe et al., 2011]. A systematic survey of the FCA-based applications is presented by Poelmans et al. [2013].

1.2 Motivation

Given a sequence database, we can obtain its description by means of classical sequential pattern mining methods, e.g. [Agrawal and Srikant, 1995]. Usually, the number of sequential patterns discovered in a sequence database is huge, and thus the pattern evaluation step is a laboured task for domain experts.

To diminish the huge number of sequential patterns, we can directly obtain a more compact set of these patterns, namely closed sequential patterns, by using existing methods, e.g. [Fournier-Viger et al., 2014]. Actually, a sequential pattern is closed if it is not contained in another sequential pattern that has the same support. Since the number of closed sequential patterns can still be quite large, a better option is to directly extract a more compact set of such sequential patterns, precisely closed partially-ordered patterns (cpo-patterns), by means of existing methods, e.g. [Pei et al., 2006]. Indeed, a cpo-pattern summarises a set of closed sequential patterns that coexist in the same analysed sequences, and, besides, it has a graphical

representation as a DAG that facilitates its evaluation.

To sum up, by choosing to discover cpo-patterns in a sequence database, on the one hand we obtain fewer patterns without loss of information; on the other hand, we help the pattern evaluation step thanks to their graphical representations. However, there are still some limitations of the existing methods for extracting cpo-patterns that we try to address in this thesis:

1. the evaluation step is not an easy task for domain experts since the discovered cpo-patterns are unorganised; thus, the experts should manually figure out how these cpo-patterns relate to each other;

2. domain experts do not have a global view of the discovered cpo-patterns; therefore, they can overlook pertinent cpo-patterns during the evaluation step;

3. some interesting cpo-patterns cannot be found since taxonomies over sequence-building items are not used;

4. the discovered cpo-patterns exploit only the order on itemsets from the analysed sequences, and thus the cpo-patterns do not capture the particularities hidden in these sequences.

1.3 Contributions

In this thesis we present an approach for enhancing the analysis of sequential data within the framework of *Relational Concept Analysis* (RCA), which is an extension of FCA. We have decided to rely on RCA rather than on FCA since the explored data are relational data and relations are not natively supported by FCA. In addition, we cope simultaneously with different types of relations, e.g. temporal/spatial and qualitative. Thus, the analysed sequences are built from a set of items that have associated qualitative values.

Basically, we devise a comprehensive KDD approach, namely **R**elational **C**oncept **A**nalysis for Exploring **Seq**uential Data (RCA-Seq, [Nica et al., 2017]), that exploits the relational structure of sequential data. Indeed, RCA classifies sets of objects described by attributes and relations, allowing the discovery of hierarchies of patterns. Figure 1.1 depicts the schema of RCA-Seq that is a fivefold approach:

1. *data preprocessing*: relying on domain knowledge, the data collected from a sequence database are prepared to be explored. Then, these data are remodelled in order to build the RCA input according to a proposed data model;

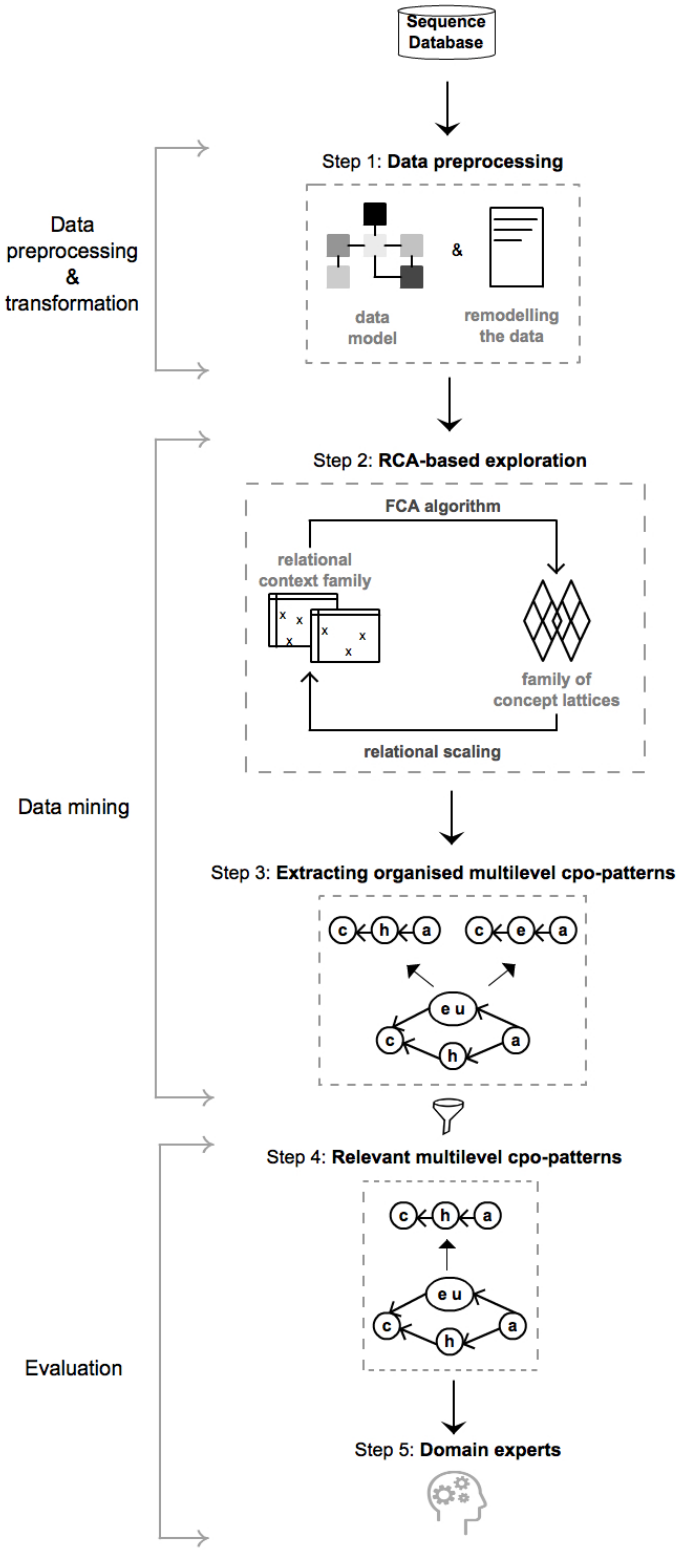


Figure 1.1: Overview of the RCA-SEQ approach [Nica et al., 2016b]

2. *RCA-based exploration*: the preprocessed data are encoded into the RCA input (i.e. relational context family). Iteratively, an FCA algorithm and a relational scaling mechanism (which highlights the relations between the considered objects) are applied to the RCA input in order to derive interconnected concept lattices (i.e. family of concept lattices);

3. *extracting organised multilevel cpo-patterns*: the concept lattices are navigated in order to extract a hierarchy of multilevel cpo-patterns;

4. *relevant multilevel cpo-patterns*: the discovered multilevel cpo-patterns are filtered based on various measures of interest;

5. *domain experts*: the relevant multilevel cpo-patterns are evaluated in order to draw valuable insights.

In the following, we outline the key contributions of this thesis:

– **the proposal and formalisation of a novel problem, namely directly extracting cpo-patterns that are implicitly organised into a hierarchy**, which is a more difficult task than only enumerating cpo-patterns discovered in sequential data. Indeed, existing works, e.g. [Cellier et al., 2011], have demonstrated that a hierarchical order on already extracted patterns helps in understanding the obtained knowledge, and, besides, it provides a quick way to navigate to interesting patterns;

– **the RCA-Seq approach that is the first attempt to explore sequential data by means of RCA**. In addition, it is a multi-relational data mining [Džeroski, 2003] approach that looks for regularities in sequential data whose sequences are built from multiple tables out of a relational database. Indeed, the itemsets of a sequence are instances from different tables. Hence, these itemsets are defined and ordered according to inter-table and intra-table relations. RCA yields concept lattices (one for each considered table) that are interconnected through various relations. Consequently, domain experts can find regularities hidden in the analysed sequential data by navigating amongst the interrelated concepts of these lattices;

– **a generic data model that allows the conversion of various qualitative sequential data** (e.g. medical or hydro-ecological) **into the RCA input**. This model also underpins the navigation of the obtained concept lattices;

– **an algorithm** CPOHrchy **that automatically navigates the obtained concept lattices in order to extract a hierarchy of cpo-patterns**. Since the RCA output is complex, i.e. it contains several lattices whose concepts are interrelated through various relations, we

automate its navigation. This algorithm relies on the structure and the properties of the RCA output in order to directly obtain the minimal representations of cpo-patterns without involving post-processing;

– **the extraction of multilevel cpo-patterns, namely concrete, abstract and hybrid, without specifically preprocessing the original sequential data**. Indeed, RCA allows to discover a partial order on items, and thus abstract and hybrid cpo-patterns are obtained rather than only concrete (standard, [Casas-Garriga, 2005]) cpo-patterns. In addition, the abstract cpo-patterns highlight general regularities of the analysed sequential data, while the hybrid cpo-patterns emphasise simultaneously general and specific regularities of these data;

– **the proposal of hierarchies of multilevel cpo-patterns with two generalisation levels**. Precisely, the generalisation regarding, firstly, the structure of cpo-patterns (e.g. the number of items, vertices and/or edges); secondly, the accuracy of items (e.g. from abstract to defined);

– **the proposal of weighted cpo-patterns that are more informative than standard cpo-patterns**. Indeed, additional information regarding the repetitive occurrences of specific itemsets in the analysed sequences can be discovered by exploiting the "richness" of the RCA output. Therefore, by means of weighted cpo-patterns we help the evaluation step by capturing and explicitly showing not only the order on itemsets (as standard cpo-patterns do), but also their different roles in the analysed sequences through new statistical measures;

– **the proposal of measures of interest for selecting and filtering concepts/cpo-patterns**. Usually, the number of generated concepts/cpo-patterns is quite large, but only a few of them are likely to be relevant for domain experts. Therefore, we propose to deal with the "concept explosion" problem by means of a new distribution index of a formal concept that makes use of the information encoded into the objects of the concept extent in order to determine if this concept is pertinent. Furthermore, we propose to filter the discovered cpo-patterns based on their accuracies;

– **a study of the RCA-SEQ approach adaptability**. We show that the proposed approach can be easily adapted to: (i) integrate a user-defined taxonomy over sequence-building items, and thus to obtain cpo-patterns containing items from different levels of this taxonomy; (ii) extract cpo-patterns with user-defined constraints on the order relations on itemsets from the analysed sequences; (iii) explore simple sequential data (i.e. the items do not have associated qualitative values); (iv) explore heterogeneous sequential

data (i.e. an itemset contains subsets of items from different domains) in order to obtain hierarchies of multilevel heterogeneous cpo-patterns.

The contributions of this thesis are presented and explained by means of an illustrative medical example. Then, these contributions are assessed through several quantitative statistics and qualitative interpretations resulting from experiments carried out on various hydro-ecological datasets. The hydro-ecological data were collected during two interdisciplinary research projects, namely Fresqueau[1] and REX[2].

1.4 Thesis Structure

In Chapter 2 we present the state of the art and the theoretical underpinnings of this thesis, i.e. sequential pattern mining and Formal Concept Analysis.

In Chapter 3 we present the first two steps of the RCA-Seq approach, namely the data preprocessing and the RCA-based exploration of sequential data. A generic data model is proposed. Then, relying on this model we explain how to encode a sequence database into the RCA input. In addition, the obtained RCA output is explained and analysed.

In Chapter 4 we present the third step of RCA-Seq, precisely the direct extraction of a hierarchy of multilevel cpo-patterns from the obtained RCA output. The structure and the properties of the RCA output are discussed. Then, we present an algorithm that automatically extracts multilevel cpo-patterns. In addition, a complexity analysis of RCA-Seq is given.

In Chapter 5 we present the fourth step of RCA-Seq, namely new measures of interest for guiding domain experts. The "richness" of the RCA output is exploited to compute the distribution index of a formal concept, to extract weighted cpo-patterns and to categorise the obtained multilevel cpo-patterns.

In Chapter 6 we discuss the adaptability of RCA-Seq. A user-defined taxonomy over sequence-building items, and, besides, user-defined constraints on the order relations on itemsets are pushed deep into the RCA-based exploration step. Then, we present how to explore simple sequential data and heterogeneous sequential data.

In Chapter 7 we present the application context of this thesis, i.e. hydro-ecology. We explain how to preprocess hydro-ecological data in order to apply the RCA-Seq approach. Then, we describe and discuss the results obtained from experiments carried out on various hydro-ecological datasets.

In Chapter 8 we conclude and give some perspectives of this thesis.

[1] http://engees-fresqueau.unistra.fr/presentation.php?lang=en

[2] http://obs-rhin.engees.eu

2 Preliminaries and State of the Art

Contents

2.1 Introduction

In the following sections, we present the definitions and principles of the theoretical underpinnings of this thesis, namely *sequential pattern mining* and *Formal Concept Analysis*, in order to contextualise and familiarise the reader with the terminology used in the next chapters. In addition, we mention the most relevant related work for this thesis.

2.2 Pattern Mining

An important subfield of data mining is *pattern discovery* and, in this thesis, we are particularly interested in mining patterns from sequential data.

2.2.1 Sequential Patterns

Sequential patterns were defined by Agrawal and Srikant [1995] as an extension of frequent itemsets and represent regularities hidden in a sequence database. *Discovering sequential patterns* is a data mining task whose aim is to obtain relevant subsequences from a set of analysed sequences. Usually, a subsequence is relevant if it occurs in many analysed sequences.

Formally, let $\mathcal{I} = \{I_1, I_2, ..., I_m\}$ be a fixed set of *items*. An *itemset* $IS = (I_{j_1}...I_{j_k})$, where $I_{j_i} \in \mathcal{I}$ and $I_{j_i} \neq I_{j_l} \forall i \neq l$, is a non-empty unordered set of items. An itemset IS with k items is referred to as *k-itemset*. We denote by $\mathcal{IS}$ the set of all itemsets built from $\mathcal{I}$.

Definition 2.1 (Sequence). *A* sequence $S = \langle IS_1 IS_2 ... IS_p \rangle$, *where* $IS_j \in \mathcal{IS}$, *is a non-empty ordered list of itemsets.*

In a sequence S the itemsets are ordered under a binary relation, denoted by $\leq_{IS}$, which is total, antisymmetric and transitive. Therefore, for any two distinct itemsets IS_α and IS_β in S it is possible to determine if IS_α precedes IS_β ($IS_\alpha \leq_{IS} IS_\beta$) or IS_β precedes IS_α ($IS_\beta \leq_{IS} IS_\alpha$). It is worthwhile to mention that the binary relation $\leq_{IS}$ can be a temporal relation (e.g. New Year's Eve 2015 *is preceded by* New Year's Eve 2013), a topological relation (e.g. Polygon1 *contains* Polygon2), a directional relation (e.g. (Canada US) *is north of* Mexico), a part of relation (e.g. cytoplasm *is part of* cell) or any other order in which related itemsets follow each other.

An item can occur only once in an itemset, but can occur several times in different itemsets of the same sequence. A sequence S with p itemsets is referred to as a *p-sequence*. The length of a sequence $S = \langle IS_1 IS_2 ... IS_p \rangle$, denoted by $l(S)$, is the total number of items in S.

$$l(S) = \sum_{i=1}^{p} |IS_i| \tag{2.1}$$

Definition 2.2 (Subsequence). *A sequence $S = \langle IS_1 IS_2 ... IS_p \rangle$ is a* subsequence *of another sequence $S' = \langle IS'_1 IS'_2 ... IS'_q \rangle$, denoted by $S \preceq_s S'$, if $p \leq q$ and if there are integers $j_1 < j_2 < ... < j_k < ... < j_p$ such that $IS_1 \subseteq IS'_{j_1}, IS_2 \subseteq IS'_{j_2}, ..., IS_p \subseteq IS'_{j_p}$.*

Definition 2.3 (Sequence Database). *A* sequence database, *denoted by $\mathcal{D}_\mathcal{S} = \{S_1, S_2, ..., S_n\}$, is a set of sequences, where each sequence has a unique identifier.*

In Tab. 2.1 is an illustrative example of a sequence database $\mathcal{D}_\mathcal{S}$ that contains three sequences $S1$, $S2$ and $S3$ built from the set of items $\mathcal{I} = \{a, b, c, d\}$. For example, $S1 = \langle (a)(b\ c)(d) \rangle$ is a 3-sequence that contains the (a) 1-itemset, the $(b\ c)$ 2-itemset and the (d) 1-itemset. A sequence $S' = \langle (a)(c) \rangle$ is a subsequence of $S1$, denoted by $S' \preceq_s S1$, since $(a) \subseteq (a)$, $(c) \subseteq (b\ c)$ and the order on itemsets is preserved. Moreover, $S' \preceq_s S3$, and therefore we can say that subsequence S' occurs often in $\mathcal{D}_\mathcal{S}$ since it is contained in 2 out of 3 analysed sequences.

Table 2.1: Example of a sequence database $\mathcal{D}_\mathcal{S}$

Sequence Id	Sequence
$S1$	$\langle (a)(b\ c)(d) \rangle$
$S2$	$\langle (b)(c\ d) \rangle$
$S3$	$\langle (a)(b\ c)(a) \rangle$

Definition 2.4 (Subsequence Support). *Let $\mathcal{D}_\mathcal{S}$ be a sequence database. The* support *of a subsequence S', denoted by $Support(S')$, represents the total number of sequences in $\mathcal{D}_\mathcal{S}$ that contain S'.*

$$Support(S') = |\{S \in \mathcal{D}_\mathcal{S} | S' \preceq_s S\}| \tag{2.2}$$

Furthermore, the *frequency* of S', denoted by $Freq(S')$, is the relative number of sequences in $\mathcal{D}_\mathcal{S}$ that contain S'.

$$Freq(S') = \frac{Support(S')}{|\mathcal{D}_\mathcal{S}|} \tag{2.3}$$

To illustrate this, in Tab. 2.1 subsequence $S' = \langle (a)(c) \rangle$ has $Support(S') = |\{S1, S3\}| = 2$ and $Freq(S') = \frac{|\{S1,S3\}|}{|\{S1,S2,S3\}|} \approx 0.67$.

Definition 2.5 (Sequential Pattern). *Given a sequence database $\mathcal{D}_\mathcal{S}$, a user-defined minimum support θ and a subsequence S'. S' is a frequent subsequence in $\mathcal{D}_\mathcal{S}$ according to θ if $Support(S') \geq \theta$. A frequent subsequence is called a* sequential pattern.

In Tab. 2.2 is listed the complete set of sequential patterns discovered in the sequence database given in Tab. 2.1 when the minimum support $\theta = 2$. For instance, sequential pattern $P1 = \langle (b) \rangle$ is a frequent subsequence since it is contained in $S1$, $S2$ and $S3$, i.e. $Support(P1) =$

Table 2.2: All sequential patterns discovered in the sequence database from Tab. 2.1 when $\theta = 2$

Unique Identifier	Sequential Pattern	Set of Sequences	Closed	Maximal
$P1$	$\langle(b)\rangle$	$\{S1, S2, S3\}$	✓	
$P2$	$\langle(c)\rangle$	$\{S1, S2, S3\}$	✓	
$P3$	$\langle(a)\rangle$	$\{S1, S3\}$		
$P4$	$\langle(d)\rangle$	$\{S1, S2\}$		
$P5$	$\langle(b\ c)\rangle$	$\{S1, S3\}$		
$P6$	$\langle(a)(b)\rangle$	$\{S1, S3\}$		
$P7$	$\langle(a)(c)\rangle$	$\{S1, S3\}$		
$P8$	$\langle(a)(b\ c)\rangle$	$\{S1, S3\}$	✓	✓
$P9$	$\langle(b)(d)\rangle$	$\{S1, S2\}$	✓	✓

$|\{S1, S2, S3\}| = 3 \geq \theta$. However, subsequence $S' = \langle(b)(c)\rangle$ is not a sequential pattern since it is contained only in $S2$, i.e. $Support(S') = |\{S2\}| = 1 \not\geq \theta$.

It is noted, in Tab. 2.2, that 9 sequential patterns are discovered. This number is quite large compared with the number of the analysed sequences (3 in Tab. 2.1). Indeed, if a sequence database $\mathcal{D}_\mathcal{S}$ contains a sequential pattern $P = \langle IS_1 IS_2 ... IS_q\rangle$ (where the itemsets have distinct items), then $\mathcal{D}_\mathcal{S}$ contains at most

$$2^{l(P)} - 1 \tag{2.4}$$

frequent subsequences of P as well. For example, sequential pattern $P9 = \langle(b)(d)\rangle$ is obtained, but also its frequent subsequences $P1 = \langle(b)\rangle$ and $P4 = \langle(d)\rangle$. Since $P4$ and $P9$ have the same support $Support(P4) = Support(P9) = |\{S1, S2\}| = 2$, and, besides, $P4$ can be discovered from $P9$, Yan et al. [2003] proposed a more concise representation of sequential patterns, namely closed sequential patterns.

Definition 2.6 (Closed Sequential Pattern). *Given a sequence database $\mathcal{D}_\mathcal{S}$ and a sequential pattern P. P is* closed *if there is no sequential pattern P' in $\mathcal{D}_\mathcal{S}$ such that $P \preceq_s P'$ and $Support(P) = Support(P')$.*

Let us consider again the $P4$ and $P9$ sequential patterns that have the same support. Since there exists no P such that $P9 \preceq_s P$, $P9$ is closed, while $P4$ is not closed. By analysing the Closed column in Tab. 2.2, we notice that there are only 4 closed sequential patterns extracted in our example, i.e. the number of extracted patterns is decreased by 56% without loss of information. Indeed, by discovering closed sequential patterns, the set of the most representative sequential patterns is obtained. In addition, the complete set of sequential patterns can be recovered from these closed sequential patterns.

Nevertheless, when the analysed database contains long sequences, the set of closed sequential patterns is still too large. To address this problem, Luo and Chung [2005] introduced

a more concise representation, namely maximal sequential pattern. Here, we recall its definition, but this pattern will not be explored further in this thesis since we focus on closed sequential patterns.

Definition 2.7 (Maximal Sequential Pattern). *Given a sequence database $\mathcal{D}_\mathcal{S}$ and a sequential pattern P. P is* maximal *if there is no sequential pattern P' in $\mathcal{D}_\mathcal{S}$ such that $P \preceq_s P'$.*

By analysing the Maximal column in Tab. 2.2, there are only 2 sequential patterns that are maximal, and thus the number of patterns is decreased by 78%. For example, $P9$ is maximal since there is no P such that $P9 \preceq_s P$.

2.2.2 Heterogeneous Sequential Patterns

Egho et al. [2014] have proposed an approach for mining heterogeneous sequences, where a sequence contains itemsets whose items are from distinct domains. In addition, an item can be an atomic item from a partially ordered set or it can be a subset of items from an unordered set. In the following, we formalise a generalisation of a heterogeneous sequence, precisely, we consider that its itemsets contain other itemsets.

Suppose now that there is a *partial order* (i.e. a reflexive, antisymmetric and transitive binary relation) on the set of items $\mathcal{I}$, denoted by $(\mathcal{I}, \leq)$. We say that $(\mathcal{I}, \leq)$ is a *poset*.

Definition 2.8 (Multilevel Itemset). *A* multilevel itemset $IS_{ml} = (I_{j_1}...I_{j_k})$, *where $I_{j_i} \in \mathcal{I}$ and $\nexists I_{j_i}, I_{j_{i'}} \in IS_{ml}$ such that $I_{j_i} \leq I_{j_{i'}}$, is a non-empty and unordered set of items that can be at different levels of granularity (i.e. items from different levels of poset $(\mathcal{I}, \leq)$).*

We denote by $\mathcal{IS}_{ml}$ the set of all multilevel itemsets built from $(\mathcal{I}, \leq)$. The partial order on the set of all multilevel itemsets $(\mathcal{IS}_{ml}, \subseteq_{ml})$ is defined as follows: $IS_{ml} \subseteq_{ml} {IS'}_{ml}$ if $\forall I_j \in IS_{ml}, \exists I_{j'} \in IS'_{ml}, I_{j'} \leq I_j$ and $\forall I_l \neq I_j, \exists I_{l'} \neq I_{j'}$ such that $I_{l'} \leq I_l$. The order on the *multilevel sequences*, i.e. sequences that contain multilevel itemsets, is defined accordingly.

To illustrate this, let us consider $\mathcal{I}_1 = \{a, b, c, d, e, Consonants, Vowels, Letters\}$ a set of items and $(\mathcal{I}_1, \leq)$ a partial order depicted in Fig.2.1, where an edge represents the binary relation *is-a*, denoted by $\leq$.

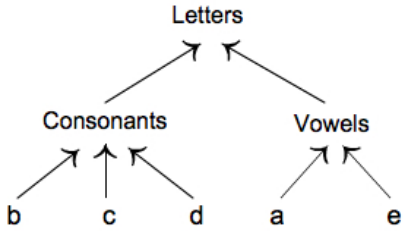


Figure 2.1: An example of a partial order on $\mathcal{I}_1 = \{a, b, c, d, e, Consonants, Vowels, Letters\}$

For example, $a \leq \mathit{Vowels}$ designates that the letter "a" is a vowel. Let be two itemsets $(a\ b\ c)$ and $(a\ \mathit{Consonants})$, then $(a\ \mathit{Consonants}) \subseteq_{ml} (a\ b\ c)$ since $a \leq a$ and $b \leq \mathit{Consonants}$ (or $c \leq \mathit{Consonants}$).

Let $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_n\}$ be a set of distinct sets of items, where $\mathcal{I}_j$ with $j \in \{1, ..., n\}$ represents a domain. We note that $\mathcal{I}_j$ can be a poset or an unordered set. Let $\mathcal{IS}_j$ be the set of all itemsets built from $\mathcal{I}_j \in \mathcal{H}$.

Definition 2.9 (Heterogeneous Itemset). *A* heterogeneous itemset $IS_{\mathcal{H}} = \{IS_1, IS_2, ..., IS_n\}$, *where $IS_j \in \mathcal{IS}_j$, is a non-empty and unordered set of itemsets built from distinct sets of $\mathcal{H}$.*

Moreover, a *multilevel heterogeneous itemset* is a set of itemsets that has at least one multilevel itemset.

Let $\mathcal{IS}_{\mathcal{H}}$ be the set of all heterogeneous itemsets built from $\mathcal{H}$. The partial order $(\mathcal{IS}_{\mathcal{H}}, \subseteq_{\mathcal{H}})$ is defined as follows: $IS_{\mathcal{H}} \subseteq_{\mathcal{H}} IS'_{\mathcal{H}}$ if $\forall IS_k \in IS_{\mathcal{H}}$, $\exists IS'_k \in IS'_{\mathcal{H}}$ such that $IS_k \subseteq IS'_k$, where $IS_k, IS'_k \in \mathcal{IS}_k$, $k \in \{1, ..., n\}$. The order on multilevel heterogeneous itemsets is defined accordingly relying on $\subseteq_{ml}$.

To illustrate this, let us consider $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2\}$, where $\mathcal{I}_1$ is partially ordered as shown in Fig. 2.1 and $\mathcal{I}_2 = \{\square, \lozenge, \triangle\}$ is an unordered set of shapes. Furthermore, let be two multilevel heterogeneous itemsets $IS_{\mathcal{H}_1} = \{(\mathit{Vowels}\ c), (\lozenge)\}$ and $IS_{\mathcal{H}_2} = \{(a\ c), (\square\ \lozenge)\}$, then $IS_{\mathcal{H}_1} \subseteq_{\mathcal{H}} IS_{\mathcal{H}_2}$ since $(\mathit{Vowels}\ c) \subseteq_{ml} (a\ c)$ (that is $a \leq \mathit{Vowels}$ and $c \leq c$) and $(\lozenge) \subseteq (\square\ \lozenge)$.

Definition 2.10 (Heterogeneous Sequence). *A* heterogeneous sequence $S_{\mathcal{H}} = \langle IS_{\mathcal{H}_1} IS_{\mathcal{H}_2} ... IS_{\mathcal{H}_r} \rangle$, *where $IS_{\mathcal{H}_i} \in \mathcal{IS}_{\mathcal{H}}$ with $i \in \{1, ..., r\}$, is a non-empty ordered list of heterogeneous itemsets.*

In addition, a heterogeneous sequence that has at least one multilevel heterogeneous itemset represents a *multilevel heterogeneous sequence* (hereinafter referred to as heterogeneous sequence). A heterogeneous sequence $S_{\mathcal{H}} = \langle IS_{\mathcal{H}_1} IS_{\mathcal{H}_2} ... IS_{\mathcal{H}_r} \rangle$ is a subsequence of another heterogeneous sequence $S'_{\mathcal{H}} = \langle IS'_{\mathcal{H}_1} IS'_{\mathcal{H}_2} ... IS'_{\mathcal{H}_q} \rangle$, denoted by $S_{\mathcal{H}} \preceq_{s_{\mathcal{H}}} S'_{\mathcal{H}}$, if $r \leq q$ and if there are integers $j_1 < j_2 < ... < j_k < ... < j_r$ such that $IS_{\mathcal{H}_1} \subseteq_{\mathcal{H}} IS'_{\mathcal{H}_{j_1}}, IS_{\mathcal{H}_2} \subseteq_{\mathcal{H}} IS'_{\mathcal{H}_{j_2}}, ..., IS_{\mathcal{H}_r} \subseteq_{\mathcal{H}} IS'_{\mathcal{H}_{j_r}}$. Accordingly, the order on multilevel heterogeneous sequences is defined.

To illustrate this, let be two heterogeneous sequences on the aforementioned $\mathcal{H} = \{\mathcal{I}_1, \mathcal{I}_2\}$:

- $S1_{\mathcal{H}} = \langle \{(a\ \mathit{Consonants}), (\square\ \lozenge)\}\ \{(\mathit{Letters}), \emptyset\} \rangle$ and
- $S2_{\mathcal{H}} = \langle \{(a\ d), (\square\ \triangle\ \lozenge)\}\ \{(a\ c), (\square)\} \rangle$,

then $S1_{\mathcal{H}} \preceq_{s_{\mathcal{H}}} S2_{\mathcal{H}}$ since

- $\{(a\ \mathit{Consonants}), (\square\ \lozenge)\} \subseteq_{\mathcal{H}} \{(a\ d), (\square\ \triangle\ \lozenge)\}$, i.e. $a \leq a$, $d \leq \mathit{Consonants}$, $(\square\ \lozenge) \subseteq (\square\ \triangle\ \lozenge)$,
- $\{(\mathit{Letters}), \emptyset\} \subseteq_{\mathcal{H}} \{(a\ c), (\square)\}$, i.e. $a \leq \mathit{Letters}$ (or $c \leq \mathit{Letters}$), $\emptyset \subseteq (\square)$.

Following Def. 2.5 and Eq. 2.2, we define a heterogeneous sequential pattern as follows.

Definition 2.11 (Heterogeneous Sequential Pattern). *Given a heterogeneous sequence database $\mathcal{D}_{\mathcal{S}_\mathcal{H}}$, a user-defined minimum support θ and a heterogeneous subsequence $S'_\mathcal{H}$. $S'_\mathcal{H}$ is a frequent heterogeneous subsequence in $\mathcal{D}_{\mathcal{S}_\mathcal{H}}$ according to θ if $Support(S'_\mathcal{H}) \geq \theta$. A frequent heterogeneous subsequence is called a* heterogeneous sequential pattern.

Definition 2.12 (Closed Heterogeneous Sequential Pattern). *Given a heterogeneous sequence database $\mathcal{D}_{\mathcal{S}_\mathcal{H}}$ and a heterogeneous sequential pattern $P_\mathcal{H}$. $P_\mathcal{H}$ is* closed *if there is no heterogeneous sequential pattern $P'_\mathcal{H}$ in $\mathcal{D}_{\mathcal{S}_\mathcal{H}}$ such that $P_\mathcal{H} \preceq_s P'_\mathcal{H}$ and $Support(P_\mathcal{H}) = Support(P'_\mathcal{H})$.*

2.2.3 Closed Partially-Ordered Patterns

Closed partially-ordered patterns were introduced by Casas-Garriga [2005] in order to synthesise sets of closed sequential patterns. The closed sequential patterns from a set coexist exactly in the same sequences from a sequence database.

Formally, let $\mathcal{D}_\mathcal{S}$ be a sequence database and P, P' two sequential patterns discovered in $\mathcal{D}_\mathcal{S}$. When both P and P' are contained in the same sequences in $\mathcal{D}_\mathcal{S}$, a more concise representation of P and P' can be obtained by relying on a partial order on the set of their itemsets.

Definition 2.13 (Partially-Ordered Pattern (po-pattern)). *A* partially-ordered pattern, *called po-pattern, is a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, l)$. $\mathcal{V}$ is a set of vertices, $\mathcal{E}$ is a set of directed edges such that $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ and l is a labelling function mapping each vertex to an itemset. This structure allows to define a strict partial order on vertices u and v such that $u \neq v : u < v$ if there is a directed path from the* tail *vertex u to the* head *vertex v. However, if there is no directed path from u to v, these elements are not comparable. Each path of the graph represents a sequential pattern and the set of paths in $\mathcal{G}$ is denoted by $\mathcal{P}_\mathcal{G}$. A po-pattern is associated with the set of sequences $\mathcal{S}_\mathcal{G}$ that contain all paths of $\mathcal{P}_\mathcal{G}$.*

Figure 2.2 illustrates a few po-patterns associated with the set of sequences $\{S1, S3\}$ from the sequence database given in Tab. 2.1, except for po-pattern $\mathcal{G}_6$ (Fig.2.2f) associated with $\{S1, S2\}$ and po-pattern $\mathcal{G}_7$ (Fig. 2.2g) associated with $\{S1, S2, S3\}$. For instance, $\mathcal{G}_1$ (Fig.2.2a) is a concise representation of the $P6 = \langle(a)(b)\rangle$ and $P7 = \langle(a)(c)\rangle$ sequential patterns shown in Tab. 2.2 that coexist in the same sequences $S1$ and $S3$. Indeed, the set of itemsets in $P6$ and $P7$, denoted by $\mathcal{IS}_{P6\&P7} = \{(a), (b), (c)\}$ and the binary relation *precedes*, referred to as $<$, build the partial order $(\mathcal{IS}_{P6\&P7}, <)$. Thus, the same itemset (a) is comparable with the (b) and (c) itemsets, i.e. $(a) < (b)$ and $(a) < (c)$, while the (b) and (c) itemsets are not comparable.

Since a po-pattern is associated with a set of sequences, its support can be defined following Eq. 2.2.

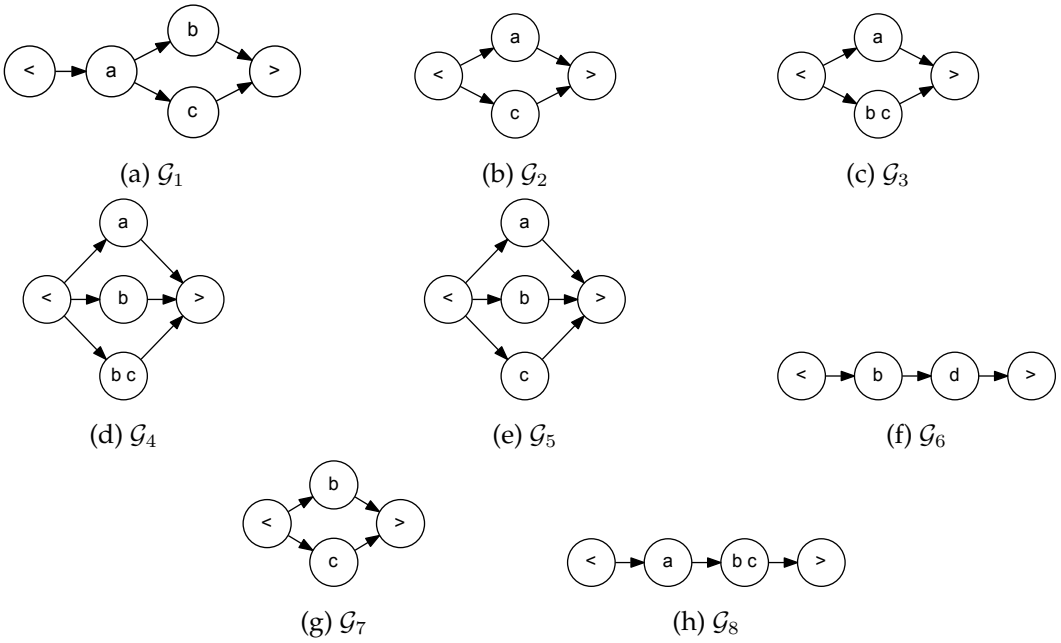


Figure 2.2: Several po-patterns that summarise the sequential patterns associated with $\{S1, S3\}$ (Tab. 2.1), except for po-pattern $\mathcal{G}_6$ associated with $\{S1, S2\}$ and po-pattern $\mathcal{G}_7$ associated with $\{S1, S2, S3\}$

Definition 2.14 (PO-Pattern Support). *Let $\mathcal{D}_\mathcal{S}$ be a sequence database, $\mathcal{G}$ a po-pattern, and $\mathcal{P}_\mathcal{G}$ its set of paths. The* support *of $\mathcal{G}$, denoted by $Support(\mathcal{G})$, represents the total number of sequences in $\mathcal{D}_\mathcal{S}$ that contain all paths in $\mathcal{P}_\mathcal{G}$.*

$$Support(\mathcal{G}) = |\mathcal{S}_\mathcal{G}| = |\{S \in \mathcal{D}_\mathcal{S} | \forall P \in \mathcal{P}_\mathcal{G}, P \preceq_s S\}| \tag{2.5}$$

Five po-patterns depicted in Fig.2.2 have the same support and are associated with the same set of sequences, precisely $Support(\mathcal{G}_1) = Support(\mathcal{G}_2) = Support(\mathcal{G}_3) = Support(\mathcal{G}_4) = Support(\mathcal{G}_5) = |\{S1, S3\}| = 2$.

Moreover, there are po-patterns that are contained in other po-patterns. For example, po-pattern $\mathcal{G}_2$ (Fig.2.2b) summarises the $P2 = \langle(c)\rangle$ and $P3 = \langle(a)\rangle$ sequential patterns from Tab. 2.2, while po-pattern $\mathcal{G}_5$ (Fig.2.2e) summarises the $P1 = \langle(b)\rangle$, $P2$, and $P3$ sequential patterns. Then, we can say that $\mathcal{G}_2$ is a sub po-pattern of $\mathcal{G}_5$.

Definition 2.15 (Sub PO-Pattern). *Given two po-patterns $\mathcal{G}$ and $\mathcal{G}'$ with $\mathcal{P}_\mathcal{G}$ and $\mathcal{P}_{\mathcal{G}'}$ their sets of paths. $\mathcal{G}'$ is a* sub po-pattern *of $\mathcal{G}$, denoted by $\mathcal{G}' \preceq_g \mathcal{G}$, if $\forall P' \in \mathcal{P}_{\mathcal{G}'}, \exists P \in \mathcal{P}_\mathcal{G}$ such that $P' \preceq_s P$.*

A set of sequential patterns can have multiple concise representations. For instance, the po-patterns $\mathcal{G}_3$ and $\mathcal{G}_4$ (Fig. 2.2c and 2.2d) synthesise the set of sequential patterns $\mathcal{P} = \{\langle(b)\rangle, \langle(a)\rangle, \langle(b\ c)\rangle\}$, i.e. $\mathcal{G}_3 \preceq_g \mathcal{G}_4$ and $\mathcal{G}_4 \preceq_g \mathcal{G}_3$. Therefore, one of these po-patterns is redundant and only the most compact representation of $\mathcal{P}$ should be discovered.

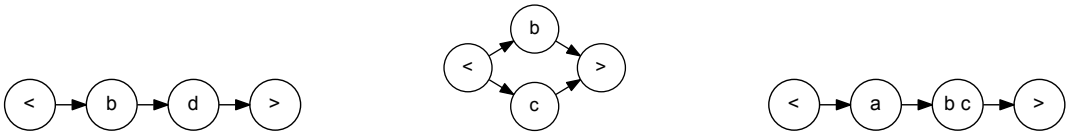
Definition 2.16 (Minimal PO-Pattern). *Given a set of sequential patterns $\mathcal{P}$ and its concise representation $\mathcal{G} = (\mathcal{V}, \mathcal{E}, l)$. The po-pattern $\mathcal{G}$ is* minimal *if there is no other concise representation $\mathcal{G}' = (\mathcal{V}', \mathcal{E}', l')$ of $\mathcal{P}$ such that $\mathcal{G}' \preceq_g \mathcal{G}$ and $\mathcal{G} \preceq_g \mathcal{G}'$ with $|\mathcal{V}'| < |\mathcal{V}|$ or $|\mathcal{E}'| < |\mathcal{E}|$.*

Now, following Def. 2.16, po-pattern $\mathcal{G}_4$ is not minimal since path $\langle(b)\rangle$ is contained in path $\langle(b\ c)\rangle$ and $|\mathcal{V}_4| > |\mathcal{V}_3|$. Thus, $\mathcal{G}_3$ is the minimal po-pattern that should be extracted and po-pattern $\mathcal{G}_4$ is redundant.

It is worthwhile to mention that the number of obtained po-patterns can explode, and therefore inspired by the closure property of sequential patterns, more representative po-patterns can be extracted.

Definition 2.17 (Closed PO-Pattern (cpo-pattern)). *Let $\mathcal{D}_\mathcal{S}$ be a sequence database and $\mathcal{G}$ a po-pattern. The po-pattern $\mathcal{G}$ is* closed*, referred to as cpo-pattern, if it is minimal and there is no po-pattern $\mathcal{G}'$ in $\mathcal{D}_\mathcal{S}$ such that $\mathcal{G} \prec_g \mathcal{G}'$ with $\mathcal{S}_\mathcal{G} = \mathcal{S}_{\mathcal{G}'}$.*

Figure 2.3 shows the set of cpo-patterns that summarise the sequential patterns given in Tab. 2.2. It is noted that there is a cpo-pattern for each distinct set of sequences, and hence the number of obtained patterns is decreased. For example, cpo-pattern $\mathcal{G}_7$ (Fig.2.3b) synthesises the $P1 = \langle(b)\rangle$ and $P2 = \langle(c)\rangle$ closed sequential patterns contained in the sequences $S1$, $S2$ and $S3$ shown in Tab. 2.1.



(a) $\mathcal{G}_6$ on $\{S1, S2\}$ (b) $\mathcal{G}_7$ on $\{S1, S2, S3\}$ (c) $\mathcal{G}_8$ on $\{S1, S3\}$

Figure 2.3: The set of cpo-patterns that synthesise the sequential patterns discovered in the sequence database shown in Tab. 2.1 when $\theta = 2$

2.2.4 Related Work

After Agrawal and Srikant [1995] had introduced the sequential pattern mining problem, researchers focused on extracting more efficiently (closed) sequential patterns/cpo-patterns, i.e. mining patterns with low execution time and low memory usage when defining a low minimum support or when dealing with very large databases. Zhao and Bhowmick [2003], Mabroukeh and Ezeife [2010], Mooney and Roddick [2013] and Fournier-Viger et al. [2017] surveyed the existing approaches and algorithms for sequential pattern mining. There are two main approaches, namely *apriori-based* and *pattern-growth*.

Initially, the apriori-based method was proposed. This relies on the candidate-generation-and-test principle. The key drawbacks of this approach are the huge number of generated

candidates (since candidates that do not exist in the analysed database can be generated) and the repeated full scans of the analysed database in order to evaluate the support of these candidates. The pattern-growth method avoids completely the candidate generation step by constructing recursively longer frequent subsequences only from the shorter frequent ones. To this end, the analysed database is compressed into a frequent pattern tree that is then partitioned in order to explore small amounts of data.

In the following, we briefly recall the main algorithms for mining (closed) sequential patterns/cpo-patterns. In addition, since exploring sequential data with such algorithms can generate a huge number of patterns that makes difficult their evaluation by domain experts, we recall a few measures of interest proposed for ranking and filtering patterns.

2.2.4.1 Discovering (Heterogeneous) Sequential Patterns

Algorithm AprioriAll, proposed by Agrawal and Srikant [1995], relies on the Apriori property [Agrawal and Srikant, 1994]: *if a sequential pattern P is not frequent in a sequence database $\mathcal{D}_\mathcal{S}$, then $\forall P' \in \mathcal{D}_\mathcal{S}$ such that $P \preceq_s P'$, P' is not frequent*. Briefly, the algorithm spans three steps. Firstly, the frequent itemsets are found. Secondly, the original sequences are transformed into sequences of frequent itemsets. Finally, the transformed sequences are mined iteratively to discover sequential patterns. The last step begins with the frequent itemsets that are used to generate new possible candidates (sequences of frequent itemsets). These candidates, excepting the ones that are not frequent, are the input of the next iteration. The complete set of sequential patterns is obtained when there is no new candidate or there is no frequent candidate. The same authors introduced algorithm GSP [Srikant and Agrawal, 1996] that outperforms AprioriAll due to the *hash-tree* data structure used to organise the generated candidate sequences.

Algorithm SPADE, proposed by Zaki [2001], is an apriori-based approach that encodes a sequence database into a vertical *id-list* database format (each item I is associated with a list of pairs (S, id_{IS}), where S uniquely identifies a sequence and id_{IS} uniquely identifies an itemset of S where I occurs). All sequential patterns are discovered in only three scans of the vertical database. Firstly, the frequent 1-sequences are computed; secondly, the frequent 2-sequences. Finally, based on joins of the id-lists of pairs and on Lattice Theory [Davey and Priestley, 1990], the search space of the candidate sequences (decomposed into sub-lattices) is generated and all sequential patterns are enumerated using a breadth-first/depth-first search strategy. Fournier-Viger et al. [2014] presented an improved version of SPADE, namely CM-Spade, that the authors claim it to be the fastest algorithm.

Algorithm PrefixSpan, proposed by Pei et al. [2001], is an efficient pattern-growth approach that relies on the FreeSpan algorithm [Han et al., 2000]. The main advantage of this algorithm is the confined search space (projected database) used to look for frequent sub-

sequences in each step, i.e. no candidate subsequence has to be generated or tested if it is not in a projected database. Briefly, a sequence database is scanned to find the frequent 1-itemsets. Then, the sequence database is divided according to the number of obtained frequent 1-itemsets into projected databases. An α-*projected database* consists in the postfix of the subsequences whose prefix is the frequent 1-itemset (α). Each α-*projected database* is scanned once to obtain the sequential patterns with 2 items and having the prefix (α). These new sequential patterns are used to partition again the α-*projected database*. The same idea is executed recursively until the projected databases are empty or no more sequential patterns can be obtained.

Algorithm SPAM, introduced by Ayres et al. [2002], is a mixture of the techniques employed by the aforementioned algorithms: GSP, SPADE and FreeSpan. Empirically it is shown that SPAM is mainly efficient when the mined sequential patterns are very long. The authors claim that SPAM is the first algorithm that uses the depth-first search strategy for mining sequential patterns. CM-Spam [Fournier-Viger et al., 2014] is an improved version of SPAM.

Since the number of obtained sequential patterns can be quite large, Yan et al. [2003] introduced CloSpan, i.e. the first algorithm that generates the complete set of closed sequential patterns from a sequence database. Basically, the algorithm relies on PrefixSpan but it generates a more compact set of sequential patterns (when a frequent α-subsequence is found its α-projected database is not mined if all possible descendants of α-subsequence have been discovered before). Then, the compact set is post-pruned to eliminate non-closed sequential patterns.

The bottleneck of CloSpan is the space used to record the compact set of sequential patterns, which are needed for pattern closure checking. A solution to this problem is given by Wang and Han [2004] who proposed the BIDE algorithm that does not keep track of the discovered closed sequential patterns. The authors proposed an adapted technique for checking the pattern closure. Moreover, this algorithm mines sequences of items, but the authors show how it can be extended to sequences of itemsets. ClaSP [Gomariz et al., 2013], CM-ClaSP [Fournier-Viger et al., 2014] and CloFast [Fumarola et al., 2016] are more recent algorithms for mining closed sequential patterns based on the vertical database format. It is shown that these algorithms outperform CloSpan and BIDE.

In addition, the aforementioned sequential pattern mining algorithms consider only the order on itemsets in the analysed sequences and treat all the itemsets uniformly. To capture more particularities hidden in the analysed data, Srikant and Agrawal [1996] added time constraints in advance, and thus a sequential pattern is extracted only if it admits a max-gap and a min-gap between adjacent itemsets. Pei et al. [2002] pushed various constraints, e.g. time-interval and gap information between items, into the mining process to limit the results.

Chen et al. [2003] proposed to extract time-interval sequential patterns that reveal the time interval between successive items, and, besides, these time intervals are explicitly shown in the patterns. To capture the time interval between all the pairs of items in the extracted patterns, Hu et al. [2009] introduced the multi-time-interval sequential patterns. Chang [2011] proposed to find weighted sequential patterns by pushing a time-interval weight measure (the weight of a sequence derived from the time intervals of the sequence itemsets) into the mining process. Furthermore, in [Kim et al., 2007] and [Yun, 2008] more informative sequential patterns are obtained by pushing pre-assigned quantitative information, which are recorded in the analysed database, into the mining process.

However, these algorithms extract sequential patterns whose items are homogeneous, and therefore cannot be applied to mine heterogeneous sequences, i.e. sequences whose items are different in nature. To our knowledge, Pinto et al. [2001] proposed the first work for exploring heterogeneous sequential data, which are called multidimensional sequential data. A multidimensional sequence takes the form $(d_1, d_2, ..., d_m, S)$, where S is a sequence of itemsets and d_i represents the i^{th} type of information associated with S. The authors proposed three approaches for extracting multidimensional sequential patterns from such data that rely on the PrefixSpan algorithm and/or the BUC algorithm [Beyer and Ramakrishnan, 1999].

A key drawback of such multidimensional sequences is the additional information that is constant for all itemsets of sequence S. Plantevit et al. [2010] proposed to discover multidimensional sequential patterns in multidimensional databases. A multidimensional sequence is defined as an ordered list of multidimensional items. A multidimensional item takes the form $(d_1, d_2, ..., d_n)$, where d_k is an item of the k^{th} dimension. Furthermore, each considered dimension is represented at different levels of granularity by means of partial orders, and hence multilevel sequential patterns can be discovered, as explained in [Srikant and Agrawal, 1996]. Plantevit et al. proposed algorithm M3SP that searches for multidimensional and multilevel sequential patterns in two steps. First, the most specific frequent multidimensional items, referred to as *maf-sequences*, are found. Second, the maf-sequences are used to remodel the original multidimensional sequences, and then these sequences are mined by using the SPADE algorithm.

Nevertheless, Egho et al. [2014] highlighted a limitation of M3SP, i.e. the multidimensional items do not allow itemsets whose items are of k^{th} dimension. The proposed algorithm MMISP tackles this issue by considering complex and heterogeneous sequences, where a sequence contains *elementary sequences*, i.e. itemsets whose items can be of two types: atomic and different in nature taken from user-defined taxonomies or subsets of unordered sets of items.

2.2.4.2 Discovering CPO-Patterns

Casas-Garriga [2005] presented the first algorithm for discovering cpo-patterns in a sequence database. The author focuses more on the formalisation of the cpo-pattern notion, and therefore does not provide an efficient algorithm for mining cpo-patterns. The proposed method spans two steps. First, the set of closed sequential patterns is extracted by using an existing algorithm, namely CloSpan [Yan et al., 2003] or BIDE [Wang and Han, 2004]. Second, the obtained closed sequential patterns are post-processed in order to build cpo-patterns.

Pei et al. [2006] studied the problem of mining cpo-patterns in string databases. The efficient algorithm FreCPO was proposed to extract cpo-patterns only from sequences of items without repetitive items. This constraint limits the algorithm applicability in real-life sequence databases, e.g. sequences that contain several occurrences of the same items, at different timestamps, alongside other items. FreCPO is a pattern-growth approach that searches cpo-patterns in a depth-first manner.

Fabrègue et al. [2015] highlighted a key drawback of the approach proposed by Casas-Garriga [2005], i.e. building cpo-patterns from already extracted closed sequential patterns yields only a subset of the complete set of cpo-patterns. The authors proposed the OrderSpan algorithm, based on pattern-growth approach, that directly extracts cpo-patterns from sequences of itemsets that can contain repetitive items. The algorithm is a twofold approach. Firstly, the prefix-tree that covers all sequences from a database $\mathcal{D}_\mathcal{S}$ is built, and then recursively frequent sub-prefix-trees on subsets of $\mathcal{D}_\mathcal{S}$ are extracted. Secondly, since the prefix property of sequences is used to mine the complete set of sub-prefix-trees, the associated cpo-patterns contain redundant vertices and edges, and hence a pruning and merging step is necessary.

Furthermore, Mannila et al. [1997] proposed to mine frequent episodes in a single long input-sequence, where an episode is formalised as a DAG.

2.2.4.3 Filtering and Ranking Sequential Patterns

To cope with the "pattern explosion" problem, there are two approaches. First, pushing constraints into the mining process, e.g. Garofalakis et al. [1999] proposed to use regular expressions as user-defined constraints in order to prune patterns during the mining process.

Second, measures of interest, e.g. δ-freeness [Hébert and Crémilleux, 2005] and cosine interest [Cao et al., 2014], can be used to rank or filter the generated patterns, and therefore facilitating their evaluation by domain experts. For example, Fabrègue et al. [2014] introduced the *generalised growth rate* of a cpo-pattern $\mathcal{G}$, which is discovered in a dataset, with respect to the maximal frequency of $\mathcal{G}$ in other datasets. This measure of interest is used to discriminate the same cpo-patterns extracted from different datasets. The *growth rate* measure was

inspired by emerging pattern mining [Dong and Li, 1999].

Geng and Hamilton [2006] surveyed existing interestingness measures that can be applied to all types of patterns since these measures rely mainly on the *support* measure. For example, the *confidence* measure [Agrawal et al., 1993] is used to filter association rules discovered in a transaction database. An association rule is a logical implication $X \rightarrow Y$, where X (antecedent) and Y (consequent) are sets of items and $X \cap Y = \emptyset$. The confidence of such an association rule determines how often the items in Y appear in transactions containing the items in X. Therefore, only association rules with high values of confidence can be selected to be evaluated by domain experts.

2.3 Formal Concept Analysis

Formal Concept Analysis (FCA, [Barbut and Monjardet, 1970] and [Ganter and Wille, 1999]) is a mathematical framework used for data analysis and knowledge discovery. FCA encodes binary data into a formal context and yields a hierarchy of conceptual abstractions.

Definition 2.18 (Formal Context). *A* formal context K *is a 3-tuple* (G, M, I)*, where* G *is a set of objects,* M *is a set of attributes and* $I \subseteq G \times M$ *is a binary relation that specifies which objects have which attributes.*

A formal context $\mathtt{K_1} = (\mathtt{G_1}, \mathtt{M_1}, \mathtt{I_1})$ is shown in Tab. 2.3 by using a cross table, i.e. the rows are the objects $\mathtt{G_1} = \{\mathtt{g1}, \mathtt{g2}, \mathtt{g3}, \mathtt{g4}, \mathtt{g5}\}$, the columns are the attributes $\mathtt{M_1} = \{\mathtt{m1}, \mathtt{m2}, \mathtt{m3}, \mathtt{m4}\}$ and a cross from a cell identified by a pair $(g_i, m_j) \in \mathtt{I_1}$ signifies that object $g_i \in \mathtt{G_1}$ has attribute $m_j \in \mathtt{M_1}$.

Table 2.3: Example of a formal context

$\mathtt{K_1}$	m1	m2	m3	m4
g1	×	×		
g2	×		×	
g3		×		×
g4			×	×
g5	×	×		×

Two derivation operators, both denoted by $'$, are defined for $X \subseteq G$ and $Y \subseteq M$ as follows:

$$' : 2^G \rightarrow 2^M, X' = \{m \in M | \forall g \in X, (g, m) \in I\}$$

$$' : 2^M \rightarrow 2^G, Y' = \{g \in G | \forall m \in Y, (g, m) \in I\}$$

where 2^G and 2^M are the power sets of G and M, respectively. The $'$ operators define a Galois connection between 2^G and 2^M. The set X' is the set of all attributes in M shared by the objects in X. Similarly, Y' is the set of all objects in G that have the attributes in Y. The composition operators $''$ are closure operators since they are idempotent, monotonous and extensive. X and Y, such that $X = X''$ and $Y = Y''$, are closed sets.

Definition 2.19 (Formal Concept). *A* formal concept C *derived from a formal context* $K = (G, M, I)$ *is a pair* (X, Y)*, where* $X \subseteq G$ *and* $Y \subseteq M$*, such that* $X' = Y$ *and* $Y' = X$*. The set of objects* X *is called the* extent *of* C*, while the set of attributes* Y *is called the* intent *of* C*.*

Using the $\mathtt{K}_1$ formal context (Tab. 2.3), if we consider the set of objects $X_1 = \{\mathtt{g1}\}$, then $X_1' = Y_1 = \{\mathtt{m1}, \mathtt{m2}\}$. Since $Y_1' = \{\mathtt{g1}, \mathtt{g5}\} = X_2$ and $X_1 \neq X_2$, (X_1, Y_1) is not a formal concept, while (X_2, Y_1) is a formal concept.

Definition 2.20 (Formal Concept Support). *The* support *of a concept* $C = (X, Y)$ *is defined as the cardinality of* X*.*

To illustrate this, the support of concept (X_2, Y_1) is $|X_2| = |\{\mathtt{g1}, \mathtt{g5}\}| = 2$.

Let $\mathcal{C}_K$ be the set of all formal concepts derived from a formal context $K = (G, M, I)$. Let $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ be two concepts from $\mathcal{C}_K$. The *concept generalisation order*, referred to as $\preceq_K$, is defined by $C_1 \preceq_K C_2$ if $X_1 \subseteq X_2$ ($\Leftrightarrow Y_2 \subseteq Y_1$). In this case, C_1 is called a *subconcept* of C_2 and C_2 a *superconcept* of C_1. If $C_1 \prec_K C_2$ and there is no C_3 such that $C_1 \prec_K C_3 \prec_K C_2$, then C_1 is a *lower neighbour* of C_2, denoted by $C_1 \lhd_K C_2$ and C_2 is an *upper neighbour* of C_1, denoted by $C_2 \rhd_K C_1$ [Roth et al., 2008].

The set $\mathcal{C}_K$ ordered by $\preceq_K$ forms a complete lattice, denoted by $\mathcal{L}_K = (\mathcal{C}_K, \preceq_K)$, which is called the *concept lattice* of the formal context K [Wille, 1982]. We denote by $\top(\mathcal{L}_K)$ the concept from $\mathcal{C}_K$ whose extent has all the objects in G and by $\bot(\mathcal{L}_K)$ the concept from $\mathcal{C}_K$ whose intent has all the attributes in M. Lattice $\mathcal{L}_K$ is represented by a Hasse diagram where the vertices are the concepts in $\mathcal{C}_K$ and the edges are defined by the relation $\lhd_K$.

Figure 2.4 illustrates the Hasse diagram[1] of lattice $\mathcal{L}_{\mathtt{K}_1}$ derived from $\mathtt{K}_1$ (Tab. 2.3). Each concept is represented by a box structured from top to bottom as follows: concept name, simplified intent and simplified extent. The representation of the lattice is simplified as every attribute/object is top-down/bottom-up inherited. Thus an attribute/object is shown only in the highest/lowest concept where it appears. For example, in Fig. 2.4, the `CK1_6` concept has the intent {`m1`,`m2`}, where the attributes `m1` and `m2` are inherited from the `CK1_9` and `CK1_8` concepts, respectively. The `CK1_6` concept extent is {`g1`,`g5`}, where object `g5` is inherited from the `CK1_1` concept.

[1] created with RCAExplore tool (`http://dolques.free.fr/rcaexplore`)

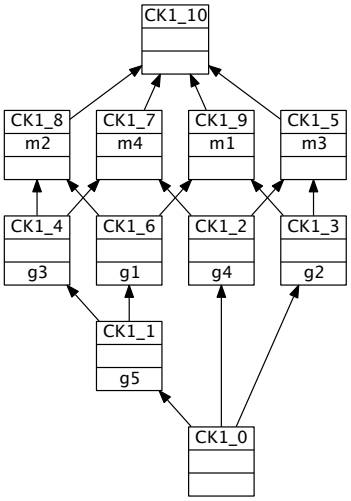


Figure 2.4: The Hasse diagram of the $\mathcal{L}_{\mathtt{K}_1}$ concept lattice derived from the $\mathtt{K}_1$ formal context (Tab. 2.3)

2.3.1 Computing Formal Concepts and Concept Lattices

Many studies have been dedicated to compute all formal concepts derived from a formal context and the associated concept lattice. Kuznetsov and Obiedkov [2002] classified existing algorithms into batch algorithms and incremental ones. Given a formal context, batch algorithms (e.g. [Ganter, 1984], [Bordat, 1986] and [Kuznetsov, 1993]) build the set of formal concepts and its Hasse diagram from scratch. In contrast, incremental algorithms (e.g. [Norris, 1978], [Godin et al., 1995], [Carpineto and Romano, 1996] and [Merwe et al., 2004]) compute at the i^{th} step the set of formal concepts or the Hasse diagram for the i first objects of a formal context. For example, the NextClosure algorithm [Ganter, 1984] uses a linear order on the set of objects and generates formal concepts in the lexicographical order of their extents. At each step a current object is considered and a generated formal concept is unique (i.e. the concept is derived for the first time) if its extent comprises no predecessor of the current object. Bordat [1986] proposed a level-wise algorithm that uses a top-down strategy to build formal concepts and the corresponding lattice. First, this algorithm finds all maximal subsets of objects; second, builds the corresponding concepts, and then computes new maximal subsets of the subsets generated at the first step. The process is repeated for the new maximal subsets of objects. A tree structure is used for fast storing and to retrieve concepts. Valtchev and Missaoui [2001] introduced an algorithm that divides a formal context into two parts by splitting the set of objects/attributes. Then, both Hasse diagrams obtained for these two parts are assembled into a global lattice. This approach is suitable for parallel computing.

Since the number of formal concepts can be exponential in the size of the input context, Godin and Mili [1993] introduced the *attribute-object-concept poset* (AOC-poset) that is a sub-

hierarchy of the obtained concept lattice that comprises only highest concepts introducing an attribute and lowest concepts introducing an object. There are a few algorithms proposed to compute such reduced lattices, e.g. [Berry et al., 2014].

Stumme [2002] proposed the TITANIC algorithm that allows to compute an *iceberg concept lattice* from a formal context. An iceberg concept lattice consists in all frequent concepts derived from a formal context. Let us mention that a concept is frequent if its support is greater than or equal to a user-defined minimum support θ.

Several tools are available online for computing formal concepts and concept lattices, e.g. CONEXP[2] and TOSCANAJ [3].

2.3.2 Conceptual Scaling

There are many cases when the analysed data contain objects related to attributes that can take several values. Wille [1992] formalised such data as a many-valued context (G, M, W, I), where G is a set of objects, M is a set of attributes, W is a set of attribute values and $I \subseteq G \times M \times W$ such that $(g, m, v) \in I$ and $(g, m, w) \in I$ always imply $v = w$. Let us note that (g, m, w) indicates that object g has value w for attribute m.

For each many-valued attribute $m \in M$ a formal context can be derived by means of *conceptual scaling*. Various types of scaling exist [Ganter and Wille, 1999]. This thesis relies on elementary scales, e.g. *nominal scaling* and *ordinal scaling*. The nominal scales are used to scale a multi-valued attribute whose values mutually exclude each other. Therefore, a partition of the objects into extents is obtained. The ordinal scales are used to scale a multi-valued attribute whose values are ordered and each value implies the weaker ones.

2.3.3 FCA Extensions

There are several extensions of FCA that allow to manipulate complex data (e.g. graphs, intervals and logical formulae) or n-ary relations between objects. In the following, we briefly state the main ideas of several extensions that are relevant in our work.

Power Context Family (PCF, [Wille, 1997] and [Kötters, 2016]) is a family of formal contexts linked by their sets of objects. To our knowledge, it was the first attempt to embed arbitrary n-ary relations between objects (at the context level) in FCA. Let us note that FCA naturally handles no relation. Formally, a power context family is a n-tuple $(K_1, K_2, ..., K_n)$, $n \geq 2$ with $K_i = (G_i, M_i, I_i)$ such that $G_i \subseteq (G_1)^i$, $i \in \{1, ..., n\}$. Therefore, K_1 is the formal context that describes a set of objects and K_n is the formal context describing n-ary relations that link n of these objects. For each formal context a lattice is built, and thus the navigation of the lattice built from K_1 does not consider the relational properties.

[2] `http://conexp.sourceforge.net/`
[3] `http://toscanaj.sourceforge.net/`

In Logical Concept Analysis (LCA, [Ferré and Ridoux, 2000]) an object is described by a logical formula instead of a set of attributes as in classical FCA. LCA considers a logical context K as a 3-tuple $(G, \mathcal{L}, d)$, where G is a set of objects, $\mathcal{L}$ is a logic (lattice) that describes the domain and d is a mapping that associates a formula in $\mathcal{L}$ to each object in G. A logical concept C is a pair (X, f), where the extent $X \subseteq G$ is the set of objects whose description is subsumed by f; the intent $f \in \mathcal{L}$ is the most precise formula that subsumes all descriptions of the objects of X. The set of all logical concepts derived from K can be ordered and a lattice $\mathcal{L}_K$ is obtained.

Pattern Structures (PS, [Ganter and Kuznetsov, 2001]) use a pattern, instead of a set of attributes, as object description. This extension can be applied directly to complex data without involving a conversion of these complex data into binary ones. A pattern structure is a 3-tuple $(G, (D, \sqcap), \delta)$, where G is a set of objects, $(D, \sqcap)$ is a semi-lattice of potential object descriptions and δ is a mapping that associates a description in $(D, \sqcap)$ with each object in G. From this pattern structure are derived pattern concepts, which can be ordered according to the inclusion on extents into a lattice of pattern concepts.

Graph-FCA (G-FCA, [Ferré, 2015]) is an extension of FCA where objects are substituted with n-tuples of objects. The input of G-FCA is a knowledge graph (e.g. conceptual graphs or RDF graphs), called graph context, formalised as (G, M, I), where G is a set of objects, M is a set of attributes, and $I \subseteq G^* \times M$ is an incidence relation that relates k-tuples of objects from G^* and attributes from M. From this graph context are derived graph concepts with projected graph patterns as intents and object relations as extents. These graph concepts can be organised into a graph concept lattice.

Finally, this thesis focuses on the Relational Concept Analysis extension of FCA, and therefore we recall its theoretical aspects in the following.

2.3.4 Relational Concept Analysis

Relational Concept Analysis (RCA, [Rouane-Hacene et al., 2013]) was devised to explore multi-relational data. Indeed, RCA classifies sets of objects described by attributes and relations, allowing to discover knowledge patterns and implication rules in relational datasets by applying iteratively FCA to a *relational context family* (the RCA input).

Definition 2.21 (Relational Context Family (RCF)). *A* relational context family *is a pair* $(\mathcal{K}, \mathcal{R})$*:*

- $\mathcal{K} = \{K_i\}_{i \in [1,n]}$ *is a set of formal contexts* $K_i = (G_i, M_i, I_i)$*;*
- $\mathcal{R} = \{R_j\}_{j \in [1,m]}$ *is a set of relational contexts* $R_j = (G_k, G_l, r_j)$*, where* $r_j \subseteq G_k \times G_l$ *is a binary relation with* $k, l \in [1, n]$*,* $G_k = dom(r_j)$ *is the domain of the relation and* $G_l = ran(r_j)$ *is the range of the relation.*

Briefly, an RCF spans several categories of objects described within formal contexts and relations between these objects. To illustrate this, let us consider the $\mathtt{K1} = (\mathtt{G}_1, \mathtt{M}_1, \mathtt{I}_1)$ and $\mathtt{K2} = (\mathtt{G}_2, \mathtt{M}_2, \mathtt{I}_2)$ formal contexts shown in Fig. 2.5a and 2.5b, respectively. Using these two formal contexts and the relational context R1 shown in Fig. 2.5c, we build the RCF $(\{\mathtt{K1}, \mathtt{K2}\}, \{\mathtt{R1}\})$. R1 defines the relation $\mathtt{r1} \subseteq \mathtt{G}_1 \times \mathtt{G}_2$ between the objects of $\mathtt{G}_1 = \{\mathtt{a1}, \mathtt{a2}, \mathtt{a3}, \mathtt{a4}\}$ and $\mathtt{G}_2 = \{\mathtt{b1}, \mathtt{b2}, \mathtt{b3}\}$, e.g. $(\mathtt{a1}, \mathtt{b1}) \in \mathtt{r1}$.

K1	m1	m2	m3
a1	×	×	
a2	×		×
a3		×	
a4	×	×	

(a) K1

K2	m1′	m2′	m3′
b1	×	×	
b2			×
b3	×		

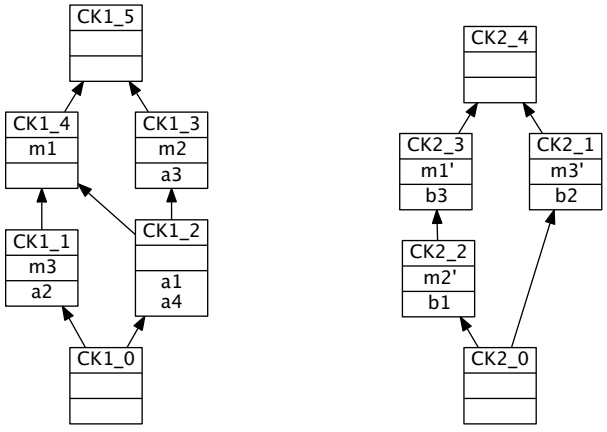
(b) K2

R1	b1	b2	b3
a1	×	×	
a2	×		×
a3	×	×	
a4	×		×

(c) R1

Figure 2.5: Illustrative example of a relational context family

For each formal context in an RCF, an *initial lattice* is built using any classical FCA algorithm. For example, $\mathcal{L}_{\mathtt{K1}}$ (Fig. 2.6a) and $\mathcal{L}_{\mathtt{K2}}$ (Fig. 2.6b) represent respectively the initial lattices built for the $\mathtt{K}_1$ and $\mathtt{K}_2$ formal contexts.



(a) $\mathcal{L}_{\mathtt{K1}}$ (b) $\mathcal{L}_{\mathtt{K2}}$

Figure 2.6: The initial lattices $\mathcal{L}_{\mathtt{K1}}$ and $\mathcal{L}_{\mathtt{K2}}$ built respectively for the formal contexts $\mathtt{K}_1$ and $\mathtt{K}_2$

RCA relies on a *relational scaling mechanism* that is used to transform a relation $r_j \subseteq G_k \times G_l$ into a set of *relational attributes* that extends the K_k formal context, which describes the set of objects $G_k = dom(r_j)$, to a K_k^+ scaled context.

Definition 2.22 (Relational Attribute). *Given a binary relation $r_j \subseteq G_k \times G_l$, a formal context K_l describing the $G_l = ran(r_j)$ set of objects and the set of concepts $\mathcal{C}_{K_l}$ derived from K_l. A* relational attribute *takes the syntactic form $qr_j(C)$, where q is a scaling quantifier, r_j is the scaled relation and $C = (X, Y) \in \mathcal{C}_{K_l}$.*

The relational attribute $qr_j(C)$ highlights a relation between $g \in G_k$ and the objects of X based on r_j.

Several scaling quantifiers used in RCA are shown in Tab. 2.4.

Table 2.4: Several quantifiers for relational scaling mechanism [Rouane-Hacene et al., 2013]. $r_j \subseteq G_k \times G_l$ is a binary relation, $g \in G_k$ and $C = (X, Y) \in \mathcal{C}_{K_l}$, where $\mathcal{C}_{K_l}$ is derived from K_l whose set of objects is G_l

Quantifier Name	Relational Attribute	Condition
Universal	$\forall r_j(C)$	$r_j(g) \subseteq X$
Existential	$\exists r_j(C)$	$r_j(g) \cap X \neq \emptyset$
Universal strict	$\forall\exists r_j(C)$	$r_j(g) \subseteq X$ and $r_j(g) \neq \emptyset$
Qualified cardinality restriction (max)	$\geq_n r_j(C)$	$r_j(g) \subseteq X$ and $\lvert r_j(g)\rvert \geq n$
Qualified cardinality restriction (min)	$\leq_n r_j(C)$	$r_j(g) \subseteq X$ and $\lvert r_j(g)\rvert \leq n$

Definition 2.23 (Relational Extension of a Formal Context). *Given a binary relation $r_j \subseteq G_k \times G_l$, two formal contexts $K_k = (G_k, M_k, I_k)$ and $K_l = (G_l, M_l, I_l)$, the set of concepts $\mathcal{C}_{K_l}$ derived from K_l and a scaling quantifier q. The* relational extension *of K_k, denoted by $K_k^{r_j}$, is a 3-tuple $(G_k^{r_j}, M_k^{r_j}, I_k^{r_j})$ where:*

- $G_k^{r_j} = G_k$;
- $M_k^{r_j} = \{qr_j(C) | C \in \mathcal{C}_{K_l}\}$;
- $I_k^{r_j} = \{(g, qr_j(C)) | g \in G_k, C = (X, Y) \in \mathcal{C}_{K_l}, g \text{ is connected by } r_j \text{ and } q \text{ to objects of } X\}$.

$I_k^{r_j}$ depends on the chosen scaling quantifier, e.g. when the $\exists$ quantifier is used $I_k^{r_j} = \{(g, \exists r_j(C)) | g \in G_k, C = (X, Y) \in \mathcal{C}_{K_l}, r_j(g) \cap X \neq \emptyset\}$.

Furthermore, the relational extension of the formal context K_k when n relations r_{j_i} with $i \in \{1, ..., n\}$ are considered, referred to as $K_k^{r_{j_n}}$, is a 3-tuple $(G_k^{r_{j_n}}, M_k^{r_{j_n}}, I_k^{r_{j_n}})$ where:

- $G_k^{r_{j_n}} = G_k$;
- $M_k^{r_{j_n}} = \bigcup_{i=1}^{n} M_k^{r_{j_i}}$;
- $I_k^{r_{j_n}} = \bigcup_{i=1}^{n} I_k^{r_{j_i}}$.

Definition 2.24 (Scaled Context). *Given a formal context $K_k = (G_k, M_k, I_k)$ and a relational extension $K_k^{r_j} = (G_k^{r_j}, M_k^{r_j}, I_k^{r_j})$ of K_k. The* scaled context *K_k^+ of K_k is a 3-tuple (G_k^+, M_k^+, I_k^+) where:*

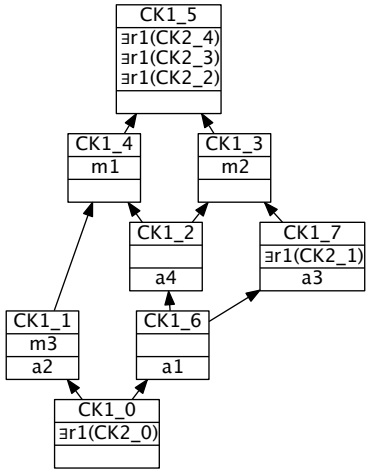
- $G_k^+ = G_k$;
- $M_k^+ = M_k \cup M_k^{r_j}$;
- $I_k^+ = I_k \cup I_k^{r_j}$.

To illustrate these, `K1` (Fig. 2.5a) is upgraded with relational attributes built using concepts from $\mathcal{L}_{\texttt{K2}}$ (Fig. 2.6b), and thus the $\texttt{K1}^+$ scaled context shown in Fig. 2.7a is obtained.

Scaled context $\texttt{K1}^+$ was obtained by applying the existential scaling to the $\texttt{r}_1$ binary relation. For instance, the a3 object has the relational attribute $\exists\texttt{r1}(\texttt{CK2\_1})$ since the extent of CK2_1 contains b2 and $(\texttt{a3}, \texttt{b2}) \in \texttt{r1}$. The FCA algorithm is again applied to the upgraded RCF $(\{\texttt{K1}^+, \texttt{K2}\}, \{\texttt{R1}\})$. A family of lattices is generated comprising two lattices: lattice $\mathcal{L}_{\texttt{K1}^+}$ shown in Fig. 2.7b built from $\texttt{K1}^+$ and lattice $\mathcal{L}_{\texttt{K2}}$ given in Fig. 2.6b. Let us note that $\mathcal{L}_{\texttt{K2}}$ is unchanged since the $\texttt{G}_2$ set of objects is not the domain of any relation in our illustrative example. These two lattices can be navigated following the concepts used to build relational attributes. For example, $\exists\texttt{r1}(\texttt{CK2\_1})$ of the CK1_7 concept intent (Fig. 2.7b) allows us to navigate from lattice $\mathcal{L}_{\texttt{K1}^+}$ to concept CK2_1 in lattice $\mathcal{L}_{\texttt{K2}}$.

$\texttt{K1}^+$	m1	m2	m3	$\exists$r1(CK2_4)	$\exists$r1(CK2_3)	$\exists$r1(CK2_2)	$\exists$r1(CK2_1)
a1	×	×		×	×	×	×
a2	×		×	×	×	×	
a3		×		×	×	×	×
a4	×	×		×	×	×	

(a) $\texttt{K1}^+$



(b) $\mathcal{L}_{\texttt{K1}^+}$

Figure 2.7: (a) the scaled context of K1; (b) the lattice built from $\texttt{K1}^+$

The RCA process, depicted in Fig. 2.8, firstly, consists in applying an FCA algorithm (Sect. 2.3.1) to each formal context of an RCF in order to obtain the initial lattices. Then, FCA is applied iteratively to each formal context extended by the relational attributes built with the concepts previously learnt. The RCA output is obtained when a *fix point* is found, i.e. the families of lattices of two consecutive steps are isomorphic and the formal contexts are unchanged. This fix point is always found since the process is monotonic and bounded.

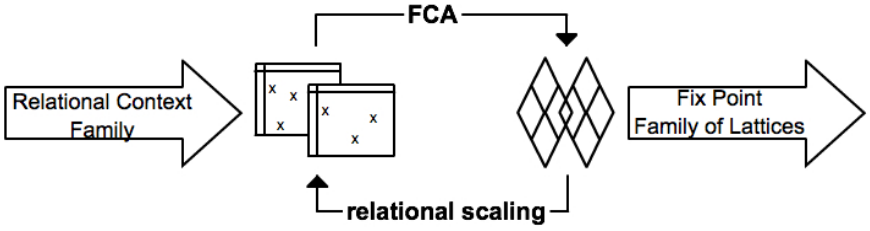


Figure 2.8: The schema of the RCA process

In our illustrative example (Fig. 2.5), the family of lattices obtained by applying FCA to the upgraded RCF $(\{\texttt{K1}^+, \texttt{K2}\}, \{\texttt{R1}\})$ represents the fix point. Therefore, the RCF is not again upgraded since there is no new learnt concept.

2.3.5 Related Work

In the following, we discuss the FCA-based and the multi-relational data mining (MRDM, [Džeroski, 2003]) approaches for mining sequential data. We mention as well the existing RCA-based works. In addition, we present some methods to reduce the complexity of a concept lattice.

2.3.5.1 FCA-Based Approaches for Exploring Sequential Data

There are various related FCA approaches used to explore sequential data. Wolff [2001] introduced Temporal Concept Analysis where objects are characterised by a date and a state (i.e. a set of attributes). The data are merged into a single context and the resulting concept lattice is analysed thanks to the date element in the concepts. Therefore, the temporal relations between concepts are revealed manually by domain experts. This approach was used to analyse sequential data about crime suspects [Poelmans et al., 2010a].

Ferré [2007] proposed to analyse strings in order to find linear orders, i.e. substrings. The method is applied to a string context built from the titles of all published papers at an international conference in a given period of time. The aim is to compute the complete set of maximal substrings.

In [Casas-Garriga, 2005], closed sequential patterns are mined by using the BIDE [Wang and Han, 2004] or CloSpan [Yan et al., 2003] algorithms, and then regrouped in a lattice similar to a concept lattice obtained with FCA.

Recently, Buzmakov et al. [2016] have analysed sequential data by using the PS extension of FCA. From multidimensional and multilevel sequential data a sequential pattern structure is built. To overcome the large number of concepts in the pattern concept lattice, the projections of the sequential pattern structure are used. Therefore, these mathematical projections significantly decrease the number of patterns (i.e. patterns from the $(D, \sqcap)$ sequential meet-semilattice) depending on the motivation behind the analysis step. The approach was applied to medical data.

FCA can classify and filter complex data (e.g. graphs and sequences) based on its extensions. For instance, the set of cpo-patterns, denoted by D, obtained by using a cpo-pattern mining algorithm, e.g. [Pei et al., 2006] or [Fabrègue et al., 2015], can be combined with the intersection operation on graphs $\sqcap$ in order to build a pattern structure $(G, (D, \sqcap), \delta)$, where G is the set of objects described by the obtained cpo-patterns through relation δ.

In addition, Cellier et al. [2011] explained how LCA can be used to organise already extracted patterns in a concept lattice. Similarly, Egho et al. [2011] combined the sequential pattern mining and FCA domains to explore a database of heterogeneous sequences. Heterogeneous sequential patterns are extracted relying on algorithm M3SP [Plantevit et al., 2010]. Then, FCA is used to organise the obtained sequential patterns. The formal context is built by taking patients as objects and sequential patterns as attributes. The approach is applied to medical data.

2.3.5.2 Multi-Relational Approaches for Exploring Sequential Data

Multi-relational data can be mined with propositional data mining algorithms by transforming these data into a single table (task usually achieved by means of propositionalization [Kramer et al., 2001]). In contrast, MRDM respects the multi-relational nature of such data. For example, Jacobs and Blockeel [2001] presented an approach to find frequent shell scripts in shell logs. This work is seen as a relational pattern discovery task in sequential data and is based on system Warmr [Dehaspe and Toivonen, 1999]. Indeed, the shell commands induce a sequence, according to their execution order, and each command can be related to one or more parameters. Esposito et al. [2009] proposed an Inductive Logic Programming (ILP, [Muggleton, 1991]) algorithm for discovering first-order sequential patterns in multi-dimensional relational sequences (e.g. sequences where spatial and temporal information coexist).

Ferreira et al. [2015] proposed the MuSer framework to explore multi-relational sequential data, where the order relation on itemsets is temporal. To this end, the multi-relational temporal data are converted into a set of heterogeneous sequences, one for each object of interest from a target table. Such a sequence can include intra-table and/or inter-table relations within the temporal data. Then, the obtained set of sequences is mined using a classical sequential pattern mining algorithm, e.g. [Pei et al., 2001] or [Yan et al., 2003]. The extracted sequential patterns are filtered and only the most interesting ones are used to enlarge (as attributes) the original multi-relational database (the target table) by specifying which object of interest is characterised by a particular selected pattern. Finally, from the enlarged database a classification model is induced based on an ILP algorithm.

Moreover, related approaches exist in data stream mining, where a data stream represents a sequence of instances generated and gathered continually. For instance, Silva and Antunes [2015] introduced the Star FP-Stream method, which joins the MRDM and data streaming techniques for discovering frequent itemsets in large star schemas [Kimball and Ross, 2002]. A star schema consists in a fact (central) table interrelated via foreign keys with dimensional tables. The same authors proposed the Star FP-Growth [Silva and Antunes, 2010] algorithm for mining frequent multi-relational itemsets in a relational database mod-

elled as a star schema. This pattern-growth algorithm builds for each dimension a FP-tree structure of all frequent patterns. Then, these FP-trees are combined based on the fact table into a Super FP-tree structure on which the FP-Growth [Han et al., 2004] algorithm is applied to list all multi-relational patterns. Silva and Antunes [2014] used this approach to analyse hepatitis data.

2.3.5.3 A Brief Survey on RCA-Based Works

Various works from software engineering rely on RCA. Dao et al. [2004], Arévalo et al. [2006], and Huchard et al. [2007] applied RCA to reorganise hierarchies of classes from UML models. Moha et al. [2008] focused on the identification and semi-automatic correction of design defects from object-oriented software systems. Saada et al. [2012] used RCA to learn model transformation rules from transformation examples. In [Azmeh et al., 2011a], the relations between abstract tasks are used to classify relevant Web services to instantiate the tasks.

Several works in ontology engineering are based on RCA. Bendaoud et al. [2008] combined FCA and RCA for building and refining domain ontologies. Rouane-Hacene et al. [2011] proposed an approach for ontologies restructuring. In [Shi et al., 2011], an effective method for reengineering semantic wikis is proposed.

Azmeh et al. [2011b] focused on the navigation of the RCA output and on reducing its complexity. To this end, multi-relational data are encoded into the RCA input based on a user-defined query. The query is seen as a DAG that specifies the order on relations. The obtained family of concept lattices is navigated, following the given order, to obtain the objects that satisfy this query.

Codocedo and Napoli [2014] proposed to combine the RCA and PS extensions of FCA. RCA is adapted to integrate a description of G_1, a set of source objects with descriptors (coming from a pattern structure $(G_1, (D, \sqcap), \delta)$) and relational attributes to a set of concepts on a target formal context (G_2, M_2, I_2). For a relation $r \subseteq G_1 \times G_2$, the relational attributes are built using the classical quantifiers, e.g. $\exists$ or $\forall\exists$. This is formalised as "heterogeneous pattern structure". An application to the Information Research domain is described, where the source objects are documents, the descriptors are vectors of intervals of Latent Variable values, the target objects are terms grouped into concepts when they have the same meaning (represented by a synset) and the relation r connects documents to their included terms. Latent Variables abstract hidden topics spread over the documents.

Dolques et al. [2015] proposed an adaptation of RCA to explore relations in a guided way in order to increase the pertinence of the results. To this end, at each step the user can select dynamically the formal contexts that are considered and/or the quantifiers used for relational scaling. In [Dolques et al., 2016], AOC-posets [Godin and Mili, 1993] are used rather than concept lattices to reduce the complexity of the RCA output. This variant is

applied to discover rules in hydro-ecological data. Furthermore, Dolques et al. [2014] showed that RCA uses the relational scaling mechanism to transform the multi-relational data into a singe table similarly to propositionalization approaches.

There are several tools that can be used to apply RCA as follows: Galicia[4], eRCA[5] and RCAExplore[6].

2.3.5.4 Filtering and Ranking Formal Concepts

A well-known problem of the FCA-based approaches is the exponential number of concepts that can be derived in the worst-case scenario from a formal context. Recently, Dias and Vieira [2015] have surveyed the techniques for concept lattice reduction and proposed to group them into: redundant information removal, simplification and selection. Our work is concerned with the selection techniques.

A popular measure used in FCA for ranking concepts is the stability index [Kuznetsov, 2007] and its new estimates [Buzmakov et al., 2014]. Two types of stability are defined by Kuznetsov et al. [2007], namely extensional and intensional. The extensional stability indicates how a concept extent depends on particular attributes of a formal context. The intensional stability indicates the probability of preserving the concept intent when removing some objects of a formal context. Jay et al. [2008] used iceberg lattices and stability index in social healthcare network analysis. Klimushkin et al. [2010] introduced two new measures of interest, precisely probability and separation of concepts, and discussed how they can be combined with stability index. In addition, the authors showed that stability is more reliable for selecting formal concepts derived from noisy data. Recently, Buzmakov et al. [2016] have used stability index and projections of sequential pattern structures to select relevant heterogeneous medical sequential patterns. The projections of sequential pattern structures are used to introduce user-defined constraints in the mining process, e.g. for heterogeneous medical sequences some additional information can be ignored during the mining process.

Formica [2008] and Alqadah and Bhatnagar [2011] defined the similarity measures of formal concepts in order to cluster formal concepts without relying on human domain expertise. Melo et al. [2011] used visualisation techniques to enhance the readability of concept lattices. Trees derived from a concept lattice are extracted based on measures of interest, e.g. stability, support and confidence (estimates how likely an object which has an attribute set A, also has an attribute set C [Ganter and Wille, 1999]) of formal concepts.

Belohlavek and Trnecka [2013] proposed to compute the degree to which a formal concept belongs to a basic level [Rosch, 1988]. A basic level is seen as a fuzzy set [Zadeh, 1965]. Five

[4] http://www.iro.umontreal.ca/~galicia/
[5] https://code.google.com/p/erca/
[6] http://dolques.free.fr/rcaexplore

basic level metrics, i.e. similarity, cue validity, category feature collocation, category utility and predictability, are discussed in order to rank the formal concepts derived from a formal context. Therefore, only the formal concepts with high values of such metrics are evaluated by domain experts.

Dias and Vieira [2010] proposed a different approach to reduce the complexity of a concept lattice, precisely junction based on object similarity. A formal context is preprocessed replacing groups of similar objects by representative objects. To this end, based on domain knowledge a weight is assigned to each attribute of the formal context and it is used to compute the similarity between objects. A similarity matrix is created from which clusters of similar objects are extracted. These clusters are used to remodel the original formal context from which a concept lattice is built using a classical FCA algorithm.

2.4 Summary

In this chapter, we have presented the theoretical underpinnings of this thesis, namely sequential pattern mining, FCA and RCA. In addition, we have discussed already proposed approaches for exploring sequential data.

We have outlined several FCA-based and MRDM approaches that deal with sequential data. Most of these approaches rely on propositional algorithms to extract sequential patterns (rather than cpo-patterns). The FCA-based approaches organise already discovered patterns into a hierarchy, e.g. a pattern concept lattice.

3 Relational Analysis of Sequential Data

Contents

3.1 Introduction

In this chapter, we present the first two steps of the RCA-Seq approach, precisely the data preprocessing and the RCA-based exploration of sequential data. To this end, firstly, we explain how to clean and preprocess the raw data to obtain sequences. Secondly, we transform the generated sequences into the input of RCA based on a general data model. Lastly, we explain the concept lattices from the RCA output and we show how domain experts can leverage the "richness" of these results, i.e. their hierarchical nature and the various captured information, in order to discover interesting and useful patterns.

3.2 Running Example

Patterns hidden in sequential medical data about patients and their medical histories can provide valuable knowledge for physicians. Here, we propose to study the symptoms (e.g. fever, headache, fatigue, and cough) that indicate the presence of viruses (e.g. influenza and hepatitis) in patients. The symptoms and viruses are detected by medical examinations and viral tests, respectively.

Table 3.1 shows an illustrative example of medical data from last year, where we focus on influenza virus. We consider that these data are exported from a relational database.

Table 3.1: Illustrative example of medical data

Patient	Date	Medical Examination		Viral Test
		COUGH	FEVER	Influenza
P1	25/09	–	high	–
	26/09	moderate	–	–
	27/09	moderate	high	–
	28/09	–	–	A
	02/10	high	moderate	–
	05/10	–	high	–
	07/10	–	–	B
	17/11	high	–	–
	18/11	–	–	A
	28/12	–	–	A
P2	08/02	moderate	–	–
	09/02	–	–	A
	13/05	high	–	–
	14/05	moderate	high	–
	15/05	–	–	A
	20/10	moderate	–	–
	25/10	–	–	B
P3	03/02	–	moderate	–
	10/04	high	–	–
	11/04	moderate	–	–
	12/04	–	high	–
	13/04	–	–	A

Physicians try to assess the cough and fever symptoms felt by patients to better understand how to identify in advance the outbreak of influenza A or B virus and to distinguish between the influenza A and B outbreaks. The symptoms can be moderate or high, while the influenza virus can be of type A or B. Thus, in this example we deal with *qualitative medical data*. For instance, patient P3 underwent four medical examinations and did a viral test. The first medical examination was on February 3rd when patient P3 experienced moderate fever.

The second medical examination was on April 10^{th} when the same patient P3 experienced high cough. The third medical examination was on April 11^{th} when the same patient P3 experienced moderate cough. The fourth medical examination was on April 12^{th} when patient P3 experienced high fever. Then, on April 13^{th}, patient P3 was diagnosed with influenza A virus.

Let us mention that only pertinent medical data are considered to recognise influenza outbreaks. We suppose that physicians focus on the viral tests done after at least one medical examination. They are also interested in the medical examinations undergone by patients within 10 days before their viral tests.

3.3 Data Preprocessing

To discover patterns in such medical data (Tab. 3.1), which are already discretized, we apply a data cleaning process. Then, the obtained data are transformed into sequential data from which qualitative sub-datasets are built based on the type of the diagnosed influenza virus.

3.3.1 Data Cleaning

The data cleaning process relies on the two aforementioned requirements:

- *only the medical examinations undergone by patients within* 10 *days prior to their viral tests are analysed*. Therefore, the medical examination undergone by patient P3 on February 3^{rd} is not considered since there is no pertinent viral test done by P3;
- *only the viral tests done after at least one medical examination are analysed*. Therefore, the viral test done by patient P1 on December 28^{th} is not considered since there is no medical examination undergone by P1 within 10 days before this test.

The cleaned medical data are transformed into qualitative sequential sub-datasets as we detail in the following.

3.3.2 Building Qualitative Sequential Sub-Datasets

A *patient sequence*, as shown in Fig. 3.1, consists in a chronologically ordered set of medical examinations undergone by the same patient and a corresponding viral test that ends the sequence.

A medical examination represents a *non-target itemset* of symptoms, while a viral test represents a *target 1-itemset* (set of only one item) comprising the studied *item of interest*, i.e. the virus that infected the patient. A viral test points to a patient sequence and is formalised as a 3-tuple $(Patient, Date, Result)$. The pair $(Patient, Date)$ uniquely identifies the viral

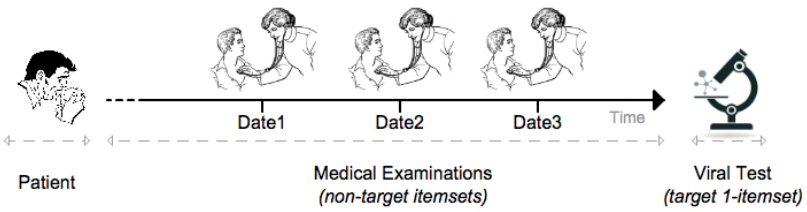


Figure 3.1: Patient sequence

test; $Date$ designates the time when the viral test was done by $Patient$ and is formatted as $Day/Month$; $Result$ is a target 1-itemset. A medical examination is similarly defined, but $Result$ is a non-target itemset. A patient can undergo several medical examinations and do several viral tests. Let us note that, first, the non-target itemsets and the target 1-itemset in a patient sequence are ordered according to a temporal relation. Second, a pair $(Patient, Date)$ is referred to as a *temporal object* since it contains a temporal information. Finally, each item in an itemset has associated a qualitative value.

For example, in Tab. 3.1, (P1, 28/09) is a temporal object that identifies the viral test (P1, 28/09, $(\texttt{Influenza}_{\texttt{A}})$), where P1 is a patient, September 28^{th} is the time when the viral test was done by P1 and $(\texttt{Influenza}_{\texttt{A}})$ is the target 1-itemset specifying the influenza A virus that infected P1. Similarly, the (P1, 27/09) temporal object identifies the medical examination (P1, 27/09, $(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})$) undergone by patient P1 on September 27^{th} before the viral test identified by (P1, 28/09) and $(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})$ is the non-target itemset of symptoms felt by P1. The items `Influenza`, `COUGH` and `FEVER` have associated the qualitative values `A`, `moderate` and `high`, respectively.

To obtain patient sequences from the medical data shown in Tab. 3.1, we order temporally the physical examinations for each distinct patient according to the values from the Date column. Then, for each patient, we cut the obtained sequence out in patient sequences based on a user-defined *time window* (here, 10 days before a viral test).

To illustrate this, let us analyse the medical data of patient P1 depicted in Fig. 3.2. There are 6 medical examinations (pairs coloured in black) and 3 viral tests (pairs in gray) ordered temporally. Three patient sequences are obtained for P1, i.e. one for each viral test.

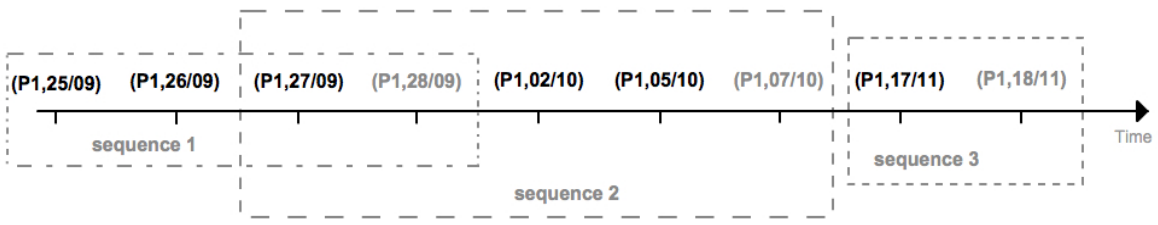


Figure 3.2: The sequences of patient P1

For instance, the temporal objects (P1, 27/09), (P1, 02/10), (P1, 05/10) and (P1, 07/10) illustrated in Fig. 3.2 constitute the sequence $S6$ of P1 (Tab. 3.2), i.e. $\langle(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}}\ \texttt{FEVER}_{\texttt{moderate}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$. All the patient sequences obtained from the medical data given in Tab. 3.1 are shown in Tab. 3.2.

Table 3.2: The sequential dataset obtained from Tab. 3.1

Sequence Id	Sequence
$S1$	$\langle(\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}})(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S2$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S3$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S4$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S5$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S6$	$\langle(\texttt{COUGH}_{\texttt{moderate}}\ \texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}}\ \texttt{FEVER}_{\texttt{moderate}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S7$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{B}})\rangle$

From the sequential data shown in Tab. 3.2, we build sub-datasets based on the diagnosed type of influenza virus. Thus, there are two sequential sub-datasets referred to as $\mathcal{D}_{\mathcal{S}fluA}$ (the patient sequences from $S1$ to $S5$) and $\mathcal{D}_{\mathcal{S}fluB}$ (the patient sequences $S6$ and $S7$).

3.3.3 Modelling Qualitative Sequential Data

Exploiting the relational nature of our sequential medical data, we propose to model the sub-datasets as shown in Fig. 3.3. This data model is used to build the RCA input, as we explain in Sect. 3.4.1. There are four rectangles, one for each set of objects (analogous to a table from a relational database) we manipulate, as follows: viruses (V), symptoms (S), viral tests (VT) and medical examinations (ME). The set of viruses contains only one object (item) `Influenza` and the set of symptoms contains two objects `COUGH` and `FEVER`.

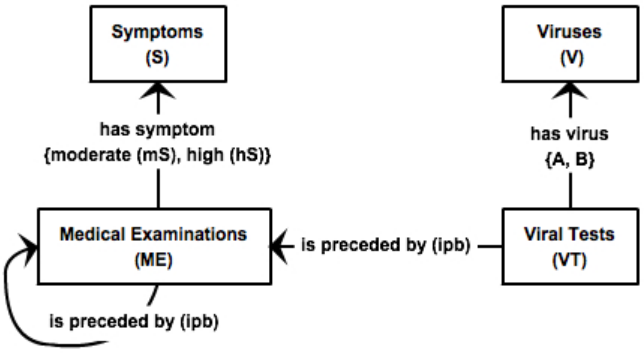


Figure 3.3: The modelling of the sequential medical data shown in Tab. 3.2 [Nica et al., 2016b]

Viral tests are linked to viruses by a qualitative binary relation, namely *has virus A* or *has virus B*, that is an inter-table relation since it links distinct types of objects. Similarly, medical

examinations are linked to symptoms by inter-table qualitative relations, precisely *has symptom*, differentiated by the type of the identified symptom, e.g. *moderate* (mS) or *high* (hS). Viral tests/medical examinations and medical examinations are linked by a temporal binary relation *is preceded by* (ipb) that associates a viral test/medical examination with a medical examination if the viral test/medical examination is preceded in time by the medical examination. The temporal relation between a viral test and a medical examination is an inter-table relation, while the temporal relation between medical examinations is an intra-table one since it relates the same type of objects. There is no temporal binary relation between viral tests since our aim is to study the symptoms that help physicians to prognosticate the influenza A or B virus.

Let us mention that we actually propose a *general data model* shown in Fig. 3.4 that allows to encode any qualitative sequential data (where a sequence resembles the one depicted in Fig. 3.1) into the RCA input. For example, the development of a football player skills prior to an upcoming match leading to a sequence of training sessions followed by a player evaluation when the squad role of the player is assigned.

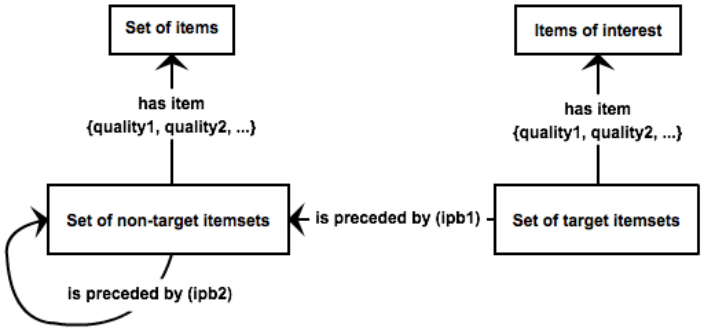


Figure 3.4: General data model for exploring qualitative sequential data with RCA

3.4 Exploration of Qualitative Sequential Data Using RCA

The exploration of qualitative sequential data with RCA spans two main steps: building the RCA input (an RCF), and then applying the RCA process. In the following, we exemplify how to explore such data by using only sub-dataset $\mathcal{D}_{\mathcal{S}fluA}$.

3.4.1 Building the RCA Input

In order to encode $\mathcal{D}_{\mathcal{S}fluA}$ (Tab. 3.2) into an RCF we follow the data model depicted in Fig. 3.3. As explained in Sect. 3.3.3, the set of viruses contains only one object `Influenza` and the set of symptoms contains two objects `COUGH` and `FEVER`. Below, we detail how to create the set of viral tests (that represents the objects of interest) and the one of medical examinations.

Analysing the patient sequences in $\mathcal{D}_{\mathcal{S}fluA}$, it is noted that a similar itemset can correspond to several viral tests or medical examinations. For example, the $(\texttt{Influenza}_\texttt{A})$ target 1-itemset occurs in all 5 analysed sequences, but it corresponds to different viral tests done by patients. Similarly, the $(\texttt{COUGH}_\texttt{high})$ non-target itemset occurs in the $S2$, $S4$ and $S5$ sequences, but it corresponds to different medical examinations undergone by patients. Furthermore, if we suppose that our sub-dataset comprises all the sequences in Tab. 3.2, we notice that a medical examination can occur in different patient sequences. For instance, the medical examination identified by (P1, 27/09) (Fig. 3.2) occurs in two sequences $S1$ and $S6$. Consequently, since in the RCA input we encode the temporal links between these itemsets, we decide to uniquely identify inter-sequence and intra-sequence each occurrence of an itemset. Table 3.3 illustrates how we propose to remodel our *patient sequences of itemsets* as *patient sequences of unique identifiers* (UIDs).

Table 3.3: $\mathcal{UID}_{\mathcal{S}fluA}$: sub-dataset of sequences of UIDs

Sequence
⟨(P1,25/09)_Seq1 (P1,26/09)_Seq1 (P1,27/09)_Seq1 (P1,28/09)_Seq1⟩
⟨(P1,17/11)_Seq2 (P1,18/11)_Seq2⟩
⟨(P2,08/02)_Seq3 (P2,09/02)_Seq3⟩
⟨(P2,13/05)_Seq4 (P2,14/05)_Seq4 (P2,15/05)_Seq4⟩
⟨(P3,10/04)_Seq5 (P3,11/04)_Seq5 (P3,12/04)_Seq5 (P3,13/04)_Seq5⟩

Formally, let $\mathcal{D}_\mathcal{S}$ be a sequential dataset and $S_i \in \mathcal{D}_\mathcal{S}$ a sequence of itemsets. We model S_i as $\langle \texttt{IS1\_Seq}i\ \texttt{IS2\_Seq}i ... \texttt{IS}p\texttt{\_Seq}i\ \texttt{Seq}i \rangle$, that is, a sequence of UIDs. Let $\mathcal{UID}_\mathcal{S}$ be the set of all such sequences of UIDs derived from the $\mathcal{D}_\mathcal{S}$ sequences. $\texttt{Seq}i$ is the UID of the *target itemset* and it uniquely identifies the sequence S_i. We define $G_M = \{\texttt{Seq}i\}_{i\in[1,n]}$, where $n = |\mathcal{UID}_\mathcal{S}|$, as the set of all the target itemset UIDs in $\mathcal{UID}_\mathcal{S}$. $\texttt{IS}j\texttt{\_Seq}i$ is the UID of a *non-target itemset* and specifies the sequence $\texttt{Seq}i$ that owns the itemset. We define $G_T = \{\texttt{IS}j\texttt{\_Seq}i\}_{i\in[1,n];j\in[1,p]}$ as the set of all the non-target itemset UIDs in $\mathcal{UID}_\mathcal{S}$, where p is the number of itemsets (except the target itemset) in sequence $\texttt{Seq}i$. The function $getS : G_M \cup G_T \rightarrow \mathcal{D}_\mathcal{S}$ maps a target/non-target itemset UID to the original sequence that owns the itemset. The function $getIS : G_M \cup G_T \rightarrow \mathcal{IS}$ maps a UID to the corresponding target/non-target itemset.

For instance, $\mathcal{UID}_{\mathcal{S}fluA}$ (Tab. 3.3) is a sequential dataset of UIDs, where G_M is the set of all the viral test UIDs, while G_T is the set of all the medical examination UIDs. The patient sequence of UIDs ⟨(P1,17/11)_Seq2 (P1,18/11)_Seq2⟩ is derived from the patient sequence of itemsets $getS(\texttt{(P1,18/11)\_Seq2}) = S2 = \langle(\texttt{COUGH}_\texttt{high})(\texttt{Influenza}_\texttt{A})\rangle$ shown in Tab. 3.2. The UID (P1,18/11)_Seq2 uniquely identifies the target 1-itemset $getIS(\texttt{(P1,18/11)\_Seq2}) = (\texttt{Influenza}_\texttt{A})$ that ends $S2$. Similarly, (P1,17/11)_Seq2 uniquely identifies the non-target itemset $getIS(\texttt{(P1,17/11)\_Seq2}) = (\texttt{COUGH}_\texttt{high})$ owned by the $getS(\texttt{(P1,17/11)\_Seq2}) = S2$ patient sequence.

Relying on Tab. 3.2 and 3.3 and following the data model depicted in Fig. 3.3, we encode $\mathcal{D}_{\mathcal{S}fluA}$ into the RCA input shown in Tab. 3.4.

Table 3.4: RCF that encodes sub-dataset $\mathcal{D}_{\mathcal{S}fluA}$ (Tab. 3.2); formal contexts: KS, KVT and KME; qualitative relational contexts: RmS and RhS; temporal relational contexts: RME-ipb-ME and RVT-ipb-ME

KS	COUGH	FEVER
COUGH	×	
FEVER		×

KVT
(P1,28/09)_Seq1
(P1,18/11)_Seq2
(P2,09/02)_Seq3
(P2,15/05)_Seq4
(P3,13/04)_Seq5

KME
(P1,25/09)_Seq1
(P1,26/09)_Seq1
(P1,27/09)_Seq1
(P1,17/11)_Seq2
(P2,08/02)_Seq3
(P2,13/05)_Seq4
(P2,14/05)_Seq4
(P3,10/04)_Seq5
(P3,11/04)_Seq5
(P3,12/04)_Seq5

RmS	COUGH	FEVER
(P1,25/09)_Seq1		
(P1,26/09)_Seq1	×	
(P1,27/09)_Seq1	×	
(P1,17/11)_Seq2		
(P2,08/02)_Seq3	×	
(P2,13/05)_Seq4		
(P2,14/05)_Seq4	×	
(P3,10/04)_Seq5		
(P3,11/04)_Seq5	×	
(P3,12/04)_Seq5		

RhS	COUGH	FEVER
(P1,25/09)_Seq1		×
(P1,26/09)_Seq1		
(P1,27/09)_Seq1		×
(P1,17/11)_Seq2	×	
(P2,08/02)_Seq3		
(P2,13/05)_Seq4	×	
(P2,14/05)_Seq4		×
(P3,10/04)_Seq5	×	
(P3,11/04)_Seq5		
(P3,12/04)_Seq5		×

RME-ipb-ME	(P1,25/09)_Seq1	(P1,26/09)_Seq1	(P1,27/09)_Seq1	(P1,17/11)_Seq2	(P2,08/02)_Seq3	(P2,13/05)_Seq4	(P2,14/05)_Seq4	(P3,10/04)_Seq5	(P3,11/04)_Seq5	(P3,12/04)_Seq5
(P1,25/09)_Seq1										
(P1,26/09)_Seq1	×									
(P1,27/09)_Seq1	×	×								
(P1,17/11)_Seq2										
(P2,08/02)_Seq3										
(P2,13/05)_Seq4										
(P2,14/05)_Seq4						×				
(P3,10/04)_Seq5										
(P3,11/04)_Seq5								×		
(P3,12/04)_Seq5								×	×	

RVT-ipb-ME	(P1,25/09)_Seq1	(P1,26/09)_Seq1	(P1,27/09)_Seq1	(P1,17/11)_Seq2	(P2,08/02)_Seq3	(P2,13/05)_Seq4	(P2,14/05)_Seq4	(P3,10/04)_Seq5	(P3,11/04)_Seq5	(P3,12/04)_Seq5
(P1,28/09)_Seq1	×	×	×							
(P1,18/11)_Seq2				×						
(P2,09/02)_Seq3					×					
(P2,15/05)_Seq4						×	×			
(P3,13/04)_Seq5								×	×	×

The cross tables KS (symptoms), KVT (viral tests) and KME (medical examinations) represent formal contexts. There is no formal context of viruses since we focus on a specific virus, and thus all viral tests detect the influenza A virus. Therefore, KVT has no column since by default each viral test represents the target 1-itemset ($\texttt{Influenza}_\texttt{A}$). KME has no column since a medical examination is described only by using the *has symptom* qualitative relations and the rows represent the UIDs of the medical examinations from Tab. 3.3. The RVT-ipb-ME (viral test *ipb* medical examination) and RME-ipb-ME (medical examination *ipb* medical examination) cross tables represent temporal relational contexts since both define temporal relations. The RmS (medical examination detects a *moderate* symptom) and RhS (medical examination detects a *high* symptom) cross tables represent qualitative relational contexts since

both define qualitative relations. For example, RVT-ipb-ME has viral tests as rows and medical examinations as columns. A cross indicates a link between objects, e.g. the cell identified by the (P2,15/05)_Seq4 viral test UID and the (P2,13/05)_Seq4 medical examination UID contains a cross since both are undergone by the same patient P2 and the medical examination precedes the viral test, as shown in Tab. 3.3.

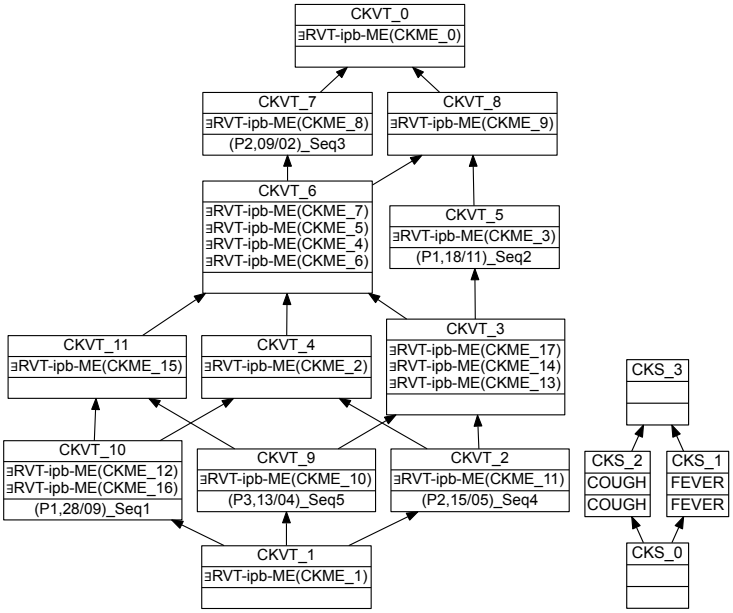
3.4.2 Applying the RCA Process

RCA is applied to the RCF shown in Tab. 3.4 and the family of concept lattices (the RCA output) depicted in Fig. 3.5 is obtained after four iterations. The simplified representations of these concept lattices are shown. There is a concept lattice for each formal context as follows: $\mathcal{L}_{\texttt{KVT}}$ (viral tests), $\mathcal{L}_{\texttt{KS}}$ (symptoms) and $\mathcal{L}_{\texttt{KME}}$ (medical examinations). Actually, $\mathcal{L}_{\texttt{KVT}}$ is an iceberg concept lattice since a user-defined minimum support θ is used to discover frequent concepts from the lattice of the objects of interest.

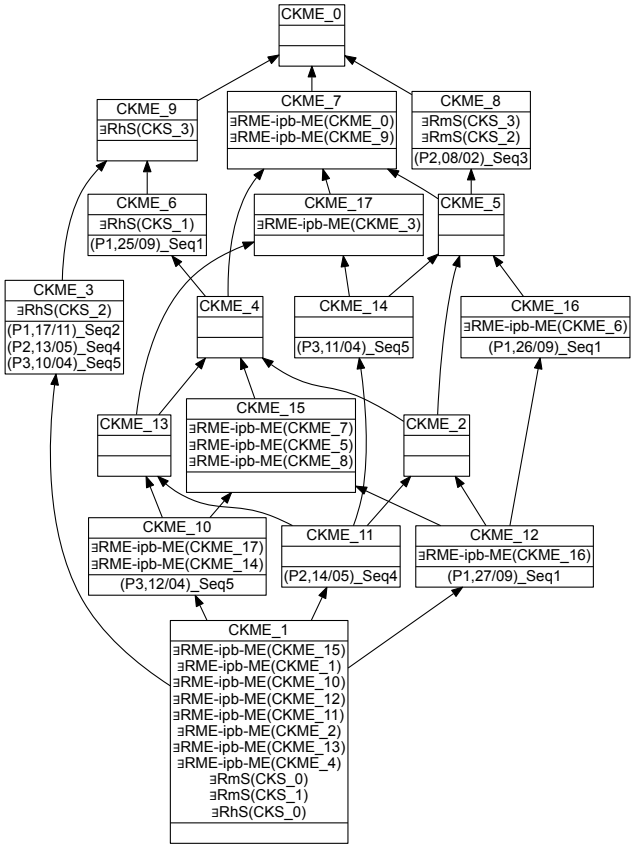
$\mathcal{L}_{\texttt{KVT}}$ is considered as the *main lattice* since it describes the temporal links between viral tests (target 1-itemsets) and medical examinations (non-target itemsets). $\mathcal{L}_{\texttt{KME}}$ is considered as the *temporal lattice* since it describes the temporal links between medical examinations. The $\mathcal{L}_{\texttt{KVT}}$ and $\mathcal{L}_{\texttt{KME}}$ concept lattices are modified during the iterative steps due to the qualitative and temporal relations that have respectively as domain the set of objects of KVT and the one of KME. The concept intents of these two lattices contain *temporal* and/or *qualitative* relational attributes derived by the relational scaling mechanism. It is worthwhile to mention that the relational scaling mechanism relies on the *existential quantifier* ($\exists$) since our objective is to capture all the relations between the analysed objects. For instance, the relational attribute $\exists$RhS(CKS_1) of the CKME_6 concept intent in lattice $\mathcal{L}_{\texttt{KME}}$ is a qualitative one since it highlights the qualitative relation *has symptom high* and allows us to navigate from lattice $\mathcal{L}_{\texttt{KME}}$ to lattice $\mathcal{L}_{\texttt{KS}}$. In contrast, the relational attribute $\exists$RVT-ipb-ME(CKME_7) of the CKVT_6 concept intent in $\mathcal{L}_{\texttt{KVT}}$ is a temporal one since it highlights the temporal relation *is preceded by* and allows us to navigate from lattice $\mathcal{L}_{\texttt{KVT}}$ to lattice $\mathcal{L}_{\texttt{KME}}$.

3.4.3 Analysing Relational Conceptual Structures

The RCA output is composed of relational conceptual structures (concept lattices) whose concepts can be navigated by domain experts (physicians in our running example) to obtain regularities from the analysed sequential data. It is worthwhile to mention that, firstly, a concept extent from the main or temporal lattice gathers all the instances from a set of objects (table) that share a set of intra-table and inter-table relations, which allow the navigation of the RCA output. Secondly, the navigation amongst the concept lattices follows the relations given by the proposed data model (Fig. 3.3). Finally, the generalisation order on concepts



(a) $\mathcal{L}_{\mathtt{KVT}}$ (the main lattice) (b) $\mathcal{L}_{\mathtt{KS}}$



(c) $\mathcal{L}_{\mathtt{KME}}$ (the temporal lattice)

Figure 3.5: The fix point of the RCF given in Tab.3.4: (a) the simplified lattice of viral tests; (b) the simplified lattice of symptoms; (c) the simplified lattice of medical examinations

can guide domain experts by highlighting how the obtained regularities relate to each other.

To illustrate these, let us consider the navigation paths depicted in Fig 3.6. From left to right there are excerpts out of the lattice of viral tests, the lattice of medical examinations and the lattice of symptoms (shown in Fig. 3.5). Let us recall that the analysed dataset is composed of 5 viral tests and 10 medical examinations. Thus, since our illustrative example is small, we use a minimum support $\theta = 1$ (number of sequences) for the lattice of viral tests. We consider that a regularity occurs *often* if its support is greater than or equal to 7 medical examinations or 4 viral tests; *sometimes* if its support is between 4 and 7 medical examinations or between 2 and 4 viral tests; *rarely* if its support is less than or equal to 4 medical examinations or 2 viral tests.

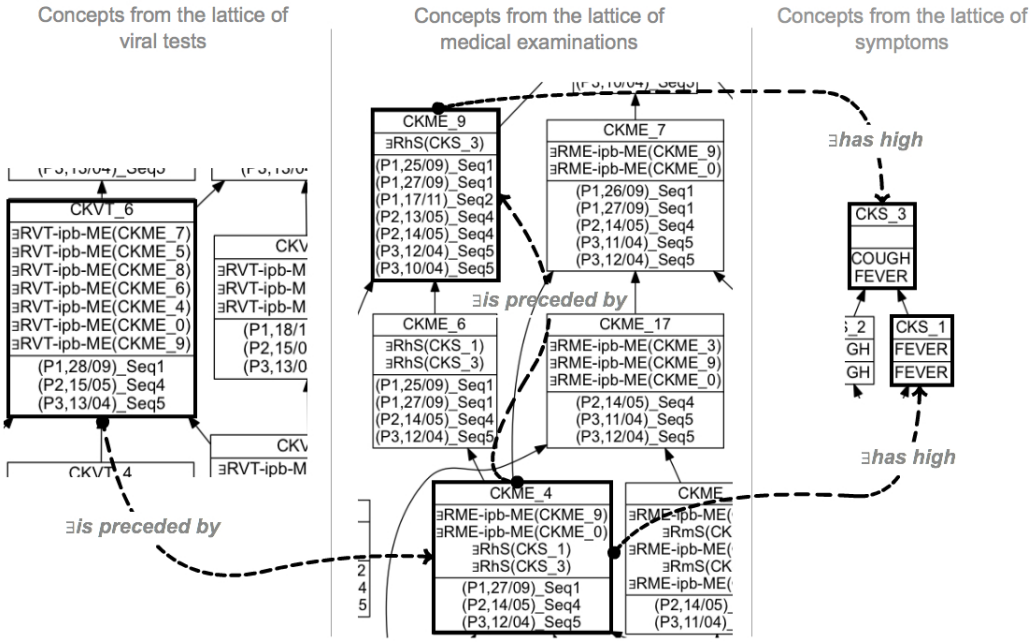


Figure 3.6: Several relations between the conceptual structures depicted in Fig. 3.5

By individually interpreting the highlighted concepts from Fig 3.6 without navigating the lattices, physicians can deduce that:

– concept CKVT_6 gathers all the viral tests ($|extent(\texttt{CKVT\_6})| = 3$) that are temporally related to at least one medical examination out of specific subsets from the analysed set of medical examinations (i.e. concept extents from lattice $\mathcal{L}_{\texttt{KME}}$);

– concept CKME_9 gathers all the medical examinations ($|extent(\texttt{CKME\_9})| = 7$) when the patients P1, P2 and P3 experienced at least one symptom with high intensity;

– concept CKME_4 gathers all the medical examinations ($|extent(\texttt{CKME\_4})| = 3$) when the

patients P1, P2 and P3 experienced at least one symptom with high intensity, but after undergoing one or more medical examinations;

- concept CKS_1 represents a specific symptom, namely fever, used to build the analysed sequences;
- concept CKS_3 represents the set of studied symptoms, precisely cough and fever.

Let us note that when we do not actually exploit the relations between concepts, the interpretation of each main or temporal concept is imprecise, and thus makes the understanding of the revealed regularity difficult. By considering the qualitative and temporal relations highlighted by the relational attributes of the concept intents, physicians can discover:

- from concept CKME_9 the frequent itemset "*often* (7 out of 10 medical examinations) *patients feel high cough and/or high fever*" revealed by exploiting the qualitative relation highlighted by the $\exists$RhS(CKS_3) relational attribute;
- from concept CKME_4 the frequent itemset "*rarely* (3 out of 10 medical examinations) *patients feel high fever symptom*" revealed by exploiting the $\exists$RhS(CKS_1) qualitative relational attribute; then, the sequential pattern "*rarely patients feel high fever symptom, but after experiencing high cough and/or high fever one or more times*" is revealed by additionally exploiting the temporal relation highlighted by the $\exists$RME-ipb-ME(CKME_9) relational attribute;
- that the frequent itemset associated with concept CKME_4 is a specialisation of the one associated with concept CKME_9 since CKME_4 $\preceq_{\texttt{KME}}$ CKME_9, where $\preceq_{\texttt{KME}}$ is the generalisation order between the concepts of lattice $\mathcal{L}_{\texttt{KME}}$;
- from concept CKVT_6 the sequential pattern "*sometimes* (3 out of 5 viral tests) *before the outbreak of influenza A virus patients feel high fever symptom, but after experiencing high cough and/or high fever one or more times*" sequential pattern is revealed by exploiting the inter-table temporal relation between viral tests and medical examinations that is highlighted by relational attribute $\exists$RVT-ipb-ME(CKME_4).

Moreover, the information encoded into the UIDs of concept extents can be considered. For example, the CKME_17 concept extent (Fig. 3.6) contains medical examinations undergone by the patients P2 and P3 and no medical examination undergone by patient P1. Then, physicians can deduce that the regularity revealed by this concept is not a global valid one in subdataset $\mathcal{D}_{\mathcal{S}fluA}$. In addition, by analysing the UIDs of the CKME_7 extent, physicians can notice that the regularity revealed by this concept is more persistent in $getS((\texttt{P1}, \texttt{26/09})\texttt{\_Seq1}) = getS((\texttt{P1}, \texttt{27/09})\texttt{\_Seq1}) = S1$ and $getS((\texttt{P3}, \texttt{11/04})\texttt{\_Seq5}) = getS((\texttt{P3}, \texttt{12/04})\texttt{\_Seq5}) = S5$,

where $\{S1, S5\} \subset \mathcal{D}_{\mathcal{S}fluA}$, since there are 2 occurrences of this regularity in each of these sequences compared with the only one occurrence in $getS((\texttt{P2}, \texttt{14/05})\_\texttt{Seq4}) = S4$. Thus, in $S4 \in \mathcal{D}_{\mathcal{S}fluA}$ this regularity can have another cause, e.g. a bacterial infection.

So far we have shown how relational conceptual structures can be analysed by physicians to discover various regularities in sequential data, which can be assessed and interpreted to understand how to recognise in advance virus outbreaks as well as to discriminate between virus types. In addition, the information encoded into the UIDs of medical examinations/viral tests can be exploited to obtain more valuable regularities.

However, navigating such relational conceptual structures in order to discover meaningful regularities is not a trivial task for domain experts since the number of concepts can be large, and, besides, these experts should move their focus from concept to concept and from lattice to lattice by considering respectively intra-table and inter-table relations. To help domain experts, in the next chapter, we propose a method to synthesise the obtained relational conceptual structures into a hierarchy of cpo-patterns.

3.5 Summary

In this chapter, we have presented a new approach for exploring sequential data, which can be gathered from multiple tables out of a relational database, within the framework of RCA. Our method respects the relational nature of sequential data and provides relational conceptual structures that can be navigated by domain experts to discover regularities from the analysed data. A general data model has been introduced that allows the conversion of various sequential data into the input of RCA.

With a small example we have illustrated the "richness" of the RCA output (which is the benefit of exploiting the relational nature of the analysed data), i.e. the interrelated concept intents via various inter-table and/or intra-table relations, the concept extents that capture more information within the analysed sequential data and the generalisation order on the discovered regularities.

4 Extraction of Hierarchies of Multilevel CPO-Patterns

Contents

4.1 Introduction

In this chapter, we present the third step of the RCA-Seq approach, namely the extraction of organised multilevel cpo-patterns. To this end, our purpose is to formalise an automatic approach for extracting a hierarchy of multilevel cpo-patterns from the RCA output obtained by exploring qualitative sequential data as explained in Sect. 3.4. To illustrate our method we use the running example from Sect. 3.2 and the RCA output depicted in Fig. 3.5.

4.2 Characteristics of the RCA Output Obtained by Exploring Sequential Data

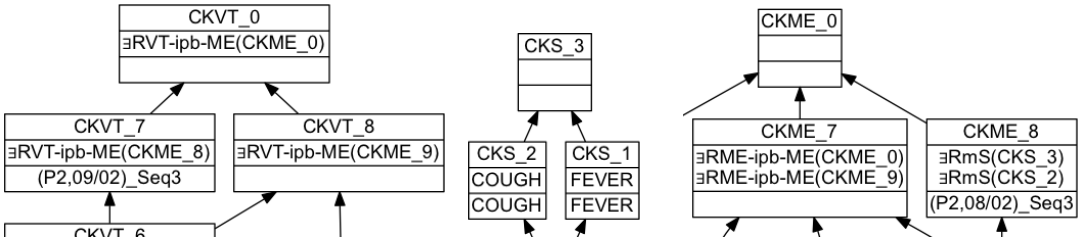
In order to extract a hierarchy of multilevel cpo-patterns from the RCA output we exploit the structure and the properties of this output.

4.2.1 Structure of the RCA Output

Relying on the data model shown in Fig. 3.4, we denote the four sets of objects: G_M the set of all the target itemset UIDs (e.g. viral tests), G_T the set of all the non-target itemset UIDs (e.g. medical examinations), G_I the set of all the items (e.g. symptoms) used to build the non-target itemsets and G_O the set of all the items of interest (e.g. viruses) used to build the target itemsets. Accordingly, the RCA output comprises four concept lattices, one for each set of objects, as follows: the *main lattice* (e.g. $\mathcal{L}_{\texttt{KVT}}$ in Fig. 3.5a), the *temporal lattice* (e.g. $\mathcal{L}_{\texttt{KME}}$ in Fig. 3.5c), the *lattice of items* (e.g. $\mathcal{L}_{\texttt{KS}}$ in Fig. 3.5b) and the *lattice of the items of interest* (e.g. the lattice of viruses).

Let us recall that in our running example we consider only one virus, e.g. influenza A, and thus we do not build the lattice of the items of interest. In addition, we note that the items used to build itemsets are atomic ones. Furthermore, the RCA-based exploration step employs a relational scaling mechanism that relies on the $\exists$ quantifier since the objective is to capture all the relations between the analysed objects.

Now, we describe the structure of the first three aforementioned concept lattices. For easy reading, we show in Fig. 4.1 excerpts from the RCA output (Fig. 3.5) obtained with our running example.



(a) $\mathcal{L}_{\texttt{KVT}}$ (b) $\mathcal{L}_{\texttt{KS}}$ (c) $\mathcal{L}_{\texttt{KME}}$

Figure 4.1: Excerpts from the RCA output (the simplified concept lattices) depicted in Fig. 3.5; (a) the lattice of viral tests, (b) the lattice of symptoms and (c) the lattice of medical examinations

Firstly, let $\mathcal{L}_{K_M} = (\mathcal{C}_{K_M}, \preceq_{K_M})$ be the main lattice whose set of concepts $\mathcal{C}_{K_M}$ is derived from a formal context $K_M = (G_M, M_M, I_M)$. G_M is the domain of a temporal relation *is preceded by*, denoted by $ipb_1 \subseteq G_M \times G_T$ (e.g. viral test ipb_1 medical examination), that defines the temporal links between the target itemsets and the non-target itemsets. A main

concept $C_M \in \mathcal{C}_{K_M}$ is a pair (X_m, Y_m) such that:

- **the intent** Y_m contains temporal relational attributes of the form $\exists ipb_1(C_T)$, e.g. the CKVT_7 concept intent (Fig. 4.1a) contains $\exists$RVT-ipb-ME(CKME_8), where C_T is a concept from the temporal lattice that describes the objects (the non-target itemset UIDs) from $ran(ipb_1) = G_T$;

- **the extent** X_m gathers all the target itemset UIDs in G_M that respect the temporal order with the G_T objects pointed by the temporal relational attributes of Y_m. We recall that the G_M objects uniquely identify the sequences in $\mathcal{D}_\mathcal{S}$ and we denote by $\mathcal{D}_m$ the set of all the sequences identified by the objects of X_m, thus $\mathcal{D}_m \subseteq \mathcal{D}_\mathcal{S}$.

Secondly, let $\mathcal{L}_{K_I} = (\mathcal{C}_{K_I}, \preceq_{K_I})$ be the lattice of items whose set of concepts $\mathcal{C}_{K_I}$ is derived from a $K_I = (G_I, M_I, I_I)$ formal context. Since the items of G_I are unordered and they mutually exclude each other, a nominal scaling was applied to obtain a partition of these items into concept extents. Therefore, lattice $\mathcal{L}_{K_I}$ is a partial order on G_I where the $\top(\mathcal{L}_{K_I})$ concept extent contains all the items, while the other concept extents, except for the $\bot(\mathcal{L}_{K_I})$ extent, contain only one item. For example, in Fig. 4.1b, the CKS_3 concept extent is composed of all surveyed items, namely FEVER and COUGH, while the extent of CKS_1 contains only the FEVER item.

Finally, let $\mathcal{L}_{K_T} = (\mathcal{C}_{K_T}, \preceq_{K_T})$ be the temporal lattice whose set of temporal concepts $\mathcal{C}_{K_T}$ is derived from a $K_T = (G_T, M_T, I_T)$ formal context. G_T is both the domain and the range of a second temporal relation *is preceded by*, denoted by $ipb_2 \subseteq G_T \times G_T$ (e.g. medical examination ipb_2 medical examination), that defines the temporal links between the non-target itemsets. Furthermore, G_T is the domain of a set of qualitative relations *has item quality*, denoted by $hi_q \subseteq G_T \times G_I$ (e.g. medical examination hi_q symptom), that define the non-target itemsets. There is a qualitative relation for each item quality (e.g. a symptom can have a moderate or high intensity, and thus we have *has item moderate* and *has item high*). Let us note that in our running example *has item* is called *has symptom*. A temporal concept $C_T \in \mathcal{C}_{K_T}$ is a pair (X_t, Y_t) such that:

- **the intent** Y_t can contain two types of relational attributes as follows: the temporal attributes of the form $\exists ipb_2(C'_T)$, e.g. $\exists$RME-ipb-ME(CKME_9) of the CKME_7 intent (Fig. 4.1c), where C'_T is a concept from the temporal lattice that describes the objects from $ran(ipb_2) = G_T$, i.e. $C'_T \in \mathcal{C}_{K_T}$; the qualitative attributes of the form $\exists hi_q(C_I)$, e.g. $\exists$RmS(CKS_2) of the CKME_8 intent (Fig. 4.1c), where C_I is a concept from the lattice of items that describes the objects (items) from $ran(hi_q) = G_I$;

- **the extent** X_t gathers all the UIDs in G_T identifying non-target itemsets that contain the

items revealed by the qualitative relational attributes of Y_t and that respect the temporal order with the G_T objects pointed by the temporal relational attributes of Y_t.

4.2.2 Properties of the RCA Output

In this section, we give some useful properties of the RCA output, which rely on its aforementioned structure, to help the extraction step of cpo-patterns. Briefly, the sequential patterns that coexist in the same sequences in a sequential dataset $\mathcal{D}_\mathcal{S}$ are revealed by navigating interrelated concept intents.

Property 4.1. *Each temporal relational attribute of a main concept intent allows to extract at least one sequential pattern. In contrast, if there is no temporal relational attribute in a main concept intent, this concept represents no sequential pattern.*

Indeed, let $C_M \in \mathcal{C}_{K_M}$ be a main concept (e.g. the CKVT_4 concept in Fig. 4.2) and $\exists ipb_1(C_T)$ a temporal relational attribute of its intent (e.g. $\exists$RVT-ipb-ME(CKME_4)), where $C_T \in \mathcal{C}_{K_T}$. If the C_T intent contains a qualitative relational attribute $\exists hi_q(C_I)$ (e.g. $\exists$RhS(CKS_1) in Fig. 4.2), where $C_I \in \mathcal{C}_{K_I}$, then C_T reveals an itemset of *qualitative values* (e.g. $\texttt{FEVER}_{\texttt{high}}$). Moreover, if the C_T concept intent contains a temporal relational attribute $\exists ipb_2(C'_T)$ (e.g. $\exists$RME-ipb-ME(CKME_9)), then C_T leads to another itemset in the sequential pattern, depending on the C'_T intent. Therefore, the temporal relational attributes reveal the *order on itemsets* in the sequential pattern and the qualitative relational attributes reveal the *itemsets* of the sequential pattern. In contrast, if the C_T intent (e.g. CKME_8 used to build $\exists$RVT-ipb-ME(CKME_8) of CKVT_4 in Fig. 4.2) contains no temporal relational attribute, the extraction of the sequential pattern is finished.

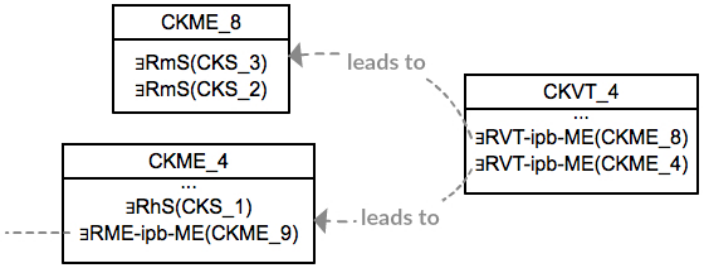


Figure 4.2: Two navigation paths beginning with the CKVT_4 main concept intent (Fig. 3.5a)

Property 4.2. *Let $C_M = (X_m, Y_m) \in \mathcal{C}_{K_M}$ be a main concept whose intent Y_m contains at least one temporal relational attribute. Then, C_M can be associated with a cpo-pattern $\mathcal{G}_m$ that summarises the set of sequential patterns derived from Y_m. Thus, $Support(\mathcal{G}_m) = |X_m|$.*

Proof. A set of sequential patterns can be transformed into po-patterns [Casas-Garriga, 2005]. A po-pattern associated with a concept is closed since the corresponding set of sequences is

maximal or equivalently the concept extent is maximal. Each element of X_m corresponds to a sequence that supports the $\mathcal{G}_m$ cpo-pattern. □

Property 4.3. *The set of cpo-patterns associated with the $\mathcal{L}_{K_M}$ main lattice is ordered according to the inclusion on extents. This order corresponds to the subsumption on graphs $\preceq_g$ (Sect. 2.2.3).*

Proof. Let $\mathcal{G}_m$ and $\mathcal{G}'_m$ be two cpo-patterns with $\mathcal{P}_{\mathcal{G}_m}$ and $\mathcal{P}_{\mathcal{G}'_m}$ their sets of paths. Suppose $\mathcal{G}_m$ (resp. $\mathcal{G}'_m$) is associated with a concept $C_M = (X_m, Y_m) \in \mathcal{C}_{K_M}$ (resp. $C'_M = (X'_m, Y'_m)$) and $X_m \subseteq X'_m$. Then $Y'_m \subseteq Y_m \leftrightarrow \forall m \in Y'_m, m \in Y_m$. Then $\forall M' \in \mathcal{P}_{\mathcal{G}'_m}, \exists M \in \mathcal{P}_{\mathcal{G}_m}, M' \preceq_s M \rightarrow \mathcal{G}'_m \preceq_g \mathcal{G}_m$. □

A naive approach to extract cpo-patterns from the RCA output navigates all the temporal and qualitative relational attributes of the interrelated concept intents. This approach generates redundant information, hence the obtained cpo-patterns need post-processing. However, two properties of the RCA output can be used to improve the extraction process. In the following, we show how to directly obtain the minimal representations of the extracted cpo-patterns by considering only the relational attributes pointing to the most specific concepts and by pruning the temporal relational attributes that can be deduced by transitivity.

Property 4.4. *Let $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ be two concepts from the same lattice $\mathcal{L}_K = (\mathcal{C}_K, \preceq_K)$ such that $C_1 \preceq_K C_2$. Let $C = (X, Y)$ be a concept whose intent has two relational attributes $\exists r(C_1)$ and $\exists r(C_2)$ (derived from the same relation* r*). Then $\exists r(C_1) \rightarrow \exists r(C_2)$.*

Proof. $\exists r(C_1) \in Y \leftrightarrow \forall g \in X, r(g) \cap X_1 \neq \emptyset$. Since $C_1 \preceq_K C_2$, $X_1 \subseteq X_2$, and thus $r(g) \cap X_2 \neq \emptyset \leftrightarrow \exists r(C_2) \in Y$. □

Thus, the relational attributes are ordered and the $\exists r(C_2)$ relational attribute is redundant in the interpretation of concept C. Moreover, let $\mathcal{Q}$ be the set of all the concepts used to introduce relation r in concept intent C; then, we can remove all the relational attributes $\exists r(C') \in Y, C' \in \mathcal{Q}$ that point to concepts which are upper covers for other ones in $\mathcal{Q}$. We recall that Roth et al. [2008] define an upper cover (or upper neighbour) C_1 of C_2 in a lattice $\mathcal{L}_K = (\mathcal{C}_K, \preceq_K)$, denoted by $C_1 \rhd_K C_2$, if $C_2 \prec_K C_1$ and there is no $C_3 \in \mathcal{C}_K$ such that $C_2 \prec_K C_3 \prec_K C_1$.

Property 4.5. *Let* ipb *be a temporal relation. Let $C = (X, Y)$, $C_1 = (X_1, Y_1)$ and $C_2 = (X_2, Y_2)$ be three concepts such that $\{\exists ipb(C_1), \exists ipb(C_2)\} \subseteq Y$ and $\exists ipb(C_2) \in Y_1$. Then $\exists ipb(C_2) \in Y$ can be deduced from $\exists ipb(C_1) \in Y$.*

Proof. Property 4.5 is directly obtained from the transitivity of the temporal relation ipb. □

4.3 From the RCA Output to a Hierarchy of Multilevel CPO-Patterns

Based on the structure and the properties of the RCA output, we introduce an algorithm that generates directly organised cpo-patterns, one for each main concept in $\mathcal{L}_{K_M}$.

4.3.1 The CPOHrchy Algorithm

Algorithms 1 and 2 illustrate our proposal for extracting a hierarchy of cpo-patterns by navigating the RCA output. Since our objective is to directly obtain organised cpo-patterns, and, besides, since there is a generalisation order on the concepts, we propose to use a 3-tuple structure for a main concept $C_M = (X_m, Y_m, \mathcal{G}_m)$. We note that $\mathcal{G}_m$ is the cpo-pattern associated with C_M and it is represented as an adjacency list of pointers to concepts. Moreover, since $\mathcal{G}_m$ is represented as an adjacency list that contains pointers to temporal concepts, we propose to use a 3-tuple structure for a temporal concept $C_T = (X_t, Y_t, v_t)$ as well. We note that v_t is the vertex derived from intent Y_t.

Algorithm 1: CPOHrchy

Input : the RCA output comprises $\mathcal{L}_{K_M} = (\mathcal{C}_{K_M}, \preceq_{K_M})$, $\mathcal{L}_{K_T} = (\mathcal{C}_{K_T} \preceq_{K_T})$ and $\mathcal{L}_{K_I} = (\mathcal{C}_{K_I}, \preceq_{K_I})$ whose concepts have the aformentioned 3-tuple structures.
Output: $\mathcal{L}_{K_M}$ whose concepts are updated with the associated cpo-patterns.

```
1  foreach C_M = (X_m, Y_m, G_m) where C_M ≠ ⊥(L_{K_M}) do
2      if Y_m has temporal relational attributes then
3          C_next ← SearchAdjacentConcepts(Y_m);
4          G_m ← initialise to (C_M, C_next);
5          Queue ← enqueue the C_next concepts and mark them as visited;
6          repeat
7              C_T = (X_t, Y_t, v_t) ← dequeue Queue;
8              v_t ← derive an itemset based on L_{K_I} and the qualitative relational attributes of Y_t
               (Sect. 4.3.2);
9              if Y_t has temporal relational attributes then
10                 C'_next ← SearchAdjacentConcepts(Y_t);
11                 G_m ← add (C_T, C'_next) to G_m;
12                 C'_next ← delete already visited concepts from C'_next;
13                 if C'_next is not empty then
14                     Queue ← enqueue the C'_next concepts and mark them as visited;
15                 end
16             else
17                 G_m ← add (C_T, {}) to G_m;
18             end
19         until Queue is empty;
20     end
21 end
```

Algorithm 1, referred to as CPOHrchy, takes as input the three lattices $\mathcal{L}_{K_M}$, $\mathcal{L}_{K_T}$ and $\mathcal{L}_{K_I}$ and its output is the main lattice $\mathcal{L}_{K_M}$ whose concepts are updated with the corresponding cpo-patterns. The three lattices are represented as sets of concepts, where for each concept its upper covers are known. For each main concept C_M whose intent has at least one temporal relational attribute, an adjacent list of pointers to the navigated concepts (i.e. the concepts that are adjacent to each navigated concept) is built in a breadth-first manner based on Properties 4.4 and 4.5. For each navigated concept is derived a vertex labelled with an itemset (detailed in Sect. 4.3.2). It is worth mentioning that $\perp(\mathcal{L}_{K_M})$ is not taken into consideration since generally this is too specific and not frequent (according to a user-defined minimum support θ).

Algorithm 2, called SearchAdjacentConcepts, shows how to derive from the temporal relational attributes of the intent of a main concept C_M the next concepts $\mathcal{C}_{next}$ that should be navigated by relying on Properties 4.4 and 4.5. This algorithm is applicable to temporal concepts (in this case ipb_1 is replaced with ipb_2) as well. **Lines [2-8]:** delete all the concepts in $\mathcal{C}_{next}$ that are upper covers for other concepts in $\mathcal{C}_{next}$, i.e. delete concepts that are not the most specific ones in $\mathcal{C}_{next}$. **Lines [9-15]:** prune all the concepts in $\mathcal{C}_{next}$ that can be deduced by navigating other ones in $\mathcal{C}_{next}$.

Algorithm 2: SearchAdjacentConcepts

Input : intent Y_m of a main concept C_M.
Output: $\mathcal{C}_{next}$ the set of the next navigated concepts.

1 $\mathcal{C}_{next} \leftarrow$ initialise to $\{C_T | (\exists ipb_1(C_T)) \in Y_m\}$;
2 **if** $|\mathcal{C}_{next}| > 1$ **then**
3 $\quad UpperCovers \leftarrow \{\}$;
4 $\quad$ **foreach** $C_T \in \mathcal{C}_{next}$ **do**
5 $\quad\quad UpperCovers \leftarrow$ add $\{C'_T | C'_T \rhd_{K_T} C_T\}$ to $UpperCovers$;
6 $\quad$ **end**
7 $\quad \mathcal{C}_{next} \leftarrow \mathcal{C}_{next} \setminus UpperCovers$;
8 **end**
9 **if** $|\mathcal{C}_{next}| > 1$ **then**
10 $\quad ToBeDeleted \leftarrow \{\}$;
11 $\quad$ **foreach** $C_T = (X_t, Y_t, v_t) \in \mathcal{C}_{next}$ **do**
12 $\quad\quad ToBeDeleted \leftarrow$ add $\{C'_T | (\exists ipb_2(C'_T)) \in Y_t\}$ to $ToBeDeleted$;
13 $\quad$ **end**
14 $\quad \mathcal{C}_{next} \leftarrow \mathcal{C}_{next} \setminus ToBeDeleted$;
15 **end**

We propose two optimisations, first, for the CPOHrchy algorithm and second, for the RCA-based exploration step:

1. since a temporal concept $C_T = (X_t, Y_t, v_t)$ can be navigated several times for distinct cpo-patterns, we process C_T only at its first navigation, i.e. SearchAdjacentConcepts is

applied only once and its result is saved for later use; similarly, v_t is computed and saved;

2. since a cpo-pattern $\mathcal{G}_m$ associated with a main concept $C_M = (X_m, Y_m)$ is discovered if $Support(\mathcal{G}_m) = |X_m| \geq \theta$ (we recall that θ is a user-defined minimum support for the main lattice), then all the navigated temporal concepts $C_T = (X_t, Y_t)$ should have $|X_t| \geq |X_m|$ (this holds since in our case each itemset of a sequence is uniquely identified intra-sequence and inter-sequence). Therefore, we diminish the navigation space by defining a minimum support $\theta' = \theta$ for the temporal lattice as well.

We highlight in Sect. 7.3.2.2 how these two optimisations improve the RCA-SEQ approach.

4.3.2 From Concepts to Vertices Labelled with Itemsets

To convert an adjacency list of pointers to concepts (i.e. the representation of a cpo-pattern obtained with CPOHrchy) to one of vertices labelled with itemsets, we analyse the qualitative relational attributes from these concept intents. It is worthwhile to mention that we apply again Property 4.4 to analyse, for the same qualitative relation, only the most specific concepts used to build the corresponding qualitative relational attributes in a concept intent.

4.3.2.1 Deriving Items

A qualitative relational attribute can be vague or defined depending on the generality or specificity of the concept it points to.

Definition 4.1 (Vague/Defined Relational Attribute). *The relational attribute $\exists r(C)$, where C is a concept in $\mathcal{L}_K = (\mathcal{C}_K, \preceq_K)$, is called* vague *if $C \equiv \top(\mathcal{L}_K)$, respectively it is called* defined *if $C \prec_K \top(\mathcal{L}_K)$.*

Relying on the partial order on items given by lattice $\mathcal{L}_{K_I}$ (e.g. the lattice of symptoms $\mathcal{L}_{\text{KS}}$ given in Fig. 4.1b) from the RCA output and on the aforementioned types of relational attributes, we define three types of items that reveal concrete (the most specific) and abstract (the most general) information from the analysed sequential data as follows:

- let C be a concept whose intent contains the $\exists hi_q(C')$ defined qualitative relational attribute, with $extent(C') = \{item\}$. The extracted item is a **concrete qualitative item**, denoted by "$item_q$", where q is the item quality. The concrete qualitative item describes a collection of objects (e.g. medical examinations) which point out the occurrence of the same concrete item having the same quality;

- let C be a concept whose intent contains the $\exists hi_q(\top(\mathcal{L}_{K_I}))$ vague qualitative relational attribute that summarises all the qualitative relational attributes representing the same

qualitative relation. The extracted item is an **abstract qualitative item**, denoted by "$?_q$". The abstract qualitative item describes a collection of objects which point out the occurrence of dissimilar items having the same quality;

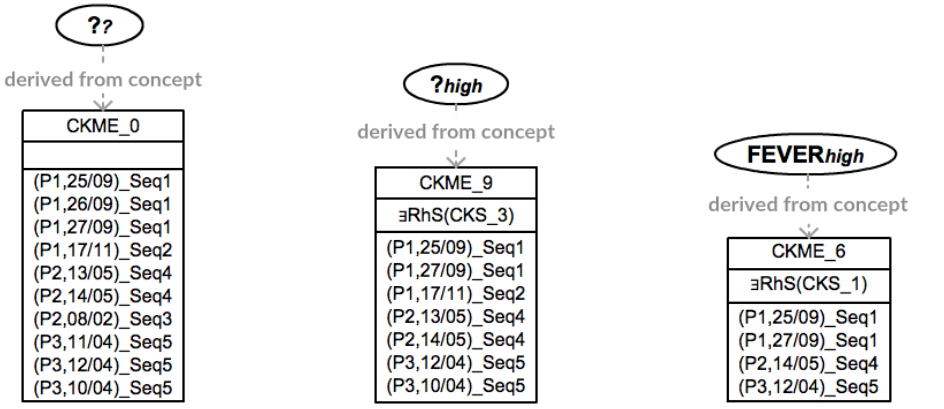
– let C be a concept whose intent has no qualitative relational attribute. Then the extracted item is an **abstract item**, denoted by "$?_?$". The abstract item describes a collection of objects which point out the occurrence of dissimilar items having dissimilar qualities.

Based on these three types of items, we are able to extract *multilevel cpo-patterns*, i.e. cpo-patterns that contain items from a poset (e.g. $\mathcal{L}_{\texttt{KS}}$ in Fig. 4.1b). It is worthwhile to mention that these three types emerge for two reasons. Firstly, we use a nominal scaling to build the formal context of items that reveals a partial order over the set of items. Thus, in our running example emerges an abstraction of the fever and cough symptoms, denoted by "?", that represents the surveyed symptoms. Secondly, we highlight the qualitative values of items, e.g. moderate or high cough, by means of qualitative binary relations. Therefore, during the iterative learning process, vague qualitative relational attributes are derived (based on the aforementioned poset and the qualitative relations) that disclose abstract or abstract qualitative items. For instance, we model the moderate and high symptoms into two different qualitative relations, i.e. *has symptom high* and *has symptom moderate*, in order to be able to discover qualitative abstractions, namely $?_{\texttt{moderate}}$ and $?_{\texttt{high}}$. In contrast, if the conceptual scaling is used to transform a many-valued attribute (e.g. FEVER that can be *high* or *moderate*) into one-value attributes (e.g. $\texttt{FEVER}_{\texttt{high}}$ and $\texttt{FEVER}_{\texttt{moderate}}$), which constitute the set of attributes of the formal context built for the medical examinations, the aforementioned qualitative abstractions cannot be revealed.

4.3.2.2 Labelling Vertices

For each concept intent in the adjacency list of pointers to concepts we derive a vertex labelled with an itemset comprising the items derived from all the qualitative relational attributes of the concept intent, as shown in Fig. 4.3.

Figure 4.3a depicts a vertex having an abstract item that characterises all the medical examinations from Tab. 3.1 (i.e. all the objects from the CKME_0 concept extent) since all of them detect at least one high or moderate symptom. Figure 4.3b shows a vertex having an abstract qualitative item characteristic to all the medical examinations from the CKME_9 extent that are described by different symptoms, but all having high intensity. Figure 4.3c illustrates a vertex having a concrete qualitative item characteristic to all the medical examinations from the CKME_6 extent that are described by a specific symptom, namely fever, that has high intensity.



(a) abstract (b) abstract qualitative (c) concrete qualitative

Figure 4.3: Deriving vertices from concept intents that contain only the relational attributes after pruning according to Property 4.4

4.4 Complexity Analysis of the RCA-Seq Approach

First, we present a time complexity analysis of RCA-Seq that is compared with the time complexity of the [Fabrègue et al., 2015] approach. Second, we present a space complexity of RCA-Seq.

4.4.1 Time Complexity

We first consider the method for extracting cpo-patterns presented in [Fabrègue et al., 2015]. In the following, $\mathcal{D}_\mathcal{S}$ is a sequence database, $\mathcal{I}$ is a set of items, $\mathcal{IS} \subseteq 2^\mathcal{I}$ is the set of the itemsets in the sequences of $\mathcal{D}_\mathcal{S}$ and l is the maximum length of the sequences in $\mathcal{D}_\mathcal{S}$. Let us denote by m the number of the obtained cpo-patterns. The proposed method spans two steps and its overall complexity in the worst-case scenario is $O(m \cdot 2 \cdot (2 \cdot |\mathcal{I}|)^l)$. If we build a hierarchy of these results in a post-processing step and since the patterns are closed and already associated with their supporting sequences, the complexity of the extra step would be 1) building the context patterns-sequences: $O(m \cdot |\mathcal{D}_\mathcal{S}|)$ and 2) building the lattice[1]: $O(m^2 \cdot |\mathcal{D}_\mathcal{S}| \cdot (m+2))$. Thus, the whole process complexity would be $O(m \cdot 2 \cdot (2 \cdot |\mathcal{I}|)^l + m^3 \cdot |\mathcal{D}_\mathcal{S}|) = O_1$.

We now consider the current RCA-Seq approach that relies on two algorithms, namely Multi-FCA (the RCA process, [Rouane-Hacene et al., 2013]) and CPOHrchy. We focus on the worst-case scenario. Following the proposed data model (Fig. 3.4), let us consider an RCF (the RCA input) that comprises the set of formal contexts $\{K_M = (G_M, M_M, I_M), K_T = (G_T, M_T, I_T), K_I = (G_I, M_I, I_I)\}$ and the associated relational contexts. The obtained fix point contains the set of concept lattices $\{\mathcal{L}_{K_M}, \mathcal{L}_{K_T}, \mathcal{L}_{K_I}\}$ built from the scaled contexts

[1] the complexity of building a lattice $\mathcal{L}$ from a context (G, M, I) is $O(|G|^2 \cdot |M| \cdot |\mathcal{L}|)$ [Merwe et al., 2004]

$\{K_M^+ = (G_M, M_M^+, I_M^+), K_T^+ = (G_T, M_T^+, I_T^+), K_I = (G_I, M_I, I_I)\}$. We denote by $|\mathcal{L}_{K_i}|$ the number of formal concepts in lattice $\mathcal{L}_{K_i}$. According to [Rouane-Hacene et al., 2013], in the worst-case scenario the overall computation time of the considered fix point is $O(n_c \cdot n_o \cdot (n_a + n_o))$, where $n_c = max(|\mathcal{L}_{K_M}|, |\mathcal{L}_{K_T}|, |\mathcal{L}_{K_I}|)$, $n_a = max(|M_M^+|, |M_T^+|, |M_I|)$ and $n_o = max(|G_M|, |G_T|, |G_I|)$.

The worst-case scenario for the CPOHrchy algorithm is when each main/temporal concept points to all concepts in $\mathcal{L}_{K_T}$. Let us denote $m = |\mathcal{L}_{K_M}|$ (each element of $\mathcal{L}_{K_M}$ reveals a cpo-pattern), $p = |\mathcal{L}_{K_T}|$ and q the number of all the qualitative relational attributes from a temporal concept intent. First, we focus on Algorithm 2. The overall computation time is $O(p)$ since we iterate throughout p concepts pointed by the temporal relational attributes of Y_m at **Lines [1, 4, 7, 11, 14]**. The other lines are $O(1)$. Second, in Algorithm 1, **Lines [2–20]** are executed m times. **Lines [3, 5]** have the complexity $O(p)$ since $\mathcal{C}_{next}$ contains p concepts pointed by the temporal relational attributes of Y_m. **Lines [6–19]** are executed p times since each temporal concept in $\mathcal{L}_{K_T}$ is visited only once and the complexity of these lines is $O(p(q+p))$. Indeed, **Lines [10, 12, 14]** are $O(p)$ since $\mathcal{C}'_{next}$ has p concepts pointed by the temporal relational attributes of Y_t, **Line [8]** is $O(q)$ and the other lines are $O(1)$. Since generally $q \leq p$, the computation time becomes $O(p^2)$. Therefore, in the worst-case scenario the overall computation time for CPOHrchy is $O(m \cdot p^2)$.

To sum up, the overall time complexity of RCA-Seq is $O(n_c \cdot n_o \cdot (n_a + n_o) + m \cdot p^2) = O_2$. To compare with the aforementioned complexity O_1, we consider that the sizes of $\mathcal{I}$ and $\mathcal{D}_\mathcal{S}$ – which correspond to sets of objects – are smaller than n_o and m, p are smaller than n_c. Then, O_1 is upper bounded by $O(n_c \cdot 2 \cdot (2 \cdot n_o)^l + n_c^3 \cdot n_o)$, while O_2 is upper bounded by $O(n_c \cdot n_o \cdot (n_a + n_o) + n_c^3)$. Finally, since l is generally greater than 3, the complexity of RCA-Seq is better than the complexity of the approach described in [Fabrègue et al., 2015] combined with a lattice building step.

4.4.2 Space Complexity

The RCA-Seq approach extracts multilevel cpo-patterns by navigating concept intents from the $\mathcal{L}_{K_M}$ and $\mathcal{L}_{K_T}$ lattices. Thus, the worst-case scenario is when $\mathcal{L}_{K_M}$, $\mathcal{L}_{K_T}$ and $\mathcal{L}_{K_I}$ contain respectively $2^{|M_M^+|}$, $2^{|M_T^+|}$ and $2^{|M_I|}$ concepts.

Firstly, we discuss the RCA-based exploration step. The space complexity of the RCA input is $O(|G_M| \cdot |M_M| + |G_T| \cdot |M_T| + |G_I| \cdot |M_I| + |G_M| \cdot |G_T| + |G_T|^2 + h \cdot (|G_T| \cdot |G_I|))$ with h the number of the considered qualitative relations. The auxiliary space concerns the upgraded RCA input and the built lattices, which are as well the output of this step. Thus, the space complexity of the upgraded RCA input becomes $O(|G_M| \cdot |M_M^+| + |G_T| \cdot |M_T^+| + |G_I| \cdot |M_I| + |G_M| \cdot |G_T| + |G_T|^2 + h \cdot (|G_T| \cdot |G_I|))$. The space complexity of the main lattice $\mathcal{L}_{K_M}$ is $O(2 \cdot (2^{|M_M^+|} + |M_M^+| \cdot 2^{|M_M^+|-1}))$ with $2 \cdot 2^{|M_M^+|}$ the space used for the unique identifiers

of concepts and the number of objects from each main concept extent; $2 \cdot |M_M^+| \cdot 2^{|M_M^+|-1}$ the space used for all concept intents and the upper covers of concepts. Indeed, $\binom{|M_M^+|}{k} \cdot k$ is the number of attributes from all concept intents of size k, and thus the number of attributes from all the concept intents becomes $\sum_{k=0}^{|M_M^+|} \binom{|M_M^+|}{k} \cdot k = |M_M^+| \cdot 2^{|M_M^+|-1}$. Accordingly, the space complexity of the RCA output (i.e. the lattices $\mathcal{L}_{K_M}$, $\mathcal{L}_{K_T}$ and $\mathcal{L}_{K_I}$) is $O(2 \cdot (2^{|M_M^+|} + |M_M^+| \cdot 2^{|M_M^+|-1} + 2^{|M_T^+|} + |M_T^+| \cdot 2^{|M_T^+|-1} + 2^{|M_I|} + |M_I| \cdot 2^{|M_I|-1})) = O_\alpha$.

Secondly, CPOHrchy has O_α as the space complexity of the input. The auxiliary space is $O(3 \cdot |M_T^+| + (\mathcal{V}, \mathcal{E}))$ with $(\mathcal{V}, \mathcal{E})$ the space used for $\mathcal{G}_m$. The space complexity of the CPOHrchy output is $O_\alpha + O(2^{|M_M^+|} \cdot (\mathcal{V}, \mathcal{E})) + O(2^{|M_T^+|})$, i.e. the main lattice is updated with a cpo-pattern for each concept and the temporal lattice is updated with a vertex for each concept.

4.5 Application to the Running Example

To illustrate our approach, let us consider that we want to extract cpo-pattern $\mathcal{G}_{\texttt{CKVT\_10}}$ associated with the CKVT_10 main concept from the lattice of viral tests $\mathcal{L}_{\texttt{KVT}}$ (Fig. 3.5a). For easy reading, we give in Tab. 4.1 the intents of the navigated concepts.

Table 4.1: The concept intents navigated to extract the $\mathcal{G}_{\texttt{CKVT\_10}}$ cpo-pattern

Concept	Relational Attributes	
	Temporal	Qualitative
CKVT_10	∃RVT-ipb-ME(CKME_16) ∃RVT-ipb-ME(CKME_9) ∃RVT-ipb-ME(CKME_15) ∃RVT-ipb-ME(CKME_0) ∃RVT-ipb-ME(CKME_12) ∃RVT-ipb-ME(CKME_7) ∃RVT-ipb-ME(CKME_5) ∃RVT-ipb-ME(CKME_8) ∃RVT-ipb-ME(CKME_6) ∃RVT-ipb-ME(CKME_2) ∃RVT-ipb-ME(CKME_4)	
CKME_12	∃RME-ipb-ME(CKME_16) ∃RME-ipb-ME(CKME_9) ∃RME-ipb-ME(CKME_0) ∃RME-ipb-ME(CKME_7) ∃RME-ipb-ME(CKME_5) ∃RME-ipb-ME(CKME_8) ∃RME-ipb-ME(CKME_6)	∃RmS(CKS_3) ∃RmS(CKS_2) ∃RhS(CKS_1) ∃RhS(CKS_3)
CKME_16	∃RME-ipb-ME(CKME_9) ∃RME-ipb-ME(CKME_0) ∃RME-ipb-ME(CKME_6)	∃RmS(CKS_3) ∃RmS(CKS_2)
CKME_6		∃RhS(CKS_3) ∃RhS(CKS_1)

Following Fig. 4.4, from right to left, we start by examining all temporal relational attributes from the CKVT_10 concept intent (shown in Tab. 4.1) that are ordered according to the generalisation order $\preceq_{\texttt{KME}}$ on the concepts used to build them. Since there is only one

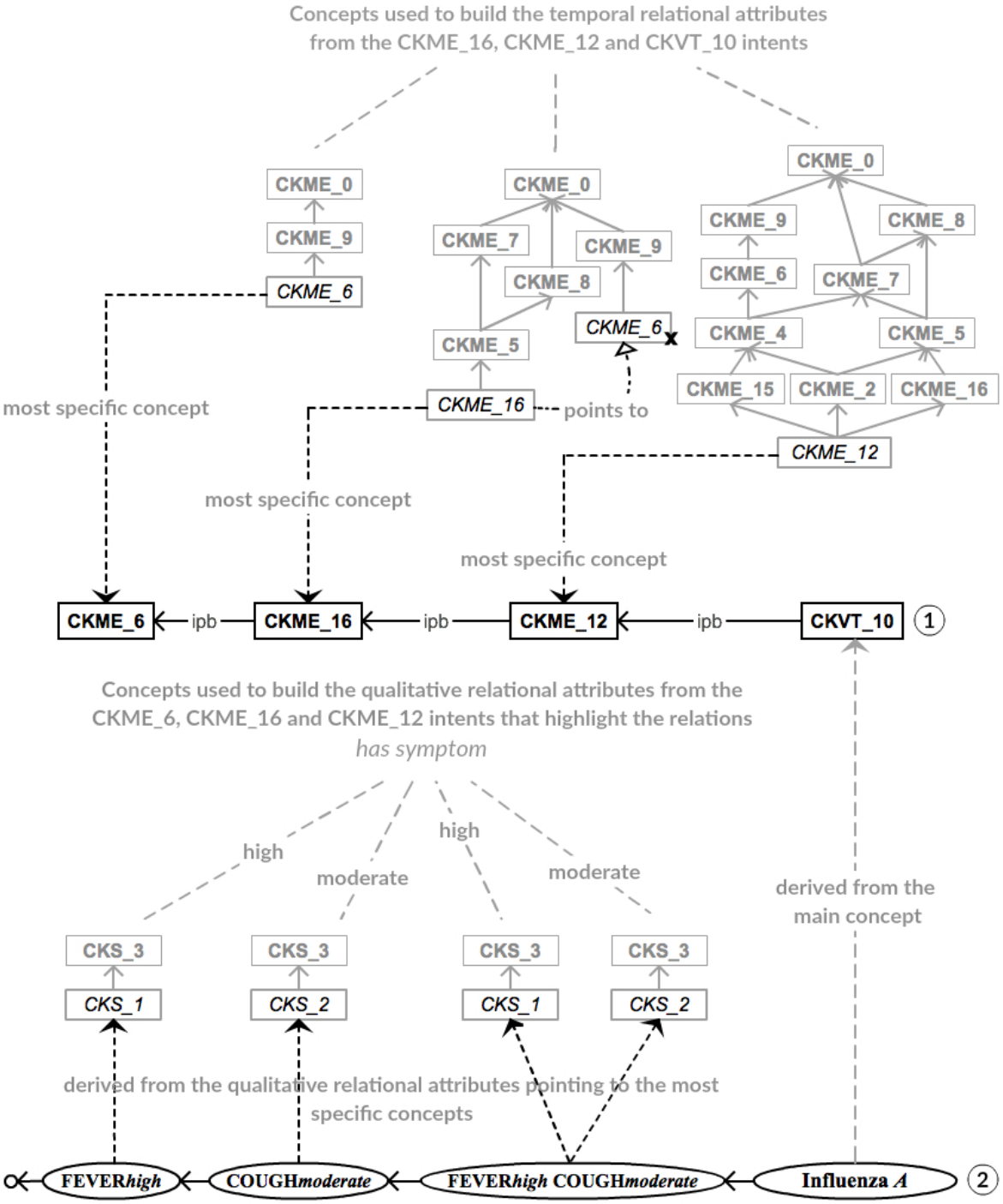


Figure 4.4: Extracting the cpo-pattern associated with the CKVT_10 main concept. ① is the set of navigated concepts and ② is the generated cpo-pattern $\mathcal{G}_{\texttt{CKVT\_10}}$. Each vertex in ② is derived from a concept in ① (from right to left)

most specific concept (i.e. $\forall C \in \{C' | (\exists \texttt{RVT-ipb-ME}(C')) \in intent(\texttt{CKVT\_10})\}, \texttt{CKME\_12} \preceq_{\texttt{KME}} C$ as illustrated in Fig. 4.4), the next navigated intent is the one of concept CKME_12 from the lattice of medical examinations $\mathcal{L}_{\texttt{KME}}$ (Fig. 3.5c). By analysing the concepts used to build all the temporal relational attributes from the CKME_12 concept intent shown in Tab. 4.1, we notice that there are two most specific concepts CKME_16 and CKME_6. Moreover, the CKME_16 intent given in Tab. 4.1 contains the $\exists$RME-ipb-ME(CKME_6) temporal relational attribute that points to concept CKME_6 (Fig. 4.4), and thus the next navigated concept is only CKME_16 since CKME_6 generates redundant information (Property 4.5). As shown in Tab. 4.1, the CKME_16 concept intent consists in three temporal relational attributes and the most specific concept used to build them is CKME_6 (Fig. 4.4). The CKME_6 intent given in Tab. 4.1 has no temporal relational attribute and thus the navigation is finished. In Fig. 4.4, ① represents the set of navigated concepts that should be converted into vertices.

To this end, we analyse the qualitative relational attributes from the navigated concept intents that enable us to extract the $\mathcal{G}_{\texttt{CKVT\_10}}$ cpo-pattern, referred to as ② in Fig. 4.4. From right to left, the vertex labelled with target 1-itemset ($\texttt{Influenza}_{\texttt{A}}$) contains the default concrete qualitative item associated with the CKVT_10 concept intent. The intent of CKME_12 shown in Tab. 4.1 contains four qualitative relational attributes that highlight two qualitative relations, precisely *has symptom high* and *has symptom moderate*. Therefore, itemset ($\texttt{FEVER}_{\texttt{high}}$ $\texttt{COUGH}_{\texttt{moderate}}$) is the label of the vertex derived from the CKME_12 concept intent. It contains the concrete qualitative item $\texttt{FEVER}_{\texttt{high}}$ derived from the most specific concept CKS_1 used to highlight the *has symptom high* relation and $\texttt{COUGH}_{\texttt{moderate}}$ derived from the most specific concept CKS_2 used to highlight the *has symptom moderate* relation. Similarly, the vertex labelled with itemset ($\texttt{COUGH}_{\texttt{moderate}}$) consists in only one concrete qualitative item derived from the $\exists$RmS(CKS_2) qualitative relational attribute of the CKME_16 intent; the vertex labelled with itemset ($\texttt{FEVER}_{\texttt{high}}$) consists in only one concrete qualitative item derived from the $\exists$RhS(CKS_1) qualitative relational attribute of the CKME_6 intent.

4.6 Analysis of a Hierarchy of Multilevel CPO-Patterns

Figure 4.5 illustrates a hierarchy of multilevel cpo-patterns, i.e. main lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 3.5a) whose concepts are upgraded with the cpo-patterns discovered in the sequential medical data given in Tab. 3.1 by using RCA-Seq. Let us mention that the bottom concept is not considered since there is no analysed sequence that contains such specific cpo-pattern. Moreover, each concept has as intent a cpo-pattern, and thus this hierarchy highlights how the extracted cpo-patterns relate to each other; each concept has as extent a set of sequence UIDs whose cardinality represents the support of the associated cpo-pattern. For example, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_7}}$ associated with concept CKVT_7, whose $Support(\mathcal{G}_{\texttt{CKVT\_7}}) = |extent(\texttt{CKVT\_7})| =$

4, has 1-itemsets of qualitative symptoms or viruses (e.g. $(\texttt{COUGH}_{\texttt{moderate}})$) as vertices and temporal relations (e.g. $(\texttt{Influenza}_{\texttt{A}})$ is preceded by $(\texttt{COUGH}_{\texttt{moderate}})$) as edges.

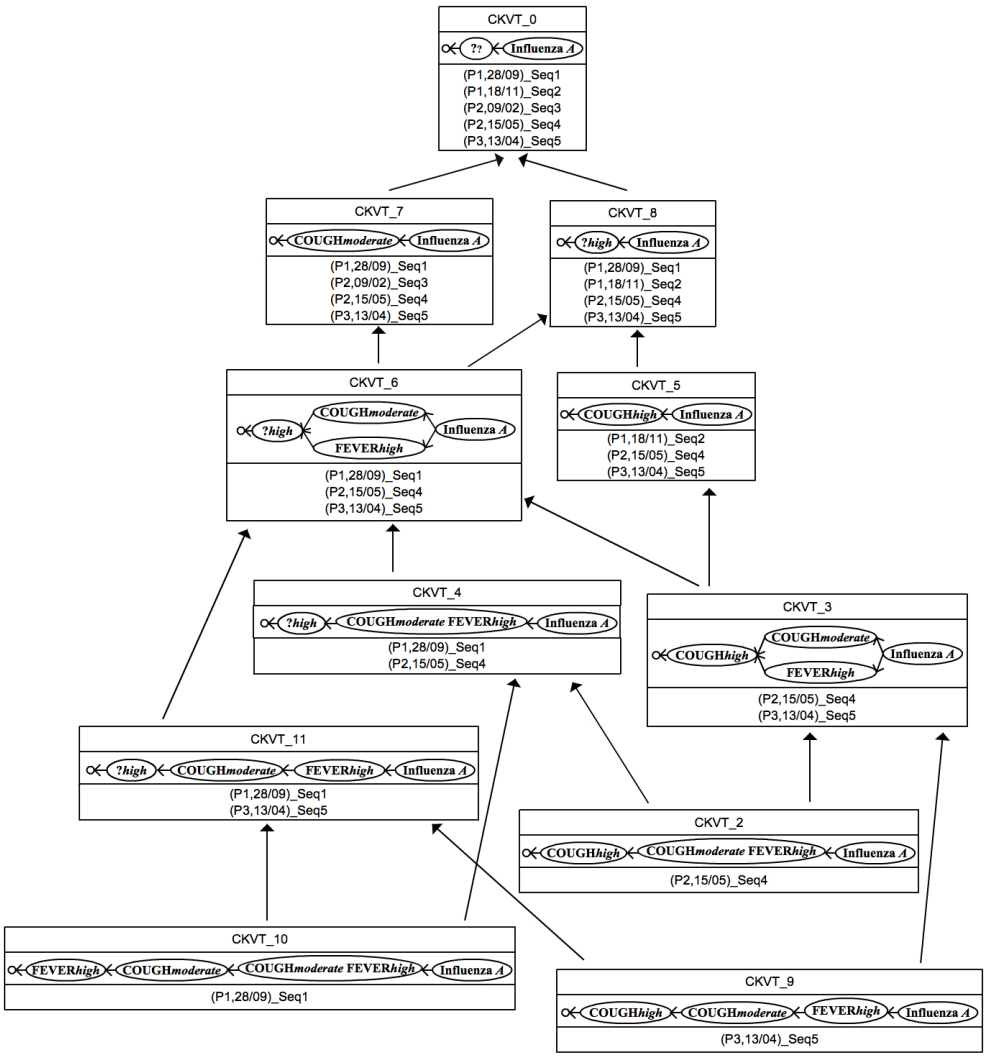


Figure 4.5: The hierarchy of multilevel cpo-patterns generated from the sequential medical data given in Tab. 3.1

The evaluation of these cpo-patterns is enhanced since physicians are guided by the relationships between the patterns. In addition, physicians can exploit two benefits of the exploration of sequential data with RCA.

Firstly, the generalisation level regarding the structure of the extracted cpo-patterns (e.g. the number of items, vertices and/or edges). For instance, the structure of the $\mathcal{G}_{\texttt{CKVT\_6}}$ cpo-pattern associated with concept $\texttt{CKVT\_6}$ is more specific than the structure of its ancestor cpo-patterns, i.e. there exists a projection from its ancestor cpo-patterns into $\mathcal{G}_{\texttt{CKVT\_6}}$. In addition, the $\mathcal{G}_{\texttt{CKVT\_4}}$ cpo-pattern associated with the $\texttt{CKVT\_4}$ concept reveals the regularity

$\{\texttt{FEVER}_{\texttt{high}}, \texttt{COUGH}_{\texttt{moderate}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$ that is a specialisation of, e.g. $\{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$ regularity revealed by $\mathcal{G}_{\texttt{CKVT\_6}}$.

Secondly, the generalisation level regarding the accuracy of items. For example, $\mathcal{G}_{\texttt{CKVT\_5}}$ associated with CKVT_5 reveals the regularity $\{\texttt{COUGH}_{\texttt{high}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$ that is a concrete specialisation of the abstract regularity $\{?_{\texttt{high}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$ revealed by $\mathcal{G}_{\texttt{CKVT\_8}}$ associated with CKVT_8.

Accordingly, when an interesting cpo-pattern is found by physicians, the evaluation can continue only by navigating amongst its descendant cpo-patterns, and thus the explored search space of patterns decreases in size. Moreover, thanks to the multilevel cpo-patterns extracted by using RCA-SEQ, abstractions of the infrequent concrete cpo-patterns can be discovered. For instance, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_3}}$ associated with concept CKVT_3 is not found for a minimum support $\theta = 3$ since $Support(\mathcal{G}_{\texttt{CKVT\_3}}) = 2 \ngeq \theta$ whereas $\mathcal{G}_{\texttt{CKVT\_6}}$ is obtained since $Support(\mathcal{G}_{\texttt{CKVT\_6}}) = 3 \geq \theta$.

However, the number of extracted cpo-patterns can be quite large, depending on the analysed sequential dataset volume and characteristics, thus complicating their evaluation and increasing the chance of overlooking interesting cpo-patterns. In our illustrative example, 5 sequences (Tab. 3.1) are analysed and by means of RCA-SEQ a hierarchy of 11 multilevel cpo-patterns is obtained. Let us note that using a minimum support $\theta = 3$ for the main lattice, we obtain only 5 multilevel cpo-patterns. In addition, there are practical cases (e.g. choosing interesting navigation paths in the hierarchy or finding global valid regularities in the analysed sequences) when the hierarchical order on cpo-patterns and the support measure are still insufficient for domain experts. For example, in Fig. 4.5, if $\mathcal{G}_{\texttt{CKVT\_7}}$ and $\mathcal{G}_{\texttt{CKVT\_8}}$ are two interesting cpo-patterns the physicians cannot decide whose descendants to analyse since both cpo-patterns have $Support(\mathcal{G}_{\texttt{CKVT\_7}}) = Support(\mathcal{G}_{\texttt{CKVT\_8}}) = 4$.

To address the aforementioned problems, in the next chapter, we propose some interestingness measures that exploit the "richness" of the RCA output obtained by exploring sequential data. Using these statistical measures, domain experts can select relevant cpo-patterns, and thus can focus only on one sub-hierarchy of cpo-patterns at a time.

4.7 Summary

In this chapter, we have devised an algorithm that automatically navigates relational conceptual structures (the RCA output) in order to extract multilevel cpo-patterns organised into a hierarchy. This algorithm relies on the structure and the properties of the RCA output.

The primary aim of our approach is to help the evaluation of the extracted set of cpo-patterns. To this end, we benefit from the fact that some cpo-patterns are naturally sub-patterns of others and we propose to extract a hierarchy of cpo-patterns where each cpo-

pattern is projected into its descendants. Consequently, when an interesting cpo-pattern is found, domain experts can continue their evaluation by focusing on the surrounding area in the hierarchy. Then, we exploit the order on items revealed by RCA and we extract multilevel cpo-patterns. Therefore, a global view of the trends of the analysed sequential data is obtained. Let us note that the CPOHrchy algorithm can be applied to any sequential data that can be modelled as depicted in Fig. 3.4.

5 Interestingness Measures for Guiding Domain Experts

Contents

5.1 Introduction

In this chapter, we present measures of interest computed by exploiting the "richness" of the RCA-Seq output that can enhance the pattern evaluation step. Firstly, we propose to cope with the "concept explosion" problem by means of a new distribution index, which makes use of the information encoded in the objects of a concept extent in order to determine the

concept relevancy. Secondly, we introduce three types of cpo-patterns that reveal "more or less" accurate information, and, besides, help in not overlooking interesting navigation paths and/or patterns in the obtained hierarchies. Finally, we present a more informative type of cpo-pattern, namely weighted cpo-pattern, that helps in better understanding the obtained pattern by capturing and explicitly showing the different roles of its itemsets in the analysed sequences.

5.2 Motivating Example

Based on the same running medical example from Sect. 3.2, let us consider that we explore the sequential data illustrated in Tab. 5.1 collected from the patients P1, P2 and P3 diagnosed as having influenza A virus. Let us suppose that physicians are interested to discover global valid cpo-patterns in the analysed sequential data, as follows:

- frequent cpo-patterns that are related to many monitored patients whose viral tests are evenly distributed amongst them;
- cpo-patterns that reveal regularities available to many analysed sequences.

Table 5.1: Illustrative sequential sub-dataset $\mathcal{D}_{\mathcal{S}fluA}$

Id	Sequence
$S1$	$\langle(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}}\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S2$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{COUGH}_{\texttt{moderate}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S3$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S4$	$\langle(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}})(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S5$	$\langle(\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S6$	$\langle(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S7$	$\langle(\texttt{FEVER}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S8$	$\langle(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S9$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$
$S10$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{A}})\rangle$

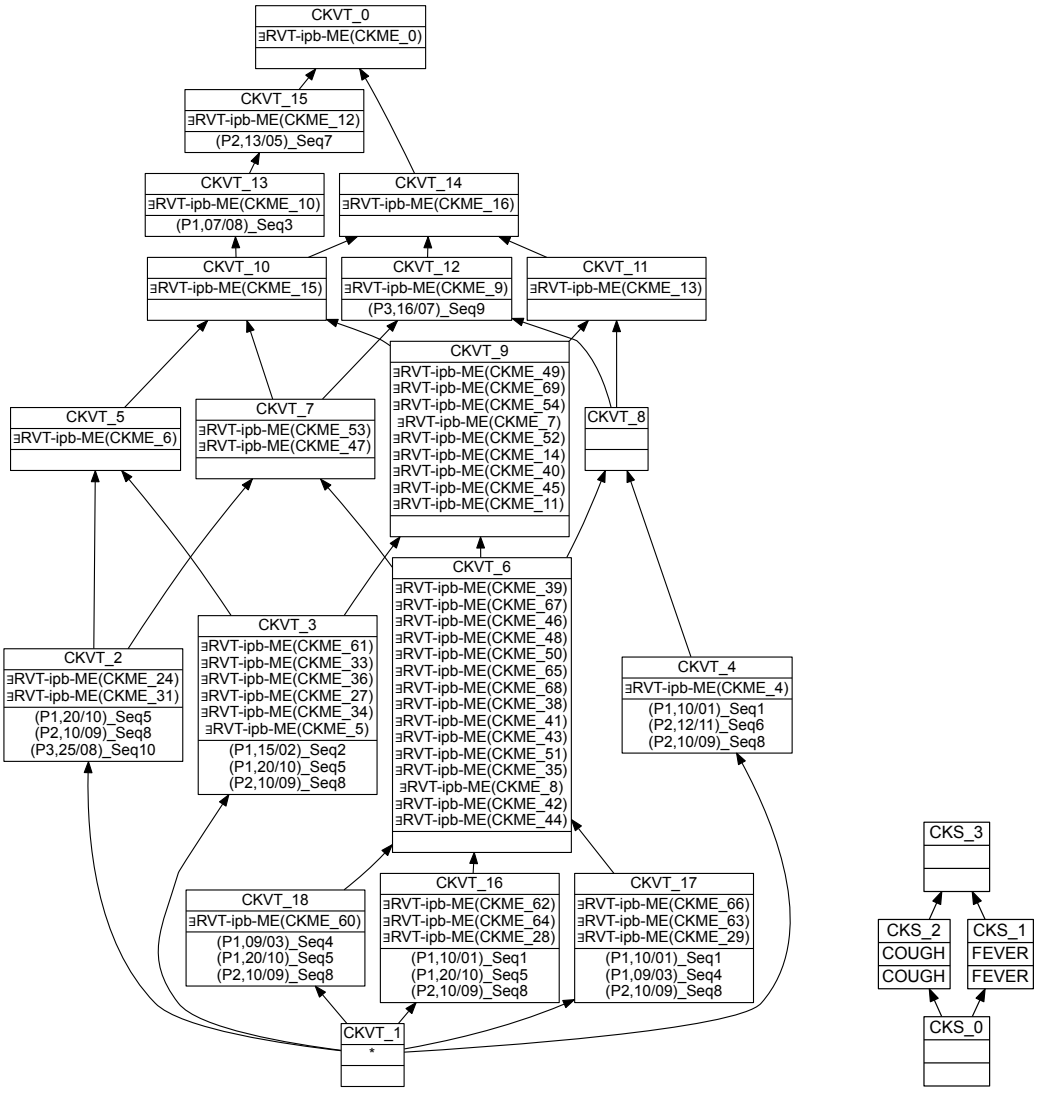
As explained in Sect. 3.4 in order to explore such data by means of RCA, firstly, we remodel the patient sequences shown in Tab. 5.1 as the sequences of UIDs given in Tab. 5.2.

Table 5.2: The patient sequences of UIDs obtained by remodelling the sequences of itemsets shown in Tab. 5.1

Sequence
⟨(P1,06/01)_Seq1 (P1,07/01)_Seq1 (P1,08/01)_Seq1 (P1,09/01)_Seq1 (P1,10/01)_Seq1⟩
⟨(P1,12/02)_Seq2 (P1,13/02)_Seq2 (P1,14/02)_Seq2 (P1,15/02)_Seq2⟩
⟨(P1,05/08)_Seq3 (P1,06/08)_Seq3 (P1,07/08)_Seq3⟩
⟨(P1,04/03)_Seq4 (P1,05/03)_Seq4 (P1,06/03)_Seq4 (P1,07/03)_Seq4 (P1,08/03)_Seq4 (P1,09/03)_Seq4⟩
⟨(P1,15/10)_Seq5 (P1,16/10)_Seq5 (P1,17/10)_Seq5 (P1,18/10)_Seq5 (P1,19/10)_Seq5 (P1,20/10)_Seq5⟩
⟨(P2,11/11)_Seq6 (P2,12/11)_Seq6⟩
⟨(P2,12/05)_Seq7 (P2,13/05)_Seq7⟩
⟨(P2,05/09)_Seq8 (P2,06/09)_Seq8 (P2,07/09)_Seq8 (P2,08/09)_Seq8 (P2,09/09)_Seq8 (P2,10/09)_Seq8⟩
⟨(P3,15/07)_Seq9 (P3,16/07)_Seq9⟩
⟨(P3,23/08)_Seq10 (P3,24/08)_Seq10 (P3,25/08)_Seq10⟩

For a sequence in Tab. 5.1, the correspondence from its itemsets to the UIDs in Tab. 5.2 is from left to right. For example, `(P1,06/01)_Seq1` from the first sequence in Tab. 5.2 is the UID of the non-target itemset $(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})$ out of $S1$ (Tab. 5.1); `(P1,10/01)_Seq1` from the same first sequence in Tab. 5.2 is the UID of the target 1-itemset $(\texttt{Influenza}_{\texttt{A}})$ out of $S1$ (Tab. 5.1).

Secondly, the corresponding RCF is built and the RCA output depicted in Fig. 5.1 and 5.2 (the simplified representations of the concept lattices) is obtained by defining $\theta = 3$ for the main lattice $\mathcal{L}_{\texttt{KVT}}$.



(a) $\mathcal{L}_{\texttt{KVT}}$ (the main lattice) (b) $\mathcal{L}_{\texttt{KS}}$

Figure 5.1: The fix point obtained for the sequential data given in Tab. 5.1: the simplified lattice of viral tests $\mathcal{L}_{\texttt{KVT}}$ and the simplified lattice of symptoms $\mathcal{L}_{\texttt{KS}}$; * is the intent of the bottom concept

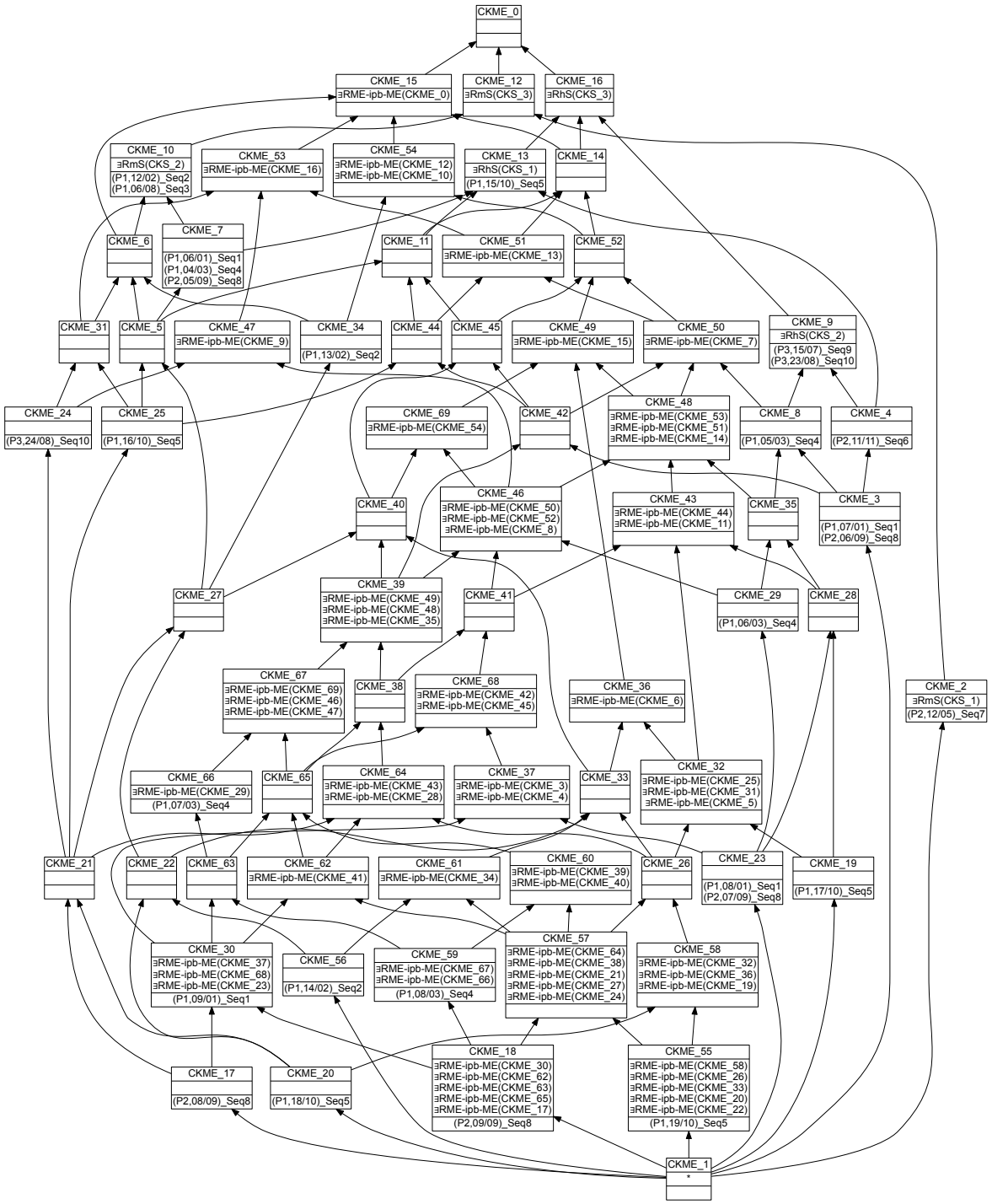
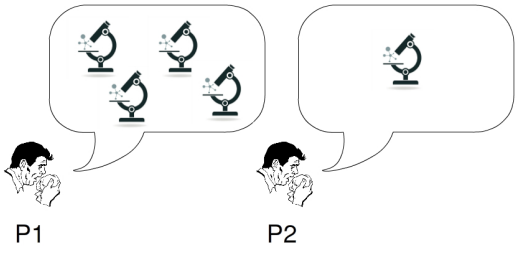


Figure 5.2: The fix point obtained for the sequential data given in Tab. 5.1: the lattice of medical examinations $\mathcal{L}_{\texttt{KME}}$ (temporal lattice); * is the intent of the bottom concept

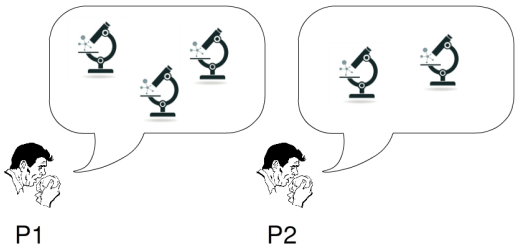
It is noted that 18 main concepts (cpo-patterns) were derived for $\theta = 3$. As already stated, if θ is decreased the number of cpo-patterns increases and their navigation becomes difficult even if these patterns are organised. Consequently, some measures of interest that can guide the evaluation step are presented in the following.

5.3 Distribution Index of a Formal Concept

A measure of interest used to select relevant concepts derived from sequential data should take into account the specificity of these concepts (indeed, the concept extents contain temporal objects) whereas well-known measures (e.g. stability index [Kuznetsov, 2007] briefly explained in Sect. 2.3.5.4) fit classical concepts. For example, Fig. 5.3 depicts the extents of two main concepts CKVT_9 and CKVT_8 (Fig. 5.1a) whose temporal objects identify viral tests (i.e. sequences).



(a) CKVT_9



(b) CKVT_8

Figure 5.3: The viral test distribution by patients for two concept extents CKVT_9 and CKVT_8 from $\mathcal{L}_{\text{KVT}}$ (Fig. 5.1a)

Both main concept extents have 5 viral tests that are done by the same patients P1 and P2, and, besides, they have the same stability index 0.38. However, if two viral tests are deleted, following the idea of stability measure, one of patient P1 and one of patient P2, then both concepts still have the same number of viral tests, but these are done by different patients (only one patient for CKVT_9 and still two patients for CKVT_8).

To highlight this difference, we introduce below an approach based on the distribution of a main concept extent that may have a better discriminant power than the stability index. In

our example, the concept distribution is different: CKVT_8 is more relevant than CKVT_9 since it better represents both monitored patients.

5.3.1 Formalisation

Let (X_m, Y_m) be a formal concept of the main lattice $\mathcal{L}_{K_M}$, then its extent X_m is a set of temporal objects – or pairs – $(Object, Date)$. If the value of $Object$ is not identical for all the pairs, then these pairs can be grouped into categories by objects. We accordingly define $\bar{X}_m$ that represents the set of distinct objects from X_m pairs: $\bar{X}_m = \{o \in O | \exists t \in T, (o,t) \in X_m\}$, where O is the set of objects and T the set of dates. In addition, we define the following measures.

Definition 5.1 (Absolute Frequency (ϕ_o)). *Let $C_M = (X_m, Y_m)$ be a main concept and* o *an object of $\bar{X}_m$. The absolute frequency of* o *in C_M, denoted ϕ_o, is equal to the number of pairs of X_m where* o *occurs.*

Definition 5.2 (Support and Richness (ρ)). *The support of a main concept (X_m, Y_m) corresponds to the number of pairs $(Object, Date)$ out of X_m. Its richness, referred to as ρ, is defined as the cardinality of $\bar{X}_m$.*

Definition 5.3 (Distribution Index (IQV)). *The distribution of a main concept (X_m, Y_m) describes the number of times each object out of $\bar{X}_m$ occurs in X_m. The Index of Qualitative Variation (IQV, [Frankfort-Nachmias and Leon-Guerrero, 2010]) is used to measure how the pairs of X_m are distributed amongst the objects. We introduce $\bar{X}_{m_\phi} = \{(o, \phi_o) \, | o \in \bar{X}_m\}$. IQV is based on the ratio of observed differences in $\bar{X}_{m_\phi}$ to the total number of possible differences within $\bar{X}_{m_\phi}$ when $\rho > 1$. If $\rho = 1$, $IQV = 0$.*

$$IQV = \frac{\rho \left(|X_m|^2 - \sum_{i=1}^{\rho} {\phi_{o_i}}^2 \right)}{|X_m|^2 \, (\rho - 1)} \tag{5.1}$$

Our choice of IQV [Frankfort-Nachmias and Leon-Guerrero, 2010] stems from the observation that the objects of $\bar{X}_m$ do not have an intrinsic ordering. IQV ranges from 0 to 1. When all pairs of X_m contain the same object, there is no diversity and $IQV = 0$. In contrast, when there are different objects and all pairs of $\bar{X}_{m_\phi}$ have equal ϕ_o, there is even distribution and $IQV = 1$.

5.3.2 Application to a Small Example

Returning to our example (Fig. 5.3):

- both main concepts have $\bar{X}_{\texttt{CKVT\_8}} = \bar{X}_{\texttt{CKVT\_9}} = \{\text{P1}, \text{P2}\}$;
- $\bar{X}_{\texttt{CKVT\_9}_\phi} = \{(\text{P1}, 4), (\text{P2}, 1)\}$;

- $\bar{X}_{\texttt{CKVT\_8}_{\phi}} = \{(\text{P1}, 3)\,, (\text{P2}, 2)\}$;
- both main concepts have the support $|X_{\texttt{CKVT\_8}}| = |X_{\texttt{CKVT\_9}}| = 5$;
- both main concepts have the richness $\rho_{\texttt{CKVT\_8}} = \rho_{\texttt{CKVT\_9}} = 2$;
- for the CKVT_8 main concept the distribution is $IQV_{\texttt{CKVT\_8}} = \frac{2\left[5^2-\left(3^2+2^2\right)\right]}{5^2(2-1)} = 0.96$ and for CKVT_9 the distribution is $IQV_{\texttt{CKVT\_9}} = \frac{2\left[5^2-\left(4^2+1^2\right)\right]}{5^2(2-1)} = 0.64$.

Hence, CKVT_8 is more relevant than CKVT_9 since its temporal objects (viral tests) are better distributed amongst the categories (patients), i.e. $IQV_{\texttt{CKVT\_8}} > IQV_{\texttt{CKVT\_9}}$.

5.4 Accuracy of a Multilevel CPO-Pattern

We recall that a multilevel cpo-pattern contains items from a poset. In Sect. 4.3.2.1, we have presented three types of items revealed by RCA-SEQ when dealing with qualitative sequential data, precisely abstract item, abstract qualitative item and concrete qualitative item. Based on the presence of such items in a multilevel cpo-pattern (except for the item of the target itemset), we introduce three types of cpo-patterns that allow us to gradually navigate from general to specific regularities without overlooking interesting ones.

Definition 5.4 (Abstract/Hybrid/Concrete CPO-Pattern). *A multilevel cpo-pattern is as follows:*

- Abstract *if it contains only abstract and/or abstract qualitative items;*
- Hybrid *if it contains both abstract and/or abstract qualitative items and concrete qualitative items;*
- Concrete *if it contains only concrete qualitative items.*

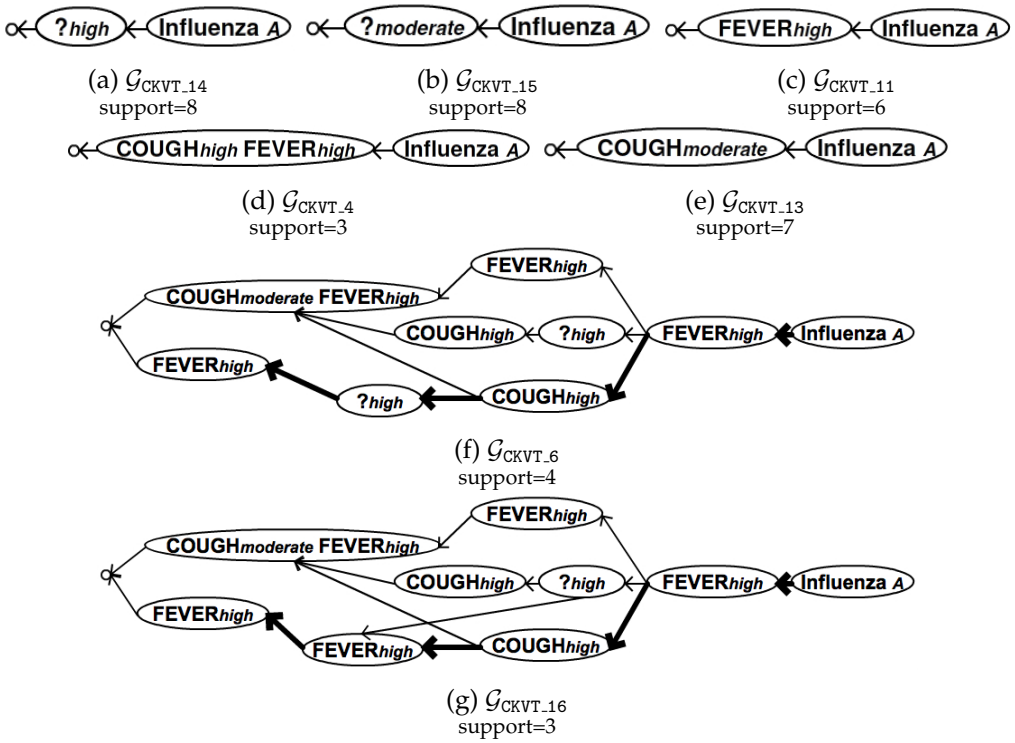
Hybrid cpo-patterns can be characterised using a measure of precision referred to as accuracy.

Definition 5.5 (Accuracy(υ)). *Let $\mathcal{I}$ be the set of items. Let $\mathcal{G}$ be a multilevel cpo-pattern and $\mathcal{I}_{\mathcal{G}}$ the multiset of items labelling the nodes of $\mathcal{G}$ ($\forall I \in \mathcal{I}_{\mathcal{G}}$, $I \in \mathcal{I}$). Let $\mathcal{I}^c_{\mathcal{G}}$ be the subset of $\mathcal{I}_{\mathcal{G}}$ containing the concrete qualitative items. The accuracy of $\mathcal{G}$ is defined as the ratio of $\mathcal{I}^c_{\mathcal{G}}$ cardinality to $\mathcal{I}_{\mathcal{G}}$ cardinality.*

$$\upsilon(\mathcal{G}) = \frac{|\mathcal{I}^c_{\mathcal{G}}|}{|\mathcal{I}_{\mathcal{G}}|} 100 \in [0\%, 100\%] \tag{5.2}$$

If $\mathcal{G}$ is abstract, $\upsilon(\mathcal{G}) = 0\%$; if $\mathcal{G}$ is concrete, $\upsilon(\mathcal{G}) = 100\%$.

To illustrate these, let us consider the multilevel cpo-patterns depicted in Fig. 5.4 associated with main concepts in $\mathcal{L}_{\texttt{KVT}}$ (Fig. 5.1a). We consider that a regularity occurs *often* if its support is greater than or equal to 7; *sometimes* if its support is between 4 and 7; *rarely* if its support is less than or equal to 4.



(a) $\mathcal{G}_{\texttt{CKVT\_14}}$ support=8

(b) $\mathcal{G}_{\texttt{CKVT\_15}}$ support=8

(c) $\mathcal{G}_{\texttt{CKVT\_11}}$ support=6

(d) $\mathcal{G}_{\texttt{CKVT\_4}}$ support=3

(e) $\mathcal{G}_{\texttt{CKVT\_13}}$ support=7

(f) $\mathcal{G}_{\texttt{CKVT\_6}}$ support=4

(g) $\mathcal{G}_{\texttt{CKVT\_16}}$ support=3

Figure 5.4: Several multilevel cpo-patterns; abstract: (a) and (b); concrete: (c), (d) and (e); hybrid: (f) and (g)

Figures 5.4a and 5.4b show two abstract cpo-patterns such that $\upsilon(\mathcal{G}_{\texttt{CKVT\_14}}) = \upsilon(\mathcal{G}_{\texttt{CKVT\_15}}) = 0\%$. For instance, $\mathcal{G}_{\texttt{CKVT\_14}}$ (Fig. 5.4a) subsumes a group of multilevel cpo-patterns that share the imprecise regularity "*often before the outbreak of influenza A virus patients feel high symptoms*" since its support is equal to 8; $\mathcal{G}_{\texttt{CKVT\_15}}$ (Fig. 5.4b) subsumes a group of multilevel cpo-patterns that share the imprecise regularity "*often before the outbreak of influenza A virus patients feel moderate symptoms*" since its support is equal to 8.

Figures 5.4c, 5.4d and 5.4e depict three concrete cpo-patterns, such that $\upsilon(\mathcal{G}_{\texttt{CKVT\_11}}) = \upsilon(\mathcal{G}_{\texttt{CKVT\_4}}) = \upsilon(\mathcal{G}_{\texttt{CKVT\_13}}) = 100\%$, that are specialisation of the aforementioned abstract cpo-patterns. For example, $\mathcal{G}_{\texttt{CKVT\_11}}$ subsumes a group of multilevel cpo-patterns that share the accurate regularity "*sometimes before the outbreak of influenza A virus patients feel high fever*" since its support is equal to 6. Besides, $\mathcal{G}_{\texttt{CKVT\_4}}$ subsumes a subgroup of the above-mentioned group, i.e. it is a specialisation of the concrete cpo-pattern $\mathcal{G}_{\texttt{CKVT\_11}}$ (Fig.5.4c). This subgroup encapsulates cpo-patterns that share the accurate regularity "*rarely before the outbreak of influenza A virus patients feel simultaneously high cough and high fever*" since its support is 3.

Figures 5.4f and 5.4g illustrate two hybrid cpo-patterns whose accuracies are: $\upsilon(\mathcal{G}_{\texttt{CKVT\_6}}) = \frac{7}{9}100 \approx 78\%$ and $\upsilon(\mathcal{G}_{\texttt{CKVT\_16}}) = \frac{8}{9}100 \approx 89\%$. The cpo-pattern $\mathcal{G}_{\texttt{CKVT\_16}}$ is a specialisation of $\mathcal{G}_{\texttt{CKVT\_6}}$. Initially, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_6}}$ can be analysed in order to gradually increase the

accuracy of the discovered regularities. Furthermore, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_16}}$ reveals that rarely ($Support(\mathcal{G}_{\texttt{CKVT\_16}}) = 3$) the concrete regularity $\{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{COUGH}_{\texttt{high}}\} \leftarrow \{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$ (highlighted path in Fig. 5.4g) is felt by patients. Besides, by analysing the highlighted path of cpo-pattern $\mathcal{G}_{\texttt{CKVT\_6}}$ (Fig. 5.4f) and taking into account that only the fever and cough symptoms are analysed, physicians can deduce that exceptionally ($|extent(\texttt{CKVT\_6})| - |extent(\texttt{CKVT\_16})| = 4-3 = 1$ viral test) patients felt the concrete regularity $\{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{COUGH}_{\texttt{high}}\} \leftarrow \{\texttt{COUGH}_{\texttt{high}}\} \leftarrow \{\texttt{FEVER}_{\texttt{high}}\} \leftarrow \{\texttt{Influenza}_{\texttt{A}}\}$. Indeed, as $\texttt{CKVT\_16} \preceq_{\texttt{KVT}} \texttt{CKVT\_6}$, the concrete highlighted path in Fig. 5.4g is supported by 3 out of the 4 sequences that support the hybrid highlighted path in Fig. 5.4f. Hence, there is only one sequence that has $\texttt{COUGH}_{\texttt{high}}$ instead of $?_{\texttt{high}}$ from the hybrid highlighted path.

5.5 Weightiness of a CPO-Pattern

The main objective of RCA-SEQ is to make easier the evaluation step of the discovered cpo-patterns in sequential data by highlighting how they relate to each other. This task is achieved by navigating only the intents of the interrelated concepts from the RCA output. Nevertheless, cpo-patterns still do not capture all the particularities hidden in the analysed sequential data. Indeed, a cpo-pattern considers only the order on itemsets in its supporting sequences, and, besides, the itemsets are treated uniformly even if their incidences differ in these sequences. In fact, previous studies showed that exploiting additional information from the analysed sequences, e.g. capturing time-intervals between adjacent itemsets in the extracted sequential patterns [Chen et al., 2003], leads to more valuable knowledge. In contrast, in this thesis, we propose to study and measure the repetitive occurrences of *preceded itemsets* in a cpo-pattern, i.e. non-target itemsets with specific predecessors. This measurement may show the non-accidental occurrence of preceded itemsets in the analysed sequences.

For example, Fig. 5.5 has at the top a set of concept intents that are navigated beginning with concept CKVT_2, which is from the main lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 5.1a), in order to extract the cpo-pattern $\mathcal{G}_{\texttt{CKVT\_2}}$ depicted at the bottom of the figure. This cpo-pattern is supported by the sequences (Tab. 5.1):

- $S5 = \langle(\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$;
- $S8 = \langle(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}}\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{A}})\rangle$;
- $S10 = \langle(\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{A}})\rangle$.

We recall that cpo-patterns preserve the order on itemsets in their supporting sequences. However, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_2}}$ is misleading since it does not encapsulate that in its supporting sequences $S5$, $S8$ and $S10$ exist only 3 occurrences of itemset $(\texttt{COUGH}_{\texttt{moderate}})$ when each occurrence is preceded by itemset $(\texttt{COUGH}_{\texttt{high}})$, while there are 4 occurrences of $(\texttt{COUGH}_{\texttt{high}})$ with no constraint on its order.

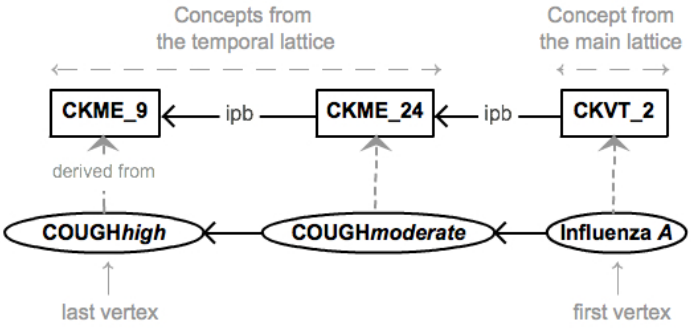


Figure 5.5: From a set of navigated concept intents to cpo-pattern $\mathcal{G}_{\texttt{CKVT\_2}}$ associated with the CKVT_2 main concept depicted in Fig. 5.1a

To address the aforementioned limitation, we propose to extract hierarchies of more informative cpo-patterns, namely *weighted cpo-patterns* (wcpo-patterns), that capture and explicitly show the different weightiness of their vertices (itemsets). In the following, we show how to capture such information by additionally navigating the interrelated concept extents.

5.5.1 From Uniform Vertices to Weighted Vertices

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, l)$ be a cpo-pattern and $\mathcal{S}_\mathcal{G}$ the set of sequences that support $\mathcal{G}$. Let $v_t \in \mathcal{V}$ be a vertex of $\mathcal{G}$, and $\mathcal{V}_t = \{v \in \mathcal{V} | v \leq v_t\}$ the set of predecessors of v_t in $\mathcal{G}$ (including v_t). Furthermore, $\mathcal{E}_t = \{(v_k, v_l) \in \mathcal{E} | v_k \in \mathcal{V}_t \text{ and } v_l \in \mathcal{V}_t\}$ is the set of edges between vertices of $\mathcal{V}_t$. $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t, l)$ associated with vertex v_t is a sub-graph of $\mathcal{G}$. $\mathcal{P}_{\mathcal{G}_t}$ is the set of paths in $\mathcal{G}_t$.

Definition 5.6 (Preceded Itemset). *Let $IS \supseteq l(v_t)$ be an itemset in a sequence $S \in \mathcal{S}_\mathcal{G}$. IS is a* preceded itemset *w.r.t. $v_t \in \mathcal{V}$, iff $\exists S_t \preceq_s S, S_t = \langle IS_1 IS_2 ... IS_p IS \rangle$ and $\forall M \in \mathcal{P}_{\mathcal{G}_t}, M \preceq_s S_t$ (i.e. there exists a subsequence of S, ending with IS, that supports $\mathcal{G}_t$).*

Our purpose is to formalise an approach for determining the weightiness of vertices (derived from concepts of the temporal lattice) that correspond to preceded itemsets. To this end, as explained in Sect. 3.4.1, let $\mathcal{D}_\mathcal{S}$ be a sequence database remodelled as $\mathcal{UID}_\mathcal{S}$ that is a database of sequences of UIDs. G_M is the set of all target itemset UIDs in $\mathcal{UID}_\mathcal{S}$, while G_T is the set of all non-target itemset UIDs.

As explained in Sect. 4.2.1, we consider two temporal relations $ipb_1 \subseteq G_M \times G_T$ and $ipb_2 \subseteq G_T \times G_T$. Let $\mathcal{L}_{K_M} = (\mathcal{C}_{K_M}, \preceq_{K_M})$ be the main lattice built from $K_M = (G_M, M_M, I_M)$. A main concept $C_M = (X_m, Y_m) \in \mathcal{C}_{K_M}$ has:

- **the intent** Y_m that consists of temporal relational attributes which are navigated to reveal cpo-pattern $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m, l_m)$ whose first vertex v_m is the one derived from C_M; v_m is labelled with a target itemset;

– **the extent** X_m that gathers all UIDs in G_M of the sequences that contain all paths in $\mathcal{G}_m$; $\mathcal{S}_{\mathcal{G}_m} = \{S \in \mathcal{D}_\mathcal{S} | \exists \texttt{Seq} \in X_m, S = getS(\texttt{Seq})\}$ is the set of sequences supporting $\mathcal{G}_m$.

We note that the range of ipb_1 temporal relation is G_T, and thus the set of vertices $\mathcal{V}_m$ contains one or more vertices v_t derived from temporal concepts and v_m vertex. Indeed, we recall that $\mathcal{L}_{K_T} = (\mathcal{C}_{K_T}, \preceq_{K_T})$ is the temporal lattice built from $K_T = (G_T, M_T, I_T)$. A temporal concept $C_T = (X_t, Y_t) \in \mathcal{C}_{K_T}$ has:

– **the intent** Y_t that may contain temporal relational attributes; Y_t is navigated to reveal $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m, l_m)$ cpo-pattern whose vertex v_t is derived from C_T; v_t vertex is labelled with itemset $l_m(v_t)$;

– **the extent** X_t gathers all UIDs in G_T that identify non-target itemsets containing itemset $l_m(v_t)$ and respect the temporal order with the UIDs pointed by the temporal relational attributes of Y_t (assuming that these exist).

We introduce $X_{t|m} = \{\texttt{IS\_Seq} \in X_t | getS(\texttt{IS\_Seq}) \in \mathcal{S}_{\mathcal{G}_m}\}$.

Proposition 5.1. *$X_{t|m}$ is the set of all UIDs that identify preceded itemsets w.r.t. $v_t \in \mathcal{V}_m$.*

Proof. Let IS be a preceded itemset w.r.t. $v_t \in \mathcal{V}_m$. Then $IS \supseteq l_m(v_t)$ and $\exists S \in \mathcal{D}_\mathcal{S}, \exists S_t \preceq_s S$ such that IS is the last itemset in S_t and S_t supports the sub-graph of v_t predecessors in $\mathcal{G}_m$, while S supports $\mathcal{G}_m$. Let Seq be the UID of S: $\texttt{Seq} \in X_m$. Furthermore, the UID of IS, i.e. IS_Seq, owns all temporal relational attributes of Y_t and is thus included in $X_{t|m}$.

Let C_T be a temporal concept revealing a vertex $v_t \in \mathcal{V}_m$. Let $\texttt{IS\_Seq} \in X_{t|m}$ be the UID of $IS = getIS(\texttt{IS\_Seq}) \supseteq l_m(v_t)$ and $S \in \mathcal{D}_\mathcal{S}$ the sequence referred by $getS(\texttt{IS\_Seq}) \in \mathcal{S}_{\mathcal{G}_m}$. $IS \in S$ and S supports cpo-pattern $\mathcal{G}_m$. We can define $S_t \preceq_s S$ the subsequence of S ending with IS. Let $\mathcal{G}_t$ be the sub-graph of v_t predecessors in $\mathcal{G}_m$: $\forall M \in \mathcal{P}_{\mathcal{G}_t}, M \preceq_s S_t$. Thus IS is a preceded itemset w.r.t. $v_t \in \mathcal{V}_m$. □

Furthermore, Y_t is navigated to extract $p < |\mathcal{C}_{K_M}| - 1$ different cpo-patterns $\mathcal{G}_m^k = (\mathcal{V}_m^k, \mathcal{E}_m^k, l_m^k)$, $k \in \{1, ..., p\}$ and X_t is the set of all UIDs that identify preceded itemsets w.r.t. $v_t^k \in \mathcal{V}_m^k$.

Definition 5.7 (Weighted CPO-Pattern (wcpo-pattern)). *Given a main concept C_M, the vertex v_m derived from C_M, the associated cpo-pattern $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m, l_m)$ and a function $w_m : (\mathcal{V}_m - \{v_m\}) \rightarrow \mathbb{R}^n_{\geq 0}$, where n is constant. A* weighted cpo-pattern *is a quadruple $(\mathcal{V}_m, \mathcal{E}_m, l_m, w_m)$ where the function w_m maps each vertex to a n-tuple of real positive numbers (vertex measures of weightiness).*

We propose three vertex measures of weightiness that represent: the *persistency* of the corresponding preceded itemset in a subset of sequences of $\mathcal{D}_\mathcal{S}$ (how many repetitions of

it are in that subset); the *overall weight* of the preceded itemset (how often it occurs) in $\mathcal{D}_\mathcal{S}$; the *specificity* of the preceded itemset in a subset of sequences of $\mathcal{D}_\mathcal{S}$ (the extent to which it belongs only to that subset).

In the following, we consider a main concept $C_M = (X_m, Y_m)$, the associated cpo-pattern $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m, l_m)$ and a vertex $v_t \in \mathcal{V}_m$ derived from a temporal concept $C_T = (X_t, Y_t)$.

Definition 5.8 (Vertex Persistency). *The* persistency *of v_t, denoted by ϖ_{v_t}, is the total number of repetitions (repetitive occurrences in the same sequence) of preceded itemsets w.r.t. v_t.*

$$\varpi_{v_t} = \frac{|X_{t|m}| - |X_m|}{|X_m|} \tag{5.3}$$

Persistency of a vertex $v_t \in \mathcal{V}_m$ measures the repetitive tendency of the corresponding preceded itemset in the subset of sequences that support $\mathcal{G}_m$. We consider that the preceded itemset characterises this subset if it is not accidental, i.e. the preceded itemset occurs repeatedly in the subset.

Definition 5.9 (Vertex Overall Weight). *The* overall weight *of v_t, denoted by ω_{v_t}, is the total number of occurrences of preceded itemsets w.r.t. $v_t^k \in \mathcal{V}_m^k$, $k \in \{1, ..., p\}$ where $p < |\mathcal{C}_{K_M}| - 1$ is the number of cpo-patterns extracted by navigating Y_t.*

$$\omega_{v_t} = |X_t| \tag{5.4}$$

Overall Weight of a vertex $v_t \in \mathcal{V}_m$ measures how numerous is the corresponding preceded itemset in all analysed sequences. Therefore, the overall weight provides an overview of the number of occurrences of the preceded itemset in the analysed dataset and it can be a reference point used in decision-making by domain experts. Using the overall weight of a vertex v_t, the *overall frequency* of v_t in $\mathcal{D}_\mathcal{S}$ can be computed by $\varphi_{v_t} = \frac{|X_t|}{|G_T|}$.

Definition 5.10 (Vertex Specificity). *The* specificity *of v_t, denoted by ς_{v_t}, is the relative number of preceded itemsets w.r.t. v_t.*

$$\varsigma_{v_t} = \frac{|X_{t|m}|}{|X_t|} 100 \in (0\%, 100\%] \tag{5.5}$$

Specificity of a vertex $v_t \in \mathcal{V}_m$ measures the extent to which the corresponding preceded itemset belongs to the subset of sequences that support $\mathcal{G}_m$. We consider that v_t is likely to be more interesting for low values of the specificity, that is, if the preceded itemset characterises the current subset, and, besides, other sequences from the analysed dataset.

Using these three measures, a vertex derived from a temporal concept can be mapped to a 3-tuple such as $(\varpi_{v_t}, \omega_{v_t}, \varsigma_{v_t})$.

5.5.2 Application to a Small Example

To illustrate how to extract a wcpo-pattern, let us examine the set of interrelated concept extents navigated to extract cpo-pattern $\mathcal{G}_{\texttt{CKVT\_17}}$ shown in Fig. 5.6, which is associated with the CKVT_17 main concept from lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 5.1a). The vertices of $\mathcal{G}_{\texttt{CKVT\_17}}$ are annotated with 3-tuples $(\varpi_{v_t}, \omega_{v_t}, \varsigma_{v_t})$.

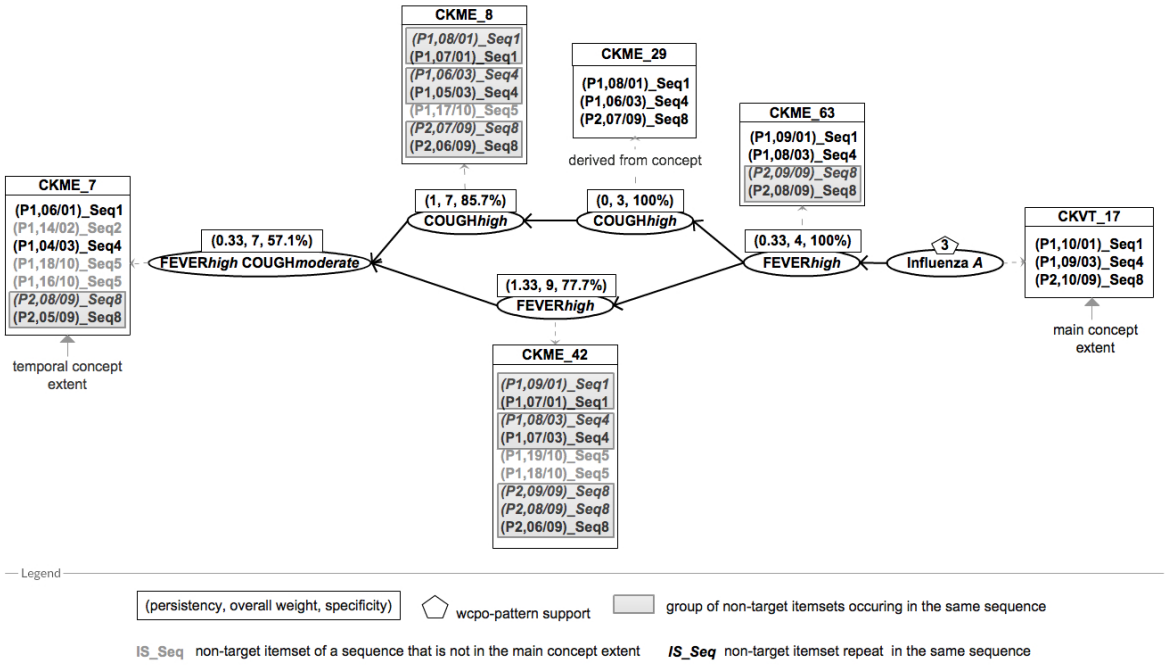


Figure 5.6: Extraction of the $\mathcal{G}_{\texttt{CKVT\_17}}$ weighted cpo-pattern associated with the CKVT_17 concept (Fig. 5.1a) by navigating concept extents

The vertex labelled with itemset $(\texttt{Influenza}_{\texttt{A}})$ is derived from the CKVT_17 main concept intent. The CKVT_17 extent comprises the sequences (each UID represents a sequence) in Tab. 5.1 that contain all paths in $\mathcal{G}_{\texttt{CKVT\_17}}$, i.e. $\mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}} = \{S1, S4, S8\}$. Thus, there are 3 distinct $(\texttt{Influenza}_{\texttt{A}})$ in Tab. 5.1 that are preceded by the itemsets $2 \times (\texttt{FEVER}_{\texttt{high}})$, $2 \times (\texttt{COUGH}_{\texttt{high}})$, and $(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}})$ in the order they appear in $\mathcal{G}_{\texttt{CKVT\_17}}$.

The vertex derived from the CKME_63 temporal concept intent is labelled with the preceded itemset $(\texttt{FEVER}_{\texttt{high}})$ and it is denoted by $v_{\texttt{CKME\_63}}$. The CKME_63 extent gathers the 4 non-target itemsets (each UID represents a non-target itemset) in Tab. 5.1 that contain $(\texttt{FEVER}_{\texttt{high}})$ and that are preceded by the itemsets $(\texttt{FEVER}_{\texttt{high}})$, $2\times(\texttt{COUGH}_{\texttt{high}})$ and $(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}})$ in the order they appear in $\mathcal{G}_{\texttt{CKVT\_17}}$. Therefore, the overall weight of $v_{\texttt{CKME\_63}}$ is $\omega_{v_{\texttt{CKME\_63}}} = 4$. Since all the itemsets in the CKME_63 extent are owned by the sequences in $\mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}$, the $v_{\texttt{CKME\_63}}$ specificity is $\varsigma_{v_{\texttt{CKME\_63}}} = \frac{4}{4}100 = 100\%$. In addition, extent CKME_63 contains a group of two non-target itemsets (P2,09/09)_Seq8 and (P2,08/09)_Seq8 (gray rectangle in Fig. 5.6)

that occur in the same sequence $getS((\texttt{P2,09/09})\texttt{\_Seq8}) = getS((\texttt{P2,08/09})\texttt{\_Seq8}) = S8 \in \mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}$. Then, extent CKME_63 contains one repetition identified by (P2,09/09)_Seq8 (italic font in Fig. 5.6), and thus the $v_{\texttt{CKME\_63}}$ persistency is $\varpi_{v_{\texttt{CKME\_63}}} = \frac{4-3}{3} = 0.33$.

Vertex $v_{\texttt{CKME\_42}}$ labelled with the preceded itemset ($\texttt{FEVER}_{\texttt{high}}$) is derived from the CKME_42 temporal concept intent and has the overall weight $\omega_{v_{\texttt{CKME\_42}}} = 9$. Indeed, the CKME_42 extent comprises the 9 non-target itemsets in Tab. 5.1 that contain itemset ($\texttt{FEVER}_{\texttt{high}}$) and that are preceded by itemset ($\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}}$). The two non-target itemsets (P1,19/10)_Seq5 and (P1,18/10)_Seq5 (gray font in Fig. 5.6) are owned by $S5 \notin \mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}$, and thus the $v_{\texttt{CKME\_42}}$ specificity is $\varsigma_{v_{\texttt{CKME\_42}}} = \frac{7}{9}100 = 77.7\%$. Since there are three groups of non-target itemsets that contain a total of 4 repetitions the persistency of $v_{\texttt{CKME\_42}}$ is $\varpi_{v_{\texttt{CKME\_42}}} = \frac{7-3}{3} = 1.33$.

Vertex $v_{\texttt{CKME\_29}}$ labelled with the preceded itemset ($\texttt{COUGH}_{\texttt{high}}$) is derived from the CKME_29 temporal concept intent. The CKME_29 concept extent gathers the 3 non-target itemsets in Tab. 5.1 that contain itemset ($\texttt{COUGH}_{\texttt{high}}$) and that are preceded by the itemsets ($\texttt{COUGH}_{\texttt{high}}$) and ($\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}}$) in the order they appear in $\mathcal{G}_{\texttt{CKVT\_17}}$. Thus, the overall weight of $v_{\texttt{CKME\_29}}$ is $\omega_{v_{\texttt{CKME\_29}}} = 3$ and its specificity is $\varsigma_{v_{\texttt{CKME\_29}}} = \frac{3}{3}100 = 100\%$. In addition, the persistency of $v_{\texttt{CKME\_29}}$ is $\varpi_{v_{\texttt{CKME\_29}}} = \frac{3-3}{3} = 0$ since there is no repetition.

Vertex $v_{\texttt{CKME\_8}}$ labelled with the preceded itemset ($\texttt{COUGH}_{\texttt{high}}$) is derived from the CKME_8 temporal concept intent and has the overall weight $\omega_{v_{\texttt{CKME\_8}}} = 7$. Indeed, the CKME_8 extent comprises the 7 non-target itemsets in Tab. 5.1 that contain itemset ($\texttt{COUGH}_{\texttt{high}}$) and that are preceded by the itemset ($\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}}$). The non-target itemset identified by (P1,17/10)_Seq5 is owned by $getS((\texttt{P1,17/10})\texttt{\_Seq5}) = S5 \notin \mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}$, and thus the $v_{\texttt{CKME\_8}}$ specificity is $\varsigma_{v_{\texttt{CKME\_8}}} = \frac{6}{7}100 = 85.7\%$. Since there are three groups of non-target itemsets that contain a total of 3 repetitions the persistency of $v_{\texttt{CKME\_8}}$ is $\varpi_{v_{\texttt{CKME\_8}}} = \frac{6-3}{3} = 1$.

Finally, vertex $v_{\texttt{CKME\_7}}$ labelled with the preceded itemset ($\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}}$) is derived from the CKME_7 temporal concept intent and has the overall weight $\omega_{v_{\texttt{CKME\_7}}} = 7$. Indeed, the CKME_7 concept extent comprises the 7 non-target itemsets in Tab. 5.1 that contain itemset ($\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{moderate}}$). Note that, since $v_{\texttt{CKME\_7}}$ vertex is not preceded by other vertices in cpo-pattern $\mathcal{G}_{\texttt{CKVT\_17}}$, there is no constraint on the order of the preceded itemset in the analysed sequences. The $v_{\texttt{CKME\_7}}$ specificity is $\varsigma_{v_{\texttt{CKME\_7}}} = \frac{4}{7}100 = 57.1\%$ since there are 3 non-target itemsets (gray font in Fig. 5.6) that are not owned by the sequences in $\mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}$. The persistency of $v_{\texttt{CKME\_7}}$ is $\varpi_{v_{\texttt{CKME\_7}}} = \frac{4-3}{3} = 0.33$ since the CKME_7 extent contains only one repetition (italic font in Fig. 5.6).

5.5.3 Enhancing Sequential Data Analysis Using Weighted CPO-Patterns

We recall that by using RCA-Seq as explained in Chapter 4, hierarchies of cpo-patterns are obtained in order to help the evaluation step by highlighting how the extracted patterns relate to each other. However, we can assume practical cases (discussed in the following) when the

order on the extracted cpo-patterns is insufficient for domain experts. To cope with these problems, we propose to use hierarchies of wcpo-patterns and to exploit the vertex measures of weightiness introduced in Sect. 5.5.1. It is worth mentioning that the persistency, overall weight and specificity of a vertex can be considered simultaneously or not depending on the motivation behind the analysis step.

Henceforth, we use the motivating example (Sect. 5.2) to illustrate three practical cases that take advantage of wcpo-patterns when physicians try to interpret the extracted medical knowledge. As these examples demonstrate, the wcpo-patterns can lead to more informative knowledge since the different importance of vertices or paths are considered.

5.5.3.1 Ranking the Vertices and Paths of a CPO-Pattern

In a cpo-pattern the vertices/paths are considered uniformly. Then, domain experts can easily be misled into thinking that all vertices/paths in a cpo-pattern have the same impact on the item of interest. For instance, let us suppose that physicians try to interpret the vertices $v_{\texttt{CKME\_7}}$, $v_{\texttt{CKME\_8}}$ and $v_{\texttt{CKME\_42}}$ of cpo-pattern $\mathcal{G}_{\texttt{CKVT\_17}}$ (Fig. 5.6) by disregarding the weightiness of vertices. Physicians find that before the outbreak of influenza A virus the patients feel high cough and high fever in any order, but after feeling simultaneously high fever and moderate cough. Since only 3 out of 10 analysed sequences support $\mathcal{G}_{\texttt{CKVT\_17}}$, physicians can infer with low confidence that:

- *"rarely the simultaneous occurrence of high fever and moderate cough can be considered as a premature sign of a possible influenza A outbreak"*;
- *"rarely high fever and high cough can be considered as early signs of influenza A outbreak"*.

However, by paying attention to the weightiness of vertices shown in Fig. 5.6, physicians discover that:

- there are $\omega_{v_{\texttt{CKME\_7}}} = 7$ simultaneous occurrences of cough moderate and high fever in the analysed dataset; $\varsigma_{v_{\texttt{CKME\_7}}} = 57.1\%$ of these occurrences are specific to only a subset of sequences ($|\mathcal{S}_{\mathcal{G}_{\texttt{CKVT\_17}}}| = 3$ patient sequences), and, besides, 42.9% are specific to other analysed sequences. Therefore, the inference *"the simultaneous occurrence of high fever and moderate cough can be a premature sign of influenza A outbreak"* may be globally valid in the analysed sequences; since the simultaneous occurrence of moderate cough and high fever in the subset of sequences is not too persistent $\varpi_{v_{\texttt{CKME\_7}}} = 0.33$, physicians can further examine, e.g. if these symptoms may be caused by a bacterial infection;
- there are $\omega_{v_{\texttt{CKME\_8}}} = 7$ occurrences of high cough preceded simultaneously by high fever and moderate cough in the analysed dataset; $\varsigma_{v_{\texttt{CKME\_8}}} = 85.7\%$ of these occurrences are specific to only the aforementioned subset of sequences. In addition, the occurrence of high cough (the $v_{\texttt{CKME\_8}}$ label) in this subset of sequences is persistent $\varpi_{v_{\texttt{CKME\_8}}} = 1$ and

the inference *"high cough can be an early sign of influenza A outbreak"* seems to be valid at least for this subset of sequences;

– in the analysed dataset there are more occurrences of the $(\texttt{FEVER}_{\texttt{high}})$ preceded itemset rather than the $(\texttt{COUGH}_{\texttt{high}})$ preceded itemset, i.e. $\omega_{v_{\texttt{CKME\_42}}} > \omega_{v_{\texttt{CKME\_8}}}$. Then, the $(\texttt{FEVER}_{\texttt{high}})$ preceded itemset is $\varsigma_{v_{\texttt{CKME\_42}}} = 77.7\%$ specific only to this subset of sequences and 22.3% specific to other analysed sequences, while the $(\texttt{COUGH}_{\texttt{high}})$ preceded itemset is only 14.3% specific to other sequences. Therefore, the inference *"high fever can be an early sign of influenza A outbreak"* seems to be valid for this subset of sequences as well as valid for other analysed sequences. Moreover, the occurrence of high fever (the $v_{\texttt{CKME\_42}}$ label) in this subset of sequences is more persistent, i.e. $\varpi_{v_{\texttt{CKME\_42}}} > \varpi_{v_{\texttt{CKME\_8}}}$. Hence, physicians can rank the paths and can infer that regularity $\{\texttt{FEVER}_{\texttt{high}}, \texttt{COUGH}_{\texttt{moderate}}\} \leftarrow \{\texttt{FEVER}_{\texttt{high}}\}$ is more pertinent to recognise influenza A outbreak.

5.5.3.2 Selecting Interesting Navigation Paths in a Hierarchy of CPO-Patterns

Usually the extracted hierarchies of cpo-patterns are very large and even if the relationships between cpo-patterns are highlighted and the support measure can be considered, their navigation is still not an easy task for domain experts. For instance, let us suppose that physicians try to navigate the hierarchy of cpo-patterns shown in Fig. 5.7 while ignoring the weightiness of vertices. This figure depicts an excerpt (with six cpo-patterns from (a) to (f)) from the hierarchy of wcpo-patterns obtained by exploring the sequential data illustrated in Tab. 5.1.

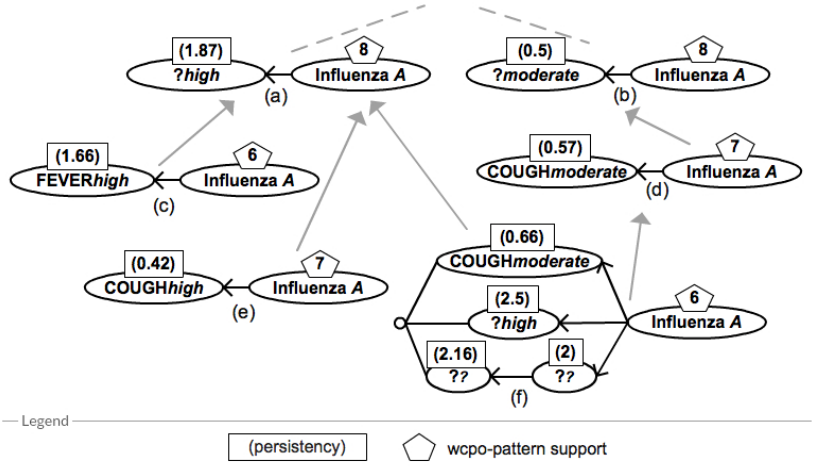


Figure 5.7: Excerpt from the hierarchy of wcpo-patterns obtained by exploring the sequential data illustrated in Tab. 5.1

Physicians can begin the navigation from the abstract cpo-patterns (a) and (b). Thus, physicians have an overview of the discovered regularities in the analysed sequences that

minimises the chance of overlooking interesting cpo-patterns. It is noted that both cpo-patterns are discovered with the same $Support = 8$ ($\pentagon$ on the first vertex) and apparently they mark out two interesting navigation paths in the hierarchy. Nevertheless, when physicians consider the persistency of vertices, their different importance is highlighted. Physicians can easily infer that the high symptoms are more probable to be signs of influenza A outbreak, i.e. the high symptoms are more persistent ($\varpi_{(?_{\texttt{high}})} = 1.87$) than the moderate symptoms ($\varpi_{(?_{\texttt{moderate}})} = 0.5$). Accordingly, physicians select the navigation path that consists in the descendant wcpo-patterns of (a) and the evaluation continues by applying the same ranking criterion.

5.5.3.3 Distinguishing the Best Sub-Dataset Supporting a CPO-Pattern

There are cases when it is useful to find out discriminant regularities for different types of the studied item of interest. Fabrègue et al. [2014] presented an approach that captures discriminant regularities for different ecological states of the aquatic ecosystem. Here, in our motivating example, we can suppose that physicians are interested in distinguishing between the outbreaks of influenza A and B by assessing the symptoms felt by patients. Usually, physicians determine that the same extracted cpo-pattern belongs rather to sub-dataset $\mathcal{D}_{\mathcal{S}fluA}$ (Tab. 5.1) or to sub-dataset $\mathcal{D}_{\mathcal{S}fluB}$ (Tab. 5.3) by relying on the support measure (or its variants, e.g. growth rate [Dong and Li, 1999]). However, there are cases when a cpo-pattern is found with equal support in both sub-datasets.

Table 5.3: Illustrative sub-dataset $\mathcal{D}_{\mathcal{S}fluB}$

Id	Sequence
$S1$	$\langle(\texttt{COUGH}_{\texttt{high}}\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S2$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S3$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S4$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{FEVER}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S5$	$\langle(\texttt{FEVER}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S6$	$\langle(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S7$	$\langle(\texttt{FEVER}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S8$	$\langle(\texttt{FEVER}_{\texttt{high}}\texttt{COUGH}_{\texttt{high}})(\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S9$	$\langle(\texttt{COUGH}_{\texttt{high}})(\texttt{Influenza}_{\texttt{B}})\rangle$
$S10$	$\langle(\texttt{COUGH}_{\texttt{moderate}})(\texttt{Influenza}_{\texttt{B}})\rangle$

For example, let us consider that physicians try to understand if the regularity given in Fig. 5.8 helps to recognise the influenza A or influenza B outbreak. Both cpo-patterns are discovered with the same $Suppport = 5$ ($\pentagon$ on the first vertex), and thus it is impossible to distinguish between them by disregarding the weightiness of vertices. In contrast, when physicians consider, for instance, the persistencies of vertices it is easily noted that the high cough and the high fever are more persistent in wcpo-pattern (a). Accordingly, physicians

can conclude that the regularity given in Fig. 5.8 is a distinguishing characteristic of the influenza A outbreak since both vertices are more significant. Moreover, the same inference is drawn by additionally considering the overall weights of the vertices.

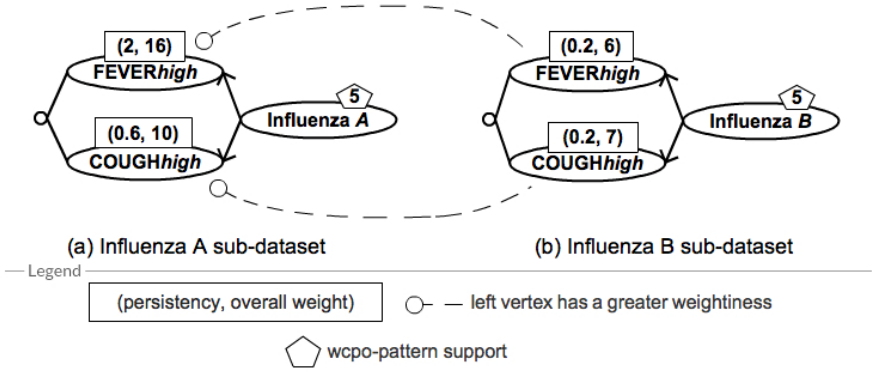


Figure 5.8: Distinguishing between the outbreaks of influenza A and B

5.6 Summary

In this chapter, we have introduced measures of interest for enhancing sequential data analysis. Mainly, we have tried to leverage the navigated concept extents rather than only the concept intents (Chapter 4). We have shown how to select the well-distributed formal concepts derived from a sequence database gathering sets of sequences, each set being associated with a distinct category (e.g. patient). Then, we have exploited the different types of items extracted by using RCA-SEQ and we have proposed three types of cpo-patterns, i.e. concrete, hybrid and abstract. Moreover, we have proposed to extract more informative patterns, precisely wcpo-patterns, that capture and explicitly show not only the order on itemsets (as standard cpo-patterns do) but also their different roles in the analysed sequences through three statistical measures, i.e. persistency, specificity and overall weight.

6 Study of the RCA-Seq Approach Adaptability

Contents

6.1 Introduction

In this chapter, we present four extensions of the RCA-Seq approach. Firstly, we consider user-defined constraints on the order relations on itemsets. Secondly, we explain how to integrate a user-defined taxonomy over sequence-building items. Then, we show how to explore simple sequential data (i.e. sequences are built from items without qualitative values). Lastly, we explain how to explore heterogeneous sequential data.

6.2 Extraction of CPO-Patterns with User-Defined Constraints on the Order Relations on Itemsets

In the RCA-SEQ approach the order on itemsets in a cpo-pattern is revealed by the relational attributes from the navigated concept intents; these relational attributes are built using the existential scaling mechanism in order to capture all the relations between the analysed itemsets. For instance, relying on the RCA output (Fig. 3.5) from the running example, in Fig. 6.1 there is a temporal link between the CKVT_7 main concept and the CKME_8 temporal concept (highlighted by the ∃RVT-ipb-ME(CKME_8) temporal relational attribute of the CKVT_7 intent) since each viral test in the CKVT_7 extent is preceded by at least one medical examination in the CKME_8 extent. Indeed, the viral tests identified by (P2,15/05)_Seq4, (P2,09/02)_Seq3 and (P3,13/04)_Seq5 are preceded respectively by 1 medical examination from the CKME_8 extent; the viral test identified by (P1,28/09)_Seq1 is preceded by 2 medical examinations. Therefore, the CKVT_7 extent gathers all the UIDs of viral tests (from the analysed data shown in Tab. 3.1) that are preceded respectively by at least one medical examination when the patient experienced a moderate cough.

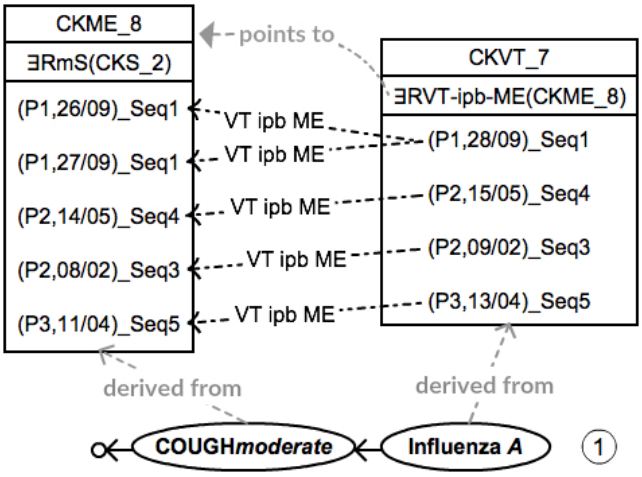


Figure 6.1: The ∃ quantifier: the concepts navigated to extract cpo-pattern $\mathcal{G}_{\texttt{CKVT\_7}}$ (①) associated with the CKVT_7 main concept from lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 3.5a). The intents contain only the relational attributes according to Properties 4.4 and 4.5

However, physicians can be interested in finding out cpo-patterns that are available for patients that frequently experience certain symptoms before a viral test. For example, physicians can look for the viral tests that are preceded by more than 50% of the associated medical examinations (from the sequences that end with these viral tests) for which the intensity of cough symptom is moderate. Using RCA-SEQ, this type of cpo-pattern can be discovered by only changing the quantifier applied to the relations encoded in the temporal relational

contexts during the iterative steps. To add the constraint *"a viral test is preceded by more than* 50% *of the associated medical examinations"*, we use the $\exists$ quantifier with a user-defined cardinality, denoted by $\exists_{>n\%}$ [Rouane-Hacene et al., 2013] where $n = 50$, that is applied only to the relations encoded in the `RVT-ipb-ME` relational context. Formally, a relational attribute $\exists_{>n\%} r(C)$, where r is a relation and $C = (X, Y)$ is a concept whose extent contains objects from $ran(r)$, describes an object $g \in dom(r)$ if $r(g) \cap X \neq \emptyset$ and $|r(g) \cap X| > \frac{n \times |r(g)|}{100}$.

To illustrate this, we apply again RCA to the RCF depicted in Tab. 3.4 by changing the $\exists$ to $\exists_{>50\%}$ quantifier only for the `RVT-ipb-ME` temporal relational context (the $\exists$ quantifier is preserved for `RME-ipb-ME`). The obtained RCA output is the same as that from Fig. 3.5 except for the $\mathcal{L}_{\texttt{KVT}}$ main lattice that has the new structure shown in Fig. 6.2. It is noted that the number of extracted cpo-patterns (associated with main concepts) is smaller, i.e. 7 cpo-patterns in comparison to 11 initial cpo-patterns (there is no extracted cpo-pattern for $\perp(\mathcal{L}_{\texttt{KVT}})$), since the criterion imposed by physicians is more restrictive. Therefore, the evaluation of the hierarchy of cpo-patterns is facilitated thanks to the smaller number of obtained cpo-patterns.

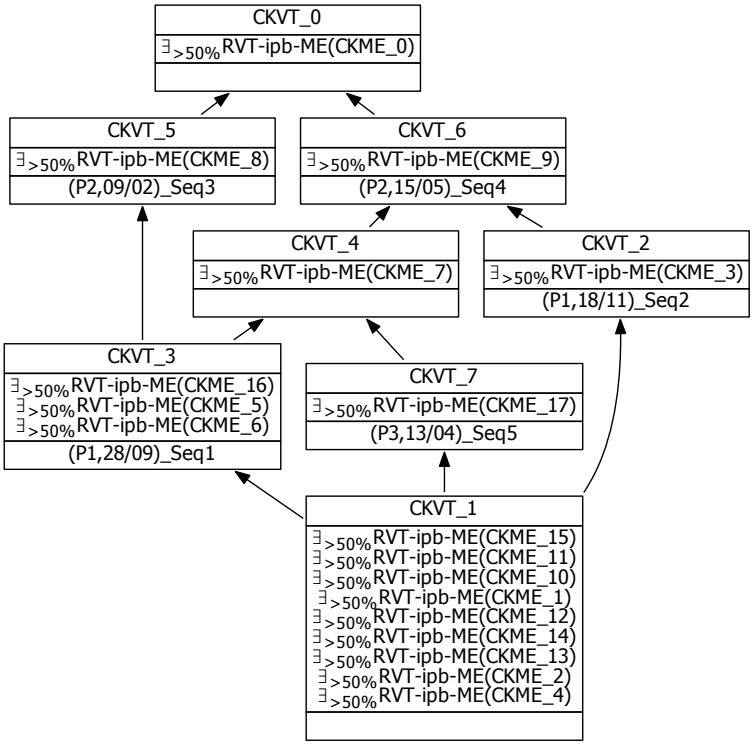


Figure 6.2: The $\mathcal{L}_{\texttt{KVT}}$ main lattice of viral tests obtained by scaling the temporal links between viral tests and medical examinations (shown in Tab. 3.4) using the $\exists_{>50\%}$ quantifier

In addition, it is noted that cpo-pattern $\mathcal{G}_{\texttt{CKVT\_7}}$ (①, Fig. 6.1) is discovered as well when using the $\exists_{>50\%}$ quantifier, precisely it is associated with the `CKVT_5` main concept in Fig. 6.2.

However, cpo-pattern $\mathcal{G}_{\texttt{CKVT\_5}}$ (①, Fig. 6.3) is less frequent than $\mathcal{G}_{\texttt{CKVT\_7}}$ ($Support(\mathcal{G}_{\texttt{CKVT\_7}}) = 4$) since there are only $Support(\texttt{CKVT\_5}) = 2$ viral tests that have more than 50% of the associated medical examinations for which the intensity of cough is moderate. Indeed, by analysing Tab. 3.1 and 3.3, the `(P1,28/09)_Seq1` viral test has 3 associated medical examinations and $2 > (50\%\ of\ 3)$ of them are gathered in the `CKME_8` concept intent (Fig. 6.3). Similarly, the `(P1,09/02)_Seq3` viral test has one associated medical examination that is gathered in the `CKME_8` intent (Fig. 6.3). In contrast, the `(P3,13/04)_Seq5` viral test (Fig. 6.1) has 3 associated medical examinations but only $1 \not> (50\%\ of\ 3)$ of them is described by the moderate cough symptom, and thus this viral test is not in the `CKVT_5` concept extent (Fig. 6.3).

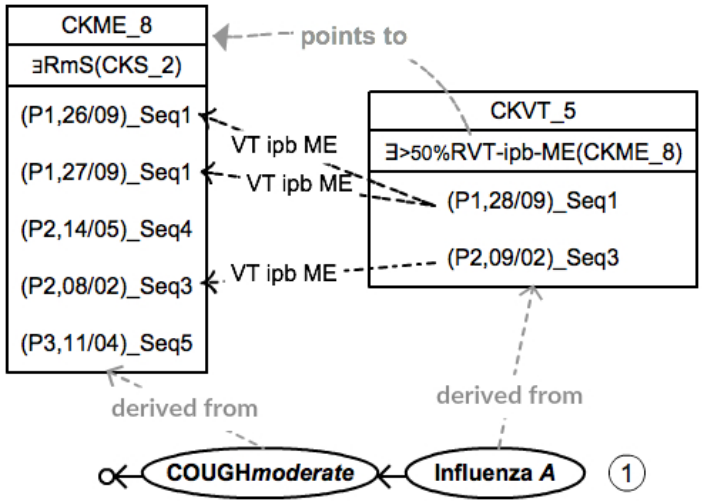
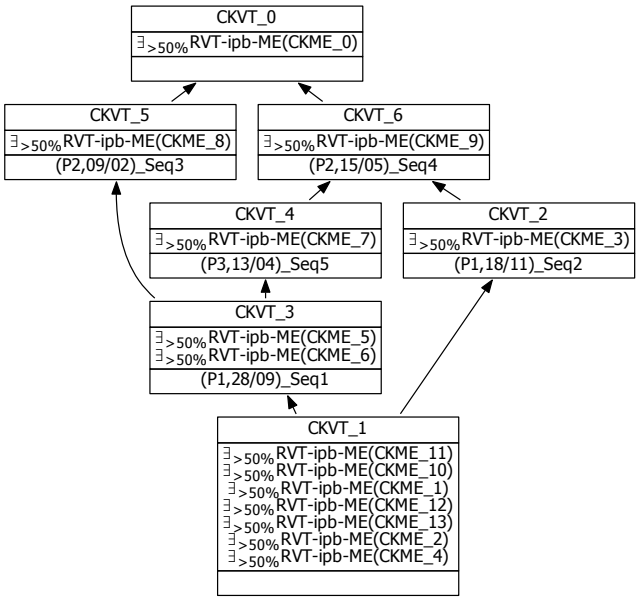


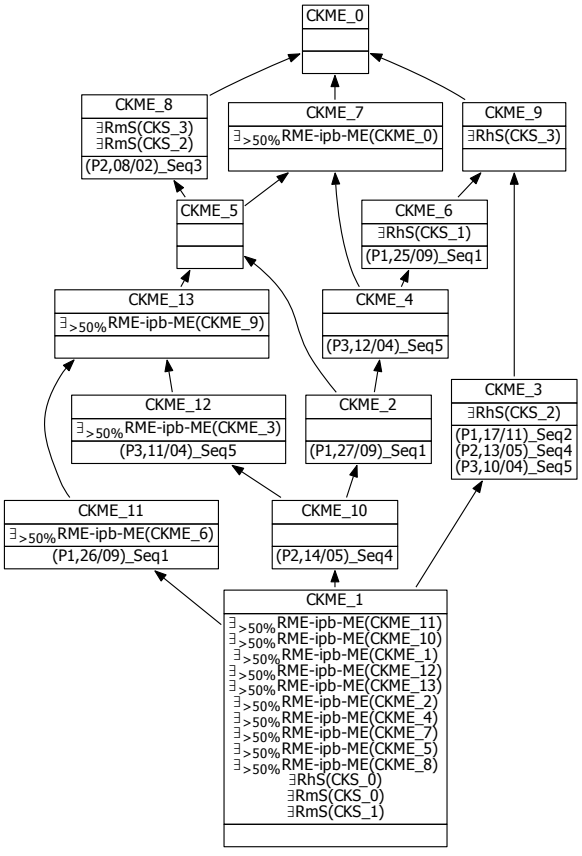
Figure 6.3: The $\exists_{>50\%}$ quantifier: the concepts navigated to extract cpo-pattern $\mathcal{G}_{\texttt{CKVT\_5}}$ (①) associated with the `CKVT_5` main concept from lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 6.2). The intents contain only the relational attributes according to Properties 4.4 and 4.5

Such constraint can be used on the temporal relations between medical examinations as well. To this end, we also change quantifier $\exists$ to $\exists_{>50\%}$ for the `RME-ipb-ME` temporal relational context and the RCA output contains the two lattices depicted in Fig. 6.4 and the lattice of symptoms shown in Fig. 3.5b. By analysing Fig. 6.4b, it is noted that the number of temporal concepts has decreased, i.e. 14 concepts in comparison to 18 concepts (Fig. 3.5c); in Fig. 6.4a the number of extracted cpo-patterns (associated with main concepts) has decreased as well, i.e. 6 cpo-patterns in comparison with 11 (Fig. 3.5a) or 7 (Fig. 6.2) cpo-patterns.

Let us mention that depending on the motivation behind the analysis, the various quantifiers presented in Rouane-Hacene et al. [2013] and their variants can be applied in the same way to any type of relation (e.g. qualitative, temporal).



(a) $\mathcal{L}_{\texttt{KVT}}$



(b) $\mathcal{L}_{\texttt{KME}}$

Figure 6.4: The $\mathcal{L}_{\texttt{KVT}}$ main lattice of viral tests and the $\mathcal{L}_{\texttt{KME}}$ lattice of medical examinations obtained by scaling the temporal links using the $\exists_{>50\%}$ quantifier

6.3 RCA-SEQ with a User-Defined Taxonomy Over the Items

RCA-SEQ reveals a taxonomy over sequence-building items due to the nominal scaling applied to encode these items into an RCF (the RCA input). This taxonomy has only two levels: first, the level comprising each atomic item and second, the level with the general item, i.e. the item that represents the set of items used to build the analysed sequences. Accordingly, the extracted multilevel cpo-patterns contain only items from these two levels.

Srikant and Agrawal [1996] proposed to integrate a user-defined taxonomy over the items in order to extract sequential patterns (rather than cpo-patterns) containing items across different levels of the taxonomy. Their method is applied to sequences whose items are atomic values. To this end, they preprocess each sequence from the database to obtain an "extended-sequence", i.e. the sequence is upgraded with the ancestors (from the taxonomy) of each item in the sequence. Thus, their algorithm GSP explores sequences that already contain the relationships between the items and their ancestors.

In contrast, RCA-SEQ can easily integrate a user-defined taxonomy in the RCF, and, besides, can extract directly organised cpo-patterns that contain items from different levels of the taxonomy without preprocessing the analysed sequences. To illustrate this, we consider the user-defined taxonomy over symptoms depicted in Fig. 6.5 (i.e. the symptoms can be described at different levels of precision) and the small sequential dataset shown in Tab. 6.1. It is worth noting that the analysed patient sequences consist in only atomic items, namely `DRY COUGH` (`DC`), `WET COUGH` (`WC`) and `FEVER` (`F`).

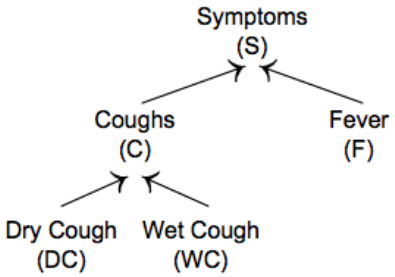


Figure 6.5: A taxonomy over the symptoms felt by patients

To explore the medical data shown in Tab. 6.1 using the taxonomy over symptoms, we follow the steps presented in Sect. 3.4.1 to build the RCF shown in Tab. 6.2. The only difference is the *ordinal scaling* [Ganter and Wille, 1999], instead of nominal scaling, used to build the `KS` formal context of symptoms. It is worthwhile to mention that the qualitative relational contexts `RmS` (has moderate symptom) and `RhS` (has high symptom) illustrated in Tab. 6.2 encode strictly the relations between the medical examinations and the specific symptoms (atomic items) as given in the analysed medical data (Tab. 6.1). For instance, the relational

context RhS encodes that on September 17th patient P1 experienced dry cough (DC) with high intensity and no extra information regarding the ancestors of dry cough (e.g. coughs (C) in Fig. 6.5).

Table 6.1: Illustrative example of medical data with atomic items from a user-defined taxonomy

Patient	Date	Medical Examination			Viral Test
		DC	WC	F	Influenza
P1	17/09	high	–	moderate	–
	19/09	–	high	–	–
	20/09	–	–	–	A
P2	15/05	–	high	moderate	–
	16/05	–	–	high	–
	17/05	–	–	–	A
P3	08/04	high	high	–	–
	09/04	–	–	–	A

Table 6.2: RCF that encodes the medical data shown in Tab. 6.1; formal contexts: KS, KVT and KME; temporal relational contexts: RME-ipb-ME and RVT-ipb-ME; qualitative relational contexts: RmS and RhS

KS	**S**	**C**	**F**	**DC**	**WC**
S	×				
C	×	×			
F	×		×		
DC	×	×		×	
WC	×	×			×

KVT
(P1,20/09)
(P2,17/05)
(P3,09/04)

KME
(P1,17/09)
(P1,19/09)
(P2,15/05)
(P2,16/05)
(P3,08/04)

RME-ipb-ME	**(P1,17/09)**	**(P1,19/09)**	**(P2,15/05)**	**(P2,16/05)**	**(P3,08/04)**
(P1,17/09)					
(P1,19/09)	×				
(P2,15/05)					
(P2,16/05)			×		
(P3,08/04)					

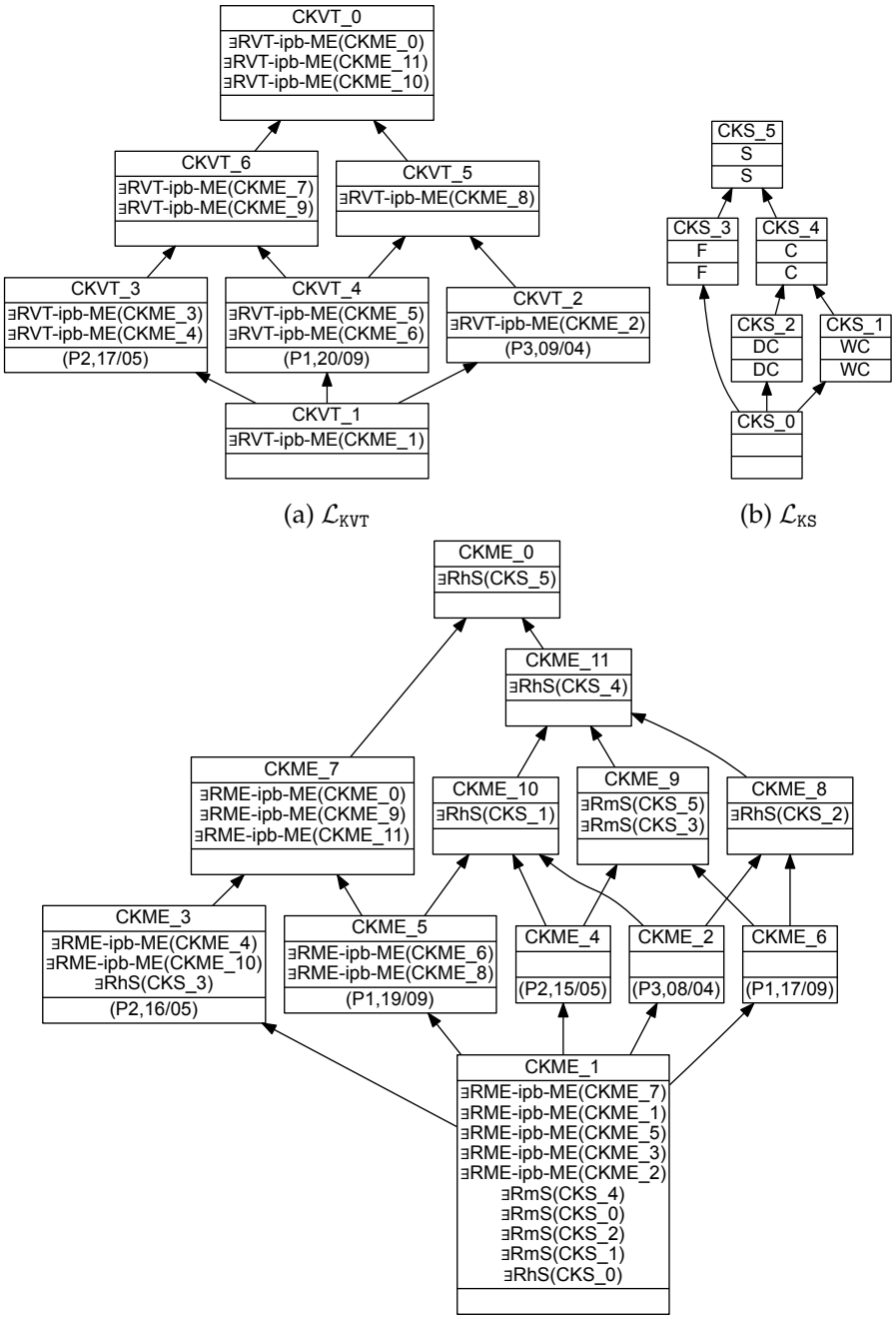
RVT-ipb-ME	**(P1,17/09)**	**(P1,19/09)**	**(P2,15/05)**	**(P2,16/05)**	**(P3,08/04)**
(P1,20/09)	×	×			
(P2,17/05)			×	×	
(P3,09/04)					×

RmS	**S**	**C**	**F**	**DC**	**WC**
(P1,17/09)			×		
(P1,19/09)					
(P2,15/05)			×		
(P2,16/05)					
(P3,08/04)					

RhS	**S**	**C**	**F**	**DC**	**WC**
(P1,17/09)				×	
(P1,19/09)					×
(P2,15/05)					×
(P2,16/05)			×		
(P3,08/04)				×	×

The RCA process is applied (as explained in Sect. 3.4.2) to the RCA input shown in Tab. 6.2 and the fix point depicted in Fig. 6.6 is obtained. Let us mention that the $\mathcal{L}_{\text{KS}}$ lattice of symptoms (Fig. 6.6b) represents the user-defined taxonomy shown in Fig. 6.5.

Using the CPOHrchy algorithm (Sect. 4.3.1), we extract a hierarchy of multilevel cpo-patterns from the obtained RCA output shown in Fig. 6.6. For instance, the multilevel cpo-



(a) $\mathcal{L}_{\texttt{KVT}}$ (b) $\mathcal{L}_{\texttt{KS}}$

(c) $\mathcal{L}_{\texttt{KME}}$

Figure 6.6: The RCA output (the simplified concept lattices) obtained by exploring the sequential data shown in Tab. 6.1 with a user-defined taxonomy over the items

pattern depicted in Fig. 6.7 is associated with the CKVT_6 main concept (Fig. 6.6a). This cpo-pattern contains qualitative items at different levels of precision, e.g. $\texttt{SYMPTOMS}_{\texttt{high}}$ (most general), $\texttt{COUGHS}_{\texttt{high}}$ (non-atomic) and $\texttt{FEVER}_{\texttt{moderate}}$ (most specific). Indeed, during the relational scaling step, RCA reveals the relationships between the medical examinations and the symptoms across different levels of the taxonomy. The conversion of the navigated concept intents into vertices relies on the simplified lattice of symptoms depicted in Fig. 6.6b.

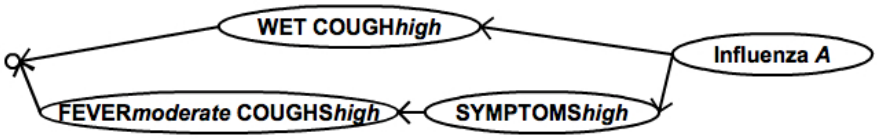


Figure 6.7: The cpo-pattern associated with concept CKVT_6 (Fig. 6.6a) that contains items across different levels of the taxonomy of symptoms shown in Fig. 6.5

Let us note that our approach with a user-defined taxonomy can be applied to explore sequences that include items across different levels of the taxonomy, as well.

6.4 Exploration of Simple Sequential Data

RCA-SEQ can be easily adapted to explore simple sequential data (i.e. the items do not have associated qualitative values). To illustrate this, let us consider the simple sequential medical data given in Tab. 6.3.

Table 6.3: Illustrative example of simple sequential medical data

Sequence Id	Sequence
$S1$	$\langle$(COUGH)(COUGH FEVER)(COUGH FEVER HEADACHE)(Influenza)$\rangle$
$S2$	$\langle$(FEVER)(Influenza)$\rangle$
$S3$	$\langle$(COUGH)(COUGH)(Influenza)$\rangle$
$S4$	$\langle$(COUGH FEVER)(Influenza)$\rangle$
$S5$	$\langle$(COUGH)(COUGH)(FEVER)(HEADACHE)(Influenza)$\rangle$

A patient sequence ends with a viral test (target itemset of viruses) that is preceded by a chronologically ordered set of medical examinations (non-target itemsets of symptoms). For example, sequence $S4$ ends with the (Influenza) target 1-itemset that contains only the influenza virus, and, besides, $S4$ contains the (COUGH FEVER) non-target itemset that has the cough and fever symptoms.

To explore the sequential data shown in Tab. 6.3, we use the data model depicted in Fig. 6.8. There are two rectangles, one for each set of objects we manipulate, as follows: viral tests (VT) and medical examinations (ME). The temporal links between viral tests/medical examinations and medical examinations are highlighted by a temporal relation *is preceded by*.

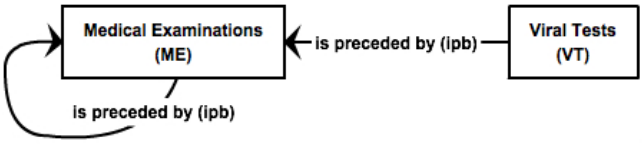


Figure 6.8: The modelling of the simple sequential data given in Tab. 6.3

To build the set of viral tests and the one of medical examinations we remodel the sequences of itemsets given in Tab. 6.3 as the sequences of UIDs shown in Tab. 6.4. For instance, the ⟨IS1_Seq2 Seq2⟩ sequence of UIDs represents the sequence $getS(\texttt{Seq2}) = S2$ given in Tab. 6.3 with $getIS(\texttt{IS1\_Seq2}) = (\texttt{FEVER})$ and $getIS(\texttt{Seq2}) = (\texttt{Influenza})$.

Table 6.4: The patient sequences of UIDs obtained by remodelling the sequences of itemsets shown in Tab. 6.3

Sequence
⟨IS1_Seq1 IS2_Seq1 IS3_Seq1 Seq1⟩
⟨IS1_Seq2 Seq2⟩
⟨IS1_Seq3 IS2_Seq3 Seq3⟩
⟨IS1_Seq4 Seq4⟩
⟨IS1_Seq5 IS2_Seq5 IS3_Seq5 IS4_Seq5 Seq5⟩

Relying on the data model depicted in Fig. 6.8 and on Tab. 6.4, we build the RCF (the RCA input) illustrated in Tab. 6.5.

Table 6.5: RCF that encodes the sequential data shown in Tab. 6.3; formal contexts: KVT and KME; temporal relational contexts: RME-ipb-ME and RVT-ipb-ME

KVT	Influenza
Seq1	×
Seq2	×
Seq3	×
Seq4	×
Seq5	×

KME	COUGH	FEVER	HEADACHE
IS1_Seq1	×		
IS2_Seq1	×	×	
IS3_Seq1	×	×	×
IS1_Seq2		×	
IS1_Seq3	×		
IS2_Seq3	×		
IS1_Seq4	×	×	
IS1_Seq5	×		
IS2_Seq5	×		
IS3_Seq5		×	
IS4_Seq5			×

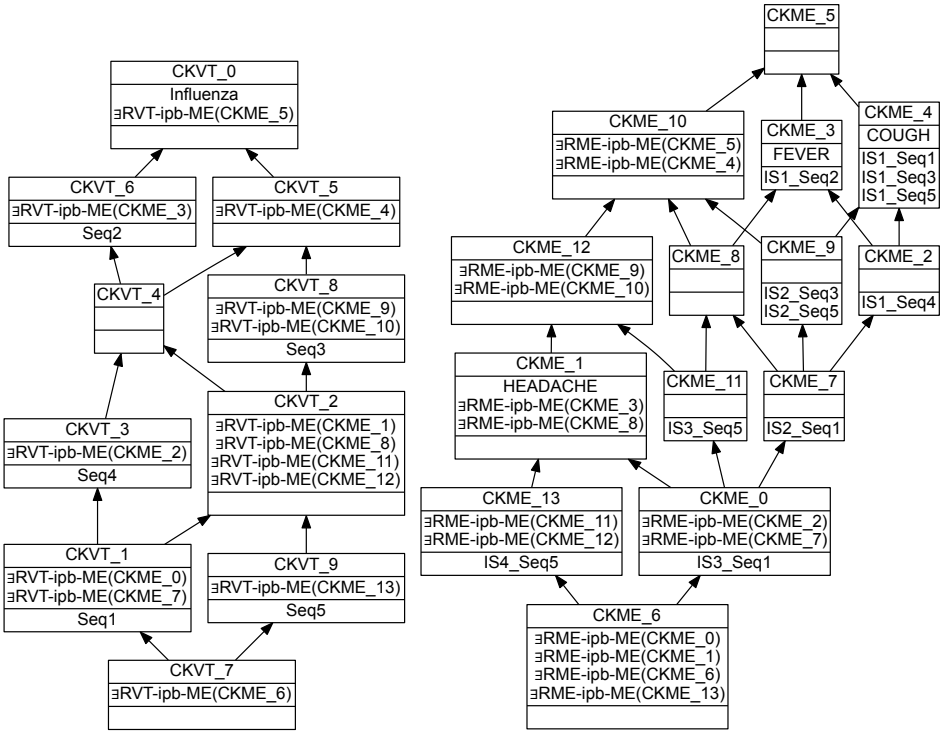
RME-ipb-ME	IS1_Seq1	IS2_Seq1	IS3_Seq1	IS1_Seq2	IS1_Seq3	IS2_Seq3	IS1_Seq4	IS1_Seq5	IS2_Seq5	IS3_Seq5	IS4_Seq5
IS1_Seq1											
IS2_Seq1	×										
IS3_Seq1	×	×									
IS1_Seq2											
IS1_Seq3											
IS2_Seq3					×						
IS1_Seq4											
IS1_Seq5											
IS2_Seq5								×			
IS3_Seq5								×	×		
IS4_Seq5								×	×	×	

RVT-ipb-ME	IS1_Seq1	IS2_Seq1	IS3_Seq1	IS1_Seq2	IS1_Seq3	IS2_Seq3	IS1_Seq4	IS1_Seq5	IS2_Seq5	IS3_Seq5	IS4_Seq5
Seq1	×	×	×								
Seq2				×							
Seq3					×	×					
Seq4							×				
Seq5								×	×	×	×

In Tab. 6.5 exist two formal contexts one for each set of objects out of the data model: KVT (viral tests) and KME (medical examinations). In addition, there are two relational contexts

one for each temporal relation out of the data model: `RVT-ipb-ME` (viral test *ipb* medical examination) and `RME-ipb-ME` (medical examination *ipb* medical examination). The formal context of viral tests encodes that each target 1-itemset contains the `Influenza` item; the formal context of medical examinations encodes that each non-target itemset contains the items `COUGH` and/or `FEVER` and/or `HEADACHE`. Therefore, the itemsets are described by means of binary attributes.

Figure 6.9 depicts the RCA output obtained by applying the RCA process to the RCF given in Tab. 6.5. Two lattices are obtained: $\mathcal{L}_{\texttt{KT}}$ (the lattice of viral tests) and $\mathcal{L}_{\texttt{ME}}$ (the lattice of medical examinations). It is noted that each concept intent can contain binary attributes and/or relational attributes. For example, the `CKVT_0` concept intent (Fig. 6.9a) contains the binary attribute `Influenza` and the relational attribute $\exists$`RVT-ipb-ME`(`CKME_5`).



(a) $\mathcal{L}_{\texttt{KVT}}$ (b) $\mathcal{L}_{\texttt{KME}}$

Figure 6.9: The RCA output (the simplified concept lattices) obtained by exploring the simple sequential data given in Tab. 6.3

As explained in Chapter 4, we extract a cpo-pattern for each concept out of the main lattice $\mathcal{L}_{\texttt{KVT}}$. Basically, we start from a main concept intent and we navigate the lattices being guided by the relational attributes pointing to the most specific concepts (Properties 4.4 and 4.5). For each navigated concept intent we derive:

– *a vertex* that is labelled with an itemset built by using the binary attributes;

– *an edge* for each temporal relational attribute.

To illustrate this, let us analyse cpo-pattern $\mathcal{G}_{\texttt{CKVT\_8}}$ (②, Fig. 6.10) extracted by navigating the interrelated concept intents ① shown in Fig. 6.10. Vertex $v_{\texttt{CKVT\_8}}$ labelled with itemset (`Influenza`) is derived from the binary attribute `Influenza` of the `CKVT_8` main concept intent. Vertex $v_{\texttt{CKME\_9}}$ labelled with itemset (`COUGH`) is derived from the binary attribute `COUGH` of the `CKME_9` intent. The edge between the $v_{\texttt{CKVT\_8}}$ and $v_{\texttt{CKME\_9}}$ vertices is derived from the temporal relational attribute ∃`RVT-ipb-ME`(`CKME_9`) of the `CKVT_8` intent that points to `CKME_9`. Similarly, vertex $v_{\texttt{CKME\_4}}$ labelled with itemset (`COUGH`) is derived from the binary attribute `COUGH` of the `CKME_4` intent. The edge between the $v_{\texttt{CKME\_9}}$ and $v_{\texttt{CKME\_4}}$ vertices is derived from the temporal relational attribute ∃`RME-ipb-ME`(`CKME_4`) of the `CKME_9` intent that points to `CKME_4`.

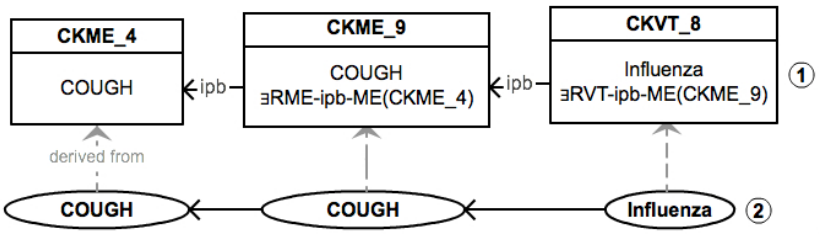


Figure 6.10: ① the interrelated concept intents navigated starting from the `CKVT_8` main concept intent; ② cpo-pattern $\mathcal{G}_{\texttt{CKVT\_8}}$ associated with the `CKVT_8` main concept from Fig. 6.9a. The intents contain only the relational attributes according to Properties 4.4 and 4.5

6.5 Exploration of Heterogeneous Sequential Data

So far, we have shown how to extract cpo-patterns of itemsets, where an itemset contains homogeneous items (i.e. items of a similar nature). Recently, Egho et al. [2014] have focused on discovering sequential patterns (rather than cpo-patterns) in complex and heterogeneous sequential data, where a sequence contains "elementary sequences" (ESs), i.e. itemsets whose items are from distinct domains. According to the authors, the complexity of these data stems from the fact that an item of an ES can be of two types: (i) *atomic item taken from a poset* or (ii) *subset of an unordered set of items*.

To explore such complex and heterogeneous data, the authors first look for frequent and specific ESs (FSESs). Second, ESs from the original sequences are replaced with FSESs. Then, FSESs are mapped to distinct integers that are used to encode the sequences obtained at the second step. Lastly, the transformed sequences are explored using a propositional algorithm, e.g. CLOSPAN [Yan et al., 2003]. Thus, the proposed algorithm MMISP does not discover directly sequential patterns in complex and heterogeneous sequential data since a preprocessing step is involved, which encodes the original heterogeneous data into sequences of homogeneous items.

In contrast, RCA-SEQ directly searches for cpo-patterns in complex and heterogeneous sequential data, and, besides, reveals how these cpo-patterns relate to each other. Moreover, our approach can be applied to sequences of ESs (in this thesis, referred to as *heterogeneous itemsets*), where an ES comprises k-itemsets (k can vary from itemset to itemset) from different domains (as presented in Sect. 2.2.2). Let us note that, a k-itemset is composed of atomic items taken from: (i) *an unordered set of items* or (ii) *a partial order over a set of items (taxonomy)*.

Consequently, we generalise the ES proposed by Egho et al. by considering its atomic items as 1-itemsets. In the following, we introduce a comprehensive KDD approach for exploring such sequential data and for extracting *hierarchies of multilevel heterogeneous cpo-patterns* (i.e. cpo-patterns whose paths are closed heterogeneous sequential patterns as described in Sect. 2.2.2) by slightly modifying the RCA-SEQ approach.

6.5.1 Motivating Example

We rely on the running example from Sect. 3.2. We recall that physicians are interested in assessing the symptoms (e.g. fever and cough) felt by patients before the outbreaks of influenza virus. The symptoms and the viruses are detected by medical examinations and viral tests, respectively.

For a *medical examination* undergone by a patient can be recorded: (i) *the experienced symptoms and their intensities (mandatory)*, (ii) *the state of vital signs (e.g. heart rate)* and (iii) *the prescribed drugs and their doses*.

For a *viral test* done by a patient can be recorded: (i) *the viruses that infected the patient (mandatory)* and (ii) *the patient category (e.g. child, infant) and gender*.

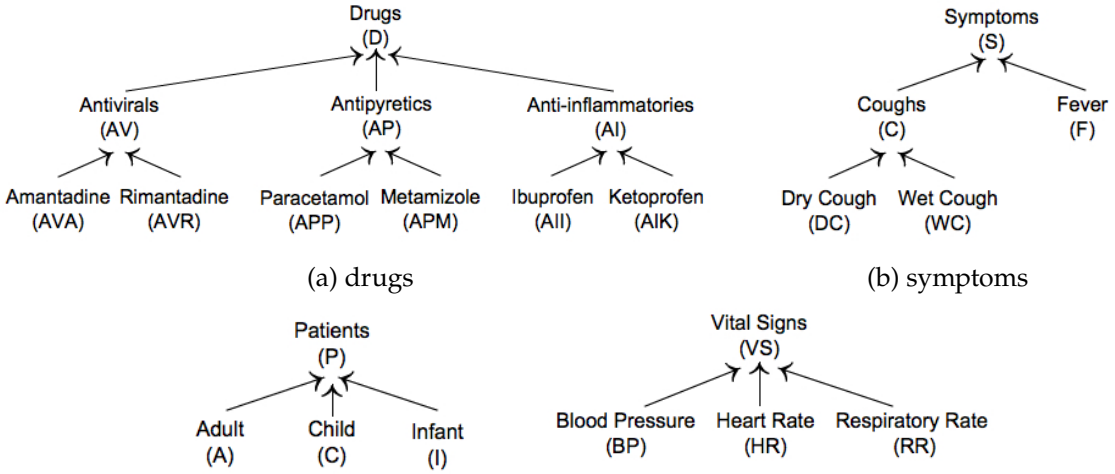
We suppose that by analysing these various collected information, physicians try to better understand patient health evolution before the outbreaks of influenza virus. The viruses that can be detected by the viral tests constitute the $\{\texttt{Influenza}_\texttt{A}, \texttt{Influenza}_\texttt{B}\}$ unordered set of atomic items. The symptoms, drugs, vital signs and patients (domains) can be described by means of taxonomies (posets) as depicted in Fig. 6.11. The sequence-building items of these sets are enumerated in Tab. 6.6.

Table 6.6: The atomic items used to build heterogeneous sequences

Set	Items
symptoms	FEVER (F), DRY COUGH (DC), WET COUGH (WC)
drugs	AMANTADINE (AVA), RIMANTADINE (AVR), IBUPROFEN (AII), PARACETAMOL (APP), METAMIZOLE (APM), KETOPROFEN (AIK)
patients	ADULT (A), CHILD (C), INFANT (I)
vital signs	BLOOD PRESSURE (BP), HEART RATE (HR), RESPIRATORY RATE (RR)

The intensity of a symptom can be *moderate* (m) or *high* (h); the state of a vital sign can be *good* (g) or *bad* (b); the prescribed dose for a drug can be *loading dose* (ld) or *maintenance*

dose (md); the gender of a patient can be *female* (f) or *male* (m). Therefore, we deal again with qualitative medical data.



(a) drugs (b) symptoms

(c) patients (d) vital signs

Figure 6.11: Taxonomies over the drugs, symptoms, patients and vitals domains

Table 6.7 shows an illustrative example of heterogeneous medical data from last year (we consider that these data are exported from a relational database). For example, the medical examination identified by the (P1,12/11) temporal object was undergone on November 12^{th} when the dry cough (DC) symptom of patient P1 was high (h) and the respiratory rate (RR) vital of P1 was good (g). It is noted that the prescribed treatment was not recorded. The viral test identified by temporal object (P1, 16/11) was done on November 16^{th} when the male (m) child (C) P1 was diagnosed with both the influenza A and B viruses.

Table 6.7: Illustrative example of heterogeneous medical data

Patient Id	Date	Medical Examination								Viral Test			
		Symptoms			Drugs			Vital Signs		Influenza		Patient	
		DC	WC	F	AVR	AVA	APP	BP	RR	A	B	A	C
P1	12/11	h	–	–	–	–	–	–	g	–	–	–	–
	13/11	–	–	m	–	–	–	–	b	–	–	–	–
	14/11	h	–	h	md	–	md	b	b	–	–	–	–
	16/11	–	–	–	–	–	–	–	–	×	×	–	m
P2	09/03	–	h	–	ld	–	ld	g	–	–	–	–	–
	13/03	–	–	–	–	–	–	–	–	×	–	–	f
P3	25/12	–	–	h	ld	–	ld	b	b	–	–	–	–
	28/12	–	–	–	–	–	–	–	–	×	–	f	–
P4	03/01	–	h	m	–	–	–	b	g	–	–	–	–
	06/01	h	–	h	–	md	ld	b	b	–	–	–	–
	08/01	–	–	–	–	–	–	–	–	×	×	–	m

6.5.2 Data Preprocessing

To obtain heterogeneous sequences as described in Sect. 2.2.2, we order temporally the examinations from Tab. 6.7 according to the Date column and we obtain the heterogeneous patient sequences given in Tab. 6.8. Note that, the information collected is specific, i.e. all itemsets contain atomic items. There are four sequences $S1$, $S2$, $S3$ and $S4$, one for each patient P1, P2, P3 and P4, respectively.

A medical examination is a heterogeneous itemset such as $\{symptoms, drugs, vital\ signs\}$, where $symptoms$, $drugs$ and $vital\ signs$ are itemsets built from the atomic items shown in Tab. 6.6. If the information about the treatment and the vital signs is not collected, the $drugs$ and $vital\ signs$ itemsets are the empty set. For example, in Tab. 6.8 the heterogeneous itemset $\{(\mathtt{F_m}), \emptyset, (\mathtt{RR_b})\}$ of sequence $S1$ describes a medical examination when patient P1 experienced a moderate fever and the respiratory rate was bad. The treatment was not recorded since drug itemset is $\emptyset$. Similarly, the $\{(\mathtt{DC_hF_h}), (\mathtt{AVA_{md}APP_{ld}}), (\mathtt{BP_b\ RR_b})\}$ heterogeneous itemset of sequence $S4$ describes a medical examination when patient P4 experienced a high dry cough and a high fever, the prescribed treatment was maintenance dose amantadine and loading dose paracetamol and the blood pressure and respiratory rate were bad.

A viral test is a heterogeneous itemset $\{viruses, patient\}$, where $viruses$ is an itemset built from an unordered set and $patient$ is an atomic item (1-itemset) taken from the poset shown in Fig. 6.11c. For instance, in Tab. 6.8 the $\{(\mathtt{Influenza_AInfluenza_B}), (\mathtt{C_m})\}$ heterogeneous itemset of sequence $S4$ describes a viral test when the male child P4 was diagnosed with both the influenza A and B viruses.

Table 6.8: Heterogeneous patient sequences obtained from Tab. 6.7

Id	Sequence
$S1$	$\langle\{(\mathtt{DC_h}), \emptyset, (\mathtt{RR_g})\}\ \{(\mathtt{F_m}), \emptyset, (\mathtt{RR_b})\}\ \{(\mathtt{DC_hF_h}), (\mathtt{AVR_{md}APP_{md}}), (\mathtt{BP_bRR_b})\}\ \{(\mathtt{Influenza_AInfluenza_B}), (\mathtt{C_m})\}\rangle$
$S2$	$\langle\{(\mathtt{WC_h}), (\mathtt{AVR_{ld}APP_{ld}}), (\mathtt{BP_g})\}\ \{(\mathtt{Influenza_A}), (\mathtt{C_f})\}\rangle$
$S3$	$\langle\{(\mathtt{F_h}), (\mathtt{AVR_{ld}APP_{ld}}), (\mathtt{BP_bRR_b})\}\ \{(\mathtt{Influenza_A}), (\mathtt{A_f})\}\rangle$
$S4$	$\langle\{(\mathtt{WC_hF_m}), \emptyset, (\mathtt{RR_gBP_b})\}\ \{(\mathtt{DC_hF_h}), (\mathtt{AVA_{md}APP_{ld}}), (\mathtt{BP_b\ RR_b})\}\ \{(\mathtt{Influenza_AInfluenza_B}), (\mathtt{C_m})\}\rangle$

6.5.3 Modelling Heterogeneous Qualitative Sequential Data

To explore the heterogeneous sequential data illustrated in Tab. 6.8, we upgrade the data model depicted in Fig. 3.3. Exploiting the relational nature of the various collected information, we propose the data model shown in Fig. 6.12. There are six rectangles, one for each set of objects we manipulate, as follows: viral tests (VT), medical examinations (ME), patients (P), symptoms (S), vital signs (VS) and drugs (D).

The relations *is preceded by* and *has symptom* are the ones explained in Section 3.3.3. Viral tests are linked to patients by the qualitative relations *undergone by* differentiated by the gen-

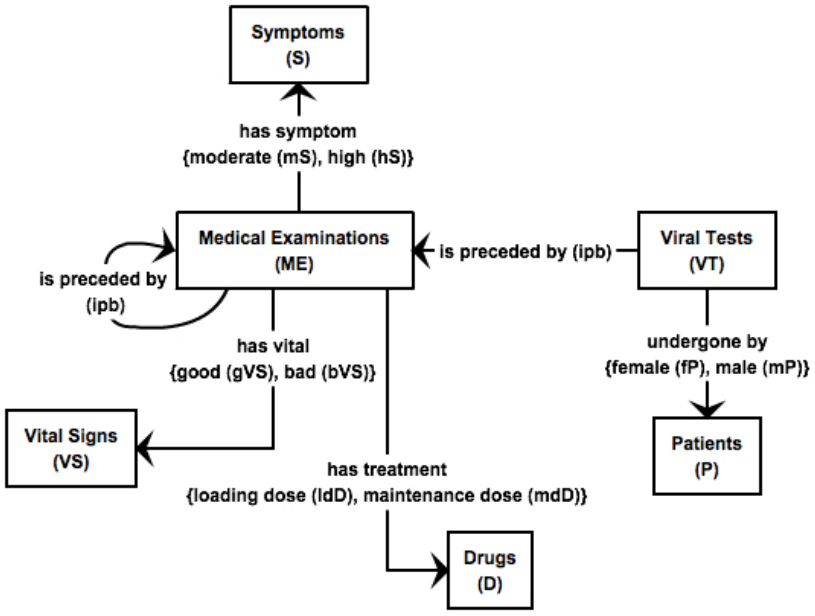


Figure 6.12: The modelling of the heterogeneous sequential medical data shown in Tab. 6.8

der of patients, i.e. *female* (fP) or *male* (mP). Medical examinations are linked to vital signs by the qualitative relations *has vital* differentiated by the state of the vitals, i.e. *good* (gVS) or *bad* (bVS). Medical examinations are linked to drugs by the qualitative relations *has treatment* differentiated by the prescribed drug dose, i.e. *loading dose* (ldD) or *maintenance dose* (mdD).

6.5.4 Relational Analysis of Heterogeneous Qualitative Sequential Data

To explore such heterogeneous data we follow the steps presented in Sect. 3.4.

6.5.4.1 Building the RCA Input

In order to constitute the set of viral tests and the one of medical examinations used to build the RCA input we remodel the heterogeneous sequences shown in Tab. 6.8 as the sequences of UIDs given in Tab. 6.9. `IS1_Seq1`, `IS2_Seq1`, `IS3_Seq1`, `IS1_Seq2`, `IS1_Seq3`, `IS1_Seq4` and `IS2_Seq4` uniquely identify the medical examinations (P1, 12/11), (P1, 13/11), (P1, 14/11), (P2, 09/03), (P3, 25/12), (P4, 03/01) and (P4, 06/01), respectively. `Seq1`, `Seq2`, `Seq3` and `Seq4` uniquely identify the viral tests (P1, 16/11), (P2, 13/03), (P3, 28/12) and (P4, 08/01), respectively.

Relying on the data model shown in Fig. 6.12 and on the dataset given in Tab. 6.9, we encode the heterogeneous data from Tab. 6.8 into the RCF illustrated in Tab. 6.10. The `KS` (symptoms), `KP` (patients), `KVS` (vital signs), `KD` (drugs), `KVT` (virus tests) and `KME` (medical examinations) cross tables represent formal contexts. Note that, the first four formal contexts

Table 6.9: The sequences of UIDs obtained by remodelling the data shown in Tab. 6.8

Sequence
⟨IS1_Seq1 IS2_Seq1 IS3_Seq1 Seq1⟩
⟨IS1_Seq2 Seq2⟩
⟨IS1_Seq3 Seq3⟩
⟨IS1_Seq4 IS2_Seq4 Seq4⟩

are built by using the ordinal scaling in order to encode the taxonomies. KVT has the set of binary attributes $\{\texttt{Influenza}_\texttt{A}, \texttt{Influenza}_\texttt{B}\}$, i.e. the formal context has two columns, one for each type of influenza virus that can be detected by a viral test. Besides, the viral tests are described by using the qualitative relations *undergone by*. KME has no column since a medical examination is described only by using the qualitative relations *has symptom*, *has treatment* and *has vital*.

The RVT-ipb-ME, RME-ipb-ME, RmS and RhS relational contexts are described in Sect. 3.4.1. The RgVS (medical examination detects a *good* vital sign), RbVS (medical examination detects a *bad* vital sign), RldD (medical examination has treatment *loading dose* drug), RmdD (medical examination has treatment *maintenance dose* drug), RfP (viral test undergone by a *female* patient) and RmP (viral test undergone by a *male* patient) cross tables represent qualitative relational contexts since they define qualitative relations.

6.5.4.2 Applying the RCA Process

RCA is applied to the RCF shown in Tab. 6.10 and the family of concept lattices depicted in Fig. 6.13 is obtained after three iterations. There is a concept lattice for each formal context as follows: $\mathcal{L}_{\texttt{KVT}}$ (viral tests), $\mathcal{L}_{\texttt{KME}}$ (medical examinations), $\mathcal{L}_{\texttt{KS}}$ (symptoms), $\mathcal{L}_{\texttt{KVS}}$ (vital signs), $\mathcal{L}_{\texttt{KD}}$ (drugs) and $\mathcal{L}_{\texttt{KP}}$ (patients). Let us note that $\mathcal{L}_{\texttt{KS}}$, $\mathcal{L}_{\texttt{KP}}$, $\mathcal{L}_{\texttt{KVS}}$ and $\mathcal{L}_{\texttt{KD}}$ correspond to the taxonomies illustrated in Fig. 6.11 and their concepts are used to describe medical examinations or viral tests by means of the qualitative relational attributes.

For example, the relational attribute $\exists\texttt{RgVS}(\texttt{CKVS\_1})$ of the CKME_15 concept intent in $\mathcal{L}_{\texttt{KME}}$ (Fig. 6.13f) is a qualitative one since it highlights the qualitative relation *has vital good*. In addition, this relational attribute describes the medical examinations gathered by the CKME_15 extent, namely IS1_Seq1 (i.e. (P1, 12/11) in Tab. 6.7) and IS1_Seq4 (i.e. (P4, 03/01) in Tab. 6.7), for which a good respiratory rate (i.e. $extent(\texttt{CKVS\_1}) = \{\texttt{RR}\}$) was measured.

6.5.5 Extracting Hierarchies of Multilevel Heterogeneous CPO-Patterns

To extract a hierarchy of multilevel heterogeneous cpo-patterns from the RCA output depicted in Fig. 6.13, we apply algorithm CPOHrchy (Sect. 4.3.1) by slightly modifying the step of converting a concept intent to a vertex. Indeed, in this case a vertex derived from a con-

Table 6.10: RCF that encodes the heterogeneous sequential data shown in Tab. 6.7; formal contexts: KS, KP, KVS, KD, KVT and KME; temporal relational contexts: RME-ipb-ME and RVT-ipb-ME; qualitative relational contexts: RmS, RhS, RfP, RmP, RgVS, RbVS, RldD and RmdD

KS	S	C	F	DC	WC
S	×				
C	×	×			
F	×		×		
DC	×	×		×	
WC	×	×			×

KP	P	A	C	I
P	×			
A	×	×		
C	×		×	
I	×			×

KVS	VS	BP	HR	RR
VS	×			
BP	×	×		
HR	×		×	
RR	×			×

KD	D	AV	AP	AI	AVA	AVR	APP	APM	AII	AIK
D	×									
AV	×	×								
AP	×		×							
AI	×			×						
AVA	×	×			×					
AVR	×	×				×				
APP	×		×				×			
APM	×		×					×		
AII	×			×					×	
AIK	×			×						×

KVT	InfluenzaA	InfluenzaB
Seq1	×	×
Seq2	×	
Seq3	×	
Seq4	×	×

KME
IS1_Seq1
IS2_Seq1
IS3_Seq1
IS1_Seq2
IS1_Seq3
IS1_Seq4
IS2_Seq4

RME-ipb-ME	IS1_Seq1	IS2_Seq1	IS3_Seq1	IS1_Seq2	IS1_Seq3	IS1_Seq4	IS2_Seq4
IS1_Seq1							
IS2_Seq1	×						
IS3_Seq1	×	×					
IS1_Seq2							
IS1_Seq3							
IS1_Seq4							
IS2_Seq4						×	

RVT-ipb-ME	IS1_Seq1	IS2_Seq1	IS3_Seq1	IS1_Seq2	IS1_Seq3	IS1_Seq4	IS2_Seq4
Seq1	×	×	×				
Seq2				×			
Seq3					×		
Seq4						×	×

RmS	S	C	F	DC	WC
IS1_Seq1					
IS2_Seq1			×		
IS3_Seq1					
IS1_Seq2					
IS1_Seq3					
IS1_Seq4			×		
IS2_Seq4					

RhS	S	C	F	DC	WC
IS1_Seq1				×	
IS2_Seq1					
IS3_Seq1			×	×	
IS1_Seq2					×
IS1_Seq3			×		
IS1_Seq4					×
IS2_Seq4			×	×	

RfP	P	A	C	I
Seq1				
Seq2			×	
Seq3		×		
Seq4				

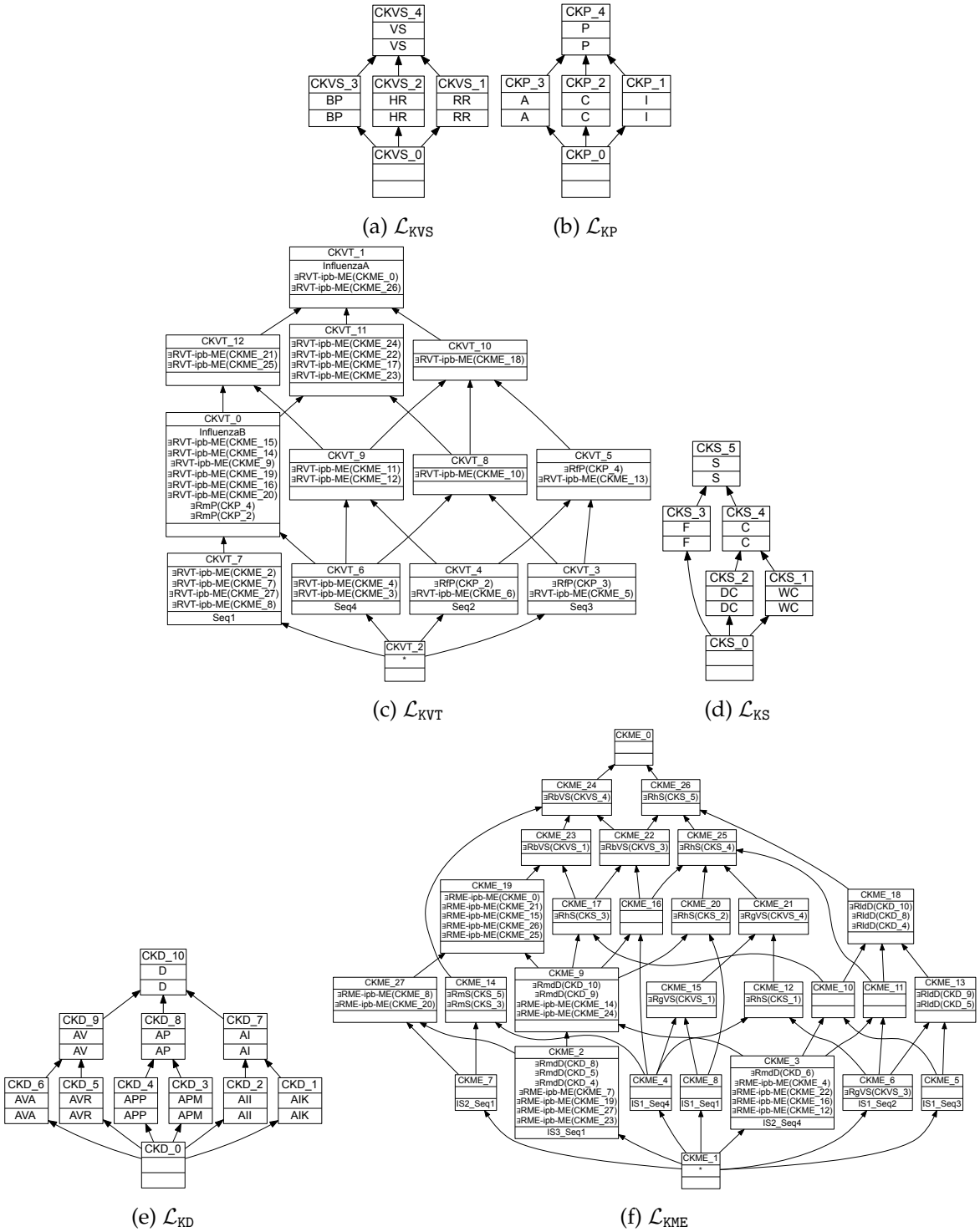
RmP	P	A	C	I
Seq1			×	
Seq2				
Seq3				
Seq4			×	

RgVS	VS	BP	HR	RR
IS1_Seq1				×
IS2_Seq1				
IS3_Seq1				
IS1_Seq2		×		
IS1_Seq3				
IS1_Seq4				×
IS2_Seq4				

RbVS	VS	BP	HR	RR
IS1_Seq1				
IS2_Seq1				×
IS3_Seq1		×		×
IS1_Seq2				
IS1_Seq3		×		×
IS1_Seq4		×		
IS2_Seq4		×		×

RldD	D	AV	AP	AI	AVA	AVR	APP	APM	AII	AIK
IS1_Seq1										
IS2_Seq1										
IS3_Seq1										
IS1_Seq2						×	×			
IS1_Seq3						×	×			
IS1_Seq4										
IS2_Seq4							×			

RmdD	D	AV	AP	AI	AVA	AVR	APP	APM	AII	AIK
IS1_Seq1										
IS2_Seq1										
IS3_Seq1						×	×			
IS1_Seq2										
IS1_Seq3										
IS1_Seq4										
IS2_Seq4					×					



(a) $\mathcal{L}_{\mathtt{KVS}}$ (b) $\mathcal{L}_{\mathtt{KP}}$

(c) $\mathcal{L}_{\mathtt{KVT}}$ (d) $\mathcal{L}_{\mathtt{KS}}$

(e) $\mathcal{L}_{\mathtt{KD}}$ (f) $\mathcal{L}_{\mathtt{KME}}$

Figure 6.13: The fix point (the simplified concept lattices) of the RCF given in Tab. 6.10: (a) the lattice of vital signs; (b) the lattice of patients; (c) the lattice of viral tests; (d) the lattice of symptoms; (e) the lattice of drugs; (f) the lattice of medical examinations; * represents the intent of a bottom concept

cept intent is actually a *heterogeneous vertex* labelled with a multilevel heterogeneous itemset. Basically, an itemset of the multilevel heterogeneous itemset is built for each set of qualitative relational attributes (which define the same qualitative relation) or for each set of binary attributes (which are from the same domain) out of the concept intent.

Therefore, for a concept intent we analyse the qualitative relational attributes, which are built using a qualitative relation hi_q and concepts from the lattice of items $\mathcal{L}_{K_I} = (\mathcal{C}_{K_I}, \preceq_{K_I})$, to derive items as follows:

- from a qualitative relational attribute $\exists hi_q(C_I)$, where $C_I \in \mathcal{C}_{K_I}$, is derived an item, denoted by "$item_q$", where $extent(C_I) = \{item\}$ and q is the item quality according to hi_q;
- if there is no qualitative relational attribute that highlights the hi_q relation and the information introduced by this relation is mandatory, then is derived an item, denoted by "$item_?$" where $extent(\top(\mathcal{L}_{K_I})) = \{item\}$, that constitutes the 1-itemset obtained for this type of information; conversely, if the information introduced by this relation is not mandatory, then no item is derived, and thus $\emptyset$ is obtained for this type of information.

To illustrate this, let us consider the CKME_6 concept intent (Fig. 6.13f) and the derived heterogeneous vertex shown in Fig. 6.14. Only the qualitative relational attributes pointing to the most specific concepts are analysed (Property 4.4). To improve the visualisation of a heterogeneous vertex, we propose to label the vertex (⬭) with the itemset (of the corresponding multilevel heterogeneous itemset) that represents the mandatory information (e.g. symptoms in Fig. 6.14), and, besides, to use other shapes to illustrate the itemsets that represent extra information (e.g. ⬭ for vital signs, ◊ for patients and ▭ for drugs).

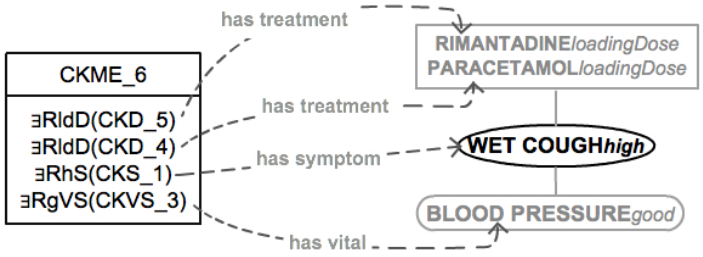


Figure 6.14: The heterogeneous vertex derived from the CKME_6 concept intent (Fig. 6.13f)

In the following, we explain how to derive the heterogeneous itemset that labels the heterogeneous vertex depicted in Fig. 6.14. The symptom itemset ($\texttt{WET COUGH}_{\texttt{high}}$) represents the label of the vertex (⬭) derived from the qualitative relational attribute $\exists\texttt{RhS}(\texttt{CKS\_1})$ of the CKME_6 concept intent since this attribute highlights the relation *has symptom high*, and, besides, $extent(\texttt{CKS\_1}) = \{\texttt{WC}\}$ (i.e. wet cough). Regarding the extra information, the drug itemset ($\texttt{PARACETAMOL}_{\texttt{loadingDose}}$ $\texttt{RIMANTADINE}_{\texttt{loadingDose}}$) is derived from the qualitative relational

attributes $\exists$RldD(CKD_4) and $\exists$RldD(CKD_5) since both highlight the relation *has treatment loading dose*, and, in addition, $extent$(CKD_4) = {APP} (i.e. paracetamol), $extent$(CKD_5) = {AVR} (i.e. rimantadine). The vital itemset (BLOOD PRESSURE$_{\texttt{good}}$) is derived from the qualitative relational attribute $\exists$RgVS(CKVS_3) since it highlights the relation *has vital good*, and, besides, $extent$(CKVS_3) = {BP} (i.e. blood pressure). Therefore, heterogeneous itemset $IS_{\mathcal{H}_{\texttt{CKME\_6}}}$ = {(WET COUGH$_{\texttt{high}}$), (PARACETAMOL$_{\texttt{loadingDose}}$ RIMANTADINE$_{\texttt{loadingDose}}$), (BLOOD PRESURE$_{\texttt{good}}$)} labels the heterogeneous vertex derived from the CKME_6 concept intent.

Now, we are able to extract a hierarchy of multilevel heterogeneous cpo-patterns from the RCA output shown in Fig. 6.13. We recall that a multilevel heterogeneous cpo-pattern is extracted for each main concept in lattice $\mathcal{L}_{\texttt{KVT}}$ (Fig. 6.13c) by navigating interrelated concept intents. For example, Fig. 6.15 depicts the set of navigated concept intents starting from the CKVT_0 main concept intent. This set of navigated concept intents is obtained as explained in Section 4.3 without any modification. Therefore, the CKVT_0 intent points to the CKME_9 intent that points to both the CKME_15 and CKME_14 intents.

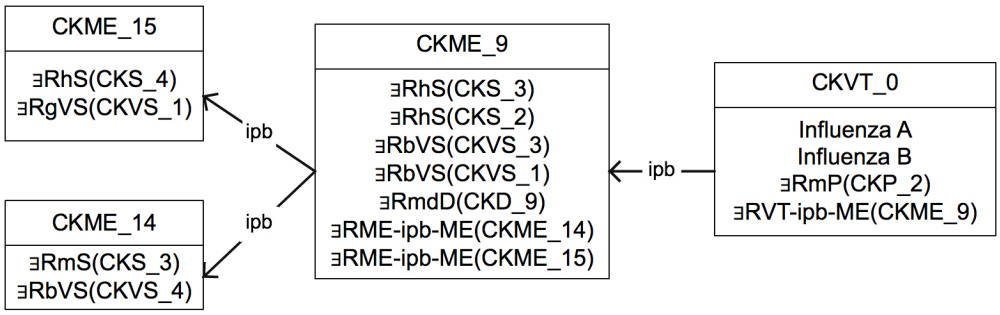


Figure 6.15: The navigated concept intents starting from the CKVT_0 main concept (Fig. 6.13c) in order to extract the $\mathcal{G}_{\texttt{CKVT\_0}}$ multilevel heterogeneous cpo-pattern

Figure 6.16 depicts the $\mathcal{G}_{\texttt{CKVT\_0}}$ multilevel heterogeneous cpo-pattern extracted by navigating the concept intents shown in Fig. 6.15. The heterogeneous vertices of cpo-pattern $\mathcal{G}_{\texttt{CKVT\_0}}$ are labelled with the following multilevel heterogeneous itemsets:

- {(Influenza$_{\texttt{A}}$ Influenza$_{\texttt{B}}$), (CHILD$_{\texttt{male}}$)} that is derived from the CKVT_0 main concept intent. The virus itemset (Influenza$_{\texttt{A}}$ Influenza$_{\texttt{B}}$) is the label of the vertex (⬭) derived from the binary attributes of the intent. Regarding the extra information, the patient itemset (CHILD$_{\texttt{male}}$) is derived from the qualitative relational attribute $\exists$RmP(CKP_2) since it highlights the relation *undergone by male*, and, besides, $extent$(CKP_2) = {C} (i.e. child);

- {(FEVER$_{\texttt{high}}$ DRY COUGH$_{\texttt{high}}$), (ANTIVIRALS$_{\texttt{maintenanceDose}}$), (BLOOD PRESSURE$_{\texttt{bad}}$ RESPIRATORY RATE$_{\texttt{bad}}$)} that is derived from the qualitative relational attributes of the CKME_9 intent. The symptom itemset (FEVER$_{\texttt{high}}$ DRY COUGH$_{\texttt{high}}$) is the label of the vertex (⬭) derived from $\exists$RhS(CKS_3) and $\exists$RhS(CKS_2) since these relational attributes highlight the relation *has symptom high*,

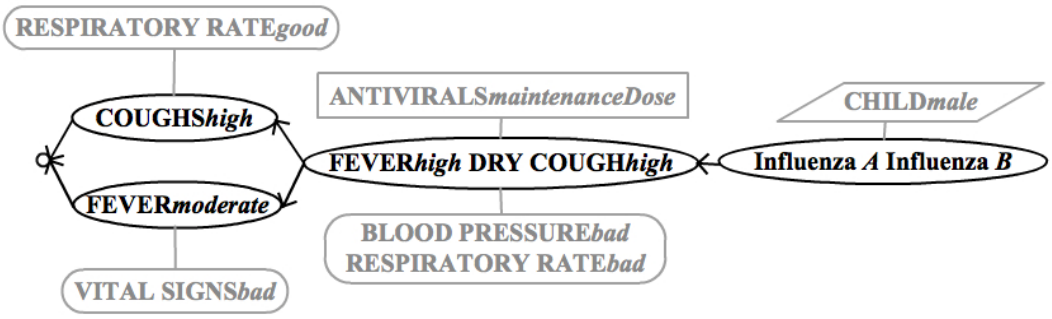


Figure 6.16: The $\mathcal{G}_{\texttt{CKVT\_0}}$ multilevel heterogeneous cpo-pattern associated with the CKVT_0 main concept (Fig. 6.13c)

and, besides, $extent(\texttt{CKS\_3}) = \{\texttt{F}\}$ (i.e. fever), $extent(\texttt{CKS\_2}) = \{\texttt{DC}\}$ (i.e. dry cough). Regarding the extra information, the drug itemset ($\texttt{ANTIVIRALS}_{\texttt{maintenanceDose}}$) is derived from $\exists\texttt{RmdD}(\texttt{CKD\_9})$ since this relational attribute highlights the relation *has treatment maintenance dose*, and, in addition, $extent(\texttt{CKD\_9}) = \{\texttt{AV}\}$ (i.e. antivirals). The vital itemset ($\texttt{BLOOD PRESSURE}_{\texttt{bad}}$ $\texttt{RESPIRATORY RATE}_{\texttt{bad}}$) is derived from $\exists\texttt{RbVS}(\texttt{CKVS\_3})$ and $\exists\texttt{RbVS}(\texttt{CKVS\_1})$ since these relational attributes highlight the relation *has vital bad*, and, besides, $extent(\texttt{CKVS\_3}) = \{\texttt{BP}\}$ (i.e. blood pressure), $extent(\texttt{CKVS\_1}) = \{\texttt{RR}\}$ (i.e. respiratory rate);

– $\{(\texttt{COUGHS}_{\texttt{high}}), \emptyset, (\texttt{RESPIRATORY RATE}_{\texttt{good}})\}$ derived from the qualitative relational attributes of the CKME_15 intent. The symptom itemset ($\texttt{COUGHS}_{\texttt{high}}$) is the label of the vertex (⬭) derived from $\exists\texttt{RhS}(\texttt{CKS\_4})$ since this relational attribute highlights the relation *has symptom high*, and, besides, $extent(\texttt{CKS\_4}) = \{\texttt{C}\}$ (i.e. coughs). Regarding the extra information, since the prescribed treatment was not collected, $\emptyset$ is obtained as the drug itemset. The vital itemset ($\texttt{RESPIRATORY RATE}_{\texttt{good}}$) is derived from the $\exists\texttt{RgVS}(\texttt{CKVS\_1})$ relational attribute that highlights the relation *has vital good*;

– $\{(\texttt{FEVER}_{\texttt{moderate}}), \emptyset, (\texttt{VITAL SIGNS}_{\texttt{bad}})\}$ that is derived from the qualitative relational attributes of the CKME_14 intent. The symptom itemset ($\texttt{FEVER}_{\texttt{moderate}}$) is the label of the vertex (⬭) derived from the $\exists\texttt{RmS}(\texttt{CKS\_3})$ relational attribute that highlights the relation *has symptom moderate*. Regarding the extra information, since the prescribed treatment was not collected, $\emptyset$ is obtained as the drug itemset. The vital itemset ($\texttt{VITAL SIGNS}_{\texttt{bad}}$) is derived from $\exists\texttt{RbVS}(\texttt{CKVS\_4})$ since this relational attribute highlights the relation *has vital bad*, and, besides, $extent(\texttt{CKVS\_4}) = \{\texttt{VS}\}$ (i.e. vital signs).

6.6 Summary

In this chapter, we have illustrated the adaptability of the RCA-Seq approach. We have shown how to push domain knowledge and user preferences into the mining process by slightly modifying our approach. Therefore, more cpo-patterns can emerge based on the user-defined taxonomies over the items. Smaller hierarchies of multilevel cpo-patterns can be extracted by considering user-defined constraints on the order relations on itemsets. Moreover, we have explained how to adapt RCA-Seq to explore simple sequential data and how to extract heterogeneous cpo-patterns from heterogeneous sequential data.

7 Hydro-Ecology as Application Context

Contents

7.1 Introduction

The RCA-Seq approach is applied to hydro-ecological data collected (from French rivers) during two interdisciplinary research projects, namely Fresqueau[1] and REX[2]. In this chapter, firstly, we briefly explain the hydro-ecological data that we have to deal with. Secondly, we show how to preprocess these data according to domain knowledge. Lastly, we present and discuss the results obtained by exploring these data.

In Europe, according to the Water Framework Directive [European Union, 2000] recommendations, a special attention should be given to preserving or restoring the good state of waterbodies. Monitoring and assessing the effect of the pollution sources or the one of the restoration processes is to be done in order to improve the domain knowledge, and, besides, to define guidelines for stakeholders.

The Fresqueau project gathered and unified databases about the north-est and south-est French waterbodies. A number of 11329 river sites (i.e. fixed points) are monitored. The collected data cover various compartments, e.g. physico-chemistry, hydro-biology, hydro-morphology and land use (as described in [Berrahou et al., 2015]). Some of these data are temporally related, e.g. a physico-chemical parameter can be measured periodically. In our experiments, we try to tackle the following issue (as in [Fabrègue et al., 2014]):

Can hydro-ecologists explain biological values from physico-chemical ones occurring in past months, and thus to improve the global assessment of the quality of the aquatic ecosystem?

Precisely, given sequential data that represent hydro-ecological sequences of biological and physico-chemical samples, we try to make sense of them by using hierarchies of multi-level cpo-patterns (obtained with RCA-Seq) that summarise the impact of physico-chemical values on biological ones. Moreover, we try to facilitate the pattern evaluation step using the measures of interest presented in Chapter 5.

[1] `http://engees-fresqueau.unistra.fr/presentation.php?lang=en`
[2] `http://obs-rhin.engees.eu`

The issue that we try to tackle is relevant for hydro-ecologists since the biological state of water determines its quality. In addition, there are several works based on data mining techniques, e.g. [Goethals et al., 2007], [Dakou et al., 2007] and [Kocev et al., 2010], that highlight the non-triviality of this task. Actually, Fabrègue et al. [2014] introduced an approach devised during the Fresqueau project for extracting cpo-patterns from sequences of biological and physico-chemical samples. However, the evaluation step of these cpo-patterns is difficult since: (i) they are unorganised, (ii) they capture only the order on itemsets from the analysed data and (iii) these cpo-patterns do not provide a global view of the extracted regularities.

The REX project collected data about the restoration projects undertaken along the Rhine river in the Alsace plain. These data are about past restoration projects, temporal evolution of the water quality (biological indicators and physico-chemical parameters) and pressures (e.g. land use). The monitored river sites induce a river site network (that can be seen as a graph of river sites linked by the spatial relation *is downstream of*). In our experiments, we try to address the following issue:

Can hydro-ecologists explain the necessity or the effect of river site restorations by assessing the quality of water and land use aspects of upstream river sites?

Precisely, by using real data from the REX project we want to show the applicability of RCA-SEQ to real-life heterogeneous sequential data.

7.2 Description of Hydro-Ecological Data

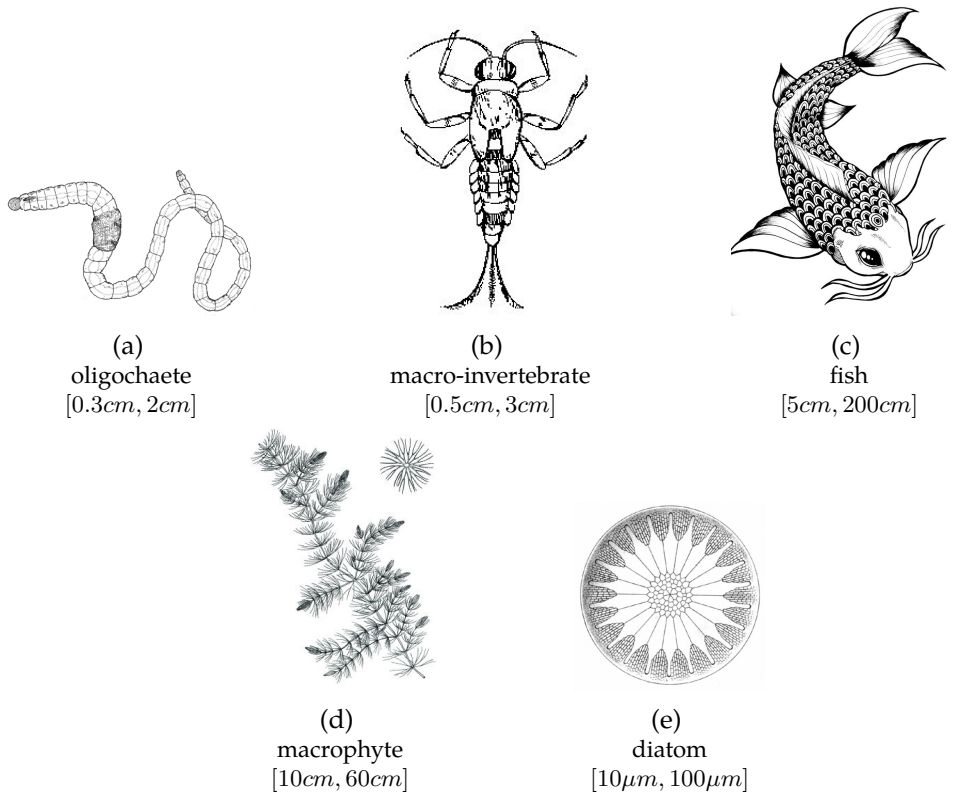
In the following, we aim to contextualise and to familiarise the reader with the hydro-ecological domain. Briefly, we present the biological, physico-chemical and land use data.

7.2.1 Biological Data

These data deal with the animals and plants living in watercourses. There are several biological groups, e.g. oligochaetes (small worms living in sediments) (Fig. 7.1a), macro-invertebrates (Fig. 7.1b), fishes (Fig. 7.1c), macrophytes (macroscopic plants living in water) (Fig. 7.1d) and diatoms (microscopic algae) (Fig. 7.1e).

In France, five biological indicators have been normalised to assess the quality of watercourses:

Standardised Global Biological Index (IBGN, [AFNOR, 2004a]) assesses the quality of watercourses by analysing macro-invertebrates. Precisely, some macro-invertebrates are sensitive to water pollution, while other ones are not. This index gives an overall estimation of the water ecosystem quality. The index score ranges from 0 to 20, where 0 represents a very bad quality of water and 20 a very good quality of water;



(a) oligochaete $[0.3cm, 2cm]$ (b) macro-invertebrate $[0.5cm, 3cm]$ (c) fish $[5cm, 200cm]$

(d) macrophyte $[10cm, 60cm]$ (e) diatom $[10\mu m, 100\mu m]$

Figure 7.1: Examples of flora and fauna and their size ranges

Fish Biotic Index (IPR, [AFNOR, 2004b]) allows to assess the chemical and physical water qualities by analysing fish species. This index measures the discrepancy between the actual fish state from a river site and the ideal fish state. An index score of 0 designates no discrepancy between the current and the ideal fish states. When the index score increases there is a more important discrepancy between the fish states;

Biological Index of Diatoms (IBD, [AFNOR, 2007]) gives an estimation of the global quality of water by analysing microscopic algae. The index score ranges from 1 to 20, where 1 represents a very bad quality of water and 20 a very good quality of water;

Oligochaete Index of Sediment Bioindication (IOBS, [AFNOR, 2002]) gives an evaluation of the sediment quality. The index score ranges from 0 to 10, where a value greater than 6 indicates a very good quality of sediments;

Biological Macrophyte Index of Rivers (IBMR, [AFNOR, 2003]) estimates the trophic level of water. Similar to IPR, IBMR measures the discrepancy between the current and the ideal state of macrophytes. The index score ranges from 0 to 20, where 20 is specific to a very good quality of the aquatic ecosystem.

In this thesis, we focus on the first three biological indicators since they are the most used

by hydro-ecologists. The values of these indicators are computed by using the biological samples taken (about) once a year from each monitored river site.

7.2.2 Physico-Chemical Data

There is a huge number of physical and chemical parameters of water, e.g. pH, temperature, nitrates, organic matter and pesticides. For example, the Fresqueau data gather more than 900 such parameters. These parameters impact the life cycle of the aquatic flora and fauna. For example, piscivorous fishes (i.e. fish feeding on fish) eat small fishes; small fishes eat macro-invertebrates that eat diatoms. Diatoms are sunlight-dependent and nutrient-dependent. Therefore, a contamination of water by excessive inputs of nutrients can disturb the food chain since it can cause the disappearance of some diatom species and/or the abundance of some other ones.

The physico-chemical data encompass two groups of parameters:

Macro-pollutants (mg/l metric unit) that naturally exist in water. Some examples are organic matter (e.g. plant residues), particulate matter (e.g. soot and dust), nitrogenous matter (e.g. ammonium (NH_4^+), Kjeldahl nitrogen (NKJ) and nitrite (NO_2^-)) and phosphorous matter (e.g. total phosphorus (P) and orthophosphate (PO_4^{3-})). However, human activity can cause an excess of macro-pollutants (e.g. nutrients) by means of e.g. agricultural practices. This high concentration of nutrients (e.g. nitrate (NO_3^-) and PO_4^{3-}) effects the aquatic ecosystem [Almasri and Kaluarachchi, 2004] and may cause e.g. eutrophication.

Micro-pollutants ($\mu g/l$ metric unit) that do not naturally exist in water. In contrast to macro-pollutants, they are toxic at very low concentrations. They contain a large number of anthropogenic and natural substances [Luo et al., 2014]. Some examples are pesticides [Bermúdez-Couso et al., 2013], pharmaceuticals [Behera et al., 2011], heavy metals (that are natural or not) and industrial chemicals.

The values of these parameters are computed by using the physico-chemical samples usually taken monthly or every two months from each monitored river site.

7.2.3 Land Use Data

All types of land use, e.g. pavements, buildings and forests, effect positively or negatively the water quality. Forests and the areas covered naturally with vegetation minimise the chance of the rainfall to become run-off, and, besides, they increase the chance of the rainfall to be soaked into the soil. Therefore, the quality of water is good in the surrounding areas. In contrast, the areas covered with pavements and buildings cause high run-off charged with various elements (e.g. metals) that lead to a bad quality of water in the surrounding areas.

In the REX project, the land use around each monitored river site is assessed within two increasing buffers, precisely $100\ m$ and $500\ m$.

7.3 Hydro-Ecological Sequential Data

We focus on hydro-ecological data collected during the Fresqueau project that comprise biological and physico-chemical samples taken at fixed points (river sites) and repeated in time. The obtained sub-datasets contain sequences whose itemsets are ordered according to temporal relations and defined according to qualitative relations.

Table 7.1 shows some measurements made at 3 river sites, i.e. *S1*, *S2* and *S3*. The measurements are made at different timestamps only for the IBGN and IPR biological indicators and five physico-chemical parameters, i.e. NKJ, NH_4^+, NO_2^-, P and PO_4^{3-}. For instance, $3.987\ mg/l$ of NH_4^+ is measured in September 2011 (09/2011) at river site *S2*. An IBGN score of 18 is measured in June 2009 (06/2009) at river site *S1*.

Table 7.1: Examples from the hydro-ecological data collected during the Fresqueau project

River Site	Month/Year	Physico-Chemical Parameters					Biological Indicators	
		NKJ	NH_4^+	NO_2^-	P	PO_4^{3-}	**IBGN**	**IPR**
S1	09/2007	6.545	–	0.989	–	4	–	–
	11/2007	2.765	3.456	0.426	0.498	0.033	–	–
	12/2007	–	–	–	–	–	5	–
	03/2009	0.500	0.033	–	0.173	–	–	–
	04/2009	–	–	0.028	–	0.039	–	–
	06/2009	–	–	–	–	–	18	–
	03/2010	–	–	–	–	–	–	40
S2	09/2011	–	3.987	–	0.768	0.304	–	–
	12/2011	–	–	–	–	–	7	–
	05/2006	3.564	–	–	0.065	–	–	–
	07/2006	–	–	–	–	0.407	–	–
	09/2006	–	–	–	–	–	9	–
S3	01/2014	1.304	1.008	0.350	–	–	–	–
	03/2014	–	–	–	–	–	–	25
	05/2014	–	–	4.000	–	–	–	–

7.3.1 Data Preprocessing

We note that these raw hydro-ecological data contain only numerical values. For exploring such data, we transform them into qualitative sequential sub-datasets (as explained in Sect. 3.3) by applying the discretization and cleaning processes based on domain knowledge.

7.3.1.1 Data Discretization

The discretization aims at converting numerical values into qualitative ones and it is based on technical reports published by French water agencies. The biological indicators and the physico-chemical parameters have five qualitative values, i.e. *very good*, *good*, *medium*, *bad* and *very bad* represented respectively by the colours *blue*, *green*, *yellow*, *orange* and *red*. However, the discretization of the biological indicators uses a different standard from that of the physico-chemical parameters.

Standard AFNOR [AFNOR, 2002, 2003, 2004b, 2007, 2004a] is used for the biological discretization. Table 7.2 shows the discretization intervals for the biological indicators. For example, an IPR score of 26 is discretized as an *orange* qualitative value; an IBD score of 17 is discretized as a *blue* qualitative value.

Table 7.2: Domain knowledge: the discretization intervals for biological indicators according to the AFNOR standard

Indicator	Blue	Green	Yellow	Orange	Red
IBGN	[20,17]	(17,13]	(13,9]	(9,5]	(5,0]
IPR	[0,7]	(7,16]	(16,25]	(25,36]	(36,∞)
IOBS	[10,6]	(6,3]	(3,2]	(2,1]	(1,0]
IBMR	[20,17]	(17,13]	(13,9]	(9,5]	(5,0]
IBD	[20,17]	(17,13]	(13,9]	(9,5]	(5,0]

Standard SEQ-eau[3] groups the physico-chemical parameters into 15 macro-parameters, e.g. PAES (particulate matter), HAP (hydrocarbons) and MINE (minerals). For instance, in Tab. 7.1 the physico-chemical parameters NKJ, NH_4^+ and NO_2^- are grouped into the NITRO macro-parameter, while the physico-chemical parameters P and PO_4^{3-} are grouped into the PHOS macro-parameter. Table 7.3 shows the discretization intervals for the physico-chemical parameters. For example, $0.989\ mg/l$ of NO_2^- is discretized as an *orange* qualitative value; $4\ mg/l$ of PO_4^{3-} is discretized as a *red* qualitative value. The qualitative value of a macro-parameter represents the worst qualitative value obtained for the measured physico-chemical parameters grouped by this macro-parameter. For instance, a *blue* qualitative value of P and a *yellow* qualitative value of PO_4^{3-} are discretized as a *yellow* qualitative value of PHOS.

Table 7.4 is obtained by applying the discretization process to the raw hydro-ecological data illustrated in Tab. 7.1. We can note that the number of values (there are less columns) is significantly small thanks to the macro-parameters.

[3] `http://rhin-meuse.eaufrance.fr/IMG/pdf/grilles-seq-eau-v2.pdf`

Table 7.3: Domain knowledge: the discretization intervals for physico-chemical macro-parameters according to the SEQ-eau standard

Macro-parameter	Parameter	Blue	Green	Yellow	Orange	Red
NITRO	NH_4^+	[0,0.1)	[0.1,0.5)	[0.5,2)	[2,5)	[5,∞)
	NKJ	[0,1)	[1,2)	[2,4)	[4,10)	[10,∞)
	NO_2^-	[0,0.03)	[0.03,0.3)	[0.3,0.5)	[0.5,1)	[1,∞)
PHOS	PO_4^{3-}	[0,0.1)	[0.1,0.5)	[0.5,1)	[1,2)	[2,∞)
	P	[0,0.05)	[0.05,0.2)	[0.2,0.5)	[0.5,1)	[1,∞)

Table 7.4: The discretized hydro-ecological data obtained from Tab. 7.1

River Site	Month/Year	Physico-chemical Macro-parameters		Biological Indicators	
		NITRO	PHOS	IBGN	IPR
S1	09/2007	orange	red	–	–
	11/2007	orange	yellow	–	–
	12/2007	–	–	orange	–
	03/2009	blue	green	–	–
	04/2009	blue	blue	–	–
	06/2009	–	–	blue	–
	03/2010	–	–	–	red
S2	09/2011	orange	orange	–	–
	12/2011	–	–	orange	–
	05/2006	yellow	green	–	–
	07/2006	–	green	–	–
	09/2006	–	–	yellow	–
S3	01/2014	yellow	–	–	–
	03/2014	–	–	–	yellow
	05/2014	red	–	–	–

7.3.1.2 Data Cleaning

The cleaning process considers only relevant data by defining several constraints based on the advices given by hydro-ecologists. Thus, the only analysed physico-chemical samples are those taken within *4 months* before a biological sample, from the same river site. If there is no physico-chemical sample, then the biological sample is not analysed.

For instance, in Tab. 7.4 the biological measurement made in March 2010 (03/2010) at river site *S1* has no physico-chemical sample during the 4 months before, and thus is not considered. In addition, the physico-chemical measurement made in May 2014 (05/2014) at river site *S3* is not analysed since there is no biological sample after it.

7.3.1.3 Building Qualitative Sequential Sub-Datasets

From the preprocessed hydro-ecological data, we build qualitative sub-datasets of sequences as explained in Sect. 3.3.2. Briefly, we order temporally the samples for each distinct river site in Tab. 7.4 according to the values from the Month/Year column. Then, for each river site, we cut the obtained sequence out in hydro-ecological sequences based on an expert-defined time window, i.e. 4 months before a biological sample. Table 7.5 depicts the obtained sequences. For example, $\langle(\texttt{NITRO}_{\texttt{yellow}}\ \texttt{PHOS}_{\texttt{green}})(\texttt{PHOS}_{\texttt{green}})(\texttt{IBGN}_{\texttt{yellow}})\rangle$ means that the simultaneous occurrence of the items $\texttt{NITRO}_{\texttt{yellow}}$ and $\texttt{PHOS}_{\texttt{green}}$ is temporally followed by item $\texttt{PHOS}_{\texttt{green}}$ that is followed by item $\texttt{IBGN}_{\texttt{yellow}}$.

Table 7.5: The hydro-ecological sequences obtained from Tab. 7.4

Id	Sequence
1	$\langle(\texttt{NITRO}_{\texttt{orange}}\ \texttt{PHOS}_{\texttt{red}})(\texttt{NITRO}_{\texttt{orange}}\ \texttt{PHOS}_{\texttt{yellow}})(\texttt{IBGN}_{\texttt{orange}})\rangle$
2	$\langle(\texttt{NITRO}_{\texttt{blue}}\ \texttt{PHOS}_{\texttt{green}})(\texttt{NITRO}_{\texttt{blue}}\ \texttt{PHOS}_{\texttt{blue}})(\texttt{IBGN}_{\texttt{blue}})\rangle$
3	$\langle(\texttt{NITRO}_{\texttt{orange}}\ \texttt{PHOS}_{\texttt{orange}})(\texttt{IBGN}_{\texttt{orange}})\rangle$
4	$\langle(\texttt{NITRO}_{\texttt{yellow}}\ \texttt{PHOS}_{\texttt{green}})(\texttt{PHOS}_{\texttt{green}})(\texttt{IBGN}_{\texttt{yellow}})\rangle$
5	$\langle(\texttt{NITRO}_{\texttt{yellow}})(\texttt{IPR}_{\texttt{yellow}})\rangle$

To analyse these sequences we build qualitative sequential sub-datasets based on the biological indicators and their qualitative values. Precisely, all hydro-ecological sequences in a sub-dataset end with the same biological indicator having the same qualitative value. A survey on these sub-datasets is relevant for hydro-ecologists since they are interested in the impact of physico-chemical macro-parameters on the behaviour of the same biological indicator, for all possible qualitative values. For example, four sub-datasets can be built from the hydro-ecological sequences given in Tab. 7.5, i.e. IBGN orange (sequences 1 and 3), IBGN blue (sequence 2), IBGN yellow (sequence 4) and IPR yellow (sequence 5).

7.3.1.4 Modelling Qualitative Sequential Data

To explore such sequential sub-datasets and to build the RCA input, the data model depicted in Fig. 7.2 is used. The four rectangles represent the four sets of objects we manipulate, as follows: biological samples, physico-chemical samples, biological indicators and physico-chemical macro-parameters. We note that the set of biological indicators contains only one indicator per sub-dataset. The links between biological/physico-chemical samples and physico-chemical samples are highlighted by the temporal binary relation *is preceded by*. This temporal relation associates one sample with another one if the first sample is preceded in time by the second one, on the same river site. There is no temporal binary relation between biological samples since in this work we evaluate the impact of physico-chemistry on biology. The biological/physico-chemical samples are described only by the qualitative bi-

nary relations *has parameter blue*, *has parameter green*, *has parameter yellow*, *has parameter orange* and *has parameter red* that link the biological/physico-chemical samples with the measured biological indicators/physico-chemical macro-parameters. For instance, in Tab. 7.4 the qualitative relation *has parameter green* links the physico-chemical sample taken at the *S1* river site in March 2009 (03/2009) with the PHOS macro-parameter.

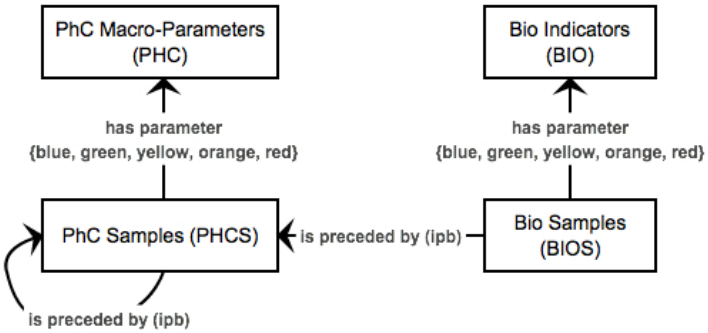


Figure 7.2: The modelling of hydro-ecological sequential data collected during the Fresqueau project [Nica et al., 2016a]. Bio and PhC stand respectively for biological and physico-chemical

7.3.2 Experiments – Performance and Quantitative Results

In this section we experimentally evaluate the RCA-Seq approach on various hydro-ecological sub-datasets. We present some quantitative statistics resulting from these experiments. First, we discuss the tools and algorithms underlying RCA-Seq, which are used to explore both the Fresqueau and REX sub-datasets. Second, we present a performance study of RCA-Seq. Then, we assess the exploration, extraction and selection steps of RCA-Seq. In addition, we empirically show that with RCA-Seq we obtain directly the minimal representations of the extracted multilevel cpo-patterns and we compare the stability index with the distribution index of a formal concept.

7.3.2.1 Tools and Algorithms

This thesis relies on the RCAExplore[4] tool (implemented in Java) that provides a user interface for manually creating/updating the RCA input and for visualising the obtained family of concept lattices. The iterative RCA process is based on the algorithm proposed by Rouane-Hacene et al. [2013]. The novelty of this tool is the interactive exploration of the data, i.e. at each iteration the user can choose the considered formal contexts, scaling quantifiers and algorithms used to build concept lattices. However, in this thesis we do not explore sequential data in an interactive way. We note that in this thesis the algorithm AddExtent [Merwe

[4] http://dolques.free.fr/rcaexplore

et al., 2004] is used to build concept lattices.

RCAExplore allows us to export the families of lattices obtained for all iterations of the RCA process to an XML file that, in this thesis, is used for further analysis. It is worthwhile to mention that the size of a generated XML file can be huge even for small datasets since the file is formatted to be easily readable by humans. Hence, since we are interested only in the fix point of the RCA process, and, besides, to avoid cases when an "OutOfMemoryError" is thrown during the construction of the XML file, we propose to modify the way in which this file is created and formatted. Precisely, we store only the family of lattices obtained in the last step and we remove the indentations, the new lines and we use maximum 3 characters for labelling the XML elements and attributes. For instance, by exploring with RCAExplore a sequential dataset with 86 sequences comprising a total number of 283 itemsets, which are built from a set of 5 items, an XML file of 25.13 GB is generated, while by using our modifications we obtain a file of 1.66 GB.

Furthermore, each hydro-ecological dataset used to validate the RCA-Seq approach is automatically preprocessed and encoded into the RCA input by means of an algorithm that we have developed and implemented in Java 8. The algorithm CPOHrchy (Sect. 4.3.1) used to extract multilevel cpo-patterns has also been developed in Java 8.

7.3.2.2 Study of the RCA-Seq Performance

The performance of the RCA-Seq approach is not our main concern. However, in the following we present a performance study regarding the execution time and the scalability of our approach. The relational scaling mechanism relies on the $\exists$ quantifier. The experiments were carried out on a MacBook Pro with 2.9 GHz Intel Core i7, 8 GB DDR3 RAM, running OS X 10.9.5.

We use two hydro-ecological sequential sub-datasets from the Fresqueau project whose characteristics, namely the number of sequences, the number of itemsets, the number of items, the average sequence length (the number of itemsets in a sequence) and the maximum sequence length, are shown in Tab. 7.6.

Table 7.6: The characteristics of two Fresqueau sub-datasets. #sequences is the number of sequences; #itemsets is the number of itemsets; #items is the number of items

Sub-dataset	#sequences	#itemsets	#items	Average sequence length	Maximum sequence length
IBD blue	1196	3012	46	2.51	7
IPR blue	1102	3077	26	2.79	8

Figure 7.3 shows how RCA-Seq scales up as the number of analysed sequences is in-

creased. To this end, we set $\theta = 20\%$ for the IPR blue sub-dataset, $\theta = 12\%$ for the IBD blue sub-dataset (where θ is the minimum support defined for the $\mathcal{L}_{K_M}$ main lattice) and we replicate the analysed sequences from 1 to 7 times for both sub-datasets. Generally, the execution time is linked to the RCA-based exploration step since the extraction one (CPOHrchy) takes ≈ 0.5 seconds. It can be observed that the execution time of RCA-SEQ scales almost linearly with the input size. For example, the execution time increases for the IBD blue sub-dataset from 156 to 278 to 430 seconds when the number of sequences is replicated from 2 to 3 to 4 times.

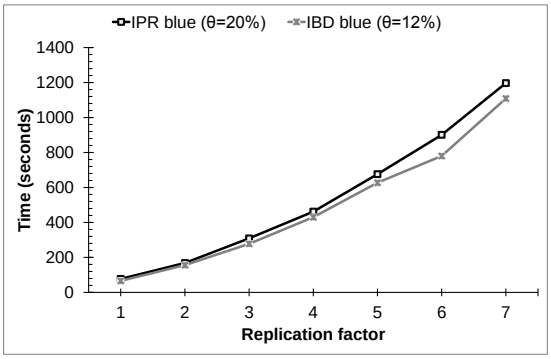
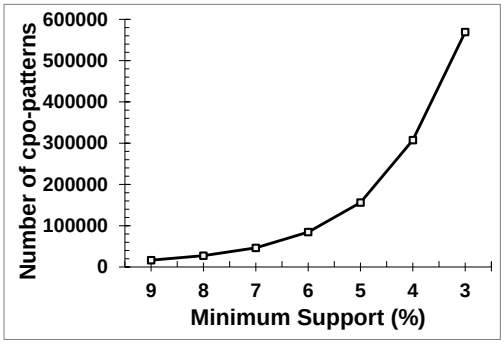


Figure 7.3: Scalability test (number of analysed sequences)

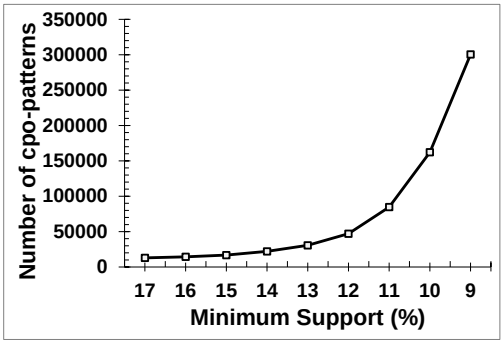
Figures 7.4a and 7.4b depict the number of the discovered multilevel cpo-patterns with respect to the minimum support θ (%) in the IBD and IPR blue sub-datasets. Even if both sub-datasets have almost the same number of sequences, the extracted number of multilevel cpo-patterns varies. For instance, 300411 multilevel cpo-patterns are discovered in the IPR blue sub-dataset with $\theta = 9\%$, while for the same minimum support in the IBD blue sub-dataset only 16525 multilevel cpo-patterns are discovered. This difference can be linked to each sub-dataset heterogeneity (e.g. the number of items, the repetitive occurrences of these items). In addition, for such small sub-datasets the number of extracted cpo-patterns is comparable to the one discovered in voluminous benchmark sub-datasets since RCA-SEQ discovers almost all combinations of the concrete and abstract items as explained in Sect. 7.3.2.4. For example, in [Fabrègue et al., 2015] is reported a number of $\approx$ 75000 cpo-patterns obtained with $\theta = 0.06\%$ in dataset *Gazelle*[5] that contains 59601 sequences built from 497 items and having an average sequence length of 2.51. In Fig. 7.4a we report a number of 569202 multilevel cpo-patterns discovered with $\theta = 3\%$ in a small sub-dataset that contains only 1196 sequences built from 46 items and having the same average sequence length of 2.51.

Figures 7.4c and 7.4d illustrate the execution time of the RCA-based exploration step with and without optimisation (as explained in Sect. 4.3) for both the IBD and IPR blue

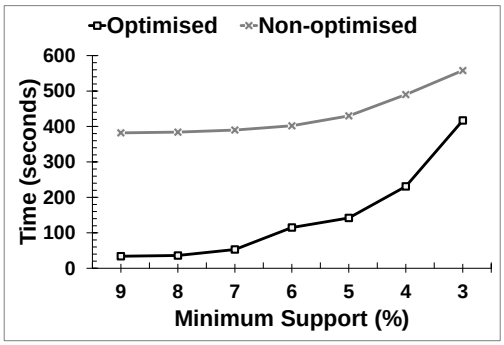
[5] `http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php`



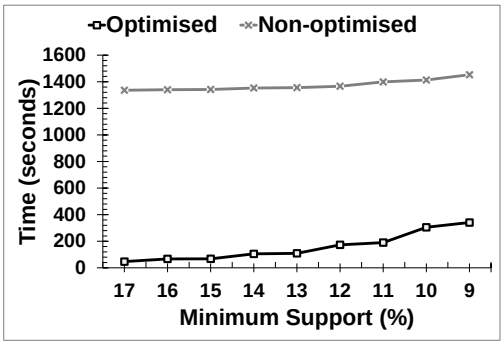
(a) Multilevel cpo-patterns (IBD blue)



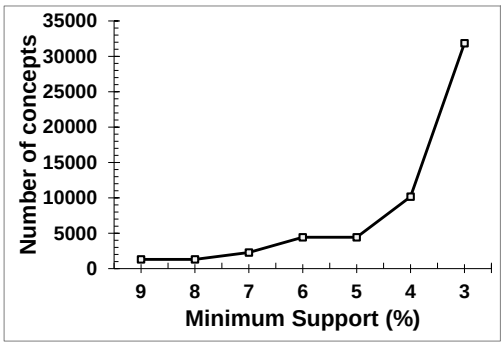
(b) Multilevel cpo-patterns (IPR blue)



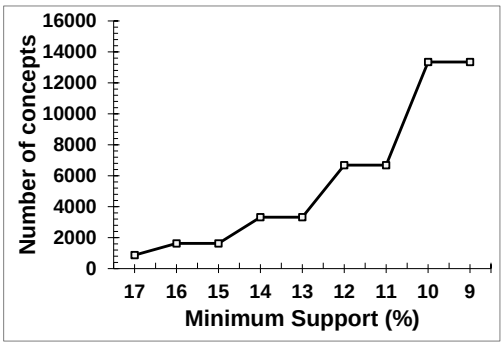
(c) RCA-based exploration (IBD blue)



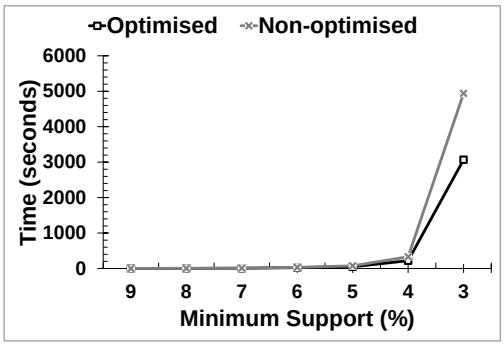
(d) RCA-based exploration (IPR blue)



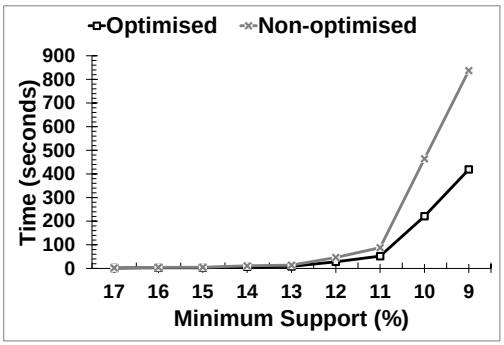
(e) $\mathcal{L}_{K_T}$ (IBD blue)



(f) $\mathcal{L}_{K_T}$ (IPR blue)



(g) CPOHrchy (IBD blue)



(h) CPOHrchy (IPR blue)

Figure 7.4: Performance evaluation; the minimum support (θ) is defined for the $\mathcal{L}_{K_M}$ main lattice; $\mathcal{L}_{K_T}$ is the temporal lattice (Sect. 4.2.1)

sub-datasets. On the horizontal axis is the minimum support θ (%). For instance, the non-optimised approach applied to the IBD blue sub-dataset (Fig. 7.4c) has no minimum support defined for the $\mathcal{L}_{K_T}$ temporal lattice and during each iterative step the relational scaling mechanism processes up to 105850 temporal concepts even if not all of them are used to extract cpo-patterns. In contrast, the optimised approach uses a minimum support $\theta' = \theta\frac{|G_M|}{|G_T|}$ for the temporal lattice, where G_M and G_T are respectively the set of objects of the K_M and K_T formal contexts (Sect. 4.2.1). Figure 7.4e shows the smaller number of derived temporal concepts for the IBD blue sub-dataset when using θ (horizontal axis) and θ'. In addition, less memory is used. For instance, when $\theta = 6\%$ ($\theta' = 3\%$) only 4429 temporal concepts are derived; $\theta = 3\%$ ($\theta' = 1\%$) then 31854 temporal concepts are derived. Thus, for $\theta = 6\%$ and $\theta = 3\%$ the optimised RCA-based exploration is respectively 3.49 and 1.33 times faster than the non-optimised one.

Similarly, for the IPR blue sub-dataset the non-optimised approach derives 933968 temporal concepts and the execution time is ≈ 9000 seconds when θ' is not defined. Thus, in Fig. 7.4d we report for the non-optimised approach the execution time obtained with $\theta' = 2\%$ that leads to only 149373 temporal concepts. Figure 7.4f illustrates a significant decrease in the number of derived temporal concepts for the IPR blue sub-dataset when using the optimised RCA-based exploration. For example, when $\theta = 15\%$ ($\theta' = 8\%$) only 1627 temporal concepts are derived; $\theta = 9\%$ ($\theta' = 5\%$) only 13348 temporal concepts are derived. Thus, for $\theta = 15\%$ and $\theta = 9\%$ the optimised RCA-based exploration is respectively 19.73 and 4.26 times faster than the non-optimised one.

Figures 7.4g and 7.4h show the computation time of the algorithm CPOHrchy with or without optimisation. In both the IBD and IPR blue sub-datasets the extraction step relies on the navigation space ($\mathcal{L}_{K_T} = (\mathcal{C}_{K_T}, \preceq_{K_T})$) depicted in Fig. 7.4e and 7.4f, respectively. When a temporal concept is navigated for distinct cpo-patterns, the non-optimised CPOHrchy searches its adjacent concepts in $\mathcal{L}_{K_T}$ and derives its itemset each time. Conversely, the optimised version of the algorithm saves the computed information for later use. In both sub-datasets it is noted that low values of θ and large temporal lattices $|\mathcal{C}_{K_T}| \geq 10000$ (Fig. 7.4e and 7.4f) slow down the extraction step. In addition, the efficiency of the CPOHrchy algorithm can be influenced by the used implementation[6], which is not currently optimised for searching in large collections.

7.3.2.3 Exploring Hydro-Ecological Sequential Data

The hydro-ecological sub-datasets obtained as explained in Sect. 7.3.1.3 are encoded into the RCA input based on the data model shown in Fig. 7.2. The relational scaling mechanism employed by the RCA process relies on the $\exists$ quantifier to derive the temporal and qual-

[6]based on Java Collection Framework and Lambda Expressions

itative links between samples and samples/biological indicators/physico-chemical macro-parameters. We note that there is no lattice of biological indicators since all the biological samples from a sub-dataset measure the same biological indicator having a specific qualitative value (e.g. a blue IBD). The RCA output comprises three lattices, one for each entity of the data model, i.e. the lattice of biological samples (the $\mathcal{L}_{K_M}$ main lattice), the lattice of physico-chemical samples (the $\mathcal{L}_{K_T}$ temporal lattice) and the lattice of physico-chemical macro-parameters (the $\mathcal{L}_{K_I}$ lattice of items). Furthermore, $\mathcal{L}_{K_M}$ is an Iceberg [Stumme, 2002] lattice since a user-defined minimum support θ (%) is used.

In Tab. 7.7 six hydro-ecological sequential sub-datasets are analysed. Each sub-dataset concerns only one biological indicator, namely IPR, IBD or IBGN, having the *yellow* or *green* qualitative value. These sub-datasets are interesting since the *yellow* (medium) qualitative value represents the threshold between the good ecological state and the bad ecological state of the aquatic ecosystem; the *green* qualitative value represents the good ecological state of the aquatic ecosystem. Other qualitative values were also analysed and we discuss them in the following sections.

Table 7.7: The results of exploring the IPR, IBD and IBGN yellow and green sub-datasets with $\theta = 0\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the lattice of physico-chemical samples ($\mathcal{L}_{K_T}$)

Sub-dataset				**RCA**	
Indicator	**Quality**	**Samples**		**Output**	
		Bio	**PhC**	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$
IPR	yellow	80	194	35699	39605
IBD				32146	20947
IBGN				9414	11580
IPR	green	69	183	26323	12102
IBD				60447	32927
IBGN				8312	15190

By analysing the quantitative results shown in Tab. 7.7, hydro-ecologists can infer some characteristics of the explored data. The results in the $\mathcal{L}_{K_M}$ column, i.e. the number of concepts from the main lattice, show that the numbers of derived concepts for the IBGN yellow and IBGN green sub-datasets are on average respectively about 3 and 5 times smaller than the number of derived concepts for the IPR and IBD yellow and green sub-datasets. These numbers reveal greater heterogeneity in the IPR and IBD sub-datasets in contrast with the IBGN sub-datasets. In addition, this heterogeneity can be linked to the fact that the IBD and IPR sequences rely on more items (46) than the IBGN sequences (26). Consequently, the cpo-patterns linking the physico-chemical macro-parameters and the IBGN biological indi-

cator are more frequent, and thus will provide more reliable knowledge of the impact of the physico-chemical macro-parameters on the medium or good ecological state of the aquatic ecosystem, as revealed by IBGN.

7.3.2.4 Analysing the Structure of the Discovered Hierarchies of Multilevel CPO-Patterns

In Tab. 7.8 three hydro-ecological sequential sub-datasets are analysed. Each sub-dataset concerns only one biological indicator, namely IBD, IPR or IBGN, having the *orange* (bad) qualitative value. We use the same input size (69 biological and 183 physico-chemical samples) in order to deepen the analysis of the extracted hierarchies of cpo-patterns. Furthermore, each analysed sub-dataset is built with 11 sequences of 2 samples, 20 sequences of 3 samples, 20 sequences of 4 samples and 18 sequences of 5 samples. By analysing the $\mathcal{L}_{K_M}$ and $\mathcal{L}_{K_T}$ columns shown in Tab. 7.8, we can notice once more that the number of concepts generated for the IPR and IBD sub-datasets is greater than for the IBGN sub-dataset.

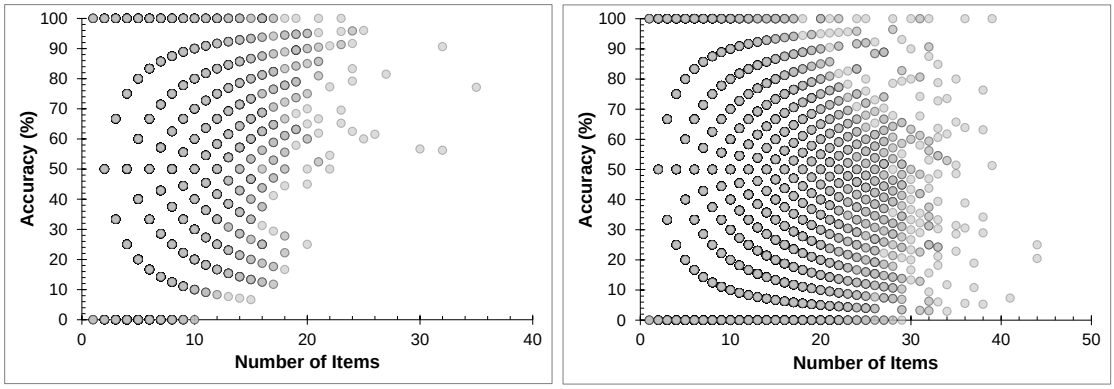
Table 7.8: The results of exploring the IBD, IPR and IBGN orange sub-datasets with $\theta = 0\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the lattice of physico-chemical samples ($\mathcal{L}_{K_T}$); column `CPO-patterns` represents the number of extracted cpo-patterns

Sub-dataset				**RCA**		CPOHrchy		
Indicator	**Quality**	**Samples**		**Output**		**CPO-patterns**		
		Bio	**PhC**	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$	**Concrete**	**Abstract**	**Hybrid**
IBD	orange	69	183	32570	20216	518	2953	29098
IPR				16973	20552	1360	253	15359
IBGN				8059	5668	570	648	6840

The `CPO-patterns` column illustrates the quite large numbers of concrete, abstract and hybrid cpo-patterns discovered in the analysed sub-datasets. In the following, we examine how the extracted hierarchies are organised according to the accuracy (Sect. 5.4), total number of items and the support of each type of cpo-pattern. We have to remind that the accuracy is in $(0\%, 100\%)$ for the hybrid cpo-patterns and it is equal to 0% and 100% for the abstract and concrete cpo-patterns, respectively.

Figure 7.5 illustrates the distribution of cpo-patterns in the extracted hierarchies for the IBGN and IBD orange sub-datasets according to their accuracies and their total number of items. The colour level of the points represents the number of cpo-patterns for a given accuracy and a given number of items. The high colour values in Fig. 7.5b are to be linked to the fact that the number of hybrid cpo-patterns is much higher for the IBD orange sub-dataset (29098) than for the IBGN orange one (6840). Conversely, the number of concrete

cpo-patterns (accuracy 100%) is greater for the IBGN orange sub-dataset than for the IBD orange one. For low values of the number of items (e.g. up to 17 items in Fig. 7.5b), we observe a symmetric distribution of the extracted cpo-patterns with respect to accuracy. It means that almost all combinations of the concrete and abstract items are represented when the cpo-patterns contain a few items. In addition, for higher values of the number of items, an asymmetric distribution of the extracted cpo-patterns is observed: there are less cpo-patterns with low accuracy (Fig. 7.5a) or with high accuracy (Fig. 7.5b). Finally, for high values of the number of items, e.g. ≥ 30 in Fig. 7.5b, we notice a sparse and irregular distribution of the extracted cpo-patterns since there are only a few (99 in Fig. 7.5b) such cpo-patterns.

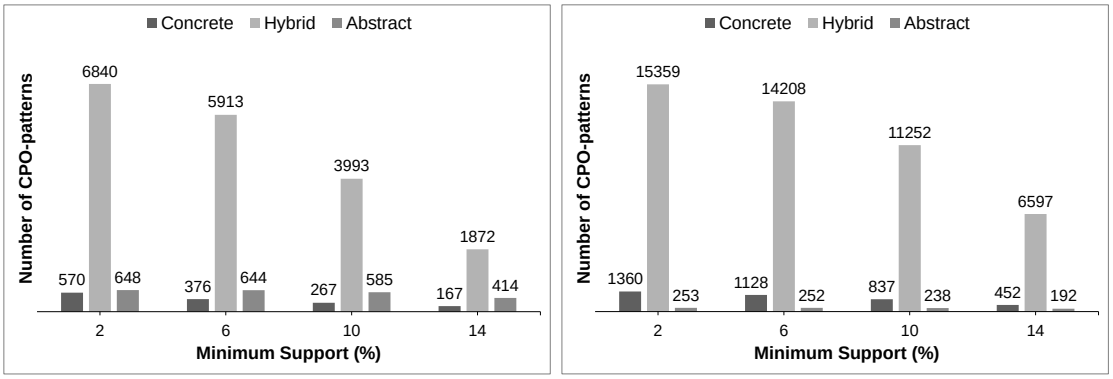


(a) IBGN orange (b) IBD orange

Figure 7.5: The distribution of cpo-patterns in the obtained hierarchies according to the cpo-pattern accuracies and their numbers of items

Figure 7.6 shows the number of concrete, hybrid and abstract cpo-patterns discovered in the IBGN and IPR orange sub-datasets for different support thresholds. The top of each bar in the histograms is labelled with the number of cpo-patterns extracted for the support threshold indicated on the horizontal axis. As expected, we can observe the decreasing trend of the number of multilevel cpo-patterns when the minimum support threshold increases. Both histograms suggest that as the minimum support increases, the number of concrete cpo-patterns decreases faster than the number of hybrid and abstract cpo-patterns. For instance, in Fig. 7.6b the number of concrete, hybrid and abstract cpo-patterns decreases on average by almost 29%, 23% and 8% respectively when the minimum support is increased by 4%. This is to be linked to the fact that the obtained hierarchies tend to concentrate the abstract cpo-patterns at the top, the hybrid ones in the middle and the concrete ones at the bottom.

As stated in Sect. 7.3.2.3, the IBGN sub-datasets provide more reliable knowledge for the assessment of the aquatic ecosystem. Hence, Tab. 7.9 details some quantitative statistics obtained for the IBGN sub-datasets (there are five sub-datasets one for each possible qualitative value of the biological indicator). A survey on these sub-datasets is relevant for hydro-



(a) IBGN orange (b) IPR orange

Figure 7.6: The number of multilevel cpo-patterns (concrete, hybrid and abstract) extracted at different support thresholds

ecologists since they are interested in the impact of the physico-chemical macro-parameters on the behaviour of the same biological indicator, for all possible qualitative values.

Table 7.9: The results of exploring the IBGN sub-datasets with $\theta = 5\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the lattice of physico-chemical samples ($\mathcal{L}_{K_T}$); column `CPO-patterns` represents the number of extracted cpo-patterns

Sub-dataset				**RCA**		CPOHrchy		
Indicator	**Quality**	**Samples**		**Output**		**CPO-patterns**		
		Bio	**PhC**	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$	**Concrete**	**Abstract**	**Hybrid**
IBGN	blue	80	164	7971	7997	284	46	7640
	green			2921	8825	342	177	2401
	yellow			6978	7706	224	694	6059
	orange			1742	2958	231	238	1272
	red			18080	20325	219	2979	14881

Figure 7.7 illustrates the distribution of the hybrid cpo-patterns extracted for all the sub-datasets given in Tab. 7.9 according to their accuracies. The height of each point in the plot represents the number of extracted hybrid cpo-patterns (in logarithmic scale) whose accuracies are in the interval indicated on the horizontal axis. We can observe that the hybrid cpo-patterns have an almost similar distribution in the 5 extracted hierarchies.

7.3.2.5 Verifying the Minimal Representations of the Extracted CPO-Patterns

In order to verify that we directly obtain the minimal representations of the extracted multilevel cpo-patterns, we rely on the merging and pruning steps presented by Fabrègue et al.

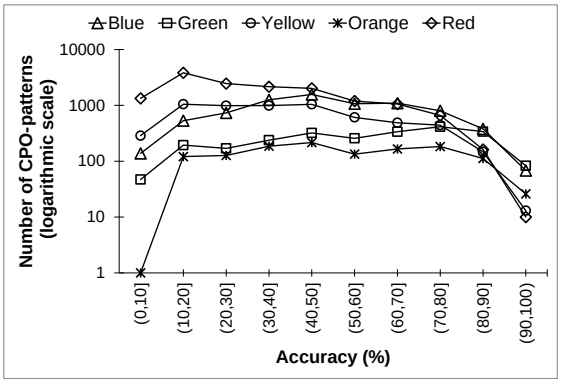
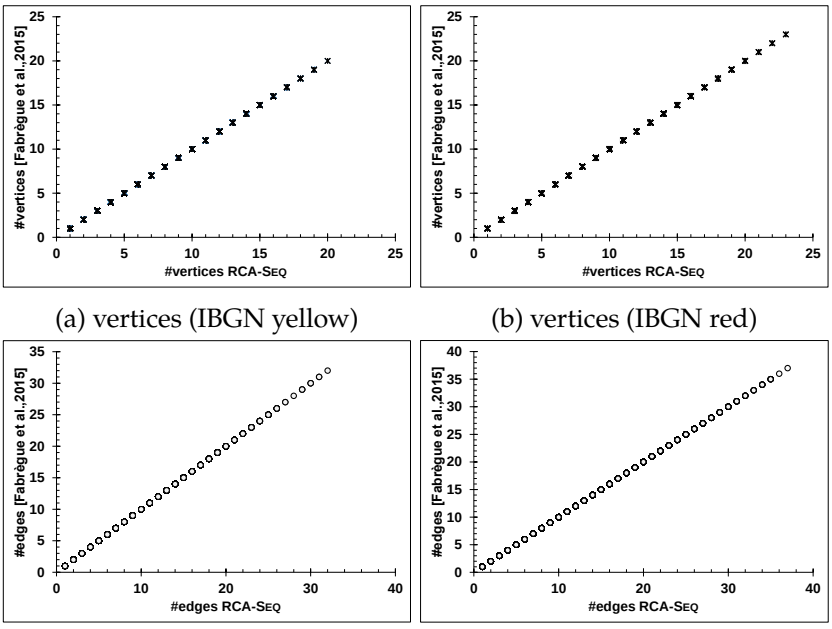


Figure 7.7: The distribution of the hybrid cpo-patterns (discovered in five distinct sub-datasets) with respect to their accuracies

[2015]. Figure 7.8 depicts the numbers of vertices and edges (before and after the merging and pruning steps) of the multilevel cpo-patterns discovered in the IBGN yellow and red sub-datasets given in Tab. 7.9.



(a) vertices (IBGN yellow) (b) vertices (IBGN red)

(c) edges (IBGN yellow) (d) edges (IBGN red)

Figure 7.8: The number of vertices (#vertices) and edges (#edges) obtained with RCA-SEQ or after merging and pruning steps [Fabrègue et al., 2015]

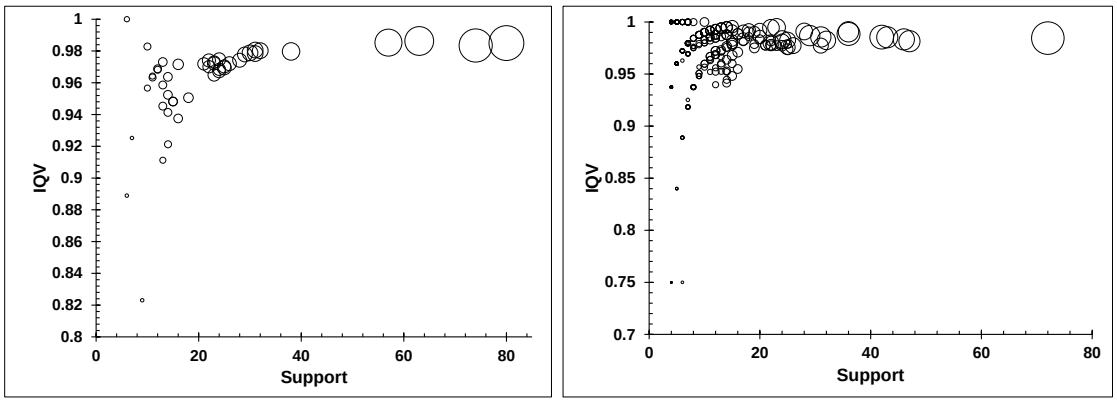
The numbers of vertices/edges obtained with RCA-SEQ are placed on the horizontal axis and the numbers of vertices/edges obtained after applying the [Fabrègue et al., 2015] steps are placed on the vertical axis. It is noted that there is no vertex or edge that should be merged or pruned, and thus by means of RCA-SEQ we obtain directly the minimal representations

of the extracted multilevel cpo-patterns without post-processing them.

7.3.2.6 Selecting Relevant CPO-Patterns

In this section, we show the results obtained only for the IBGN blue sub-dataset given in Tab. 7.9. This sub-dataset is interesting since the *blue* quality of the IBGN biological indicator represents the best ecological state of the aquatic ecosystem, and, besides, the size of the monitored geographical area (40 river sites) is appropriate for discovering global valid cpo-patterns, i.e. cpo-patterns that are available with the same frequency for many river sites.

In Tab. 7.9 the `CPO-patterns` column illustrates the quite large number of extracted concrete, abstract and hybrid cpo-patterns that should be ranked to ease their evaluation. To this end, we select relevant cpo-patterns based on the support, richness and distribution index measures (Sect. 5.3) of the associated main concepts. Figure 7.9 shows two scatter-plots of the distribution index (IQV) versus the support. A circle represents a cpo-pattern and its diameter is proportional to the richness (number of river sites) of the associated main concept.



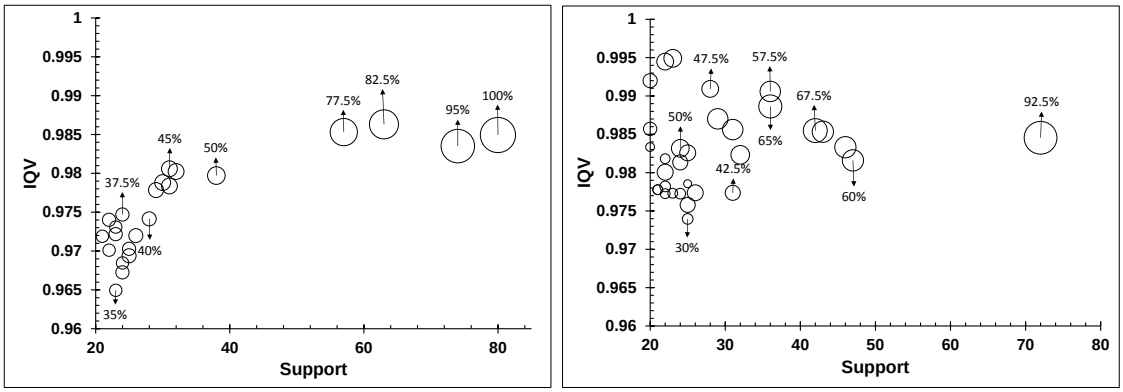
(a) abstract cpo-patterns (b) concrete cpo-patterns

Figure 7.9: CPO-patterns by the distribution index (IQV), support and richness (circle diameter) measures of the associated main concepts discovered in the IBGN blue sub-dataset (Tab. 7.9)

Using Fig. 7.9, hydro-ecologists first select a few abstract and/or concrete cpo-patterns based on high thresholds for the aforementioned measures. For example, by defining two thresholds $\tau_{IQV} = 0.96$ and $\tau_{Support} = 20$, the top-23 abstract and the top-32 concrete best-distributed and most-frequent cpo-patterns are selected.

To deepen the analysis, Fig. 7.10 is an excerpt from Fig. 7.9 that illustrates the selected top-23 abstract and top-32 concrete relevant cpo-patterns. The circle annotations represent percentages of the monitored geographical area (i.e. set of river sites). It is noted that the selected cpo-patterns cover more or less large geographical areas. Consequently, the cpo-patterns selected using the distribution index and support measures reveal frequent regu-

larities (that are well-distributed over more or less large geographical areas) of the physico-chemical macro-parameters showing a *blue* quality of the aquatic ecosystem. To select greater or smaller areas, the cpo-patterns are ranked by analysing the diameter of the circles. For instance, let us suppose that hydro-ecologists focus on the first 5 most-frequent cpo-patterns depicted in Fig. 7.10b. If they are interested in small geographical areas, the 5 cpo-patterns may be evaluated beginning from the 3 cpo-patterns that cover 60% of the monitored geographical area (a set of 40 river sites are monitored and the richness of an associated main concept is $\rho = 24$, then $\frac{24}{40}100 = 60\%$). In contrast, if they are interested in large geographical areas the cpo-pattern that covers 92.5% of the monitored area may be evaluated first.

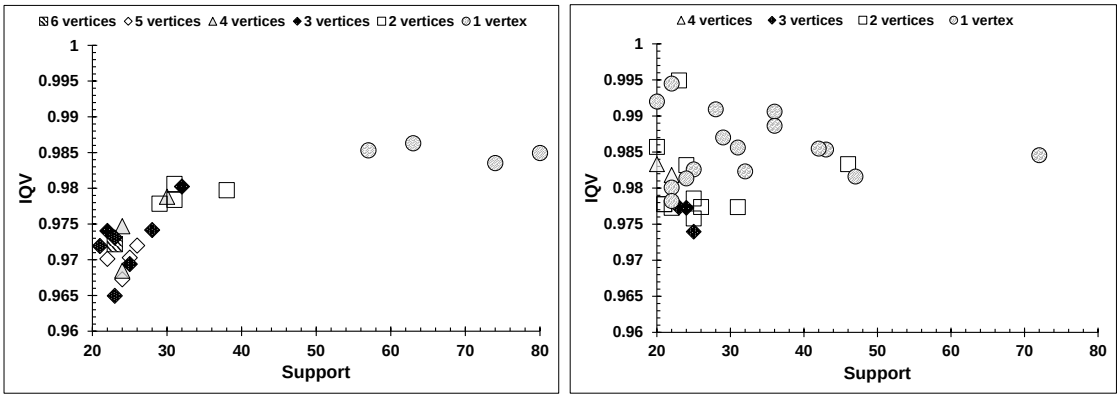


(a) abstract cpo-patterns (b) concrete cpo-patterns

Figure 7.10: The percentages of the monitored geographical area covered by the top-23 abstract and top-32 concrete relevant cpo-patterns discovered in the IBGN blue sub-dataset (Tab. 7.9)

Figure 7.11 illustrates the various numbers of vertices (except for the vertex labelled with the analysed biological indicator) of the selected top-23 abstract and top-32 concrete relevant cpo-patterns. These selected cpo-patterns are diverse in structure and reveal different frequent links between the physico-chemical macro-parameters and the biological indicator. Thus, the distribution index and support measures allow to select simple (e.g. having only one vertex) as well as more complex cpo-patterns that provide an overview of the abstract or concrete regularities from the entire monitored geographical area.

Briefly, the evaluation can continue following the hierarchy of cpo-patterns starting from the selected cpo-patterns – as we are going to detail in Sect. 7.3.3.1 – or by selecting more cpo-patterns based on lower thresholds of the aforementioned measures. For example, when $\tau_{IQV} = 0.9$ and $\tau_{Support} = 15$, the top-28 abstract and the top-54 concrete cpo-patterns are selected.



(a) abstract cpo-patterns (b) concrete cpo-patterns

Figure 7.11: The structure (the number of vertices except for the vertex labelled with the analysed biological indicator) of the top-23 abstract and top-32 concrete relevant cpo-patterns discovered in the IBGN blue sub-dataset (Tab. 7.9)

7.3.2.7 Comparing Distribution Index with Stability Index

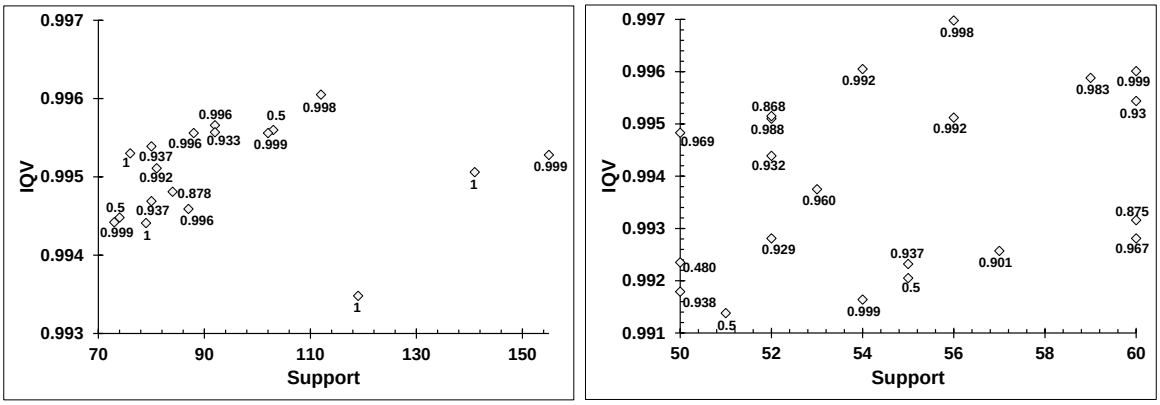
The stability index [Kuznetsov, 2007] is a well-known measure of interest that has been used in many FCA-based applications for selecting relevant formal concepts. In our case, i.e. hydro-ecological sequential data, it may show the likelihood of a cpo-pattern to still exist when several sequences that support it are ignored. In contrast, the distribution index (Sect. 5.3) of a concept shows the way in which the sequences that support the associated cpo-pattern are spread over the geographical area (the river sites from where the biological and physico-chemical samples were taken) monitored through these sequences. In the following, based on the sub-datasets given in Tab. 7.10, we try to analyse if the relevant concepts selected with the stability index can also be selected with the distribution index.

Table 7.10: The results of exploring the IBD blue and IBGN red sub-datasets with $\theta = 0\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the lattice of physico-chemical samples ($\mathcal{L}_{K_T}$)

Sub-dataset				**RCA**	
Indicator	**Quality**	**Samples**		**Output**	
		Bio	**PhC**	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$
IBD	blue	155	228	7980	1923
IBGN	red	80	155	14128	10862

Firstly, we try to analyse if the best-distributed main concepts may as well be the stable ones. Figure 7.12 shows, for different ranges of the support measure and a threshold $\tau_{IQV} = 0.99$, the best-distributed concepts (associated with the extracted cpo-patterns) from

the $\mathcal{L}_{K_M}$ main lattice discovered in the IBD blue sub-dataset (Tab. 7.10). A diamond is a main concept in $\mathcal{L}_{K_M}$; its label represents the stability of the concept. In Fig. 7.12a there are 18 best-distributed concepts for $Support \geq 70$; in Fig. 7.12b there are 21 best-distributed concepts selected for $Support \in [50, 60]$. It is noted in both scatter-plots that on average 79% of the selected best-distributed concepts have $Stability \geq 0.9$. Since only 90 out of 7980 concepts in $\mathcal{L}_{K_M}$ have $Stability \geq 0.9$, we can consider that for high values of the support measure the selected best-distributed concepts may as well be the most-stable ones.



(a) $Support \geq 70$ (b) $50 \leq Support \leq 60$

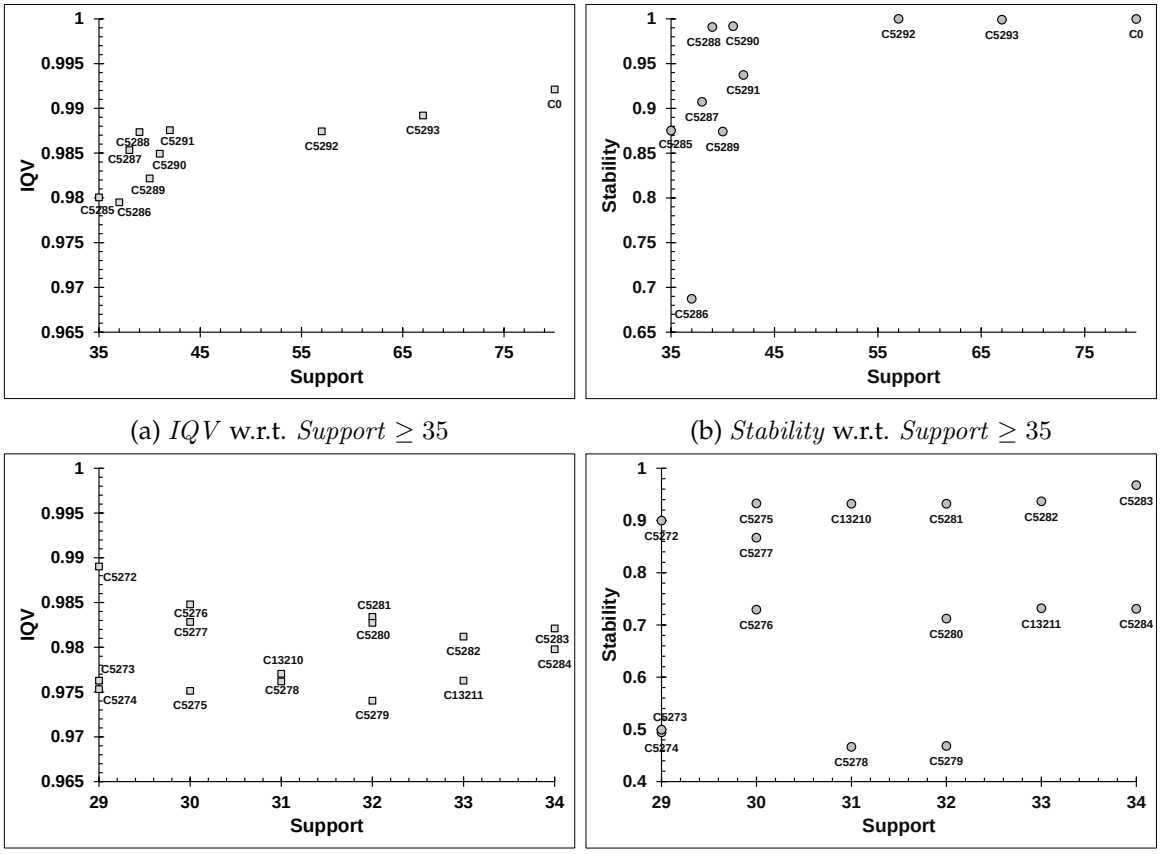
Figure 7.12: CPO-patterns by the distribution index (IQV), support and stability index (diamond labels) measures associated with the main concepts discovered in the IBD blue sub-dataset (Tab. 7.10)

Secondly, we focus on the IBGN red sub-dataset (Tab. 7.10). We try to analyse if the main concepts (associated with the extracted cpo-patterns) in $\mathcal{L}_{K_M}$ are ranked similarly with respect to the distribution or stability measure, for different ranges of the support measure. Figures 7.13a and 7.13c show two scatter-plots of the distribution index (IQV) versus the support when $Support \geq 35$ and $Support \in [29, 34]$, respectively. Similarly, Fig. 7.13b and 7.13d show two scatter-plots of the stability index versus the support when $Support \geq 35$ and $Support \in [29, 34]$, respectively. A square or a circle represents a main concept in $\mathcal{L}_{K_M}$; its label represents the UID of the main concept.

In Fig. 7.13a and 7.13b four out of five most relevant (best-distributed or most-stable) concepts are the same, precisely `C0`, `C5293`, `C5292` and `C5288`. It is noted that the distribution index helps to discriminate better these concepts, e.g. $Stability(\texttt{C0}) = Stability(\texttt{C5293}) = 0.999$ while $IQV(\texttt{C0}) = 0.992$ and $IQV(\texttt{C5293}) = 0.989$. Furthermore, 60% of the concepts are ranked identically in both figures, except for `C5290`, `C5291`, `C5285` and `C5289`.

In Fig. 7.13c and 7.13d for the same values of the support measure about 86% of the concepts are ranked in the same way, except for `C5275` and `C5276`. In addition, once more we can notice that the distribution index helps to discriminate better the concepts with approximately equal values of stability, e.g. $Stability(\texttt{C13210}) = 0.931$ and $Stability(\texttt{C5281}) = 0.932$

while $IQV(\texttt{C13210}) = 0.977$ and $IQV(\texttt{C5281}) = 0.983$.



(a) IQV w.r.t. $Support \geq 35$

(b) $Stability$ w.r.t. $Support \geq 35$

(c) IQV w.r.t. $29 \leq Support \leq 34$

(d) $Stability$ w.r.t. $29 \leq Support \leq 34$

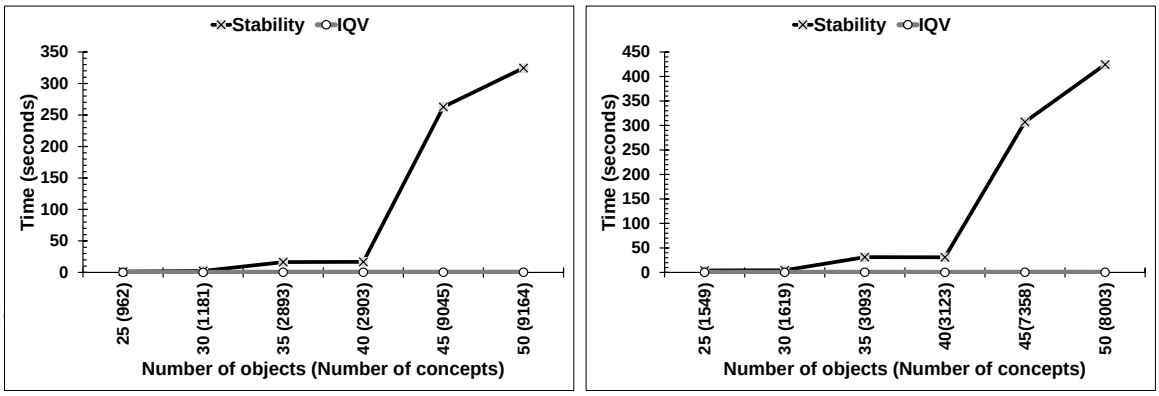
Figure 7.13: Ranking cpo-patterns by the distribution index (IQV), support and stability index measures associated with the main concepts discovered in the IBGN red sub-dataset (Tab. 7.10). The label of a square/circle represents the UID of a main concept

Furthermore, Fig. 7.14a depicts the execution time of the computation of the stability and distribution indices of the concepts in each main lattice obtained by increasing the number of analysed objects (sequences) from the IBGN red sub-dataset (Tab. 7.10). Similarly, Fig. 7.14b shows the same values for the IBD blue sub-dataset (Tab. 7.10).

Figure 7.14 shows that the computation of the stability index (as explained in [Roth et al., 2008]) is a time-consuming task in comparison to the computation of the distribution index. Indeed, Roth et al. [2008] showed that in the worst-case scenario the complexity of the computation of the stability index is quadratic in the size of the concept lattice. In contrast, in the worst-case scenario the complexity of the computation of the distribution index (using Eq. 5.1 when concept extents are already given) for all concepts in a lattice $\mathcal{L}_K$ built from a formal context $K = (G, M, I)$ is $O(m \cdot n)$ with $m = |\mathcal{L}_K|$ and $n = |G|$.

To sum up, when the analysed objects are identified by pairs $(Object, Date)$, we have

empirically observed that the distribution of concepts can be used to select relevant concepts (that tend to be stable) if the execution time is important.



(a) IBGN red (b) IBD blue

Figure 7.14: Performance study: the distribution (IQV) and stability indices of formal concepts

7.3.2.8 Pruning Irrelevant Multilevel CPO-Patterns During the Exploration Step

So far, we have shown that the number of extracted multilevel cpo-patterns as well as the number of derived formal concepts are quite large even for small hydro-ecological sub-datasets. This is due to the fact that we capture all possible temporal links between biological/physico-chemical and physico-chemical samples (itemsets) by using the $\exists$ quantifier during the RCA-based exploration step.

However, hydro-ecologists seem to be also interested in recurrent physico-chemical macro-parameters occurring in a sequence. To illustrate this, let us consider the hydro-ecological sequence depicted in Fig. 7.15.

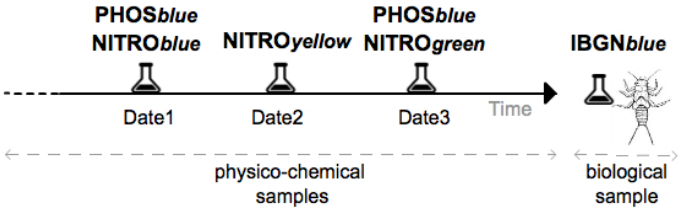


Figure 7.15: Hydro-ecological sequence

There are 3 physico-chemical samples that precede the biological one. Hydro-ecologists are interested in, e.g. discovering the qualitative values of the physico-chemical macro-parameters that occur before the blue IBGN in more than 50% of the monitored physico-chemical samples (i.e. > 1.5 samples). The regularity $\{\texttt{PHOS}_{\texttt{blue}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ (a blue IBGN

is preceded by a blue PHOS) is the only one that corresponds to this criterion in the sequence given in Fig. 7.15 since there are two physico-chemical samples for which a blue PHOS macro-parameter was measured. Contrarily, e.g. the regularity $\{\texttt{NITRO}_{\texttt{yellow}}\} \leftarrow \{\texttt{IBGN}_{\texttt{blue}}\}$ is valid only for one physico-chemical sample that precedes the biological one, and thus it should not be discovered.

To address this issue, we explore the hydro-ecological sequential data by using various quantifiers during the relational scaling mechanism as detailed in Sect. 6.2. Table 7.11 illustrates the different numbers of concepts generated for three sub-datasets, precisely IBD green, IBGN blue and IPR orange, by using the $\exists_{>n\%}$ quantifier with $n \in \{25, 50, 75\}$ for the temporal relations $ipb1 \subseteq G_M \times G_T$ (biological sample *ipb* physico-chemical sample) and $ipb2 \subseteq G_T \times G_T$ (physico-chemical sample *ipb* physico-chemical sample). We recall that G_M and G_T represent the sets of objects from the formal contexts used to build respectively the main lattice $\mathcal{L}_{K_M}$ and the temporal lattice $\mathcal{L}_{K_T}$ (Sect. 4.2.1).

Table 7.11: The results of exploring the IBD green, IBGN blue and IPR orange sub-datasets with $\theta = 10\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Relational scaling` represents the quantifier used during the relational scaling mechanism for the temporal relations $ipb1$ (biological sample *ipb* physico-chemical sample) and $ipb2$ (physico-chemical sample *ipb* physico-chemical sample); column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the one of physico-chemical samples ($\mathcal{L}_{K_T}$)

Sub-dataset				**RCA**			
Indicator	**Quality**	**Samples**		**Relational scaling**		**Output**	
		Bio	**PhC**	$ipb1 \subseteq G_M \times G_T$	$ipb2 \subseteq G_T \times G_T$	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$
IBD	green	1618	5269	$\exists_{>75\%}$	$\exists_{>75\%}$	85	22033
				$\exists_{>50\%}$	$\exists_{>75\%}$	1186	22033
				$\exists_{>50\%}$	$\exists_{>50\%}$	2275	346280
IBGN	blue	1106	2360	$\exists$	$\exists_{>75\%}$	44061	1406
				$\exists$	$\exists_{>50\%}$	90892	3209
				$\exists_{>50\%}$	$\exists_{>50\%}$	487	3209
IPR	orange	69	240	$\exists_{>50\%}$	$\exists$	1964	861146
				$\exists_{>25\%}$	$\exists_{>50\%}$	4831	16330
				$\exists$	$\exists_{>75\%}$	25142	6554

In Tab. 7.11, it is noted that:

– changing the quantifier used for $ipb1$ and at the same time preserving the one used for $ipb2$ leads to the same number of concepts for $\mathcal{L}_{K_T}$ and a different number of concepts for $\mathcal{L}_{K_M}$. The constant number of the $\mathcal{L}_{K_T}$ concepts is to be linked to the fact that $ipb2$ does not depend on the learnt concepts in $\mathcal{L}_{K_M}$. To illustrate this, for the IBD green sub-dataset by using the $\exists_{>75\%}$ quantifier for both the $ipb1$ and $ipb2$ temporal relations,

$\mathcal{L}_{K_M}$ contains 85 concepts and $\mathcal{L}_{K_T}$ contains 22033 concepts. Then, for the same sub-dataset by exploring it with the $\exists_{>50\%}$ quantifier for $ipb1$ and the same quantifier $\exists_{>75\%}$ for $ipb2$, $\mathcal{L}_{K_M}$ contains more concepts (1186), while $\mathcal{L}_{K_T}$ contains the same number of concepts;

– preserving the quantifier used for $ipb1$ and at the same time changing the one used for $ipb2$ leads to different numbers of concepts for both the $\mathcal{L}_{K_M}$ and $\mathcal{L}_{K_T}$ lattices. The variation of the number of the $\mathcal{L}_{K_M}$ concepts is to be linked to the fact that $ipb1$ depends on the learnt concepts in $\mathcal{L}_{K_T}$. To illustrate this, for the IBGN blue sub-dataset by using the $\exists$ quantifier for $ipb1$ and the $\exists_{>75\%}$ quantifier for $ipb2$, $\mathcal{L}_{K_M}$ contains 44061 concepts and $\mathcal{L}_{K_T}$ contains 1406 concepts. Then, for the same sub-dataset by exploring it with the same $\exists$ quantifier for $ipb1$ and the $\exists_{>50\%}$ quantifier for $ipb2$, $\mathcal{L}_{K_M}$ contains 90892 concepts and $\mathcal{L}_{K_T}$ contains 3209 concepts.

7.3.3 Experiments – Qualitative Assessment of the Extracted CPO-Patterns

In this section we present some qualitative interpretations resulting from experiments carried out on various hydro-ecological datasets. We highlight how hydro-ecologists are guided during the evaluation step by the obtained hierarchies of multilevel cpo-patterns, and, besides, by the weighted cpo-patterns.

7.3.3.1 Navigating a Hierarchy of Multilevel CPO-Patterns

Figure 7.16 is an excerpt from the hierarchy of cpo-patterns extracted from the IBGN blue sub-dataset given in Tab. 7.9. From (a) to (g) are the abstract cpo-patterns, from (h) to (j) are the hybrid cpo-patterns and from (k) to (s) the concrete ones.

This excerpt is subsumed by the (a) cpo-pattern that confirms the correctness of the pre-processing step, i.e. all the biological samples ($Support = 80$) from the RCA input are preceded in time by at least one physico-chemical sample. In addition, $IQV = 0.984$ shows a quite good distribution of the analysed biological samples over the monitored geographical area ($\rho = 40$ river sites). Beginning from this abstract cpo-pattern, hydro-ecologists can continue the navigation going down in the hierarchy. Both direct descendants, the (b) and (c) cpo-patterns, emphasize two well-known correspondences between the qualitative values of the physico-chemical macro-parameters and the ones of the IBGN biological indicator:

– cpo-pattern (b) highlights that $\texttt{IBGN}_{\texttt{blue}}$ is measured when it is preceded by the *blue* qualitative values of physico-chemical macro-parameters. This regularity is retrieved with 78.75% frequency of the analysed data ((b) has $Support = 63$ and the total number of sequences is equal to 80, then $Freq = \frac{63}{80}100 = 78.75\%$) and is available for 82.5% of the monitored geographical area ((b) has $\rho = 33$ and the monitored geographical area

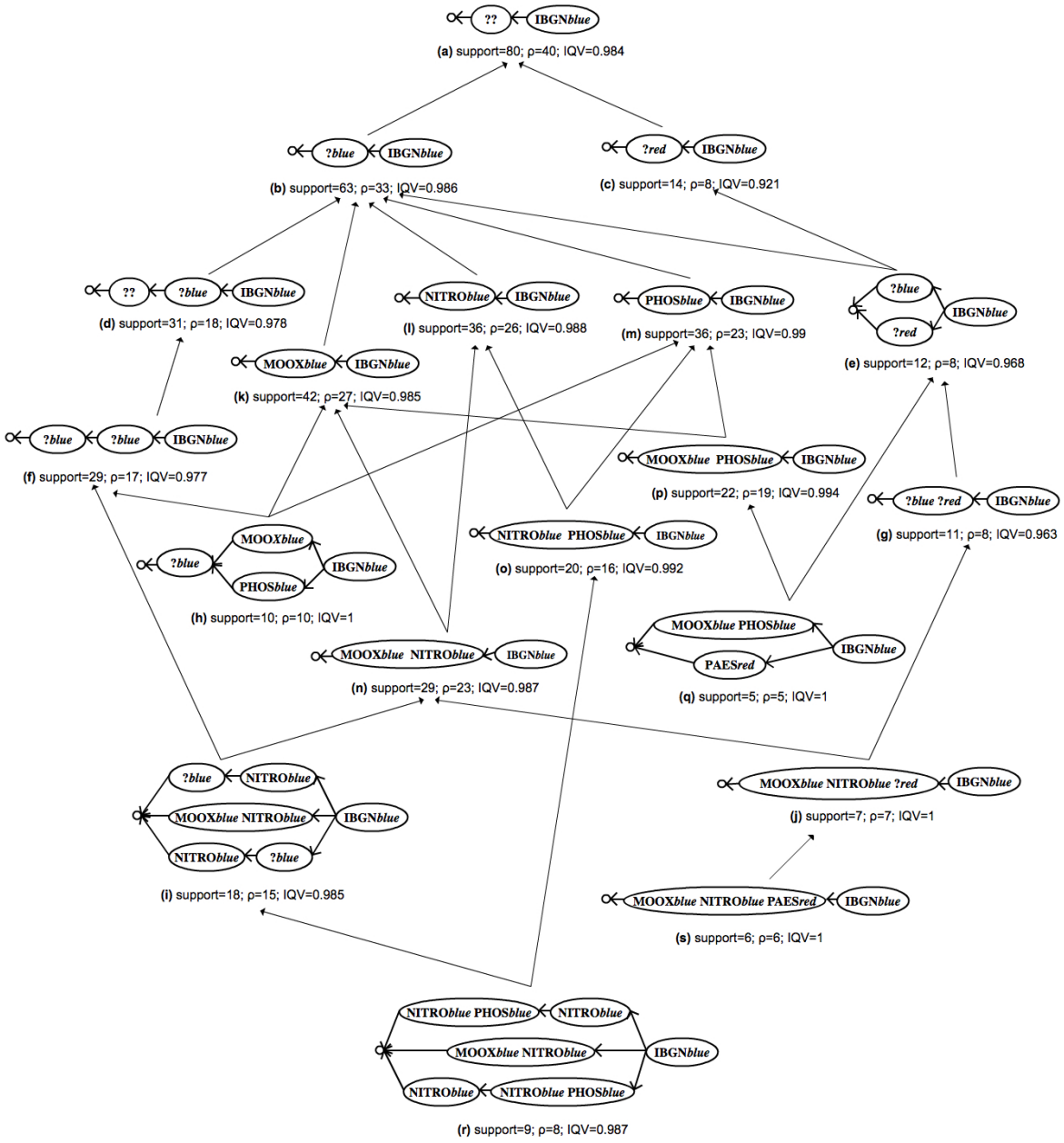


Figure 7.16: Excerpt from the hierarchy of cpo-patterns discovered in the IBGN blue sub-dataset given in Tab. 7.9. The support, richness (ρ) and distribution index (IQV) of each associated main concept are shown

contains 40 river sites, then $\frac{33}{40}100 = 82.5\%$). In addition, by analysing Tab. 7.12 it is noted that the frequency of the (b) cpo-pattern tends to decrease for negative qualitative values (yellow, orange and red) of IBGN;

– the measures associated with cpo-pattern (c) stress that the *red* physico-chemical macro-parameters are not frequently measured before $\texttt{IBGN}_{\texttt{blue}}$ since they show a degradation of the water quality and do not lead to a very good ecological state. This cpo-pattern is retrieved in 17.5% of the analysed data and covers only 20% of the monitored geographical area. As expected, in contrast with the (b) cpo-pattern, the (c) cpo-pattern has a low support and it is valid for a small percentage of the monitored geographical area. Besides, by analysing again Tab. 7.12 it is noted that the frequency of the (c) cpo-pattern decreases when the quality of IBGN increases (from red to orange, yellow and blue).

Table 7.12: The support, richness (ρ) and distribution index (IQV) of the (a) to (s) cpo-patterns (shown in Fig.7.16 but having different target qualitative values of IBGN) in the IBGN blue, green, yellow, orange and red sub-datasets given in Tab. 7.9

CPO-pattern	IBGN biological indicator														
	blue			green			yellow			orange			red		
	support	ρ	IQV	support	ρ	IQV	support	ρ	IQV	support	ρ	IQV	support	ρ	IQV
a	80	40	0.984	80	48	0.983	80	57	0.99	80	55	0.989	80	59	0.993
b	63	33	0.986	67	44	0.987	55	37	0.987	44	39	0.996	24	21	0.991
c	14	8	0.921	–	–	–	17	14	0.991	19	14	0.954	38	24	0.984
d	31	18	0.978	27	17	0.961	–	–	–	19	15	0.991	12	11	0.993
e	12	8	0.963	–	–	–	–	–	–	–	–	–	–	–	–
f	29	17	0.977	22	14	0.943	–	–	–	17	14	0.991	8	8	1
g	11	8	0.963	–	–	–	–	–	–	–	–	–	–	–	–
h	10	10	1	–	–	–	–	–	–	–	–	–	–	–	–
i	18	15	0.985	–	–	–	–	–	–	–	–	–	–	–	–
j	7	7	1	–	–	–	–	–	–	–	–	–	–	–	–
k	42	27	0.985	43	28	0.978	41	32	0.991	28	26	0.994	15	15	1
l	36	26	0.988	51	32	0.981	32	23	0.98	30	27	0.994	15	13	0.982
m	36	23	0.99	27	17	0.961	21	13	0.977	9	8	0.987	5	4	0.96
n	29	23	0.987	36	23	0.969	25	19	0.979	20	18	0.99	9	9	1
o	20	16	0.992	18	10	0.919	8	5	0.937	–	–	–	–	–	–
p	22	19	0.994	–	–	–	13	18	0.989	–	–	–	–	–	–
q	5	5	1	–	–	–	–	–	–	–	–	–	–	–	–
r	9	8	0.987	–	–	–	–	–	–	–	–	–	–	–	–
s	6	6	1	–	–	–	–	–	–	–	–	–	–	–	–

Therefore, hydro-ecologists can navigate the descendants of the (b) cpo-pattern in order to find patterns revealing the appropriate environment for the blue IBGN biological indicator that provide a very good ecological state of the aquatic ecosystem. In addition, hydro-ecologists can focus on the descendants of the (c) cpo-pattern to find out how the water quality is not impacted when red qualitative values of physico-chemical macro-parameters are measured.

By navigating the direct descendants of the (b) cpo-pattern hydro-ecologists can find three interesting concrete cpo-patterns, i.e. (k), (l) and (m). The (k) and (l) cpo-patterns have respectively 52.5% and 45% frequency and they cover respectively 67.5% and 65% of the monitored geographical area. In addition, the (k) and (l) cpo-patterns illustrate the well-known correspondence between $\texttt{IBGN}_{\texttt{blue}}$ and $\texttt{MOOX}_{\texttt{blue}}$, $\texttt{NITRO}_{\texttt{blue}}$ (that show an organic pollution).

The impact of the nutrient pollution, i.e. excessive nutrients (PHOS), on the IBGN qualitative values is a lesser-known fact highlighted by the (m) cpo-pattern with 45% frequency and covering 57.5% of the monitored geographical area.

Since $\texttt{MOOX}_{\texttt{blue}}$, $\texttt{NITRO}_{\texttt{blue}}$ and $\texttt{PHOS}_{\texttt{blue}}$ have individually relevant impact on $\texttt{IBGN}_{\texttt{blue}}$, hydro-ecologists are interested to know if their coexistence is also observed. Therefore, going down in the hierarchy the (n) cpo-pattern, which occurs with 36.25% frequency and covers 57.5% of the monitored geographical area, reveals the coexistence at the same time of $\texttt{MOOX}_{\texttt{blue}}$ and $\texttt{NITRO}_{\texttt{blue}}$; the (o) cpo-pattern, which occurs with 25% frequency and covers 40% of the monitored geographical area, reveals the simultaneous occurrence of $\texttt{NITRO}_{\texttt{blue}}$ and $\texttt{PHOS}_{\texttt{blue}}$; and the (p) cpo-pattern, which occurs with 27.5% frequency and covers 47.5% of the monitored geographical area, reveals the coexistence at the same time of $\texttt{MOOX}_{\texttt{blue}}$ and $\texttt{PHOS}_{\texttt{blue}}$. According to [Lafont et al., 2001] and [Mondy and Usseglio-Polatera, 2013], the IBGN biological indicator is sensitive to various pollutions (in particular, to macro-pollutants) without distinguishing them. Our cpo-patterns show a better answer for the organic pollution.

As aforementioned, the coexistence of $\texttt{MOOX}_{\texttt{blue}}$ with $\texttt{NITRO}_{\texttt{blue}}$ revealed by the (n) cpo-pattern is expected and indicates the absence of the organic pollution. By analysing Tab. 7.12, hydro-ecologists notice that (n) is discovered in all 5 analysed sub-datasets. Thus, hydro-ecologists can infer that only the absence of the organic pollution is not always enough to obtain a very good qualitative value of IBGN. In contrast, a surprising coexistence is revealed by the (o) and (p) cpo-patterns that still have good frequencies and indicate the absence of two pollutions, namely organic (MOOX or NITRO) and nutrient (PHOS). Furthermore, by looking over Tab. 7.12, hydro-ecologists can infer that the simultaneous absence of the organic and nutrient pollutions increases the cases when a very good qualitative value of IBGN is obtained (i.e. cpo-pattern (o) occurs only in 3 sub-datasets and cpo-pattern (p) in 2 sub-datasets).

The strong impact of the aforementioned coexistences of the *blue* physico-chemical macro-parameters is emphasized by the descendants of the (c) cpo-pattern. For instance, the (q), (j) and (s) cpo-patterns can highlight that the abiotic characteristics (the non-living chemical and physical parts) suitable for a very good ecological state of the aquatic ecosystem are not impacted by an accidental pollution (e.g. particulate matter ($\texttt{PAES}_{\texttt{red}}$)). Moreover, by analysing Tab. 7.12 it is as well noted that the (e), (g), (j), (q) and (s) cpo-patterns are not discovered for other qualitative values of IBGN.

In addition, thanks to the multilevel cpo-patterns extracted by using RCA-SEQ the hybrid (i) cpo-pattern, having 22.5% frequency, can be found when, e.g. a minimum support equal to $\theta = 15\%$ is used even if the accurate (r) cpo-pattern, which is a specialisation of (i), has only 11.25% frequency and it is not discovered.

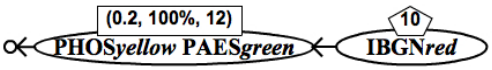
7.3.3.2 Analysing Multilevel Weighted CPO-Patterns

Table 7.13 shows some statistics regarding two hydro-ecological sub-datasets, precisely IBGN orange and IBGN red. Each analysed sub-dataset is built with 10 sequences of 2 samples, 10 sequences of 3 samples, 10 sequences of 4 samples and 15 sequences of 5 samples. In order to help hydro-ecologists to discriminate the cpo-patterns that are found simultaneously in the aforementioned sub-datasets, we extract multilevel wcpo-patterns.

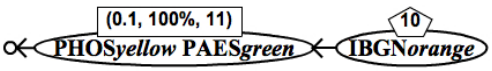
Table 7.13: The results of exploring the IBGN orange and red sub-datasets with $\theta = 10\%$. The columns `Bio` and `Phc Samples` represent respectively the number of analysed biological and physico-chemical samples; column `Output` represents the number of concepts from the lattice of biological samples ($\mathcal{L}_{K_M}$) and the one of physico-chemical samples ($\mathcal{L}_{K_T}$); column `WCPO-patterns` represents the number of extracted weighted cpo-patterns

Sub-dataset				RCA		CPOHrchy		
Indicator	Quality	Samples		Output		WCPO-patterns		
		Bio	PhC	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$	Concrete	Abstract	Hybrid
IBGN	orange	45	120	1022	1995	162	185	674
	red			8057	11138	103	1083	6870

Figures 7.17 and 7.18 depict two examples of multilevel wcpo-patterns discovered simultaneously in the IBGN red and the IBGN orange sub-datasets with the same support (⬠ on the first vertex). As defined in Sect. 5.5, each vertex is labelled with a 3-tuple $(\varpi_{v_t}, \varsigma_{v_t}, \omega_{v_t})$, i.e. $(persistency, specificity, overall\ weight)$.



(a) IBGN red



(b) IBGN orange

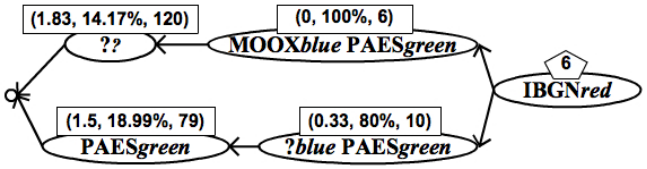
Figure 7.17: Two concrete wcpo-patterns discovered simultaneously in the IBGN red and the IBGN orange sub-datasets with the same $Support = 10$. ⬠ represents the wcpo-pattern support. The other vertices are labelled with 3-tuples $(\varpi_{v_t}, \varsigma_{v_t}, \omega_{v_t})$

The persistencies of the vertices from wcpo-patterns help hydro-ecologists to infer if the regularities are accidental or not. For instance, Fig. 7.17 reflects that if the nutrient pollution (e.g. $\texttt{PHOS}_{\texttt{yellow}}$) is too persistent in time the biocoenosis (flora and fauna) may lose its resilience capacity [Webster et al., 1983], and thus $\texttt{IBGN}_{\texttt{orange}}$ ($\varpi = 0.1$) becomes $\texttt{IBGN}_{\texttt{red}}$ ($\varpi = 0.2$). Furthermore, the overall weights of the vertices from the wcpo-patterns shown in Fig. 7.17 are quite similar in both sub-datasets. Itemset $(\texttt{PHOS}_{\texttt{yellow}}\texttt{PAES}_{\texttt{green}})$ has one extra

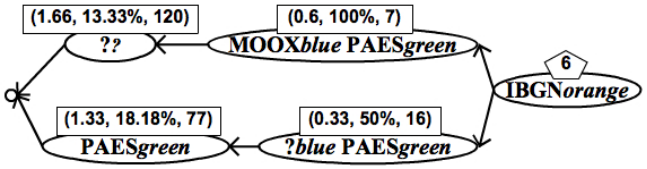
occurrence in the IBGN red sub-dataset ($\omega = 12$) than in the IBGN orange one ($\omega = 11$). Therefore, in this case the persistency and overall weight measures can be used to discriminate the cpo-patterns, but the extra occurrence may be accidental as well.

The specificities of the vertices from the wcpo-patterns shown in Fig. 7.18 seem to have a more discriminant power. By analysing the (b) wcpo-pattern, it is noted that three vertices have smaller specificity values than the values of the same three vertices from the (a) wcpo-pattern. Thus, the (b) wcpo-pattern reveals regularities available for many analysed sequences. For example, the regularity $\{\texttt{PAES}_{\texttt{green}}\} \leftarrow \{?_{\texttt{blue}}, \texttt{PAES}_{\texttt{green}}\}$ is in:

- the IBGN orange sub-dataset $\varsigma = 50\%$ specific to the 6 sequences that support the wcpo-pattern, and, besides, 50% specific to the other analysed sequences;
- the IBGN red sub-dataset $\varsigma = 80\%$ specific to the 6 sequences that support the wcpo-pattern, and, besides, only 20% specific to the other analysed sequences.



(a) IBGN red



(b) IBGN orange

Figure 7.18: Two hybrid wcpo-patterns discovered simultaneously in the IBGN red and the IBGN orange sub-datasets with the same $Support = 6$. ⬠ represents the wcpo-pattern support. The other vertices are labelled with 3-tuples $(\varpi_{v_t}, \varsigma_{v_t}, \omega_{v_t})$

Figures 7.19 and 7.20 illustrate two complex hybrid wcpo-patterns discovered respectively in the IBGN orange and red sub-datasets (Tab. 7.13) with the same $Support = 4$. These complex wcpo-patterns reflect the well-known factors that increase the impact of the physico-chemical pressures on the biological indicators, namely a strong pollution (highlighted by the qualitative values of the macro-parameters), a persistent pollution (highlighted by the persistency (ϖ) of vertices) and a combination of different pollutions (highlighted by the vertices that contain many macro-parameters, which represent distinct types of pollution).

Figure 7.19 reveals only one type of concrete pollution (organic: orange NITRO, red NITRO, red MOOX), while Fig. 7.20 reveals two types of concrete pollution (organic: orange NITRO, red MOOX and nutrient: red PHOS). Then, in Fig. 7.19 the physico-chemical pres-

sures are strong since there are 7 occurrences of the bad (orange) and very bad (red) macro-parameters. In contrast, in Fig. 7.20 the physico-chemical pressures are stronger since there are 10 occurrences of the medium (yellow), bad and very bad macro-parameters. Moreover, in Fig. 7.19 the persistent (e.g. itemsets without predecessors: $\varpi = 1.5$ for $(\texttt{MOOX}_{\texttt{red}}\texttt{PAES}_{\texttt{green}})$ and $\varpi = 0.75$ for $(\texttt{NITRO}_{\texttt{orange}}\texttt{PAES}_{\texttt{green}})$) and strong (e.g. $\texttt{NITRO}_{\texttt{red}}$ and $\texttt{MOOX}_{\texttt{red}}$) organic pollution is not enough to impact the resilience capacity of biocoenosis, and thus $\texttt{IBGN}_{\texttt{orange}}$ can be observed. In contrast, in Fig. 7.20 the persistent and strong nutrient pollution (e.g. itemsets without predecessor: $\varpi = 1.5$ for $(\texttt{PHOS}_{\texttt{red}}\texttt{PAES}_{\texttt{green}})$) augmented by the persistent and the strong coexistence of organic pollution with other ones (e.g. itemsets without predecessor: $\varpi = 0.75$ for $(?_{\texttt{red}}\texttt{NITRO}_{\texttt{orange}}\texttt{PAES}_{\texttt{green}})$ where $?_{\texttt{red}}$ may reveal various types of strong pollutions and $?$ is different from NITRO and PAES) are enough to impact the resilience capacity of biocoenosis, and thus $\texttt{IBGN}_{\texttt{red}}$ can be observed.

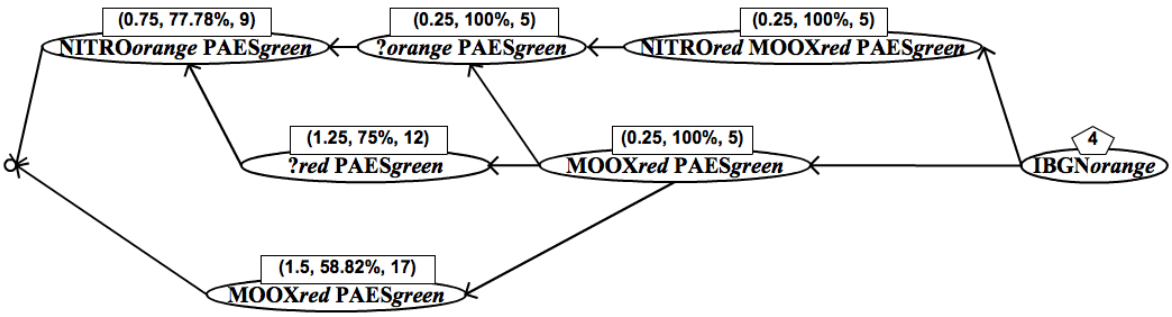


Figure 7.19: A complex hybrid wcpo-pattern discovered in the IBGN orange sub-dataset. ⬠ represents the wcpo-pattern support. The other vertices are labelled with 3-tuples $(\varpi_{v_t}, \varsigma_{v_t}, \omega_{v_t})$

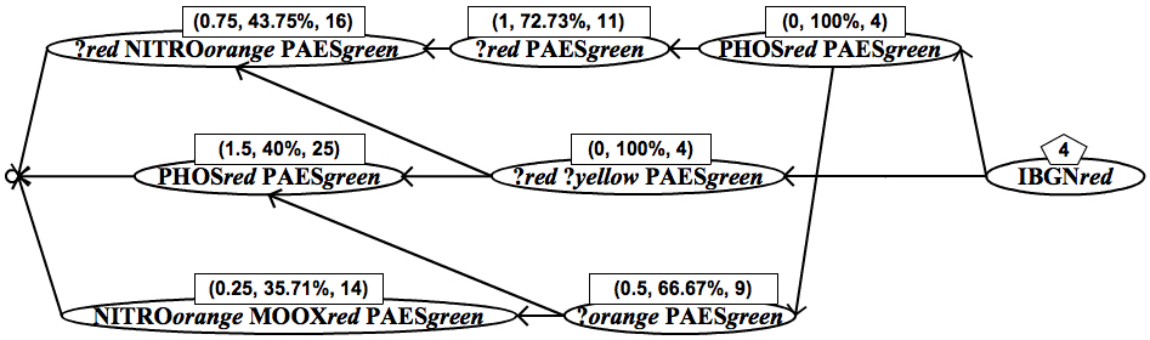


Figure 7.20: A complex hybrid wcpo-pattern discovered in the IBGN red sub-dataset. ⬠ represents the wcpo-pattern support. The other vertices are labelled with 3-tuples $(\varpi_{v_t}, \varsigma_{v_t}, \omega_{v_t})$

7.4 Hydro-Ecological Heterogeneous Sequential Data

We focus on hydro-ecological data about river restorations that were collected during the REX project. A number of 15 river sites from the Rhine river are monitored. These sites create the river site network (graph) illustrated in Fig. 7.21. The river sites are linked by a spatial relation *is downstream of*.

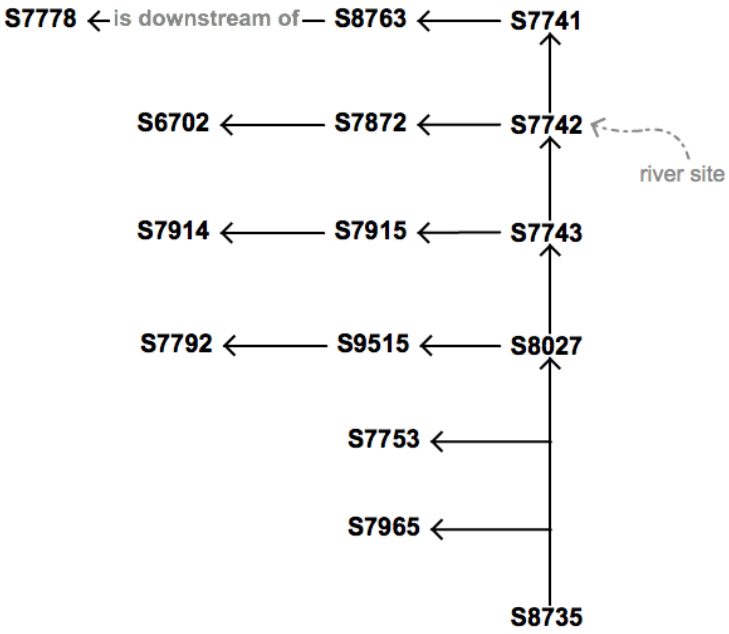


Figure 7.21: The analysed river site network

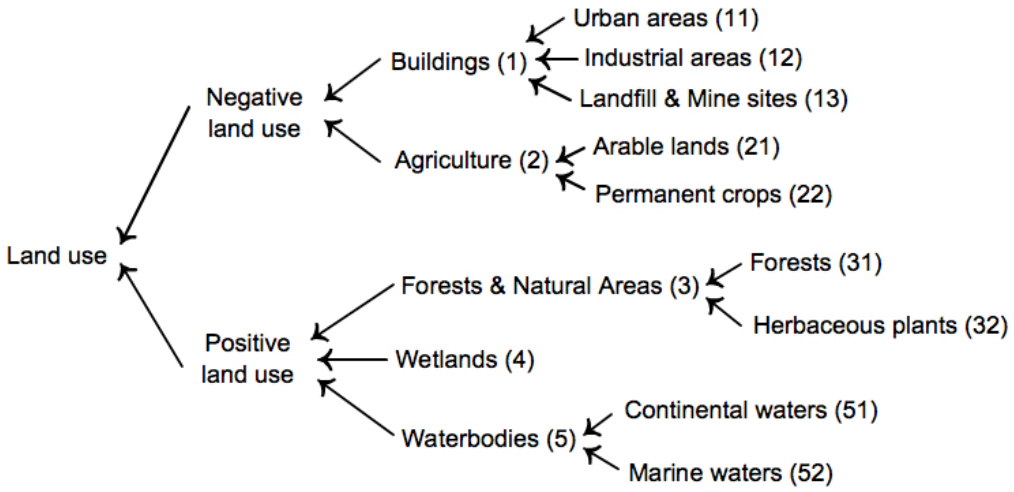
Table 7.14 shows some data gathered for three river sites given in Fig. 7.21, precisely *S7742*, *S7743* and *S7792*. We note that for a river site only the values of physico-chemical parameters are mandatory.

Table 7.14: Examples from the hydro-ecological heterogeneous data collected during the REX project: biological indicators, physico-chemical parameters (ammonium (NH_4^+), total phosphorus (P), nitrite (NO_2^-)) and types of land use

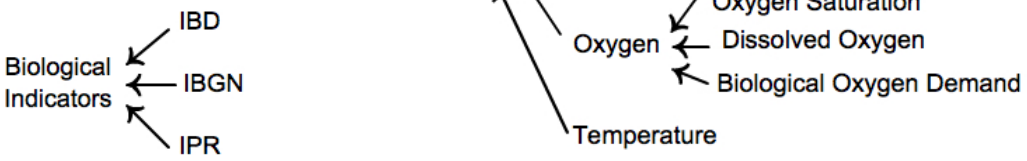
Period	River Site	Indicators/Parameters					Types of land use (%)				River Segment
		Biological		Physico-Chemical			Urban areas		Arable lands		
		IBGN	IBD	NH_4^+	P	NO_2^-	100 m	500 m	100 m	500 m	
2002 – 2005	S7792	–	orange	yellow	yellow	green	51	21	0	53	5601
	S7742	–	green	green	green	blue	0	8	0	0	20165
2006 – 2009	S7792	blue	orange	yellow	yellow	green	51	23	0	53	5601
	S7743	orange	yellow	green	green	green	0	0	0	0	19949
2010 – 2014	S7742	red	green	green	green	yellow	0	9	0	0	20165
	S7792	yellow	orange	green	yellow	green	51	23	0	53	5601

There are three monitored periods of time, i.e. 2002 – 2005, 2006 – 2009 and 2010 – 2014. For each period of time and a river site, the aggregated values are obtained from various

measurements made in this period. The data comprise three domains, precisely biological indicators, physico-chemical parameters and land use, that can be described by means of the taxonomies shown in Fig. 7.22. Let us note that the collected data concern only the atomic values from these taxonomies, e.g. urban areas, wetlands, IPR, nitrate and nitrite.



(a) land use



(b) biological indicators

(c) physico-chemical parameters

Figure 7.22: Taxonomies over land use, biological indicators and physico-chemical parameters

For example, in 2010 – 2014 period a *red* (very bad) qualitative value of IBGN and a *green* (good) one of IBD represent the overall ecological state of the river site *S7742*; the physico-chemical state is summarised by a *green* (good) qualitative value of NH_4^+, a *green* one of P, and a *yellow* (medium) qualitative value of NO_2^-; the land use state of the river site *S7742* is 9% urban areas in $500\ m$ buffer, 0% urban areas in $100\ m$ and 0% arable lands in $100\ m$ and $500\ m$. In addition, the river site *S7742* is in the river segment *20165* as shown in Fig. 7.23.

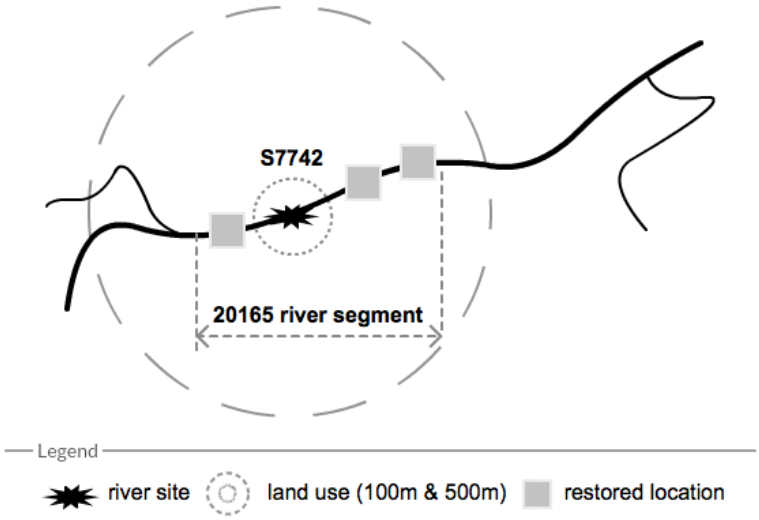


Figure 7.23: The *S7742* river site that is in the *20165* river segment

Table 7.15 shows the 12 surveyed river segments, where each river segment includes a river site from the analysed river network (Fig. 7.21).

Table 7.15: The monitored river segments and the river sites included in them (from the river network shown in Fig. 7.21)

River Segment	River Site
3163	S8763
4548	S7753
5601	S7792
6850	S6702
8614	S7914
8674	S7915
18725	S7778
19754	S7965
19949	S7743
20165	S7742
20346	S7741
26763	S7872

Table 7.16 shows the restorations undertaken on three river segments: *5601*, *19949* and *20165*. There are two types of restoration: *global* and *wetland*. During the monitored period 2002 – 2014, there can be several restorations for the same river segment. For instance, in Tab. 7.16 the *20165* river segment was restored in 3 different locations (depicted also in Fig. 7.23).

Table 7.16: Examples from the hydro-ecological data about river segments collected during the REX project

River Segment	Restoration Type	
	Wetland	Global
5601	no	yes
20165	yes	no
19949	yes	yes
20165	yes	yes
20165	yes	no

7.4.1 Data preprocessing

To explore such heterogeneous data, we preprocess them using once again the domain knowledge. Firstly, the biological and physico-chemical data in Tab. 7.14 are already discretized. The land use data are numerical values, and thus we discretize them using the qualitative values given in Tab. 7.17.

Table 7.17: Domain knowledge: the discretization intervals for the types of land use

Type of land use	low	medium	high
Buildings	$[0\%, 25\%]$	$(25\%, 52\%]$	$(52\%, 100\%]$
Agriculture	$[0\%, 25\%]$	$(25\%, 45\%]$	$(45\%, 100\%]$
Forests & Natural Areas	$[0\%, 15\%]$	$(15\%, 40\%]$	$(40\%, 100\%]$
Wetlands	$[0\%, 15\%]$	$(15\%, 40\%]$	$(40\%, 100\%]$
Waterbodies	$[0\%, 30\%]$	$(30\%, 50\%]$	$(50\%, 100\%]$

Table 7.18 shows the preprocessed data about river sites (the raw data are given in Tab. 7.14). For example, between 2010 – 2014 the surroundings of the *S7792* river site are covered with a *medium* percentage of urban areas and a *low* percentage of arable lands at $100\ m$ buffer; a *low* percentage of urban areas and a *high* percentage of arable lands at $500\ m$ buffer.

Table 7.18: The preprocessed hydro-ecological heterogeneous data (the raw data are shown in Tab. 7.14)

Period	River Site	Indicators/Parameters					Types of land use				River Segment
		Biological		Physico-Chemical			Urban areas		Arable lands		
		IBGN	IBD	NH_4^+	P	NO_2^-	$100\ m$	$500\ m$	$100\ m$	$500\ m$	
2002 – 2005	S7792	–	orange	yellow	yellow	green	medium	low	low	high	5601
	S7742	–	green	green	green	blue	low	low	low	low	20165
2006 – 2009	S7792	blue	orange	yellow	yellow	green	medium	low	low	high	5601
	S7743	orange	yellow	green	green	green	low	low	low	low	19949
2010 – 2014	S7742	red	green	green	green	yellow	low	low	low	low	20165
	S7792	yellow	orange	green	yellow	green	medium	low	low	high	5601

Secondly, the data about each river segment (Tab. 7.16) should be aggregated to obtain a global estimation of the level of the type of restoration (i.e. the number of restored loca-

tions) for the entire monitored period 2002 – 2014. To this end, we rely again on the domain knowledge and we use Tab. 7.19 to obtain the preprocessed data shown in Tab. 7.20.

Table 7.19: Domain knowledge: the levels of the restoration type by the number of undertaken restorations

	L1	**L2**	**L3**
#restorations	(0, 2]	(2, 5]	(5, ∞)

For instance, the *20165* river segment has level *L2* for the *wetland* restoration since in Tab. 7.16 there are 3 such restorations and has level *L1* for the *global* restoration since in Tab. 7.16 there is only one such restoration.

Table 7.20: The preprocessed data about the river segments shown in Tab. 7.16

River Segment	**Restoration Type**	
	Wetland	**Global**
5601	L1	L1
20165	L2	L1
19949	L1	L1

7.4.1.1 Building a Heterogeneous Sequential Dataset

For each period of time and river site in the studied river site network (Fig. 7.21), a heterogeneous itemset $\{physico\text{-}chemical\ parameters, biological\ indicators, land\ use\}$ can be built. For example, the heterogeneous itemset $\{(\mathtt{NH}^{+}_{4\ \mathtt{green}}\mathtt{P}_{\mathtt{green}}\mathtt{NO}^{-}_{2\ \mathtt{yellow}}), (\mathtt{IBGN}_{\mathtt{red}}\mathtt{IBD}_{\mathtt{green}}), (11_{\mathtt{low\_100m}}\ 11_{\mathtt{low\_500m}}21_{\mathtt{low\_100m}}21_{\mathtt{low\_500m}})\}$ is associated with river site *S7742* for the period 2010 – 2014 (Tab. 7.18). Note that `11` and `21` are respectively the identifiers of urban areas and arable lands as shown in the taxonomy depicted in Fig. 7.22a. Since between the river sites exists the spatial order *is downstream of*, we can build heterogeneous sequences. Moreover, a sequence can end with a target itemset $(restoration\ types)$ built from an unordered set, e.g. $(\mathtt{Global}_{\mathtt{L1}}\mathtt{Wetland}_{\mathtt{L1}})$ for the 19949 river segment (Tab. 7.20).

Although a dataset of heterogeneous sequences can be built for the analysed river site network, our aim is to manipulate the data as a graph that has heterogeneous itemsets as vertices and binary spatial relations as edges. Thus, we show that RCA-SEQ might be appropriate for *graph mining* [Chakrabarti and Faloutsos, 2006], as well.

7.4.1.2 Modelling Heterogeneous Sequential Data

To explore the REX heterogeneous dataset and to build the RCA input, the data model depicted in Fig. 7.24 is used. This model allows us to learn more about the ecological state of

the aquatic ecosystem by highlighting the impact of land use, water quality and river restorations.

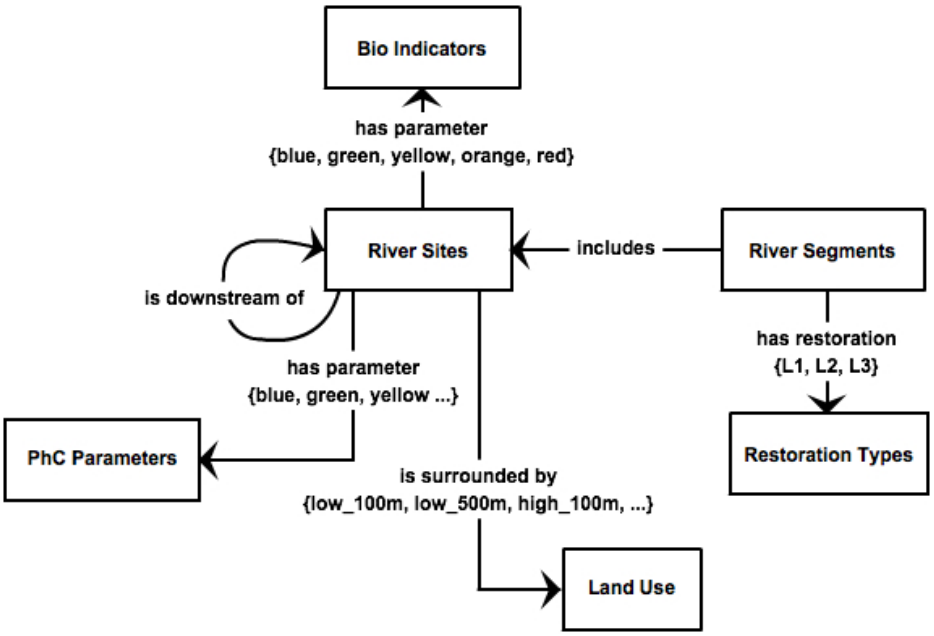


Figure 7.24: The modelling of hydro-ecological heterogeneous sequential data collected during the REX project. Bio and PhC stand respectively for biological and physico-chemical

The six rectangles represent the six sets of objects we manipulate as follows: river sites, river segments, biological indicators, physico-chemical parameters, land use and restoration types. The links between river segments and river sites are highlighted by the spatial binary relation *includes*. This spatial relation associates a river site with a river segment if the river site is in the river segment as shown in Tab. 7.15. The links between river sites are highlighted by the spatial binary relation *is downstream of*. This spatial relation is used to encode the river site network shown in Fig. 7.21. The river segments are described only by the qualitative binary relations *has restoration L1*, *has restoration L2* and *has restoration L3* that link a river segment with the type of the undertaken restoration. The river sites are described by the qualitative binary relations *has parameter blue*, *has parameter green*, *has parameter yellow*, *has parameter orange* and *has parameter red* that link the river sites with the measured biological indicators/physico-chemical parameters. Besides, the river sites are described using the spatial-qualitative binary relations *is surrounded by low_100m*, *is surrounded by low_500m*, *is surrounded by medium_100m*, *is surrounded by medium_500m*, *is surrounded by high_100m* and *is surrounded by high_500m* that indicate the types of land use around these river sites.

7.4.2 Experiments and Discussion

In this section, we present some first results obtained with the RCA-SEQ approach applied to heterogeneous sequential data collected during the REX project. A more systematic analysis should be done in the future. Table 7.21 shows the characteristics of the REX dataset, and the number of concepts generated by applying RCA-SEQ. Basically, relying on the data model depicted in Fig. 7.24 we encode into the RCA input the data collected during the entire monitored period 2002 – 2014 for the river network depicted in Fig. 7.21. The relational scaling mechanism relies on the $\exists$ quantifier. The obtained family of lattices contains the taxonomies shown in Fig. 7.22, the lattice of river segments ($\mathcal{L}_{K_M}$) and the lattice of river sites ($\mathcal{L}_{K_T}$).

Table 7.21: The results of the REX dataset exploration with $\theta = 0\%$. Column `River Site Measurements` represents the number of measurements made on the three monitored periods at the river sites shown in Fig. 7.21; column `River Segments` represents the number of monitored river segments with at least one restored location; column $\mathcal{L}_{K_M}$ represents the number of concepts from the lattice of river segments; column $\mathcal{L}_{K_T}$ represents the number of concepts from the lattice of river sites

Dataset		**RCA Output**	
River Site Measurements	**River Segments**	$\mathcal{L}_{K_M}$	$\mathcal{L}_{K_T}$
45	12	860	4554

By navigating the lattices starting from the main concepts in $\mathcal{L}_{K_M}$ we obtain a hierarchy of 859 multilevel heterogeneous cpo-patterns. Figure 7.25 depicts an excerpt from this hierarchy, precisely the organised ①, ②, ③, ④, ⑤, ⑥ and ⑦ multilevel heterogeneous cpo-patterns. A cpo-pattern is associated with a set of river segments (given in Tab. 7.22) whose number (support) is shown in ■. The restoration types of these river segments are illustrated in ▭, e.g. $\texttt{Global}_{\texttt{L1}}$ meaning that the river segments were globally restored at most 2 locations. A vertex (⬭) is associated with a set of river sites and it is labelled with physico-chemical parameters and their qualitative values. A vertex can have additional information: land use (⬭) and biological indicators (◊). In the following, we focus on the cpo-patterns ①, ④ and ⑥.

The ① cpo-pattern is associated with 11 (■ in Fig. 7.25) river segments that contain at most 2 locations that were globally restored. In addition, itemset $(\texttt{PHC}_{\texttt{blue}})$ reveals locally (i.e. in the associated river segments shown in Tab. 7.22) a very good physico-chemical state of water.

The ④ cpo-pattern is associated with 5 river segments (shown in Tab. 7.22) that contain at most 2 locations that were globally restored. Itemset $(\texttt{IBD}_{\texttt{green}})$ (◊ in Fig. 7.25) reveals locally a good ecological state of water based on the analysis of diatom species. In addition, the physico-chemical state of water is very good for the temperature, biological oxygen demand

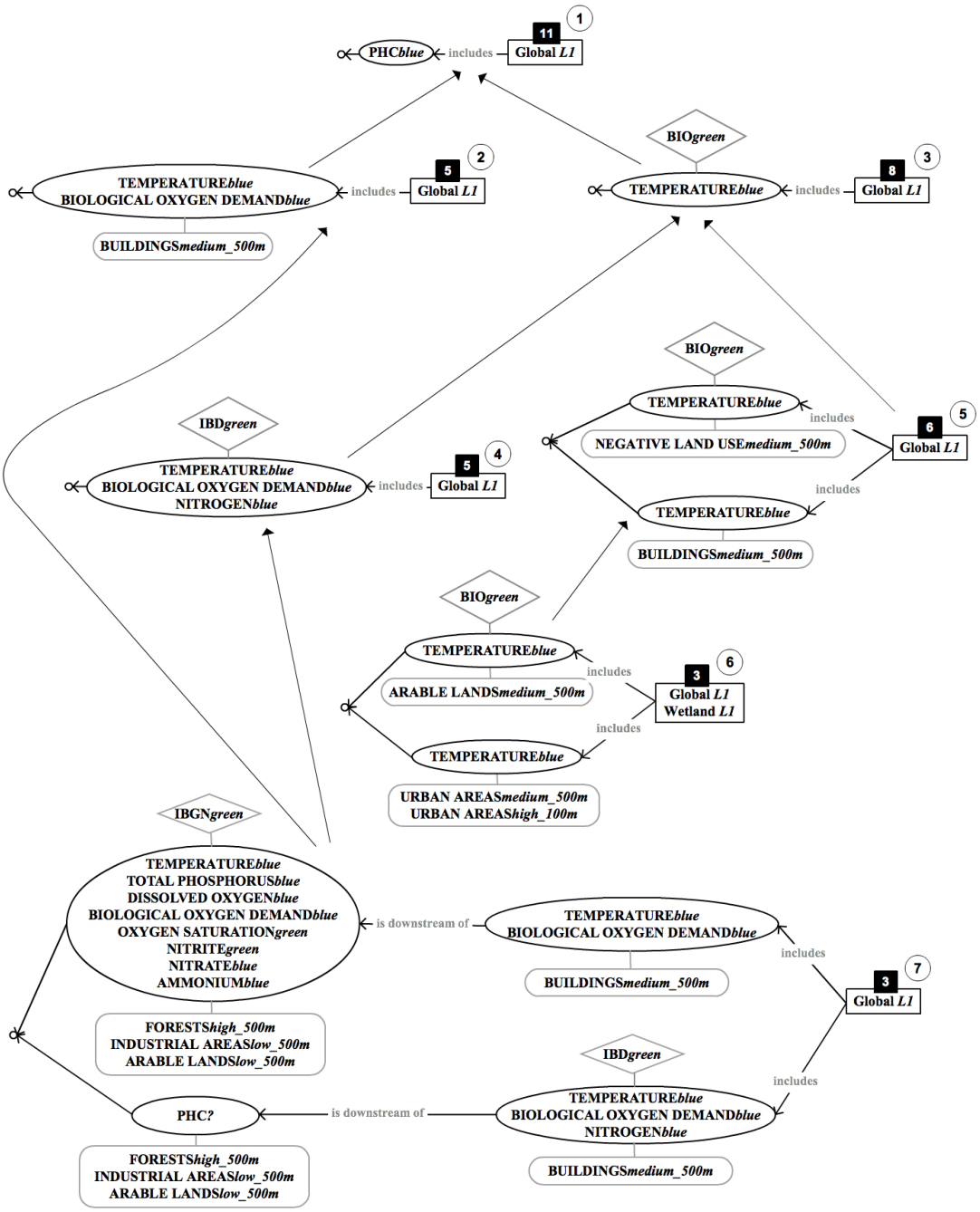


Figure 7.25: Excerpt from the hierarchy of multilevel heterogeneous cpo-patterns discovered in the REX dataset (Tab. 7.21). ①, ②, ③, ④, ⑤, ⑥ and ⑦ identify the cpo-patterns. ■ is the support (number of river segments) of a cpo-pattern; ▭ represents the types of the river segment restoration; PHC and BIO stand respectively for physico-chemical parameters and biological indicators; ⬭ represents the land use; ◊ represents the biological indicators; ⬭ represents the physico-chemical parameters

and nitrogen that represent a part of the abiotic characteristics suitable for the diatom species [Raibole M, 2011].

The ⑥ cpo-pattern, which is a more concrete specialisation of ⑤, is associated with 3 river segments (shown in Tab. 7.22) that contain at most 2 locations that had global and wetland restorations. Itemset ($\texttt{BIO}_{\texttt{green}}$) (◊ in Fig. 7.25) reveals locally a good ecological state of the aquatic ecosystem. Since BIO is an abstract item, we cannot specify the fauna and flora that underpin this regularity. In addition, itemset ($\texttt{TEMPERATURE}_{\texttt{blue}}$) reveals locally a very good physico-chemical state of the water temperature. Furthermore, locally at 500 m buffer the land use pressures of arable lands and urban areas are *medium*, while at 100 m the land use pressures of urban areas are *high*.

Table 7.22: The associated river segments of the multilevel heterogeneous cpo-patterns shown in Fig. 7.25

CPO-pattern	Associated River Segments											
	3163	**4548**	**5601**	**6850**	**8674**	**8681**	**18725**	**19754**	**19949**	**20165**	**20346**	**26763**
1	×	×	×	×	×	×	×	×	×		×	×
2	×				×		×		×			×
3	×	×		×	×	×		×	×			×
4	×			×	×				×			×
5	×	×			×			×	×			×
6		×						×				×
7					×				×			×

Figure 7.26 depicts a complex multilevel heterogeneous cpo-pattern extracted from the REX dataset. This is associated with the river segments *8674* and *19949*. The vertices Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ and Ⓕ are derived from the concepts in $\mathcal{L}_{K_T}$ whose extents (river sites) are shown in Tab. 7.23.

Table 7.23: The river sites of the extents of the concepts from which the vertices Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ and Ⓕ of the heterogeneous cpo-pattern depicted in Fig. 7.26 are derived

Vertex	Monitored Periods		
	2002 – 2005	**2006 – 2009**	**2010 – 2014**
A	–	–	S7743 S7915
B	–	S7915	S7743
C	S7915	–	S7743
D	–	S7915 S7743	–
E	–	–	S7914
F	S7914	S7914	S7914

The cpo-pattern given in Fig. 7.26 is associated with 2 river segments that contain at most 2 locations that were globally restored. Locally, in the entire monitored period 2002 – 2014

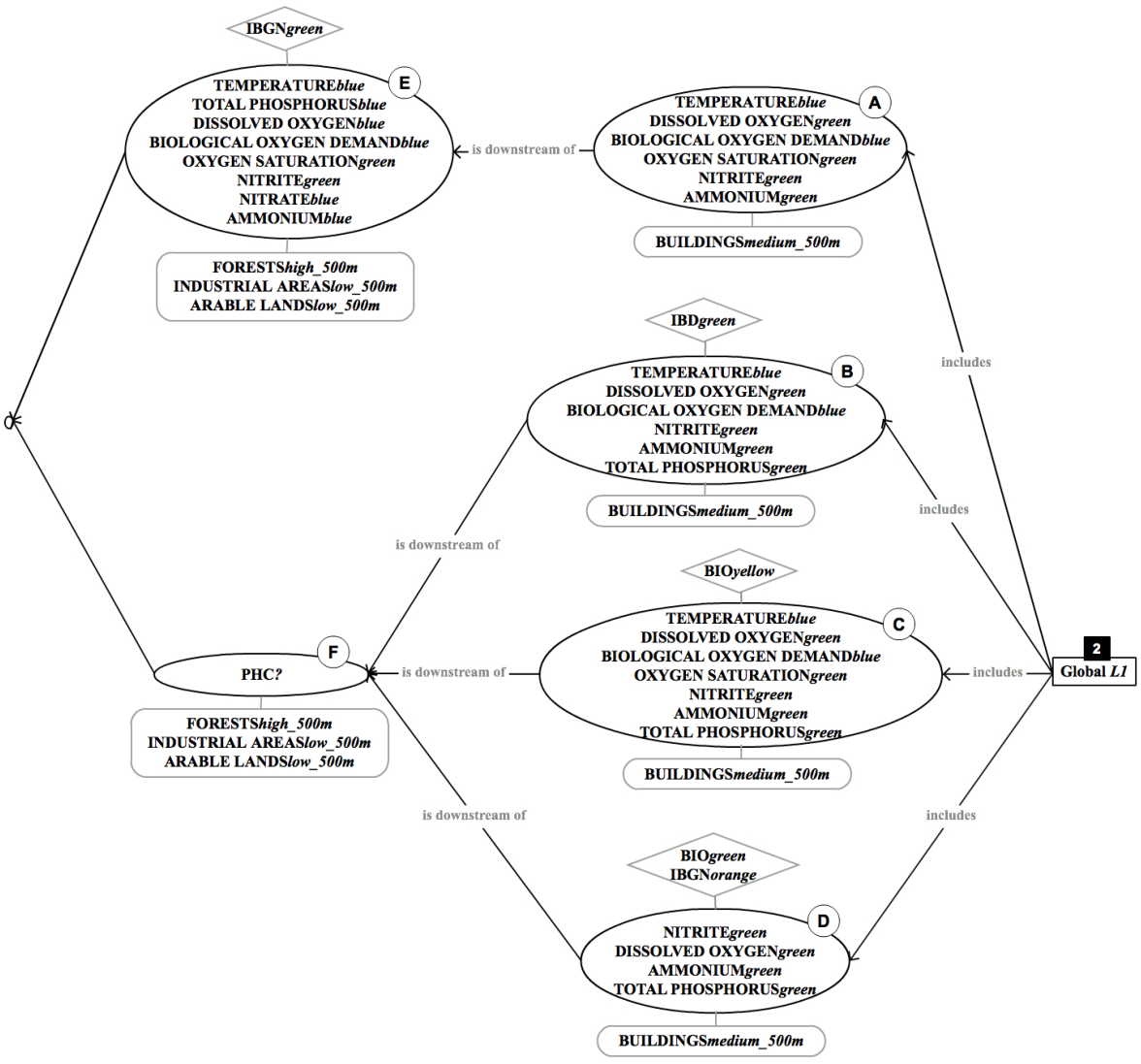


Figure 7.26: A complex multilevel heterogeneous cpo-pattern extracted from the REX dataset (Tab. 7.21). Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ and Ⓕ identify the vertices; ■ is the support (number of river segments) of the cpo-pattern; ▭ represents the types of the river segment restoration; PHC and BIO stand respectively for physico-chemical parameters and biological indicators; ⬭ represents the land use; ◊ represents the biological indicators; ⬭ represents the physico-chemical parameters

the land use pressures of buildings were *medium* at $500\ m$ buffer. In contrast, in the upstream rivers at $500\ m$ buffer on the one hand the land use pressures of industrial areas and arable lands were *low*; on the other hand, a *high* percentage of the area is covered with forests that lead to a good ecological state of the aquatic ecosystem in the surroundings. Indeed, by analysing the Ⓔ vertex, itemset $(\texttt{IBGN}_{\texttt{green}})$ (◊, Fig. 7.26) reveals a good ecological state of the aquatic ecosystem in the period 2010 – 2014 (Tab. 7.23) based on the analysis of macro-invertebrates. Moreover, the water temperature is very good; the organic matter (dissolved oxygen, biological oxygen demand and oxygen saturation) are good and very good; the nitrogenous parameters (nitrite and ammonium), which are related to the organic matter, are as well good and very good; and the nutrients (total phosphorous and nitrate) are very good.

By comparing the Ⓔ vertex with the Ⓐ, Ⓑ, Ⓒ and Ⓓ vertices, it is noted a degradation up to one level regarding the qualitative values of the physico-chemical parameters probably caused by the *medium* building pressures at $500\ m$ buffer, e.g.:

- $\texttt{AMMONIUM}_{\texttt{blue}}$ and $\texttt{DISSOLVED OXYGEN}_{\texttt{blue}}$ (very good) from Ⓔ are measured when the surroundings are covered with a low percentage of industrial areas and arable lands (i.e. the land use pressures are low), while $\texttt{AMONIUM}_{\texttt{green}}$ and $\texttt{DISSOLVED OXYGEN}_{\texttt{green}}$ (good) from Ⓐ, Ⓑ, Ⓒ and Ⓓ are measured when the surroundings are covered with a medium percentage of buildings (i.e. the land use pressures are medium);
- $\texttt{TOTAL PHOSPHORUS}_{\texttt{blue}}$ (very good) from Ⓔ is measured when in the surroundings the land use pressures are low; $\texttt{TOTAL PHOSPHORUS}_{\texttt{green}}$ (good) from Ⓑ, Ⓒ and Ⓓ is measured when in the surroundings the land use pressures are medium.

Furthermore, the cpo-pattern shown in Fig. 7.26 reflects that the biological indicators seem to be more sensitive (up to two levels of their qualitative values) to the land use pressures [Wasson et al., 2006] and [Villeneuve et al., 2015]. For instance, $\texttt{IBGN}_{\texttt{green}}$ in upstream rivers (Ⓔ) in contrast to $\texttt{BIO}_{\texttt{yellow}}$ and $\texttt{IBGN}_{\texttt{orange}}$ locally (Ⓒ and Ⓓ, respectively).

To sum up, hydro-ecologists can draw valuable insights from the heterogeneous sequential data by exploiting the "richness" (e.g. the additional information captured by the concept extents and the revealed abstract items) of the RCA-Seq results.

7.5 Summary

In this chapter, we have presented hydro-ecology as an application context of this thesis. The analysed hydro-ecological data are exported from the databases used in two interdisciplinary research projects, namely Fresqueau and REX. We have chosen these data since we collaborate with hydro-ecologists. We have explained the preprocessing of these data in order to be able to apply the RCA-Seq approach and the aspects that have been discussed in this thesis.

We have presented several interesting results discovered in these hydro-ecological (heterogeneous) sequential data with RCA-SEQ and its extensions introduced in Chapter 6. We have shown that the multilevel (heterogeneous) cpo-patterns reveal well-known correspondences as well as more surprising ones between biological indicators, physico-chemical macro-parameters and land use. We have illustrated that the hierarchical results help the pattern evaluation step by guiding the hydro-ecologists. We have highlighted how RCA-SEQ and its extensions are appropriate to make use of domain knowledge, to enumerate only multilevel cpo-patterns that answer specific questions that hydro-ecologists may have and to discover more informative patterns, i.e. wcpo-patterns.

In addition, we have presented a performance study of RCA-SEQ that underlines the usefulness of the approach for exploring small but non-trivial datasets. We have empirically verified the ability of RCA-SEQ to directly extract the minimal representations of cpo-patterns and the usefulness of the distribution index when the execution time is important.

8 Conclusions

Contents

Nowadays, in the context of the digital age, large amounts of structured data are generated and stored in order to be further harnessed by discovering valuable pieces of information relevant for stakeholders. Basically, the structured data refer to the data stored in databases.

In this thesis, we have focused on exploring sequential data and we have introduced a novel problem, i.e. to directly extract multilevel cpo-patterns implicitly organised into a hierarchy, in order to help the pattern evaluation step of the KDD process. To this end, we have devised an original and self-contained KDD approach within the RCA framework, referred to as RCA-Seq, that exploits the relational nature of sequential data, the well-founded FCA technique and the properties of the RCA output.

RCA-Seq is a multi-relational data mining approach since it looks for regularities in sequential data that are gathered from multiple tables out of a relational database. This approach spans five steps: (i) the preprocessing of the raw sequential data; (ii) the RCA-based exploration of the preprocessed sequential data; (iii) the automatic extraction of a hierarchy of multilevel cpo-patterns by navigating the RCA output; (iv) the selection of relevant multilevel cpo-patterns based on various measures of interest; (v) the pattern evaluation step done by domain experts.

The core of the RCA-Seq approach is represented by two steps: the exploration of sequential data and the extraction of multilevel cpo-patterns. Briefly, RCA builds conceptual hierarchies for each involved set of objects and iteratively highlights the relations between

their concepts through the relational scaling mechanism. Then, the multilevel cpo-patterns are extracted by actually navigating these relational conceptual hierarchies.

The primary aim of RCA-SEQ is to simplify the evaluation step of the extracted set of cpo-patterns, i.e. the synthetic description of the raw sequential data. To this end, we benefit from the fact that some cpo-patterns are naturally sub-patterns of others and we propose to extract hierarchies of cpo-patterns where each cpo-pattern is projected into its descendants. Consequently, when an interesting cpo-pattern is found the evaluation step can continue by analysing the surrounding area in the hierarchy. Then, we exploit the order on items revealed by RCA and we extract multilevel cpo-patterns without specific preprocessing. Therefore, a global view of the standard cpo-patterns (i.e. cpo-patterns built from an unordered set of items) is obtained. Next, we make use of the information encoded in the navigated concept extents. On the one hand we extract weighted cpo-patterns that capture – besides the order on itemsets – additional information hidden in the analysed sequential data; on the other hand, we propose to compute some measures of interest, e.g. the distribution index and the richness of a concept, that can be used to select pertinent concepts/cpo-patterns. In addition, we show that RCA-SEQ can be easily adapted to extract cpo-patterns with items across different levels of a user-defined taxonomy, to push user-defined constraints deep into the RCA-based exploration step or to explore heterogeneous sequential data.

The RCA-SEQ approach was applied to hydro-ecological sequential data collected during two interdisciplinary research projects, namely Fresqueau and REX. Firstly, given sequential data that represent hydro-ecological sequences of biological and physico-chemical samples, we found hierarchies of multilevel cpo-patterns that summarise the impact of physico-chemical values on biological ones. We recall that biological values determine the quality of water. By means of these cpo-patterns, we could help hydro-ecologists to check well-known correspondences between the two types of values as well as to discover lesser-known facts. Then, by using the weighted cpo-patterns and the inherent measures, we could help hydro-ecologists to discriminate the same regularities discovered for different water qualities. Moreover, by varying the quantifiers used during the relational scaling mechanism, we could help hydro-ecologists to discover regularities with constraints regarding the frequency of a specific physico-chemical state of water. Secondly, given a river network (graph) and heterogeneous sequential data collected during three distinct periods of time, we discovered a hierarchy of multilevel heterogeneous cpo-patterns that summarise the impact of the analysed ecological factors on the quality of water. These experiments are the first attempt to apply RCA-SEQ to graph mining and the obtained results are instructive.

Furthermore, it is worthwhile to mention that the RCA-SEQ approach can be applied to any data that can be modelled according to the generic data model proposed in this thesis, e.g. the trajectory of a student knowledge in a specific field leading to a sequence of test

papers followed by a final evaluation or the trajectory of a football player prior to a football game leading to a sequence of training sessions followed by a player evaluation.

8.1 Discussion

In contrast to classical sequential pattern mining methods, the problem that we have tried to tackle in this thesis is more challenging since the objective is to simultaneously enumerate the cpo-patterns from a sequential dataset and to highlight how these patterns relate to each other. In fact, this problem has been inspired by existing works, e.g. [Cellier et al., 2011] and [Egho et al., 2011], that post-process already discovered patterns in order to organise them into a hierarchy. Therefore, these existing works rely on classical sequential pattern mining methods [Fournier-Viger et al., 2017]. However, we have proposed RCA-Seq that is a self-contained approach for directly extracting a hierarchy of cpo-patterns from the given sequential dataset.

Recently, Buzmakov et al. [2016] have shown how to explore sequential data by means of FCA and pattern structures. Hence, the set of cpo-patterns extracted using classical methods, e.g. [Pei et al., 2006] and [Fabrègue et al., 2015], can be combined with the intersection operation on graph to build a pattern structure. The resulting pattern concept lattice can be compared to the hierarchy of cpo-patterns built with RCA-Seq. Let us however notice that first, in our approach the cpo-patterns are extracted and implicitly organised into a hierarchy directly from the RCA output. Second, a partial order on items is generated, and thus abstract and hybrid cpo-patterns are obtained rather than only concrete cpo-patterns as in [Pei et al., 2006] and [Fabrègue et al., 2015]. Such results can be related to [Srikant and Agrawal, 1996] where generalised sequential patterns (rather than cpo-patterns) are extracted on two steps: (i) new generalised sequences are built from the original ones based on a user-defined taxonomy over the items and (ii) the new generalised sequences are explored. Therefore, their method explores sequences that already contain the relationships between the items and their ancestors. In contrast, the RCA-Seq approach extracts multilevel cpo-patterns without a specific preprocessing of the original sequences. Indeed, RCA reveals automatically the relationships between the items and their ancestors during the relational scaling mechanism since the unordered set of items/taxonomy over the items is encoded into the RCA input based on the nominal/ordinal scaling. Moreover, RCA-Seq allows both to navigate along the sequences and to synthesise them within cpo-patterns.

To our knowledge, the existing methods [Pei et al., 2006] and [Fabrègue et al., 2015] directly extract standard cpo-patterns, i.e. patterns that consider only the order on itemsets from the analysed sequences. RCA-Seq allows to directly obtain more informative cpo-patterns by exploiting the "richness" of the RCA output, namely weighted cpo-patterns and

cpo-patterns with user-defined constraints.

Finally, Temporal Concept Analysis [Wolff, 2001] is an extension of FCA for exploring temporal data where the temporal relations between the derived concepts are actually revealed by manually analysing the dates in the concepts. The RCA-Seq approach, on the contrary, automatically reveals the temporal links between the derived concepts through the relational scaling mechanism.

8.2 Perspectives

The contributions presented in this thesis open up interesting research directions:

- *to improve* RCA-Seq *in order to be applicable to large volumes of sequential data*. In fact, since classical RCA does not cope with big datasets, the current version of RCA-Seq is not an efficiency-based approach but rather focuses on exploring small but interesting datasets, as those designed during the Fresqueau and REX projects, in order to enhance the pattern evaluation step. To address the "concept explosion" problem, it will be interesting to improve the RCA-based exploration step of sequential data by means of AOC-poset [Godin and Mili, 1993]. Indeed, Dolques et al. [2016] have shown that using AOC-poset rather than concept lattices reduces the complexity of the RCA output;

- *to avoid the "cpo-pattern explosion" by pushing measures of interest deep into the RCA-based exploration step*. Usually, the support measure is used to prune infrequent cpo-patterns. In this thesis, we have already used the iceberg lattices that exploits the support measure. However, it will be interesting to try to push the distribution index or the stability index [Kuznetsov, 2007] into the RCA-based exploration step, and thus to directly extract only relevant multilevel cpo-patterns;

- *to design a tool that interactively extract the cpo-patterns from the RCA output rather than to extract all of them at once*. To this end, the extraction and the evaluation steps of RCA-Seq may be seen as one iterative step. Precisely, relying on measures of interest (e.g. distribution index) domain experts may first select a few interesting main concepts, and thus the CPOHrchy algorithm would extract only the associated multilevel cpo-patterns. Second, the extraction of cpo-patterns may continue based on other main concepts selected by using again the measures of interest or based on the main concepts that surround the previously selected concepts. Then, the iterative step may continue in the same way. To sum up, an interactive tool may enhance the evaluation step since domain experts may gradually and systematically assess the discovered multi-level cpo-patterns rather than being overwhelmed by the potential exponential number of cpo-patterns;

- *to study the different quantifiers that can underpin the relational scaling mechanism and to analyse the emerged cpo-patterns*. In this thesis we have focused only on the $\exists$ quantifier and its variants applied to the order relations on the itemsets. It will be interesting to study more quantifiers presented in [Rouane-Hacene et al., 2013] and also to apply them to the qualitative relations used to define the itemsets. For example, the $\geq_{50\%}$ quantifier can be used to discover only the itemsets that include at least 50% out of the number of items used to build the analysed sequences;

- *to empirically compare the* RCA-Seq *approach with ad hoc methods*. In this thesis, we have presented a time complexity analysis that theoretically shows – in the worst-case scenario – the better performance of RCA-Seq compared with an ad hoc method that combines FCA and the [Fabrègue et al., 2015] approach for extracting cpo-patterns. Furthermore, it will be interesting to carry out an experimental evaluation of the RCA-Seq approach and other ad hoc methods on standard benchmark datasets.

8.3 List of Publications

International peer-reviewed journal

- Cristina Nica, Agnès Braud, Florence Le Ber: *Hierarchies of Multilevel Closed Partially-Ordered Patterns for Enhancing Sequential Data Analysis* (submitted)

International peer-reviewed conferences

- Cristina Nica, Agnès Braud, Florence Le Ber: *Hierarchies of Weighted Closed Partially-Ordered Patterns for Enhancing Sequential Data Analysis*. Proceedings of the 14^{th} International Conference on Formal Concept Analysis, ICFCA 2017, Springer, Lecture Notes in Computer Science, 138–154

- Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard, Florence Le Ber: *Exploring Temporal Data Using Relational Concept Analysis: An Application to Hydroecological Data*. Proceedings of the 13^{th} International Conference on Concept Lattices and Their Applications, CLA 2016, CEUR-WS.org, CEUR Workshop Proceedings, 299–311

- Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard, Florence Le Ber: *Extracting Hierarchies of Closed Partially-Ordered Patterns Using Relational Concept Analysis*. Proceedings of the 22^{nd} International Conference on Conceptual Structures, ICCS 2016, Springer, Lecture Notes in Computer Science, 17–30

National peer-reviewed conference and workshop

- Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard, Florence Le Ber: *L'analyse relationnelle de concepts pour la fouille de données temporelles - Application à l'étude de données hydroécologiques*. The 16ème Journées Francophones Extraction et Gestion des Connaissances, EGC 2016, Hermann-Éditions, RNTI, 267–278
- Cristina Nica, Xavier Dolques, Agnès Braud, Marianne Huchard, Florence Le Ber: *Exploration de données temporelles avec des treillis relationnels*. Atelier GAST associé à la conférence EGC, 2015

Abstract in French

Devant l'explosion actuelle et future du volume de données stockées, identifier les informations utiles, les extraire et les analyser de manière automatique implique la définition de nouvelles méthodes. C'est dans ce contexte que s'est développée l'Extraction de Connaissances à partir de Bases de Données (ECBD). Ce processus respecte généralement le schéma suivant :

1. sélection des données, étape qui consiste à sélectionner les informations pertinentes pour le problème pour lequel nous souhaitons construire de nouvelles connaissances ;

2. prétraitements, étape qui vise à nettoyer les données et à les transformer dans un format adéquat ;

3. fouille des données, qui est l'étape centrale du processus. Les données sélectionnées et prétraitées sont explorées avec un ou plusieurs algorithmes adaptés afin d'en extraire par exemple un ensemble de motifs, de règles ou un regroupement par classes ;

4. restitution, étape qui consiste à traiter les résultats en sortie des algorithmes afin de les restituer, les rendre facilement visualisables et analysables par les experts.

Ce schéma est très général et les différentes étapes peuvent varier en fonction de nombreux critères. Les domaines d'application de la fouille de données touchent à tous les secteurs. Nous pourrions même dire que partout où les données sont présentes en quantité suffisante l'extraction de connaissances est possible.

L'exploration de données séquentielles est un défi majeur dans la recherche actuelle en raison de la progression de la collecte de telles données concernant, par exemple, les comportements d'achat des clients, les examens médicaux des patients ou l'historique d'accès au Web. La découverte de motifs séquentiels est une tâche bien connue qui a pour objectif de trouver des régularités dans les données séquentielles, régularités qui peuvent être évaluées et interprétées par des experts. Différents algorithmes ont été proposés et beaucoup d'entre eux se concentrent sur l'extraction de représentations concises de motifs séquentiels (par exemple, des motifs séquentiels fermés), en réponse au fait que le nombre de motifs séquentiels

possibles peut être grand. Pour obtenir un ensemble plus restreint de ces motifs séquentiels, des algorithmes efficaces pour l'extraction directe de motifs partiellement ordonnés fermés ont été proposés. Un motif partiellement ordonné fermé résume un ensemble de motifs séquentiels fermés qui coexistent dans les mêmes séquences analysées, et il est représenté sous la forme d'un graphe orienté acyclique qui facilite l'étape d'interprétation. Toutefois, il existe certaines limitations aux approches existantes :

1. l'étape d'évaluation de motifs reste difficile parce que les motifs partiellement ordonnés fermés découverts ne sont pas organisés ; ainsi, les experts devraient déterminer manuellement comment ces motifs se rapportent les uns aux autres ;

2. les experts n'ont pas une vue globale des motifs partiellement ordonnés fermés découverts (par exemple, pour un sous-esemble de motifs découverts, il manque un motif général qui pourrait le résumer) ; ainsi, les motifs pertinents peuvent être négligés ;

3. certains motifs, plus généreaux et potentiellement intéressants, ne peuvent apparaître en l'absence d'une taxonomie sur les items ;

4. les motifs partiellement ordonnés fermés découverts n'exploitent que l'ordre des *itemsets* dans les séquences analysées et donc les motifs partiellement ordonnés fermés ne capturent pas les particularités (par exemple, l'occurrence répétitive d'un *itemset* dans un ensemble de séquences analysées) cachées dans ces séquences.

Pour résoudre ces problèmes, cette thèse présente une méthode d'exploration de données séquentielles à l'aide de l'Analyse Relationnelle de Concept (ARC), qui permet de prendre en compte des relations objet-objet par l'application itérative de l'Analyse de Concepts Formels (ACF) sur un ensemble de contextes formels. L'ACF est une méthode de classification qui s'applique à des jeux de données constitués d'objets décrits par des attributs. D'un point de vue mathématique, l'ACF permet d'extraire à partir d'une unique relation binaire objet-attribut un ensemble de concepts munis d'une structure hiérarchique appelé treillis de concepts. Un concept est constitué d'une extension et d'une intension : l'extension est l'ensemble maximal d'objets partageant le même ensemble maximal d'attributs qui constitue l'intension.

L'ARC prend en entrée une famille relationnelle de contextes, composée d'un ensemble de contextes formels et d'un ensemble de contextes relationels entre objets des contextes formels. Lors de la première itération, chaque contexte formel est utilisé pour générer un treillis. Pour les itérations qui suivent, les concepts créés à l'étape précédente sont intégrés sous forme d'attributs relationnels dans les contextes formels, pour enrichir la description des objets. En effet, grâce à une opération d'échelonnage (existentiel, par exemple) il est possible d'utiliser les relations objet-objet pour créer une relation objet-concept. Cette relation

représente le fait qu'un objet est en lien avec un ou plusieurs objets d'un concept via la relation objet-objet. Le nombre de liens requis dépend de l'opération d'échelonnage choisie. Durant l'itération, chaque contexte formel ainsi étendu permet alors de générer un nouveau treillis et les attributs rajoutés permettent potentiellement de faire émerger de nouveaux concepts par rapport à l'étape précédente. Lorsqu'aucun nouveau concept n'apparaît dans l'étape courante, le processus de l'ARC a atteint un point fixe et s'arrête.

La méthode que nous proposons, appelée RCA-Seq, est une mise en œuvre originale de l'ARC pour l'exploration de données séquentielles qualitatives (temporelles ou spatiales, par exemple), dont l'objectif principal est de faciliter l'étape d'évaluation des motifs séquentiels découverts. Une séquence analysée représente une liste d'ensembles d'*items* (*itemsets*) où un *itemset* contient des *items* auxquels sont associées des valeurs qualitatives (par exemple, une pomme rouge). L'approche proposée est un processus complet développé selon le schéma de l'ECBD, et couvre cinq étapes principales (Fig. 1.1 du manuscrit) :

1. le prétraitement des données en s'appuyant sur la connaissance du domaine ; ensuite, ces données sont remodelées afin de construire l'entrée de l'ARC selon le modèle de données proposé ;

2. l'exploration des données prétraitées ; de façon itérative, un algorithme d'ACF et un mécanisme d'échelonnage relationnel (qui met en évidence les relations entre les objets considérés) sont appliqués à l'entrée de l'ARC afin de dériver des treillis conceptuels interconnectés ;

3. l'extraction de motifs partiellement ordonnés fermés multi-niveaux organisés dans une hiérarchie en naviguant parmi les treillis interconnectés ;

4. la sélection automatique de motifs partiellement ordonnés fermés multi-niveaux pertinents en fonction de diverses mesures d'intérêt ;

5. l'évaluation de motifs par l'expert afin d'obtenir des informations pertinentes.

L'efficacité de RCA-Seq n'est pas notre principale préoccupation. En effet, comme l'ARC n'est pas destinée à l'exploration de grands jeux de données, notre approche se concentre plutôt sur l'exploration de petits jeux de données intéressants. En outre, la question que nous traitons, c'est-à-dire l'extraction directe de motifs partiellement ordonnés fermés multi-niveaux, qui sont implicitement organisés dans une hiérarchie, est nouvelle, et c'est une tâche plus difficile que l'extraction traditionnelle de motifs séquentiels.

Les contributions méthodologiques de cette thèse sont les suivantes :

– **une nouvelle méthode d'analyse relationnelle de données séquentielles qualitatives à l'aide de l'ARC**. Nous proposons un modèle général de données (Fig. 3.4 du manuscrit)

qui permet de préciser les liens entre les *itemsets* qualitatifs qui sont contenus dans les séquences analysées. Ce modèle général se compose de quatre entités : un ensemble d'*items*, un ensemble d'*itemsets* non cibles, un ensemble d'*items* d'intérêt et un ensemble d'*itemsets* cibles. Ces entités sont liées par des relations temporelles «est précédé par» et des relations qualitatives «a l'item avec la qualité». En utilisant ce modèle, diverses données séquentielles, telles que des données hydroécologiques ou médicales, peuvent être explorées à l'aide de l'ARC. Le modèle permet la conversion de données séquentielles en contextes formels et relationnels, qui représentent l'entrée de l'ARC. La richesse du résultat de l'ARC facilite l'étape d'évaluation grâce à l'organisation des concepts en hiérarchies et aux informations portées par les concepts, à savoir les objets de leurs extensions et les attributs de leurs intensions, dont les attributs relationnels qui rendent ces concepts interdépendants ;

- **l'extraction directe de hiérarchies de motifs partiellement ordonnés fermés**. Nous bénéficions du fait que certains motifs découverts sont naturellement des sous-motifs les uns des autres, et nous proposons d'extraire des hiérarchies de motifs partiellement ordonnés fermés où chaque motif est projeté dans ses descendants. Par conséquent, lorsqu'un motif partiellement ordonné fermé intéressant est trouvé, l'analyse peut continuer en explorant la zone environnante dans la hiérarchie ;

- **l'extraction de hiérarchies de motifs partiellement ordonnés fermés multi-niveaux avec deux niveaux de généralisation**. La généralisation concerne d'une part la structure des motifs (par exemple, le nombre d'items, les sommets et / ou les arêtes), et d'autre part la précision des items (par exemple, de abstrait à défini) ;

- **l'extraction de motifs partiellement ordonnés fermés multi-niveaux sans prétraiter les séquences d'origine**. Nous exploitons l'ordre sur les items révélé par l'ARC et nous extrayons des motifs partiellement ordonnés fermés multi-niveaux. Contrairement aux approches existantes pour l'extraction de motifs partiellement ordonnés fermés, on obtient au moyen de l'ARC deux nouveaux types de motifs : des motifs génériques qui représentent de façon abstraite les tendances communes des motifs séquentiels standards, et des motifs hybrides qui représentent à la fois des tendances communes et des informations spécialisées. L'existence de tels motifs et de la hiérarchie associée permet de faciliter l'analyse, en autorisant l'expert à circuler entre motifs au moyen des relations de spécialisation et de généralisation ;

- **des mesures d'intérêt pour la sélection et le filtrage de concepts et de motifs partiellement ordonnés fermés**. Nous proposons de faire face au problème de l'explosion du nombre de concepts et de motifs partiellement ordonnés fermés au moyen d'un nouvel

indice de distribution, qui utilise les informations portées par les objets d'une extension de concept afin de déterminer la pertinence du concept. De plus, nous présentons un motif partiellement ordonné fermé plus informatif, à savoir un motif partiellement ordonné fermé pondéré, qui aide à mieux comprendre le motif obtenu en capturant et en montrant explicitement les différents rôles de ses itemsets dans les séquences analysées sous-jacentes ;

– **une étude de l'adaptabilité de l'approche RCA-Seq**. Nous montrons que l'approche proposée peut être adaptée pour : (i) extraire des motifs partiellement ordonnés fermés contenant des items appartenant à différents niveaux d'une taxonomie définie par l'expert, (ii) spécifier des contraintes sur les relations d'ordre entre les itemsets de motifs partiellement ordonnés fermés découverts (par exemple, pour découvrir les motifs dont les itemsets comprennent au moins 50% des *items* utilisés pour construire les séquences analysées) et (iii) explorer des données séquentielles hétérogènes (les séquences qui sont construites en utilisant des items représentent des domaines divers).

Le manuscrit de thèse se compose de huit chapitres. Dans le deuxième chapitre, nous présentons l'état de l'art et les fondements théoriques de cette thèse : l'extraction de motifs séquentiels et l'ACF.

Dans le troisième chapitre, nous présentons les deux premières étapes de l'approche RCA-Seq : le prétraitement des données et l'exploration par l'ARC de données séquentielles. Un modèle générique de données est proposé. Ensuite, en s'appuyant sur ce modèle, nous expliquons comment encoder une base de séquences pour l'utiliser en entrée de l'ARC. En outre, la sortie de l'ARC obtenue est expliquée et analysée. Nous montrons que la navigation manuelle de la sortie de l'ARC afin de découvrir des régularités pertinentes n'est pas une tâche facile pour les experts du domaine, parce que le nombre de concepts peut être grand et, en outre, les experts doivent porter leur attention de concept en concept et de treillis à treillis en considérant des relations intra-treillis et inter-treillis.

Dans le quatrième chapitre, nous présentons l'étape suivante de RCA-Seq, précisément l'extraction directe d'une hiérarchie de motifs partiellement ordonnés fermés multi-niveaux à partir de la sortie de l'ARC obtenue. La structure et les propriétés de la sortie de l'ARC sont discutées. Ensuite, nous présentons un algorithme qui extrait automatiquement les motifs partiellement ordonnés fermés multi-niveaux. Deux optimisations de l'approche RCA-Seq sont présentées : une au niveau de l'exploration avec l'ARC et une autre au niveau de l'extraction. En outre, une analyse de complexité temporelle et spatiale de RCA-Seq est donnée.

Dans le cinquième chapitre, nous présentons de nouvelles mesures d'intérêt pour le guidage d'experts de domaine. La "richesse" de la sortie de l'ARC est exploitée pour calculer l'indice

de distribution (IQV) d'un concept formel, pour extraire les motifs partiellement ordonnés fermés multi-niveaux pondérés et pour catégoriser les motifs partiellement ordonnés fermés multi-niveaux obtenus en fonction de leur précision. De plus, nous présentons comment les motifs pondérés peuvent améliorer l'analyse des données séquentielles.

Dans le sixième chapitre, nous discutons l'adaptabilité de RCA-Seq. Une taxonomie définie par l'expert sur les éléments de construction de séquence et, en outre, les contraintes définies par l'utilisateur sur les relations d'ordre sur les itemsets sont poussées profondément dans l'étape d'exploration basée sur l'ARC. Ensuite, nous présentons comment explorer des données séquentielles hétérogènes afin d'obtenir des motifs partiellement ordonnés fermés multi-niveaux hétérogènes.

Dans le septième chapitre, nous présentons le contexte hydroécologique qui est le contexte d'application de cette thèse. Nous expliquons comment prétraiter les données hydroécologiques. Ensuite, nous décrivons et discutons les résultats obtenus à partir d'expérimentations réalisées sur différents ensembles de données hydroécologiques.

Dans le huitième chapitre, nous concluons et donnons quelques perspectives de cette thèse.

Nous avons évalué expérimentalement l'approche RCA-Seq et ses extensions sur des ensembles de données hydroécologiques collectés pendant deux projets de recherche interdisciplinaires, les projets Fresqueau[1] et REX[2]. Le projet ANR Fresqueau (2011-2015) portait sur le développement de méthodes innovantes pour l'analyse de données sur la qualité des eaux de rivière. Le projet REX (2015-2016), financé par l'École Nationale du Génie de l'Eau et de l'Environnement do Strasbourg, s'intéressait aux conditions et effets des opérations de restauration écologique menées sur différents sites de la plaine du Rhin. Les jeux de données du projet Fresqueau sont composés de séquences de valeurs, concernant des paramètres physico-chimiques et biologiques mesurés dans des stations (sites d'étude) de rivières. L'objectif est de relier les deux types de paramètres. Nous montrons sur différents jeux de données que l'approche RCA-Seq permet de mettre en évidence l'influence dans le temps des paramètres physico-chimiques sur les paramètres biologiques au moyen des hiérarchies de motifs partiellement ordonnés fermés multi-niveaux (pondérés ou non). Les jeux de données du projet REX sont composés de données hétérogènes concernant des paramètres biologiques et physico-chimiques de l'eau et l'occupation du sol relevée autour des stations de rivières (Fig. 7.23 du manuscrit). De plus, des informations concernant la restauration des tronçons de rivières sont analysées. Les stations de rivières sont intégrées dans un réseau représenté sous forme de graphe orientés (réseau des stations). Nous montrons ici que notre approche est également appropriée pour la fouille de données sur des graphes.

[1] http://engees-fresqueau.unistra.fr/presentation.php?lang=en

[2] http://obs-rhin.engees.eu

Pour le données Fresqueau on peut trouver un exemple de résultat sur la Fig. 7.16 du manuscrit. Cette figure présente une partie de la hiérarchie des motifs partiellement ordonnés fermés multi-niveaux extraits d'un jeu de données relatives à des stations de rivières pour lesquelles l'indice biologique global normalisé (IBGN) indique un très bon état. Cet indice se base sur l'analyse des macro-invertébrés présents sur la station. On peut voir sur la figure des motifs abstraits, des motifs hybrides et des motifs concrets. Chaque motif porte trois mesures : le support (le nombre de séquences qui contiennent le motif), la richesse (le nombre de stations de rivières distinctes où ont été échantillonnées les mesures des séquences qui contiennent le motif) et l'IQV (indicateur de la distribution des mesures biologiques parmi les stations de rivières). Par exemple, les deux motifs (b) et (c) soulignent deux correspondances bien connues entre les valeurs qualitatives des macro-paramètres physico-chimiques et celles de l'indicateur biologique IBGN. Le motif (b) souligne que l'IBGN bleu est mesuré lorsqu'il est précédé par des valeurs qualitatives bleues des macro-paramètres physico-chimiques. Les mesures associées au motif (c) illustrent le fait que les macro-paramètres physico-chimiques avec des valeurs qualitatives rouges ne sont pas fréquemment mesurées avant l'IBGN bleu car ils montrent une dégradation de la qualité de l'eau et ne conduisent pas à un très bon état écologique. De plus, grâce aux motifs partiellement ordonnés fermés multi-niveaux qui sont extraits en utilisant RCA-SEQ, le motif hybride (i), ayant un support de 18 peut être trouvé lorsque, par exemple un support minimum égal à 15 est utilisé même si le motif précis (r), qui est une spécialisation de (i), n'a qu'un support de 9 et n'est pas découvert.

Dans un motif partiellement ordonné fermé multi-niveaux pondéré, chaque sommet est étiqueté avec un itemset et peut être annoté avec un 3-tuple (*persistance*, *spécificité*, *poids total*) qui capture les particularités cachées dans les données analysées. La *persistance* est une mesure qui illustre le caractère répétitif d'un itemset dans chaque séquence qui supporte le motif. La *spécificité* est une mesure qui illustre l'appartenance exclusive d'un itemset aux séquences qui supportent le motif. Le *poids total* est le nombre total d'occurrences de cet itemset dans toutes les séquences analysées. Du point de vue de l'interprétation, on peut détecter *via* un motif : une forte pollution (mise en évidence par les valeurs qualitatives des macro-paramètres), une pollution persistante (mise en évidence par la mesure de la persistance des sommets) et une combinaison de pollutions différentes (soulignées par les sommets qui contiennent de nombreux macro-paramètres, qui représentent des types distincts de pollution). La Fig. 7.19 du manuscrit illustre un motif partiellement ordonné fermé multi-niveaux hybride et pondéré qui a été découvert dans 4 séquences du jeu de données associé à la valeur orange de l'IBGN. Dans cet exemple, l'itemset $(\texttt{NITRO}_{\texttt{orange}}\ \texttt{PAES}_{\texttt{green}})$ révèle la présence d'une mauvaise pollution organique (NITRO) avec une persistance de $0,75$, une spécificité de 78% et un poids global égal à 9. Un autre example, celui de la Fig. 7.19 révèle

un seul type de pollution spécifique, précisément organique (NITRO orange, NITRO rouge, MOOX rouge). Ensuite, des pressions physico-chimiques fortes peuvent être révélées par une répétition de mauvaises valeurs pour les paramètres physico-chimiques. C'est le cas dans ce dernier exemple où les pressions physico-chimiques sont fortes car il y a sept occurrences des macro-paramètres mauvais (orange) et très mauvais (rouge).

Dans un motif partiellement ordonné fermé multi-niveaux hétérogène, chaque sommet est étiqueté avec un itemset et peut être annoté avec des itemsets révélant des informations provenant de différents domaines. La Fig. 7.26 du manuscrit montre un motif partiellement ordonné fermé multi-niveaux hétérogène trouvé dans le jeu de données REX. Ce motif est associé à deux tronçons de rivières (valeur indiquée dans le carré noir sur le rectangle) qui contiennent au plus deux endroits qui ont été restaurés globalement. Les sommets Ⓐ, Ⓑ, Ⓒ, Ⓓ, Ⓔ et Ⓕ du motif sont associés aux stations de rivières. Le motif révèle que, localement (Ⓐ, Ⓑ, Ⓒ et Ⓓ), les pressions exercées par les zones bâties étaient moyennes dans un rayon de 500 m autour de ces stations. En revanche, pour les rivières en amont (Ⓔ et Ⓕ) dans un rayon de 500 m, d'une part, les pressions exercées sur la rivière par les zones industrielles bâties et les terres arables étaient faibles ($< 25\%$) ; d'autre part, un pourcentage élevé ($> 40\%$) de la zone est recouvert de forêts, qui favorisent à un bon état écologique de l'écosystème aquatique. En comparant le sommet Ⓔ avec Ⓐ, Ⓑ, Ⓒ et Ⓓ on note une dégradation concernant les valeurs qualitatives des paramètres physico-chimiques, probablement causée par les pressions liées au bâti. Par exemple, l'ammonium bleu (très bon) de Ⓔ est mesuré lorsque, dans les environs, les pressions liées à l'occupation des sols sont faibles ; l'ammonium vert (bon) de Ⓐ, Ⓑ, Ⓒ et Ⓓ est mesuré lorsque, dans les environs, les pressions liées à l'occupation des sols sont moyennes.

Pour résumer, nous avons conçu une approche dont les résultats facilitent l'évaluation des motifs partiellement ordonnés fermés découverts dans les données séquentielles comme suit :

1. les experts peuvent naviguer entre les motifs obtenus en étant guidés par les relations entre eux ;

2. les motifs plus ou moins abstraits peuvent révéler des résultats qui ne peuvent être trouvés par de motifs concrets avec des valeurs élevées du support et aussi peuvent donner un aperçu des tendances des données analysées ;

3. les motifs partiellement ordonnés fermés pondérés mettent en évidence les particularités liées aux differents itemset présents dans les séquences analysées ;

Par ailleurs, il convient de mentionner que l'approche RCA-Seq peut être appliquée à toute donnée pouvant être modélisée selon le modèle de données générique proposé, comme

par exemple la trajectoire d'un joueur de football avant un match menant à une séquence de sessions de formation suivie d'une évaluation de joueur.

Les contributions présentées dans cette thèse ouvrent plusieurs orientations de recherche. Par exemple :

- RCA-Seq peut être amélioré pour être applicable à de gros volumes de données. En fait, puisque l'ARC classique ne fait pas face aux grands ensembles de données, la version actuelle de RCA-Seq n'est pas une approche basée sur l'efficacité, mais se concentre plutôt sur l'exploration de jeux de données petits mais intéressants, comme ceux conçus lors des projets Fresqueau et REX, afin d'améliorer l'étape d'analyse des motifs. Pour résoudre le problème de l'explosion de concepts, il sera intéressant d'améliorer l'étape d'exploration des données séquentielles basée sur l'ARC au moyen des Attribute-Object-Concept-posets ;

- il sera intéressant d'utiliser des mesures d'intérêt dès la phase d'exploration basée sur l'ARC pour éviter l'explosion de motifs partiellement ordonnés fermés. Habituellement, le support est utilisé pour élaguer les motifs partiellement ordonnés fermés peu fréquents. Dans cette thèse, nous avons déjà utilisé les treillis d'Iceberg qui exploitent le support. Cependant, on peut essayer d'introduire l'indice de distribution dans l'étape d'exploration basée sur l'ARC et, par conséquent, d'extraire directement uniquement les motifs partiellement ordonnés fermés pertinents ;

- il sera intéressant de concevoir un outil qui extrait de manière interactive et itérative les motifs à partis des treillis produits par l'ARC plutôt que de les extraire tous à la fois. À cette fin, les étapes d'extraction et d'évaluation de RCA-Seq peuvent être considérées comme une étape itérative. Plus précisément, les experts peuvent s'appuyer sur des mesures d'intérêt pour sélectionner quelques concepts intéressants, pour lesquels on extrait les motifs partiellement ordonnés fermés multi-niveaux associés. L'extraction de motifs partiellement ordonnés fermés multi-niveaux peut se poursuivre en fonction d'autres concepts sélectionnés en utilisant à nouveau les mesures d'intérêt ou en fonction des concepts qui entourent les concepts sélectionnés précédemment. Ce processus itératif peut continuer de la même manière. Pour résumer, un outil interactif peut améliorer l'évaluation de motifs, car les experts du domaine peuvent évaluer de manière progressive et systématique les motifs découverts plutôt que d'être submergés par le nombre exponentiel de motifs ;

- il sera intéressant d'étudier les différents quantificateurs qui peuvent être utilisés pendant le mécanisme d'échelonnage relationnel et d'analyser les motifs partiellement ordonnés fermés qui sont alors extraits. De fait, dans cette thèse, nous nous sommes

concentrés uniquement sur le quantificateur existentiel et ses variantes appliquées aux relations d'ordre sur les itemsets ;

– il sera intéressant d'effectuer une évaluation expérimentale de RCA-Seq par comparaison à des méthodes sur des ensembles de données de référence. Dans cette thèse, nous avons présenté une analyse théorique de la complexité du temps qui montre – dans le pire des cas – la meilleure performance de RCA-Seq par rapport à une méthode ad hoc qui rejoint l'ACF et une méthode existante pour l'extraction de motifs partiellement ordonnés fermés.

List of Symbols

I	Item
$\mathcal{I}$	Set of items
$IS = (I_{j_1}...I_{j_k})$	Itemset
$\mathcal{IS}$	Set of all itemsets built from a set of items
$S = \langle IS_1...IS_p \rangle$	Sequence of itemsets
$\preceq_s$	Order on sequences
$\mathcal{D}_\mathcal{S}$	Sequence database
$(\mathcal{I}, \leq)$	Poset of items
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, l)$	Closed partially-ordered pattern (cpo-pattern)
$\mathcal{V}$	Set of vertices
$\mathcal{E}$	Set of edges
v	Vertex
$l(v)$	Function that maps vertex v to an itemset
$\mathcal{P}_\mathcal{G}$	Set of all paths in $\mathcal{G}$
$\mathcal{S}_\mathcal{G}$	Set of sequences that support $\mathcal{G}$
$\preceq_g$	Order on cpo-patterns
$K = (G, M, I)$	Formal context with the incidence relation $I \subseteq G \times M$
G	Set of objects
M	Set of attributes

$\mathcal{K}$	Set of formal contexts
$C = (X, Y)$	Formal concept
X	Extent of a formal concept
Y	Intent of a formal concept
$\mathcal{C}_K$	Set of all formal concepts derived from K
$\preceq_K$	Generalisation order on the formal concepts derived from K
$\rhd_K$	Order on formal concepts (upper neighbour)
$\lhd_K$	Order on formal concepts (lower neighbour)
$\mathcal{L}_K = (\mathcal{C}_K, \preceq_K)$	Concept lattice of K
$R = (G_k, G_l, r)$	Relational context
$r \subseteq G_k \times G_l$	Binary relation
$dom(r)$	Domain of r (G_k)
$ran(r)$	Range of r (G_l)
$\mathcal{R}$	Set of relational contexts
$(\mathcal{K}, \mathcal{R})$	Relational context family
K^+	Scaled context
$\exists r(C)$	Relational attribute with the $\exists$ quantifier
$(Object, Date)$	Temporal object
UID	Unique identifier
$\mathcal{UID}_\mathcal{S}$	Database of sequences of UIDs
`IS_Seq`	UID of a non-target itemset
`Seq`	UID of a target itemset
G_M	Set of all target itemsets in a $\mathcal{UID}_\mathcal{S}$
G_T	Set of all non-target itemsets in a $\mathcal{UID}_\mathcal{S}$
G_I	Set of all sequence-building items

$getS$	Maps the UID of a target/non-target itemset to the sequence that owns the itemset
$getIS$	Maps a UID to a target/non-target itemset
ipb	Temporal relation *is preceded by*
hi_q	Qualitative relation *has item quality*
$\mathcal{L}_{K_M}$	Main lattice built from $K_M = (G_M, M_M, I_M)$
$\mathcal{L}_{K_T}$	Temporal lattice built from $K_T = (G_T, M_T, I_T)$
$\mathcal{L}_{K_I}$	Lattice of items built from $K_I = (G_I, M_I, I_I)$
θ	User-defined minimum support for $\mathcal{L}_{K_M}$
θ'	User-defined minimum support for $\mathcal{L}_{K_T}$
$\texttt{item}_\texttt{q}$	Concrete qualitative item
$?_\texttt{q}$	Abstract qualitative item
$?_?$	Abstract item
$extent(C)$	Set of objects from the extent of C
IQV	Index of Qualitative Variation
ρ	Richness of a formal concept
ϕ_o	Absolute frequency of an object $o \in X$
$\bar{X}$	Set of distinct objects in extent X, which contains temporal objects
$\bar{X}_\phi$	Distribution of a formal concept whose extent X contains temporal objects
$\upsilon(\mathcal{G})$	Accuracy of cpo-pattern $\mathcal{G}$
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, l, w)$	Weighted cpo-pattern
ϖ_v	Persistency of vertex v
ω_v	Overall weight of vertex v
ς_v	Specificity of vertex v

Bibliography

AFNOR (2002). *Qualité de l'eau : détermination de l'Indice Oligochètes de Bioindication des Sédiments (IOBS). Norme Française NF T90-390.* Association Française de NORmalisation.

AFNOR (2003). *Qualité de l'eau : détermination de l'Indice Biologique Macrophytique en Rivière (IBMR). Norme Francaise NF T90-395.* Association Française de NORmalisation.

AFNOR (2004b). *Qualité de l'eau : détermination de l'Indice poissons rivière (IPR). Norme Française NF T90-344.* Association Française de NORmalisation.

AFNOR (2007). *Qualité de l'eau : détermination de l'Indice Biologique Diatomées (IBD). Norme Française NF T90-354.* Association Française de NORmalisation.

AFNOR (Mars 2004a). *Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). Norme Française NF T90-350*. Association Française de NORmalisation.

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216. ACM.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499. Morgan Kaufmann Publishers Inc.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In *Int. Conference on Data Engineering*, pages 3–14.

Agrawal, S. and Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708 – 713. Knowledge-Based and Intelligent Information & Engineering Systems 19th Annual Conference, KES-2015, Singapore, September 2015 Proceedings.

Al-Twijri, M. I. and Noaman, A. Y. (2015). A new data mining model adopted for higher institutions. *Procedia Computer Science*, 65:836 – 844. International Conference on Communications, management, and Information technology (ICCMIT'2015).

Almasri, M. N. and Kaluarachchi, J. J. (2004). Assessment and management of long-term nitrate pollution of ground water in agriculture-dominated watersheds. *Journal of Hydrology*, 295(1–4):225 – 245.

Alqadah, F. and Bhatnagar, R. (2011). Similarity measures in formal concept analysis. *Annals*

of Mathematics and Artificial Intelligence, 61(3):245–256.

Amani, F. A. and Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24:32 – 58.

Andrews, S. (2017). Making use of empty intersections to improve the performance of cbo-type algorithms. In *Formal Concept Analysis: 14th International Conference, ICFCA 2017, Rennes, France, June 13-16, 2017, Proceedings*, pages 56–71, Cham. Springer International Publishing.

Ansari, S., Kohavi, R., Mason, L., and Zheng, Z. (2001). Integrating e-commerce and data mining: Architecture and challenges. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 27–34. IEEE Computer Society.

Arévalo, G., Falleri, J.-R., Huchard, M., and Nebut, C. (2006). Building abstractions in class models: Formal concept analysis in a model-driven approach. In *Model Driven Engineering Languages and Systems: 9th International Conference, MoDELS. Proceedings*, pages 513–527, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ayres, J., Flannick, J., Gehrke, J., and Yiu, T. (2002). Sequential pattern mining using a bitmap representation. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 429–435. ACM.

Azmeh, Z., Driss, M., Hamoui, F., Huchard, M., Moha, N., and Tibermacine, C. (2011a). Selection of composable web services driven by user requirements. In *2011 IEEE International Conference on Web Services*, pages 395–402.

Azmeh, Z., Huchard, M., Napoli, A., Rouane Hacene, A. M., and Valtchev, P. (2011b). Querying Relational Concept Lattices. In *CLA: Concept Lattices and their Applications*, pages 377–392.

Barbut, M. and Monjardet, B. (1970). *Ordre et classification : algebre et combinatoire*. Hachette.

Behera, S. K., Kim, H. W., Oh, J.-E., and Park, H.-S. (2011). Occurrence and removal of antibiotics, hormones and several other pharmaceuticals in wastewater treatment plants of the largest industrial city of korea. *Science of The Total Environment*, 409(20):4351 – 4360.

Belohlavek, R. and Trnecka, M. (2013). Basic level in formal concept analysis: Interesting concepts and psychological ramifications. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pages 1233–1239. AAAI Press.

Bendaoud, R., Napoli, A., and Toussaint, Y. (2008). Formal concept analysis: A unified framework for building and refining ontologies. In *Knowledge Engineering: Practice and Patterns: 16th International Conference, EKAW. Proceedings*, pages 156–171, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bermúdez-Couso, A., Fernández-Calviño, D., Álvarez Enjo, M. A., Simal-Gándara, J., Nóvoa-Muñoz, J. C., and Arias-Estévez, M. (2013). Pollution of surface waters by metalaxyl and

nitrate from non-point sources. *Science of The Total Environment*, 461–462:282 – 289.

Berrahou, L., Lalande, N., Serrano, E., Molla, G., Berti-Équille, L., Bimonte, S., Bringay, S., Cernesson, F., Grac, C., Ienco, D., Le Ber, F., and Teisseire, M. (2015). A quality-aware spatial data warehouse for querying hydroecological data. *Computers & Geosciences*, 85, Part A:126–135.

Berry, A., Gutierrez, A., Huchard, M., Napoli, A., and Sigayret, A. (2014). Hermes: a simple and efficient algorithm for building the aoc-poset of a binary relation. *Annals of Mathematics and Artificial Intelligence*, 72(1):45–71.

Beyer, K. and Ramakrishnan, R. (1999). Bottom-up computation of sparse and iceberg cube. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, pages 359–370. ACM.

Birkhoff, G. (1967). Lattice theory. In *Colloquium Publications*, volume 25. Amer. Math. Soc., 3. edition.

Bordat, J. P. (1986). Calcul pratique du treillis de galois d'une correspondance. *Mathématiques et Sciences Humaines*, 96:31–47.

Braud, A., Nica, C., Grac, C., and Le Ber, F. (2011). A lattice-based query system for assessing the quality of hydro-ecosystems. In A Napoli, V. V., editor, *CLA 2011*, pages 265–277. INRIA Nancy-Grand-Est et LORIA.

Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S. O., Napoli, A., and Raïssi, C. (2016). On mining complex sequential data by means of FCA and pattern structures. *International Journal of General Systems*, 45:135–159.

Buzmakov, A., Kuznetsov, S. O., and Napoli, A. (2014). Scalable estimates of concept stability. In *Proceedings of the 12th International Conference on Formal Concept Analysis*, volume 8478 of *ICFCA'14*, pages 157 – 172. Springer.

Cao, J., Wu, Z., and Wu, J. (2014). Scaling up cosine interesting pattern discovery: A depth-first method. *Inf. Sci.*, 266:31–46.

Carpineto, C. and Romano, G. (1996). A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, 24(2):95–122.

Casas-Garriga, G. (2005). Summarizing sequential data with closed partial orders. In *2005 SIAM Int. Conference on Data Mining*, pages 380–391.

Cellier, P., Ferré, S., Ducassé, M., and Charnois, T. (2011). Partial orders and logical concept analysis to explore patterns extracted by data mining. In *Int. Conf. on Conceptual Structures for Discovering Knowledge*, pages 77–90. Springer.

Chakrabarti, D. and Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1).

Chang, J. H. (2011). Mining weighted sequential patterns in a sequence database with a time-interval weight. *Know.-Based Syst.*, 24(1):1 – 9.

Chen, J. (2010). An updown directed acyclic graph approach for sequential pattern mining. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):913–928.

Chen, Y.-L., Chiang, M.-C., and Ko, M.-T. (2003). Discovering time-interval sequential patterns in sequence databases. *Expert Systems with Applications*, 25(3):343 – 354.

Chowdhury Farhan, A., Syed Khairuzzaman, T., and Byeong-Soo, J. (2010). A novel approach for mining high-utility sequential patterns in sequence databases. *Electronics and Telecommunications Research Institute Journal*, 32(5):676–686.

Codocedo, V. and Napoli, A. (2014). A proposition for combining pattern structures and relational concept analysis. In *Formal Concept Analysis: 12th Int. Conf., ICFCA, Proceedings*, pages 96–111. Springer.

Codocedo, V. and Napoli, A. (2015). Formal concept analysis and information retrieval – a survey. In *Formal Concept Analysis: 13th International Conference, ICFCA 2015, Nerja, Spain, June 23-26, 2015, Proceedings*, pages 61–77. Springer International Publishing.

Dakou, E., D'heygere, T., Dedecker, A. P., Goethals, P. L. M., Lazaridou-Dimitriadou, M., and De Pauw, N. (2007). Decision tree models for prediction of macroinvertebrate taxa in the river axios (northern greece). *Aquatic Ecology*, 41(3):399–411.

Dao, M., Huchard, M., Hacène, M. R., Roume, C., and Valtchev, P. (2004). Improving generalization level in uml models iterative cross generalization in practice. In *Conceptual Structures at Work: 12th International Conference on Conceptual Structures, ICCS. Proceedings*, pages 346–360. Springer Berlin Heidelberg.

Davey, B. A. and Priestley, H. A. (1990). *Introduction to lattices and order*. Cambridge university press.

Dehaspe, L. and Toivonen, H. (1999). Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36".

Dias, S. M. and Vieira, N. J. (2010). Reducing the size of concept lattices: The jbos approach. In *Proceedings of the 7th International Conference on Concept Lattices and Their Applications, CLA 2010.*, pages 80–91. CEUR-WS.org.

Dias, S. M. and Vieira, N. J. (2015). Concept lattices reduction: Definition, analysis and classification. *Expert Systems with Applications*, 42(20):7084 – 7097.

Dolques, X., Le Ber, F., Huchard, M., and Grac, C. (2016). Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. *International Journal of General Systems*, 45(2):187–210.

Dolques, X., Le Ber, F., Huchard, M., and Nebut, C. (2015). *Advances in Knowledge Discovery and Management: Volume 5*, chapter Relational Concept Analysis for Relational Data Exploration, pages 57–77. Springer International Publishing.

Dolques, X., Mondal, K. C., Braud, A., Huchard, M., and Le Ber, F. (2014). RCA as a data transforming method: a comparison with propositionalisation. In *Proceedings of the 12th*

International Conference on Formal Concept Analysis, number 8478 in ICFCA'14, pages 112–127. Springer.

Dong, G. and Li, J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '99, pages 43–52. ACM.

Džeroski, S. (2003). Multi-relational data mining: An introduction. *SIGKDD Explor. Newsl.*, 5(1):1–16.

Egho, E., Jay, N., Raïssi, C., Ienco, D., Poncelet, P., Teisseire, M., and Napoli, A. (2014). A contribution to the discovery of multidimensional patterns in healthcare trajectories. *Journal of Intelligent Information Systems*, 42(2):283–305.

Egho, E., Jay, N., Raïssi, C., and Napoli, A. (2011). A FCA-based analysis of sequential care trajectories. In *The Eighth International Conference on Concept Lattices and their Applications - CLA 2011*, pages 1–11.

Esposito, F., Di Mauro, N., Basile, T. M. A., and Ferilli, S. (2009). Multi-dimensional relational sequence mining. *Fundam. Inf.*, 89(1):23–43.

European Union (2000). Directive 2000/60/ec of the European parliament and of the council of 23 october 2000 establishing a framework for community action in the field of water policy. *Official Journal*, OJ L 327:1–73.

Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., and Teisseire, M. (2014). Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics*, 24:210–221.

Fabrègue, M., Braud, A., Bringay, S., Le Ber, F., and Teisseire, M. (2015). Mining closed partially ordered patterns, a new optimized algorithm. *Know.-Based Syst.*, 79:68–79.

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. American Association for Artificial Intelligence.

Ferré, S. (2007). The efficient computation of complete and concise substring scales with suffix trees. In *Formal Concept Analysis*, pages 98–113. Springer.

Ferré, S. (2015). A proposal for extending formal concept analysis to knowledge graphs. In *Formal Concept Analysis: 13th International Conference, ICFCA 2015, Proceedings*, pages 271–286. Springer International Publishing.

Ferré, S. and Ridoux, O. (2000). A logical generelization of formal concept analysis. Technical report, Unité de recherche INRIA Rennes.

Ferreira, C. A., Gama, J., and Costa, V. S. (2015). Exploring multi-relational temporal databases with a propositional sequence miner. *Progress in Artificial Intelligence*, 4(1):11–20.

Formica, A. (2008). Concept similarity in formal concept analysis: An information content

approach. *Know.-Based Syst.*, 21(1):80–87.

Fournier-Viger, P., Gomariz, A., Campos, M., and Thomas, R. (2014). Fast vertical mining of sequential patterns using co-occurrence information. In *Advances in Knowledge Discovery and Data Mining: Proceedings of the 18th Pacific-Asia Conf., PAKDD, Part I*, pages 40–52. Springer.

Fournier-Viger, P., Lin, J. C., Kiran, R. U., Koh, Y. S., and Thomas, R. (2017). A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1):54–77.

Frankfort-Nachmias, C. and Leon-Guerrero, A. (2010). *Social Statistics for a Diverse Society*, chapter Measures of Variability. SAGE Publications.

Fumarola, F., Lanotte, P. F., Ceci, M., and Malerba, D. (2016). Clofast: Closed sequential pattern mining using sparse and vertical id-lists. *Knowl. Inf. Syst.*, 48(2):429–463.

Ganter, B. (1984). Two basic algorithms in concept analysis. Technical report fb4- preprint no. 831, TH Darmstadt.

Ganter, B. and Kuznetsov, S. O. (2001). Pattern structures and their projections. In *Conceptual Structures: Broadening the Base: 9th International Conference on Conceptual Structures, ICCS 2001, Proceedings*, pages 129–142. Springer Berlin Heidelberg.

Ganter, B. and Wille, R. (1989). Conceptual scaling. In *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences*, pages 139–167. Springer US.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis: Mathematical Foundations*. Springer.

Garofalakis, M. N., Rastogi, R., and Shim, K. (1999). Spirit: Sequential pattern mining with regular expression constraints. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pages 223–234. Morgan Kaufmann Publishers Inc.

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3).

Godin, R. and Mili, H. (1993). Building and maintaining analysis-level class hierarchies using galois lattices. In *Proceedings of the 8th Annual Conf. on Object-oriented Programming Systems, Languages, and Applications*, OOPSLA'93, pages 394–410. ACM.

Godin, R., Missaoui, R., and Alaoui, H. (1995). Incremental concept formation algorithms based on galois (concept) lattices. *Computational Intelligence*, 11(2):246–267.

Goethals, P. L. M., Dedecker, A. P., Gabriels, W., Lek, S., and De Pauw, N. (2007). Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology*, 41(3):491–508.

Gomariz, A., Campos, M., Marin, R., and Goethals, B. (2013). Clasp: An efficient algorithm for mining frequent closed sequences. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Proceedings, Part I*, pages 50–61. Springer.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., 3rd edition.

Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U., and Hsu, M.-C. (2000). Freespan: Frequent pattern-projected sequential pattern mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pages 355–359. ACM.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8(1):53–87.

Hébert, C. and Crémilleux, B. (2005). Mining frequent δ-free patterns in large databases. In *Proceedings of the 8th International Conference on Discovery Science*, DS'05, pages 124–136. Springer-Verlag.

Hu, Y.-H., Huang, T. C.-K., Yang, H.-R., and Chen, Y.-L. (2009). On mining multi-time-interval sequential patterns. *Data & Knowledge Engineering*, 68(10):1112 – 1127.

Huchard, M., Hacene, M. R., Roume, C., and Valtchev, P. (2007). Relational concept discovery in structured datasets. *Annals of Mathematics and Artificial Intelligence*, 49(1):39–76.

Jacobs, N. and Blockeel, H. (2001). From shell logs to shell scripts. In *Inductive Logic Programming: 11th Int. Conf., ILP 2001, Proceedings*, pages 80–90. Springer Berlin Heidelberg.

Jay, N., Kohler, F., and Napoli, A. (2008). Analysis of social communities with iceberg and stability-based concept lattices. In *Proceedings of the 6th International Conference on Formal Concept Analysis*, ICFCA'08, pages 258–272. Springer-Verlag.

Jothi, N., Rashid, N. A., and Husain, W. (2015). Data mining in healthcare – a review. *Procedia Computer Science*, 72:306 – 313. The Third Information Systems International Conference 2015.

Kim, C., Lim, J.-H., Ng, R. T., and Shim, K. (2007). Squire: Sequential pattern mining with quantities. *Journal of Systems and Software*, 80(10):1726 – 1745.

Kimball, R. and Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., 2nd edition.

Klimushkin, M., Obiedkov, S., and Roth, C. (2010). Approaches to the selection of relevant concepts in the case of noisy data. In *Formal Concept Analysis*, pages 255–266. Springer.

Kocev, D., Naumoski, A., Mitreski, K., Krstić, S., and Džeroski, S. (2010). Learning habitat models for the diatom community in lake prespa. *Ecological Modelling*, 221(2):330 – 337.

Kötters, J. (2016). Intension graphs as patterns over power context families. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, CLA 2016.*, pages 203–216. CEUR-WS.org.

Kramer, S., Lavrač, N., and Flach, P. (2001). Propositionalization approaches to relational data mining. In *Relational Data Mining*, pages 262–291. Springer Berlin Heidelberg.

Kuznetsov, S., Obiedkov, S., and Camille, R. (2007). Reducing the representation complexity of lattice-based taxonomies. In *Proceedings of the 15th International Conference on Conceptual Structures*, ICCS'07, pages 241–254. Springer Berlin Heidelberg.

Kuznetsov, S. O. (1993). A fast algorithm for computing all intersections of objects in a finite semi-lattice. *Automatic Documentation and Mathematical Linguistics*, 14:11–21.

Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):101–115.

Kuznetsov, S. O. and Obiedkov, S. (2002). Comparing performance of algorithms for generating concept lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14:189–216.

Lafont, M., Camus, J.-C., Fournier, A., and Sourp, E. (2001). A practical concept for the ecological assessment of aquatic ecosystems: application on the river dore in france. *Aquatic Ecology*, 35(2):195–205.

Lehmann, F. and Wille, R. (1995). A triadic approach to formal concept analysis. In *Proceedings of the Third International Conference on Conceptual Structures: Applications, Implementation and Theory*, ICCS '95, pages 32–43. Springer-Verlag.

Liao, V. C.-C. and Chen, M.-S. (2013). Efficient mining gapped sequential patterns for motifs in biological sequences. *BMC Systems Biology*, 7(4).

Luo, C. and Chung, S. M. (2005). Efficient mining of maximal sequential patterns using multiple samples. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 415–426.

Luo, Y., Guo, W., Ngo, H. H., Nghiem, L. D., Hai, F. I., Zhang, J., Liang, S., and Wang, X. C. (2014). A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *Science of The Total Environment*, 473–474:619 – 641.

Mabroukeh, N. R. and Ezeife, C. I. (2010). A taxonomy of sequential pattern mining algorithms. *ACM Comput. Surv.*, 43(1):3:1–3:41.

Mannila, H., Toivonen, H., and Inkeri Verkamo, A. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289.

Melo, C., Le-Grand, B., Aufaure, M.-A., and Bezerianos, A. (2011). Extracting and visualising tree-like structures from concept lattices. In *Proceedings of the 2011 15th International Conference on Information Visualisation*, IV '11, pages 261–266. IEEE Computer Society.

Merwe, D., Obiedkov, S. A., and Kourie, D. G. (2004). Addintent: A new incremental algorithm for constructing concept lattices. In *Eklund P. (eds) Concept Lattices. ICFCA 2004. Lecture Notes in Computer Science*, volume 2961. Springer, Berlin, Heidelberg.

Moha, N., Rouane Hacene, A. M., Valtchev, P., and Guéhéneuc, Y.-G. (2008). Refactorings of design defects using relational concept analysis. In *Formal Concept Analysis: 6th International Conference, ICFCA 2008. Proceedings*, pages 289–304, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mondy, C. P. and Usseglio-Polatera, P. (2013). Using conditional tree forests and life history traits to assess specific risks of stream degradation under multiple pressure scenario.

Science of The Total Environment, 461:750 – 760.

Mooney, C. H. and Roddick, J. F. (2013). Sequential pattern mining – approaches and algorithms. *ACM Comput. Surv.*, 45(2):19:1–19:39.

Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, 8(4):295–318.

Nica, C., Braud, A., Dolques, X., Huchard, M., and Ber, F. L. (2016a). Exploring temporal data using relational concept analysis: An application to hydroecological data. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, CLA 2016.*, pages 299–311. CEUR-WS.org.

Nica, C., Braud, A., Dolques, X., Huchard, M., and Le Ber, F. (2016b). Extracting hierarchies of closed partially-ordered patterns using relational concept analysis. In *Graph-Based Representation and Reasoning: 22nd International Conference on Conceptual Structures, ICCS 2016, Proceedings*, pages 17–30. Springer.

Nica, C., Braud, A., and Le Ber, F. (2017). Hierarchies of multilevel closed partially-ordered patterns for enhancing sequential data analysis. –. (submitted).

Norris, E. M. (1978). An algorithm for computing the maximal rectangles in a binary relation. *Revue Roumaine de Mathématiques Pures et Appliquées*, 23(2):243–250.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings of the 17th Int. Conf. on Data Engineering*, ICDE'01, pages 215–224. IEEE Computer Society.

Pei, J., Han, J., and Wang, W. (2002). Mining sequential patterns with constraints in large databases. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, CIKM '02, pages 18–25. ACM.

Pei, J., Wang, H., Liu, J., Wang, K., Wang, J., and Yu, P. S. (2006). Discovering frequent closed partial orders from strings. *IEEE Transactions on Knowledge and Data Engineering*, 18(11):1467–1481.

Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., and Dayal, U. (2001). Multi-dimensional sequential pattern mining. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 81–88. ACM.

Pitarch, Y., Ienco, D., Vintrou, E., Bégué, A., Laurent, A., Poncelet, P., Sala, M., and Teisseire, M. (2015). Spatio-temporal data classification through multidimensional sequential patterns: Application to crop mapping in complex landscape. *Eng. Appl. of AI*, 37:91–102.

Plantevit, M., Laurent, A., Laurent, D., Teisseire, M., and Choong, Y. W. (2010). Mining multidimensional and multilevel sequential patterns. *ACM Trans. Knowl. Discov. Data*, 4(1):4:1–4:37.

Poelmans, J., Elzinga, P., Viaene, S., and Dedene, G. (2010a). A Method based on Temporal Concept Analysis for Detecting and Profiling Human Trafficking Suspects. In *Artificial*

Intelligence and Applications, AIA 2010, Innsbruck, Austria, pages 1–9.

Poelmans, J., Elzinga, P., Viaene, S., and Dedene, G. (2010b). Formal concept analysis in knowledge discovery: A survey. In *Proceedings of the 18th International Conference on Conceptual Structures: From Information to Intelligence*, ICCS'10, pages 139–153. Springer-Verlag.

Poelmans, J., Ignatov, D. I., Kuznetsov, S. O., and Dedene, G. (2013). Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 40(16):6538 – 6560.

Pokou, Y. J. M., Fournier-Viger, P., and Moghrabi, C. (2016). Authorship attribution using small sets of frequent part-of-speech skip-grams. In *The International Florida Artificial Intelligence Researcg Society Conference, FLAIRS 2016, Proceedings*, pages 86–91.

Raibole M, S. Y. (2011). Impact of physico-chemical parameters on microbial diversity: Seasonal study. *Curr World Environ*, 6(1):71–76.

Rosch, E. (1988). Principles of categorization. In Collins, A. and Smith, E. E., editors, *Readings in Cognitive Science, a Perspective From Psychology and Artificial Intelligence*, pages 312–22. Morgan Kaufmann Publishers.

Roth, C., Obiedkov, S., and Kourie, D. G. (2008). On succint representation of knowledge community taxonomies with formal concept analysis. *International Journal of Foundations of Computer Science*, 19(02):383–404.

Rouane-Hacene, M., Huchard, M., Napoli, A., and Valtchev, P. (2013). Relational concept analysis: Mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, 67(1):81–108.

Rouane-Hacene, M., Valtchev, P., and Nkambou, R. (2011). Supporting ontology design through large-scale fca-based ontology restructuring. In *Conceptual Structures for Discovering Knowledge: 19th International Conference on Conceptual Structures, ICCS. Proceedings*, pages 257–269, Berlin, Heidelberg. Springer Berlin Heidelberg.

Saada, H., Dolques, X., Huchard, M., Nebut, C., and Sahraoui, H. (2012). Generation of operational transformation rules from examples of model transformations. In *Model Driven Engineering Languages and Systems: 15th International Conference, MODELS. Proceedings*, pages 546–561, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shi, L., Toussaint, Y., Napoli, A., and Blansché, A. (2011). Mining for reengineering: An application to semantic wikis using formal and relational concept analysis. In *Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web: Research and Applications - Volume Part II*, ESWC'11, pages 421–435. Springer-Verlag.

Silva, A. and Antunes, C. (2010). Pattern mining on stars with fp-growth. In *Modeling Decisions for Artificial Intelligence: 7th Int. Conf., MDAI 2010, Proceedings*, pages 175–186. Springer Berlin Heidelberg.

Silva, A. and Antunes, C. (2014). Finding multi-dimensional patterns in healthcare. In *Ma-*

chine Learning and Data Mining in Pattern Recognition: 10th Int. Conf., MLDM 2014, Proceedings, pages 361–375. Springer International Publishing.

Silva, A. and Antunes, C. (2015). Multi-relational pattern mining over data streams. *Data Mining and Knowledge Discovery*, 29(6):1783–1814.

Srikant, R. and Agrawal, R. (1996). Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17. Springer-Verlag.

Stumme, G. (2002). Efficient data mining based on formal concept analysis. In *Database and Expert Systems Applications: 13th International Conference, DEXA 2002, Proceedings*, pages 534–546. Springer Berlin Heidelberg.

Stumpfe, D., Lounkine, E., and Bajorath, J. (2011). *Chemoinformatics and Computational Chemical Biology*, chapter Molecular Test Systems for Computational Selectivity Studies and Systematic Analysis of Compound Selectivity Profiles, pages 503–515. Humana Press.

Valtchev, P. and Missaoui, R. (2001). Building concept (galois) lattices from parts: Generalizing the incremental methods. In *Proceedings of the 9th Conference on Conceptual Structures*, ICCS'01, pages 290–303. Springer-Verlag.

Ventos, V. and Soldano, H. (2005). Alpha galois lattices: An overview. In *Formal Concept Analysis: Third International Conference, ICFCA 2005, Lens, France, February 14-18, 2005. Proceedings*, pages 299–314. Springer Berlin Heidelberg.

Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferréol, M., and Valette, L. (2015). Can we predict biological condition of stream ecosystems? a multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. *Ecological Indicators*, 48:88–98.

Wang, J. and Han, J. (2004). Bide: Efficient mining of frequent closed sequences. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 79–90. IEEE Computer Society.

Wasson, J., Villeneuve, B., Mengin, N., Pella, H., and Chandesris, A. (2006). Quelle limite de " bon état écologique " pour les invertébrés benthiques en rivières ? Apport des modèles d'extrapolation spatiale reliant l'indice biologique global normalisé à l'occupation du sol. *Ingénieries - E A T*, 1(47):3–15.

Webster, J. R., Gurtz, M. E., Hains, J. J., Meyer, J. L., Swank, W. T., Waide, J. B., and Wallace, J. B. (1983). *Stream Ecology: Application and Testing of General Ecological Theory*, chapter Stability of Stream Ecosystems, pages 355–395. Springer US.

Wermelinger, M., Yu, Y., and Strohmaier, M. (2009). Using formal concept analysis to construct and visualise hierarchies of socio-technical relations. In *2009 31st International Conference on Software Engineering - Companion Volume*, pages 327–330.

Wille, R. (1982). Restructuring lattice theory: An approach based on hierarchies of concepts.

In Rival, I., editor, *Ordered Sets: Proceedings of the NATO Advanced Study Institute held*, pages 445–470. Springer Netherlands.

Wille, R. (1992). Concept lattices and conceptual knowledge systems. *Computers & Mathematics with Applications*, 23(6-9):493–515.

Wille, R. (1997). *Conceptual graphs and formal concept analysis*, pages 290–303. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wolff, K. E. (2001). Temporal Concept Analysis. In *ICCS-01 Workshop on Concept Lattice for KDD, 9th Int. Conference on Conceptual Structures*, pages 91–107.

Yan, X., Han, J., and Afshar, R. (2003). Clospan: Mining closed sequential patterns in large datasets. In *In SDM*, pages 166–177.

Yun, U. (2008). A new framework for detecting weighted sequential patterns in large sequence databases. *Know.-Based Syst.*, 21(2):110–122.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353.

Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1):31–60.

Zhao, Q. and Bhowmick, S. S. (2003). Sequential pattern mining: A survey. Technical report, Nanyang Technological University, Singapore.

Ziebarth, S., Chounta, I.-A., and Hoppe, H. U. (2015). Resource access patterns in exam preparation activities. In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Proceedings*, pages 497–502. Springer International Publishing.

Cristina NICA

Exploring Sequential Data with Relational Concept Analysis

Résumé

Aujourd'hui, de grandes quantités de données séquentielles sont générées et stockées afin d'être exploitées pour découvrir de précieuses informations. De nombreuses méthodes d'extraction de motifs séquentiels ont été proposées pour découvrir des motifs potentiellement utiles et compréhensibles qui décrivent les données séquentielles analysées. Ces travaux se sont concentrés sur l'énumération efficace de tous les motifs ou de formes plus concises, comme les motifs partiellement ordonnés fermés (cpo-motifs), ce qui rend leur évaluation laborieuse pour les experts du domaine, car leur nombre peut être assez important. Face à ce problème, nous proposons une approche nouvelle, qui consiste à extraire directement des cpo-motifs multi-niveaux qui sont implicitement organisés dans une hiérarchie. À cette fin, nous proposons une méthode originale et autonome dans la cadre de l'Analyse Relationnelle de Concepts (ARC), appelée RCA-SEQ, qui exploite la nature relationnelle des données séquentielles ainsi que la structure et les propriétés des treillis issus de l'ARC. RCA-SEQ comporte cinq étapes : (1) le prétraitement des données brutes ; (2) l'exploration par l'ARC des données prétraitées ; (3) l'extraction automatisée d'une hiérarchie de cpo-motifs multi-niveaux par navigation des treillis issus de l'ARC ; (4) la sélection de cpo-motifs multi-niveaux pertinents en fonction de diverses mesures d'intérêt ; (5) l'évaluation des motifs par les experts du domaine. En outre, nous montrons que l'approche RCA-SEQ peut être facilement adaptée pour extraire des motifs plus informatifs (des cpo-motifs pondérés), pour intégrer une taxonomie définie par l'utilisateur ou pour explorer des données séquentielles hétérogènes. L'approche a été testée sur deux jeux de données décrivant des hydrosystèmes.

Mots-Clés : Données séquentielles, Analyse Relationnelle de Concepts, motifs partiellement ordonnés fermés, motifs multi-niveaux, hiérarchie de motifs, mesures d'intérêt

Abstract

Nowadays, large amounts of sequential data are generated and stored in order to be further harnessed by discovering valuable pieces of information. Many sequential pattern mining methods have been proposed to discover potentially useful and understandable patterns that describe the analysed sequential data. These works have focused on efficiently enumerating all the patterns or concise representations, such as closed partially-ordered patterns (cpo-patterns), that makes their evaluation a laboured task for domain experts since their number can be quite large. To address this issue, we propose a new approach, that is to directly extract multilevel cpo-patterns implicitly organised into a hierarchy. To this end, we devise an original and self-contained method within the Relational Concept Analysis (RCA) framework, referred to as RCA-SEQ, that exploits the relational nature of sequential data and the structure and properties of the lattices from the RCA output. RCA-SEQ spans five steps: (1) the preprocessing of the raw data; (2) the RCA-based exploration of the preprocessed data; (3) the automatic extraction of a hierarchy of multilevel cpo-patterns by navigating the lattices from the RCA output; (4) the selection of relevant multilevel cpo-patterns based on various measures of interest; (5) the pattern evaluation done by domain experts. In addition, we show that the RCA-SEQ approach can be easily adapted to extract more informative patterns (the weighted cpo-patterns), to integrate a user-defined taxonomy or to explore heterogeneous sequential data. Two hydro-ecological datasets have been used to asses RCA-SEQ.

Keywords : Sequential data, Relational Concept Analysis, closed partially-ordered patterns, multilevel patterns, hierarchy of patterns, measures of interest