

**Proceedings of the workshop on
Analyzing Real Data with Formal Concept Analysis
(RealDataFCA'2021)**

June 29, 2021

<https://icfca2021.sciencesconf.org/page/realdatafca2021>

**in conjunction with the 16th International Conference on
Formal Concept Analysis (ICFCA 2021)**

Edited by

Agnès Braud *

Xavier Dolques *

Rokia Missaoui **

* ICube, Université de Strasbourg - CNRS, France

** LARIM, Université du Québec en Outaouais, Canada

Workshop organizers

Agnès Braud, ICube, Université de Strasbourg - CNRS, France
Xavier Dolques, ICube, Université de Strasbourg - CNRS, France
Rokia Missaoui, LARIM, Université du Québec en Outaouais, Canada

Program Committee

Jaume Baixeries, Universitat Politècnica de Catalunya, Spain
Sadok Ben Yahia, Tallinn University of Technology, Estonia
Karell Bertet, Laboratory L3I, University of La Rochelle, France
Agnès Braud, ICube, Université de Strasbourg - CNRS, France
Peggy Cellier, IRISA/INSA Rennes, France
Pablo Cordero, Universidad de Málaga, Spain
Miguel Couceiro, LORIA (CNRS - Inria - Université de Lorraine), Nancy, France
Diana Cristea, Babes-Bolyai University, Romania
Christophe Demko, Université de La Rochelle, France
Xavier Dolques, ICube, Université de Strasbourg - CNRS, France
Peter Eklund, Deakin University, Australia
Manuel Enciso, Universidad de Málaga, Spain
Sébastien Ferré, Université de Rennes 1, CNRS, IRISA, France
Jessie Galasso, Geodes, DIRO, Université de Montréal, Canada
Marianne Huchard, LIRMM, Université de Montpellier et CNRS, France
Mohamed Hamza Ibrahim, Zagazig university, Egypt
Dmitry Ignatov, National Research University Higher School of Economics, Moscow, Russia
Mehdi Kaytoue, Infologic R&D, France
Léonard Kwuida, Bern University of Applied Sciences, Switzerland
Florence Le Ber, ICube, Université de Strasbourg/ENGEES - CNRS, France
Engelbert Mephu Nguifo, LIMOS, Blaise Pascal University – CNRS, France
Rokia Missaoui, LARIM, Université du Québec en Outaouais, Canada
Emilio Muñoz-Velasco, Universidad de Málaga, Spain
Amedeo Napoli, LORIA (CNRS - Inria - Université de Lorraine), Nancy, France
Sergei Obiedkov, National Research University Higher School of Economics, Moscow, Russia
Manuel Ojeda-Aciego, Universidad de Málaga, Spain
Uta Priss, Ostfalia University, Germany
Christian Sacarea, Babes-Bolyai University, Romania
Henry Soldano, LIPN, Université Paris 13, France
Francisco José Valverde-Albacete, Universidad Carlos III de Madrid, Spain

CONTENTS

Preface	4
How to provide light to COVID data by means of FCA	5
<i>Domingo López-Rodríguez, Pablo Cordero, Manuel Enciso and Ángel Mora</i>	
Analyzing water monitoring data with RCA-based approaches	13
<i>Xavier Dolques, Agnès Braud, Corinne Grac and Florence Le Ber</i>	
Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results	20
<i>Johanna Saoud, Alain Gutierrez, Marianne Huchard, Pascal Marnotte, Pierre Silvie and Pierre Martin</i>	
Analyzing the composition of remedies in ancient pharmacopeias with FCA	28
<i>Agnès Braud, Xavier Dolques, Pierre Fechter, Nicolas Lachiche, Florence Le Ber and Véronique Pitchon</i>	
Mining the Groceries Database using Triadic Concept Analysis	36
<i>Pedro H. B. Ruas, Rokia Missaoui, Mark A. J. Song and Léonard Kwuida</i>	
Leveraging Formal Concept Analysis to Improve n-fold validation in Multilabel Classification	44
<i>Francisco J. Valverde-Albacete and Carmen Peláez-Moreno</i>	

PREFACE

Research in Formal Concept Analysis (FCA) is growing both at the theoretical and practical settings, and this theory has been successfully used and even extended in many disciplines such as mathematics, computer science (CS), bioinformatics, engineering, sociology, and so on. In CS, studies are concentrated on numerous topics like data analysis, knowledge representation and discovery, software engineering, information retrieval, social network analysis, to name a few.

The objective of the workshop is to bring together researchers who are using FCA for real-life applications, and are interested to share their experience and exchange ideas with FCA community members. It aims at getting an overview of the diversity of applications of FCA on real data, and providing a forum to discuss the strengths, limitations and challenges of using FCA, as well as the successful and unsuccessful experiences with this theory. Feedback from application domain experts is also welcome.

Six papers were submitted. Each one of them was reviewed by four program committee members, and six papers were accepted and presented at the workshop. A discussion session followed the talks.

Authors of accepted papers will be invited to submit an extended version of their contribution in a special issue of an international journal.

The organizers of the workshop would like to thank all the authors for their contributions and the organizers of ICFCA'2021 for their support. Their warm thanks go also to the program committee members for their careful review of the submissions and their useful comments and suggestions, as well as to the three session chairs. Finally, the success of this event was possible thanks to the participation of forty-eight attendees.

July 2021

Agnès Braud,

Xavier Dolques,

Rokia Missaoui

How to provide light to COVID data by means of FCA^{*}

Domingo López-Rodríguez^[0000-0002-0172-1585], Pablo Cordero^[0000-0002-5506-6467], Manuel Enciso^[0000-0002-0531-4055], and Ángel Mora^[0000-0003-4548-8030]

Universidad de Málaga, Málaga, Spain
{dominlopez,pcordero,enciso,amora}@uma.es

Abstract. COVID data are usually presented in a non-structured format and mainly focused on healthy issues (incidence, mortality, etc). At the same time, Governments have designed a set of measures to deal with the Pandemic. In addition, several institutions have studied the economical effects of the situation in each country. In this work, we combine these three data sources and illustrate how Formal Concept Analysis can become a useful tool to discover relationships among these three views of the situation: health, politics and economy. Our aim is to provide an implication-driven approach to discover knowledge behind the data.

Keywords: Formal Concept Analysis · Implications · Covid.

1 Introduction


During 2020 and so far in 2021, the world has been paralysed by the pandemic. The devastating effects of COVID-19 in terms of loss of human life, the limitations in every country worldwide and the economic downturn constitute a dramatic landmark in the history of humankind.

From the very beginning, individual researchers, universities and institutions began to accumulate data on the virus incidence oriented to measure the stress of the health infrastructures and the impact of the Pandemic on mortality. Simultaneously, governments around the world have implemented several restrictions to minimise the incidence of the virus. Each country has taken different measures depending on their situation and, as usual, taking into account the political effects of these measures on the citizens [15].

Data scientists have conducted extensive, fast and effective work, showing the power of these methods and tools and providing some valuable guidelines to support the fight against the virus and check its evolution. Most of these

* Partially supported by the projects TIN2017-89023-P (Spanish Ministry of Economy and Competitiveness, including FEDER funds), PGC2018-095869-B-I00 (Spanish Ministry of Science and Innovation), and the Junta de Andalucía projects UMA2018-FEDERJA-001, UMA18-FEDERJA-158 and UMA-CEIATECH-24 co-funded by the European Regional Development Fund.

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

works have been oriented for diagnoses (using image processing or tracking techniques) or prediction and forecast (time series analysis or machine learning methods) [8]. Other authors have developed complex mathematical models to explain the virus behaviour (nonlinear fractional-order differential equations [1], Chaos theory [14], etc.).

To the best of our knowledge, easy and direct relationships looking at the evolution curve of the number of infected persons and deaths can be extracted if we insert in the evolution curve the major measures taken by governments such as the total closure of companies and the confinement of people to their homes. These classical and quantitative relationships can be considered a good but basic approach. We are not aware of studies that allow studying all the elements involved in the three-dimensional data (health, politics and economy). We need to address the following problem: how to relate all the measures taken to a greater or lesser extent with the changes in the evolution of the number of infected persons, deaths, admissions to intensive care units and the interaction of all this with the economic downturn.

The first step is to build a comprehensive dataset from reliable sources.

Regarding the disease dimension and its impact, there exist many sources collecting up-to-date data. Mainly, we have used the data from the European Centre for Disease Prevention and Control (ECDC), an agency of the European Union¹ because of its trustworthiness. To complete our study, we have integrated economy and politics measures data. On the one hand, we have collected Eurostat data about the Economy. We have used the repository that this institution has mainly created to measure the impact of the crisis in Europe². On the other hand, we have collected the measures designed for all the EU countries in the different waves of the pandemic as well as the number of cases and deceases, hospital occupancy and other related variables³.

Formal concept analysis [4] (FCA) has shown to be a strong and solid area to deal with the complete data life-cycle, from the data extraction, knowledge representation and management. However, this solid development has still to be fruitful in real applications, as others areas are. FCA has also been used to address COVID Pandemic. In [6], the authors built a concept lattice to navigate, providing successful search in COVID resource platforms. In [2] the author characterised a set of attributes of the vaccines and created a concept lattice to provide a hierarchical landscape of the current status of phase 3 vaccines. However, few works show FCA's power to design a data-driven approach for this issue. In this work, we propose to extract valuable knowledge from public data using the so-called implications (if-then rules). These rules can be used to discover the interesting relationship among the three dimensions involved in the Pandemic. Such relationships are not merely cause-effect rules, and they provide a complete portrait of the semantic behind the data.

¹ <https://www.ecdc.europa.eu/en>

² <https://ec.europa.eu/eurostat/web/COVID-19/data>

³ <https://www.ecdc.europa.eu/en/COVID-19/data>

In this work, we are using `fcaR` software developed in our research group, *Malaga-FCA-group*, using R language. We propose this tool as a useful tool to approach real problems in FCA and illustrate the benefits of this solid framework. See <https://github.com/Malaga-FCA-group/fcaR> for more information.

The paper is organised as follows. In section 2 we describe how to address the import, processing and data curation in this real problem and how to integrate such data to be further used with FCA methods. In Section 3 we explore the discovery of knowledge extracted from the data (using implications) of COVID to establish how to act better in the future to reduce the incidence of the virus and minimise economic losses. We aim to illustrate how implications can be considered a suitable tool to help in decision-making tasks. Section 4 proposes some conclusions and future works.

2 Combination of open data into one integrated dataset

As we have explained in the introduction, we have downloaded the open data regarding health and measures from UE institutions available in the aforementioned sources. ECDC publishes weekly updates. Briefly, we have collected data about the following information: data on 14-day notification rate of new COVID-19 cases and deaths, data on hospital and ICU admission rates and occupancy for COVID-19, data on country response measures to COVID-19, data on Harmonised unemployment rates and data on Gross Domestic Product (GDP)

In this first study, we also have introduced the time dimension to infer conclusions about the Pandemic evolution. Thus, we take the second and third quarters (Q2, Q3) of 2020 as the study periods ⁴, because of the harshness of the incidence of the virus during these months and because all the EU countries in these two periods have presented significant differences in their situation.

On the other hand, these variables are both numerical and dates, then we have opted to perform conceptual scaling after normalizing the attributes. We summarize the data pre-processing we have done to achieve the formal context being the target of the application of the FCA methods:

- Each measure adopted in the EU countries in one specific quarter is stored in one column. For instance, `MEAS_Q2_MasksMandatoryClosedSpaces` means that in the second quarter (Q2), masks were mandatory in all closed spaces.
- We use the NOT token to show the absence of a given measure⁵. For instance, `MEAS_NOT_Q2_AdaptationOfWorkplace` means that in the second quarter, the country did not take any action on the adaptation of workplaces.
- For each quarter (Q2 and Q3), we have discretized the COVID data (deaths and cumulative cases in percent wrt population), ICU and hospitalization data, and excess mortality data. For instance, `COV_Q2_cases_low` means that

⁴ Remark that we use somehow a temporal information, but we do not make use of temporal FCA [16] to ease a simpler management in the future.

⁵ Note that in this paper we do not introduce the generalization of FCA to consider negative attributes [12,10]. We will address this issue in future works.

in quarter two the country had a low level of COVID confirmed cases, `COV_Q3_ExcessMortalityChange_decreased` means that in Q3 was detected a decrease in the excess of mortality in the country.

- For economical data, we have collected information about **changes** in *Unemployment rates* and *GDP*.

3 Knowledge representation and its interpretation

After the pre-processing task, we got an adequate binary formal context to be explored with FCA. This formal context has 175 attributes and 25 objects. The dataset has been imported in R using the `fcaR` package. From this point, we performed the following steps:

- First, we remove *constant* columns, corresponding to attributes (measures or health or economic parameters) that are present either in none or in all the objects (the 25 countries in the EU).
- Then, we perform clarification on the resulting formal context to identify equivalent attributes, that is, attributes a, b such that $a \Rightarrow b$ and $b \Rightarrow a$.
- We compute the Duquenne-Guigues basis of implications and inspect them to get a complete knowledge representation of the data allowing further symbolic manipulation of the coronavirus situation in the EU countries.
- In the last step, we apply the Simplification Logic for implications [13] to remove *attribute* redundancies in the above basis, keeping an equivalent set of implications with lower size (number of attributes in each implication).

The Duquenne-Guigues basis of this formal context has 4086 implications. The average size of its LHS and RHS is 11.297 and 1.594 attributes, respectively. After using the simplification logic to remove redundancies, the resulting implication set has 4.472 and 1.188 attributes in the average size in LHS and RHS. The decrease of the implication size eases its reading and interpretation.

In the next subsections, we present the main findings obtained from the simplified basis of implications. Our goal is to illustrate how these implications can be used to get a useful interpretation of the model.

3.1 Common strategies and parameters

Some measures weren't taken by any country in the management of the pandemic. Their interpretation depends on the measure or effect and provides useful information even though its uniform absence seems to induce no information.

For instance, in the 3rd quarter, no country imposed a ban on all events or restrictions on private gatherings or closed restaurants, hotels and entertainment venues. Concerning economic parameters, the unemployment rate wasn't stable in both quarters. It remarks that the strong lockdown measures were not needed in the Q3 period.

It can be observed that the GDP in the second quarter decreased in all countries whereas it increased in the third quarter (summer). And, since there is

a common value for all the objects (countries), this progress occurs irrespective of the restriction measures taken.

3.2 Equivalent attributes

We named equivalent attributes those ones that play the same role, and they can be characterized by using attribute reduction and clarification methods [7].

The attributes `NOT_Q2_EntertainmentVenues` and `Q3_StayHomeRiskG` are found to be equivalent. This means that every country that did not close entertainment venues in the 2nd quarter (Q2) of 2020, in the 3rd quarter (Q3), they had to issue a recommendation to stay home for risk groups (older adults, disabled people or people with other pathologies).

Also, having a high number of deaths in Q2 was equivalent to a high number of deaths in Q3. This is interesting since no matter the measures adopted if the pandemic has led to a high number of deaths during the first stage, in the summer, the number of deaths remains at a high level, and it was also greater than Europe's average.

A high hospitalization occupancy ratio is shown to be equivalent to the ICU occupation ratio during Q2. This data illustrates that the evolution of patients admitted in the hospitals is not very promising. This is well-known for the health experts that insistently alert us about the risks of this disease, and the FCA analysis also confirms it.

3.3 Implications provide useful interpretation

To illustrate how to get useful knowledge from the implications, we have selected some implications whose interpretation may provide a deeper insight on the pandemic management strategies and the consequences of taking restriction measures. Since the implication set is large, we only show some easily interpretable and with a low number of attributes, and with positive support. In the following, we will simplify attributes' names to allow for a better readability:

- Some implications on the influence of Q2 measures in Q2 COVID cases or deaths (short-term effects in health):

<code>NOT_Q2_MassGather50</code>	\Rightarrow	<code>Q2_cases</code> are high
<code>Q2_AdaptationOfWorkplace</code>	\Rightarrow	<code>Q2_cases</code> are medium
<code>Q2_RestaurantsCafes, Q2_MasksMandatoryAllSpaces</code>	\Rightarrow	<code>Q2_deaths</code> are medium

These implications can be interpreted as follows: the countries that didn't constraint mass gatherings of more than 50 people in the Q2 ended this period with a number of COVID cases greater than Europe's average. By contrast, the countries that promoted adaptation of workplaces ended with a number of cases equivalent to the average. However, such promising impact of the virus can also be provided by other measures, and the same situation appears in the countries that did close restaurants and where masks were mandatory in all spaces.

- Influence of Q2 measures in Q2 economy (short-term effects in economy):

NOT_Q2_WorkplaceClosures \Rightarrow Q2_UnemploymentChange is increased
 NOT_Q2_StayHomeOrder \Rightarrow Q2_UnemploymentChange is increased

These implications have a potentially *counterintuitive* interpretation. Countries that did not close workplaces or didn't order a strict lockdown had their unemployment rate increased; that is, not taking those measures didn't prevent a rise in the unemployment rates. It is important to highlight this relationship that has been discovered with the help of FCA tools.

- Influence of Q2 measures in Q3 COVID (mid-term effects in health):

NOT_Q2_StayHomeRiskG, NOT_Q3_MasksVoluntaryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased

Those countries that didn't recommend lockdown for risk groups in Q2 and didn't impose masks in all spaces in Q3, had their number of hospitalizations rise from Q2 to Q3.

- Influence of Q2 measures in Q3 economy (mid-term effects in economy):

NOT_Q2_BanOnAllEvents \Rightarrow Q3_UnemploymentChange is decreased
 NOT_Q2_StayHomeOrder \Rightarrow Q3_UnemploymentChange is decreased
 Q2_PrivateGatheringRestrictions \Rightarrow Q3_UnemploymentChange is decreased

The unemployment decreased in those countries that, in Q2, did not ban all events, nor ordered a strict lockdown nor had restrictions on private gatherings.

- Influence of Q3 measures in Q3 COVID:

NOT_Q2_StayHomeRiskG, NOT_Q3_MasksVoluntaryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased
 NOT_Q2_StayHomeRiskG, NOT_Q3_MasksMandatoryAllSpaces
 \Rightarrow Q3_ChangeInHospital is increased
 Q3_StayHomeRiskG
 \Rightarrow Q3_inICU is medium, Q3_ChangeInHospital is decreased,
 Q3_ExcessMortality is equivalent

The first two implications tell us that not recommending risk groups to stay home and not suggesting the use of masks in all spaces led to an increase in hospitalizations during summer. The countries that recommended risks groups to stay home ended with an average ICU occupancy rate, a decrease in hospitalizations and a normalization of the mortality excess (that is, the number of deaths in the country is equivalent to the previous years).

3.4 Closure of attributes to find common properties

To complete the information, we have studied the closure of some interesting attributes. In particular, we want to answer the question ‘*What had in common the countries with a low and a high hospital occupancy rate in the second quarter?*’

To answer this question, we will compute the closures of the corresponding attributes (`Q2_inHospital` is low and `Q2_inHospital` is high).

- The closure of the first attribute is the following `Q2_ClosHigh`, `Q2_ClosPrim`, `Q2_MassGather50`, `Q2_MassGatherAll`, `Q2_OutdoorOver1000`, `Q2_EntertainmentVenues`, `NOT_Q2_ClosureOfPublicTransport`, `NOT_Q3_MasksMandatoryAllSpaces`, `NOT_Q3_IndoorOver50`, `NOT_Q3_OutdoorOver50`, `NOT_Q3_Teleworking`.

This means that countries that ended Q2 quarter with low occupancy had closed high and primary schools and imposed restrictions on mass gatherings and outdoor. However, in the third quarter, they could relax some restrictions on gatherings because of the low occupancy.

- The high occupancy closure provides the following set of attributes: `Q2_ClosSec`, `NOT_Q2_HotelsOtherAccommodation`, `NOT_Q2_ClosureOfPublicTransport`, `NOT_Q3_Teleworking`, `Q2_ExcessMortality is more`, `Q3_ExcessMortalityChange is decreased`.

This result means that they didn’t have a common strategy and the only interesting consequence is that in Q3, the mortality excess was decreased.

4 Conclusions

In this work, we have used FCA to address the problem to illustrate the relationships among the government measures, the economic impact and the health situation in the Pandemic. We have combined real public data from all countries in the EU stored in different data sets, establishing a data pre-processing task. We have made a knowledge representation through the implications using the `fcaR` package. Reducing the original Duquenne-Guigues basis using the Simplification logic eases a further examination of the set of implications, allowing to deduce valuable insights about the COVID disease, its impact and the measures that can be used to deal with the virus.

As future works, we plan to collect more data. We intend to consider all the countries worldwide and collect more attributes for the evolution of the data in 2021. A more comprehensive conceptual scaling should be performed to manage such large dataset. In addition, it is of interest the use of generalized attributes [11] to handle the increasing number of attributes.

To provide more information about the new data, we will also build the concept lattice that provides other helpful information and a graphical representation of the information, essential for further navigation among the closure set of attributes. Since the number of concepts increases when more data were considered, we will need a filtering method to select those which are most likely to provide promising information. Stable concepts may provide more insights in

this delicate issue [9]. Moreover, we have to study the methods and techniques to develop some kind of conceptual exploration [5].

Finally, we will also address the negation issue (the relation between contrary attributes) by means of mixed (negative and positive) attributes [3].

References

1. Idris Ahmed, Isa Abdullahi Baba, Abdullahi Yusuf, Poom Kumam, and Wiyada Kumam. Analysis of caputo fractional-order model for covid-19 with lockdown. *Advances in Difference Equations*, 2020(1):1–14, 2020.
2. Javier Dario Burgos-Salcedo. A comparative analysis of clinical stage 3 covid-19 vaccines using knowledge representation. *medRxiv*, 2021.
3. P. Cordero, M. Enciso, A. Mora, and J. M. Rodríguez-Jiménez. Inference of mixed information in formal concept analysis. *Studies in Computational Intelligence*, 796:81–87, 2019.
4. B. Ganter and R. Wille. *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin, 1996.
5. Bernhard Ganter and Sergei A. Obiedkov. *Conceptual Exploration*. Springer, 2016.
6. Fei Hao and Doo-Soon Park. Conavigator: A framework of fca-based novel coronavirus covid-19 domain knowledge navigation. *Human-centric Computing and Information Sciences*, 11(6), 2021.
7. Jan Konecny. On attribute reduction in concept lattices: Methods based on discernibility matrix are outperformed by basic clarification and reduction. *Inf. Sci.*, 415:199–212, 2017.
8. Utku Kose, Deepak Gupta, Victor de Albuquerque, and Ashish Khanna, editors. *Data Science for COVID-19 Volume 1. Computational Perspectives*. Academic Press, 2021.
9. S. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49:101–115, 2007.
10. Sergei O. Kuznetsov and Artem Revenko. Interactive error correction in implicative theories. *International Journal of Approximate Reasoning*, 63:89–100, 2015.
11. Léonard Kwuida, Rokia Missaoui, Abdélilah Balamane, and Jean Vaillancourt. Generalized pattern extraction from concept lattices. *Annals of Mathematics and Artificial Intelligence*, 72(1):151–168, 2014.
12. Rokia Missaoui, Lhouari Nourine, and Yoan Renaud. Computing implications with negation from a formal context. *Fundamenta Informaticae*, 115(4):357–375, December 2012.
13. Angel Mora, Pablo Cordero, Manuel Enciso, Inmaculada Fortes, and Gabriel Aguilera. Closure via functional dependence simplification. *International Journal of Computational Mathematics*, 89(4):510–526, 2012.
14. O Postavaru, SR Anton, and A Toma. Covid-19 pandemic and chaos theory. *Mathematics and computers in simulation*, 181:138–149, 2021.
15. Massimo Pulejo and Pablo Querubín. Electoral concerns reduce restrictive measures during the covid-19 pandemic. *Journal of Public Economics*, 198:104387, 2021.
16. Jan Triska and Vilém Vychodil. Logic of temporal attribute implications. *Ann. Math. Artif. Intell.*, 79(4):307–335, 2017.

Analyzing water monitoring data with RCA-based approaches

Xavier Dolques¹, Agnès Braud¹, Corinne Grac^{2,3}, and Florence Le Ber¹

- (1) Université de Strasbourg, CNRS, ENGEES, ICube UMR 7357, F67000 Strasbourg
{dolques,agnes.braud}@unistra.fr florence.leber@engees.unistra.fr
(2) Université de Strasbourg, CNRS, LIVE UMR 7362, F67000 Strasbourg
(3) ENGEES, F67000 Strasbourg
corinne.grac@engees.unistra.fr

Abstract. This paper is a short feedback on a collaborative research work by computer scientists and hydroecologists. We have applied Relational Concept Analysis on complex data about running water characteristics (physical, biological and chemical parameters), to answer various questions. Two approaches are presented and discussed: the first one extracts patterns from temporal data, the second one extracts rules from a multi-relational dataset.


1 Introduction

For almost ten years, we have been working on watercourse monitoring data and, among other approaches, we have exploited Relational Concept Analysis (RCA) [8]. We have focused on temporal characteristics, that are inherent in monitoring data, but also on relations between the different parts of the system being monitored (biological as well as physical and chemical characteristics of the watercourses). We have developed RCA-based approaches to extract temporal patterns from sequential data, or to extract rules from complex relational data, results being always assessed by domain experts. This paper is a short feedback on this collaborative research work. It is organized as follows. Section 2 describes the datasets, Sect. 3 gives a short explanation on RCA functioning. The approaches used are described in Sect. 4 and discussed in Sect. 5.

2 Data characteristics

Data were collected during the ANR Fresqueau project (2011-2015)¹ and organized into a database (60 million records and 20 gigabytes). The study area covers 161,100 km² in the east of France, for the 2000-2010 period. We collected five categories of water data from 21 different sources (mainly public data banks or research projects): 1) river quality data; 2) sampling site data; 3) hydrographic network data; 4) human activities data and 5) driving data.

¹ http://dataqual.engees.unistra.fr/fresqueau_presentation_gb
RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

River quality data are temporal data, and the most complex part of the data as detailed below. They are divided into three sub-categories: physical (e.g., the dimensions and shape of the river bed, characteristics of the substrate), physico-chemical (like pH, nitrate and phosphate, pesticides in the water or sediments) and biological data (i.e. lists of fauna and flora taxa, and metrics on these taxa, e.g., total abundance, diversity, and biological indices) for four groups: macroinvertebrates, diatoms, macrophytes and fishes. Biological indices are defined according to French standards (e.g., the French macroinvertebrate index, IBGN [1] for invertebrates). Taxa are associated to their life traits (like the type of respiration, or the habitat preference).

The other four categories of data are mostly geographic data. Sampling site data give the location of the sites where the river quality data were sampled and their main features. A sampling site is considered as a point. Hydrographic data comprise the different segments of running waters or waterbodies, the size of watersheds, administrative regions, and additional information on the hydrographic network. Human activity data allow to estimate anthropogenic pressures on running waters: land use, impediments to flow, location of discharges. Driving data concern forcing or context variables such as climate (for instance, average atmospheric temperature, precipitations), flows, geology or administrative information. They allow to characterize the environment of the running waters and sampling sites.

Building a database integrating data coming from different sources was not easy. For example, concerning taxa, although the data are based on a standard, their identifiers may evolve over time so that it is difficult to combine data from different years. Besides, geometric data may be imprecise: it may be hard to determine whether a sample site is placed on a watercourse or another one.

Then these data are highly heterogeneous in their values (quantitative continuous or discrete, semi-quantitative or qualitative), temporal variability (frequency and duration of sampling) and topological structure (with a geometry or not). They may be simple measures of a parameter (e.g., a pH value) or a complex index using different metrics (e.g., IBGN) or based on expert knowledge. For instance physico-chemical measures are collected 4 to 6 times per year, while biological samples are done at most once a year, and physical measures even more rarely. Moreover, some sample sites may require stronger monitoring, based on more parameters, in particular pesticides, so that some parameters are less abundant in the database. Let us also notice that depending on the area, taxa may differ.

We have proposed some approaches based on RCA in order to deal with some of these problems when analyzing the data, starting from questions asked by experts.

3 RCA basics

Relational Concept Analysis [8] is an extension of Formal Concept Analysis [6] which considers relational data, formalized within a *relational context family*

Table 1. Relational Context Family example.

object-attribute contexts				object-object contexts			
taxons	≤1 year	> 1 year		taxon- -Presence	Atheri- -cidae	Bithy- -nia	Boreob- -della
Athericidae	x			BREI0001			x
Bithynia	x	x		BRUN001	x	x	
Boreobdella		x		FECH001	x		x

stations	small watercourse	fresh, running watercourse	phreatic stream
BREI0001	x	x	
BRUN001			x
FECH001		x	

(RCF), $\mathbf{F} = \{\mathbf{K}, \mathbf{R}\}$ where \mathbf{K} is a set of object-attribute contexts (each context corresponding to an object category) and \mathbf{R} is a set of object-object contexts (relations between objects of various categories).

The principle of RCA consists in integrating object-object relations as new attributes (called *relational attributes*) in the formal contexts of \mathbf{K} thanks to scaling quantifiers, such as the existential (*exist*) or universal strict (*exist+forall*) scaling quantifiers. It produces iteratively a set of concept lattices (one lattice per object category) interconnected through relational information. The concepts in a given lattice group objects according to the shared attributes and to the connections they have with objects of another category. The result is a family of concept lattices where concepts of a lattice are linked to concepts of other lattices. We illustrate this with a small example (from [4]). Table 1 introduces two object-attribute contexts, one about taxa and their life traits, one about river sites and their physical characteristics, and an object-object context linking taxa to the sites where they have been found.

First, the RCA process builds lattices on the two object-attribute contexts, **stations** and **taxons** (Fig. 1). Then the **stations** context is extended by relational attributes linking **stations** objects to **taxons** concepts, based on the **taxonPresence** context. For example, $\exists \text{taxonPresence} : \text{Concept}_2$ means that at least one object of **Concept_2** in the **taxons** lattice is present on each **stations** object that owns this attribute. The lattice built on this extended context is shown in Fig. 1 (right). In this example, the process stops here since there is only one (one-way) relation linking the two contexts.

4 Questions and Approaches

For domain experts, the general question is to link physical and physico-chemical data to biological ones, the first ones giving an instantaneous information, the second one giving a long term integrative information. It covers more specific questions, for example, how can values of biological indices be explained by

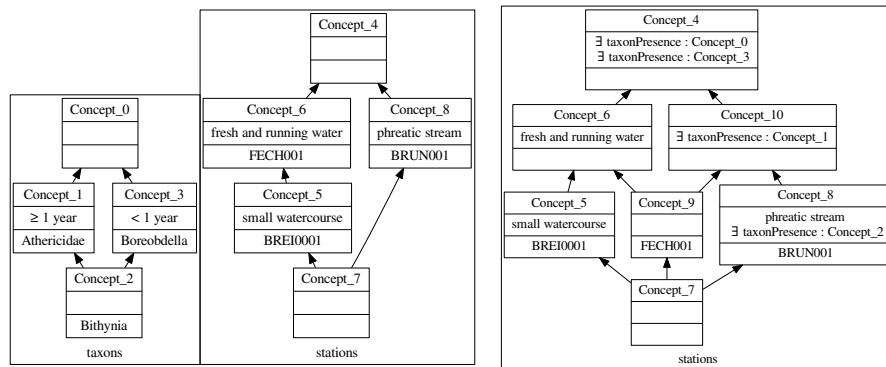


Fig. 1. Lattices of the two object-attribute contexts of Table 1 – initialization (left) and first step (right) of RCA – Only **stations** lattice changes

the preceding successive measures of physico-chemical parameters? What is the relation between the values of physico-chemical or physical parameters and the life traits of taxa living in a site? The first question raises the problem of dealing with temporality and it has been undertaken with a pattern mining approach [5] and then by RCA [9]. Moreover, working on biological quality requires to overcome the difficulty of analyzing sites with different taxa. This problem has been tackled by working with biological indices in the pattern mining approach, one of their aims being to make biological quality comparable between sites. For the second question, it has been tackled by working on life traits, and the question has been undertaken by a RCA-based rule mining approach [3].

Analyzing sequences of physico-chemical and biological measures. In [9], we focused on sequences of physico-chemical measures (6 measures per year) ending with biological samples (one per year). The selection and preprocessing of the data were done under the supervision of domain experts. Physico-chemical measures were discretized into qualitative scales. Biological samples were synthesized into qualitative indices (five levels from red to blue, corresponding to quality). The question was to explain the biological quality wrt the physico-chemical quality assessed during the last months. Sequences were encoded into an RCF, according to the schema shown in Fig. 2: each rectangle corresponds to an object-attribute context, while the arrows correspond to object-object contexts.

Then data are processed as follows. Firstly, RCA is applied to the RCF in order to obtain a family of concept lattices. Secondly, the interrelated concepts from the RCA result, are navigated to extract a set of sequential patterns for each concept of the **BioSamples** lattice. A pattern is actually a directed graph, where the various paths represent sequences of parameter values preceding a biological sample. Iceberg lattices [11] have been used to select patterns with the highest support (i.e. represented in many sample sites). Figure 3 shows an example of an extracted pattern: it summarizes a set of sequences of physico-chemical

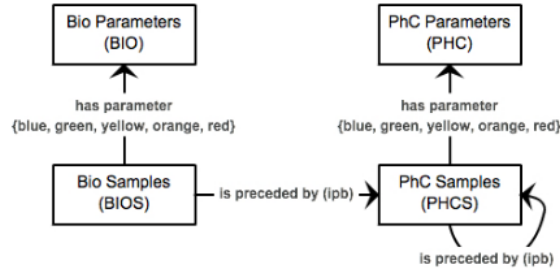


Fig. 2. Modeling sequences of physico-chemical and biological samples [9]

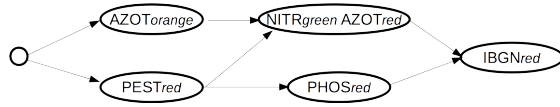


Fig. 3. An example of a sequential pattern

parameter values measured before a biological index (IBGN) with red value. The pattern is read as follows: in all sequences, an orange value for AZOT (nitrogen except nitrate) and a red value for PEST (pesticides) have been measured before a green value for NITR (nitrate) and a red value for AZOT occurring at the same moment. Also, a red value for PHOS (phosphorous) has been measured after the red value for PEST. According to expert domains, this pattern can be interpreted as follows: the quality values of physico-chemical parameters are consistent with the biological value, macroinvertebrates being sensitive to high rates of pesticides and ammonium.

Extracting rules linking taxa life traits and site physical and physico-chemical characteristics. In [3], we tried to connect physical characteristics of sample sites, physico-chemical parameters, and the traits of taxa living in sample sites. We therefore developed an RCA-based method using AOC-posets to deal with large datasets. This approach allowed to provide a reasonable number of concepts and to extract meaningful implication rules (association rules whose confidence is 1). In order to offer more flexibility on the quantification of taxa on sample sites than with the existing *exist* and *forall* scaling operators, new scaling operators were defined and experimented, providing different semantics for the rules. Data were modeled as shown in Fig. 4. Detailed temporal information (physico-chemical parameters) was aggregated into annual values and then discretized into five levels. Taxa numbers were also discretized (taxa weakly to highly represented). An example of rule is given below, where a percent-quantifier is used [3]: such a quantifier allows to build relational attributes with for example, the form $\exists \forall_{>n\%} r(C)$; an object owning this attribute is linked to at least $n\%$ of C objects

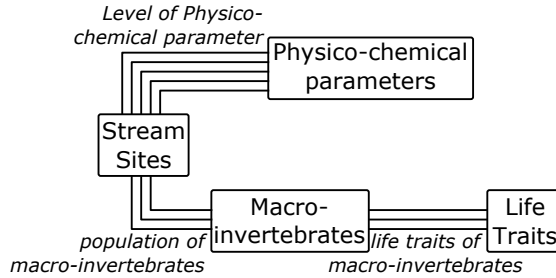


Fig. 4. Modeling the links between physico-chemical parameters and the life traits of taxa (macro-invertebrates) living in sample sites [3]

with the r relation [2].

$$\begin{aligned}
 & S_{>50\%} \text{ high_population}(\exists \text{strong_affinity}(\text{slow current})) \\
 & \rightarrow \exists \text{bad_state}(\text{hydrology})
 \end{aligned}$$

This rule means that a sample site having more than half of its highly represented taxa preferring slow current, has a bad hydrological state. According to the experts, it corresponds to small disconnected phreatic streams, with almost no current, that are specific of the Alsace plain.

5 Discussion and Conclusion

In the following we describe some problems we have faced with the data and novelties that stemmed from these works. Note that all these experiments led to new proposals to make the use of RCA easier [10].

First of all, ecosystems are very complex entities, influenced by many parameters. We have tried to integrate many of these parameters in our database, but handling all of them in a single analysis is not really possible, and even for experts to fully understand such results involving many information types. We have thus worked on subproblems, handling a smaller but consistent part of parameters, based on an expert question. Also, data like biological indices already integrate several parameters in one measure. Their definition has been proposed by groups of specialists and it is pertinent to use them when studying water quality, in particular to overcome the difference of taxa between areas, but at the same time they are not raw data and represent a kind of bias.

Regarding the pattern mining approach, given the sequence set of biological and physico-chemical samples, we found hierarchies of multilevel cpo-patterns that summarize the impact of physico-chemistry to biology. The hierarchical representation allows to enhance the analysis of the extracted set of sequential patterns. Nevertheless, there are too many patterns, and relevant interestingness measures must be chosen. Another problem is due to the irregular distribution of biological index values, leading to more or less frequent, complex, and informative patterns for the different values. Regarding the rule mining approach:

using AOC-posets causes loss of concepts, and some interesting rules will thus not be found. Other techniques should be explored for processing the complexity of this relational dataset.

With respect to classical approaches in the hydroecological domain, our approaches are original since most work are based on statistical analysis or supervised machine learning methods (see e.g., [7, 12]).

References

1. AFNOR: Qualité de l'eau : détermination de l'Indice Biologique Global Normalisé (IBGN). NF T90-350 (1992)
2. Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Generalization effect of quantifiers in a classification based on relational concept analysis. *Knowl.-Based Syst.* **160**, 119–135 (2018)
3. Dolques, X., Le Ber, F., Huchard, M., Grac, C.: Performance-friendly rule extraction in large water data-sets with AOC posets and relational concept analysis. *Int. J. General Systems* **45**(2), 187–210 (2016)
4. Dolques, X., Le Ber, F., Huchard, M., Nebut, C.: Relational Concept Analysis for Relational Data Exploration. *Adv. in Know. Disc. and Manag.* **5**, 55–77 (2016)
5. Fabrègue, M., Braud, A., Bringay, S., Grac, C., Le Ber, F., Levet, D., Teisseire, M.: Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics* **24**, 210–221 (2014)
6. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical Foundations*. Springer Verlag (1999)
7. Goethals, P.L., Dedecker, A., Gabriels, W., Lek, S., Pauw, N.: Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology* **41**(3), 491–508 (2007)
8. Hacene, M.R., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell.* **67**(1), 81–108 (2013)
9. Nica, C., Braud, A., Dolques, X., Le Ber, F., Huchard, M.: Exploring temporal data using relational concept analysis – an application to hydroecology. In: *Concept lattices and their applications (CLA 2016)*. pp. 1–13. CEUR Workshop Proc. (2016)
10. Ouzerdine, A., Braud, A., Dolques, X., Huchard, M., Le Ber, F.: Adjusting the exploration flow in Relational Concept Analysis – An experience on a watercourse quality dataset. *Adv. in Know. Disc. and Manag.* **9** (2021)
11. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. *Data & Know. Eng.* **42**(2), 189 – 222 (2002)
12. Villeneuve, B., Souchon, Y., Usseglio-Polatera, P., Ferréol, M., Valette, L.: Can we predict biological condition of stream ecosystems? a multi-stressors approach linking three biological indices to physico-chemistry, hydromorphology and land use. *Ecol. Indic.* **48**, 88–98 (2015)

Explicit versus Tacit Knowledge in Duquenne-Guigues Basis of Implications: Preliminary Results

Johanna Saoud¹, Alain Gutierrez¹, Marianne Huchard¹, Pascal Marnotte²,
Pierre Silvie^{2,3}, and Pierre Martin²

¹ LIRMM, Univ Montpellier, CNRS, Montpellier, France
johanna.saoud@etu.umontpellier.fr,

{marianne.huchard,alain.gutierrez}@lirmm.fr

² CIRAD, UPR AIDA, F-34398 Montpellier, France

AIDA, Univ Montpellier, CIRAD, Montpellier, France

{pascal.marnotte,pierre.silvie,pierre.martin}@cirad.fr

³ PHIM Plant Health Institute, Montpellier University,
IRD, CIRAD, INRAE, Institut Agro, Montpellier, France


Abstract. Formal Concept Analysis (FCA) comes with a range of relevant techniques for knowledge analysis, such as conceptual structures or implications. The Duquenne-Guigues basis of implications provides a cardinality minimal set of non-redundant implications. The concern of a domain expert is to discover new knowledge within this implication set. The objective of this prospective paper is to collect and discuss the different patterns of implications extracted from a dataset on plants used in medical care or consumed as food. We identify 16 patterns combining 3 types of knowledge elements (KE). The patterns highlight redundant KEs, or KEs of little interest, in particular, those corresponding to plant taxonomy, as it is familiar knowledge for the experts. Removing these KEs from the implications would make them tacit. We suggest a post-process for cleaning up the implications before reporting them to the experts. In addition, we discuss the different patterns and how an implication classification based on patterns could help the experts.

Keywords: Formal Concept Analysis · Duquenne-Guigues basis · Implication Rules · Life Sciences Knowledge Base · One Health

1 Introduction

Formal Concept Analysis (FCA) is a mathematical framework based on lattice theory which aims to formalize the notion of concept [6]. It gives foundations for a large range of methods for knowledge processing and knowledge discovery [12]. These methods include the construction of formal concepts and their ordering in a concept lattice or in restricted sub-structures of the lattice. For a domain expert (e.g. pathologist, entomologist), navigating through a complex lattice to extract knowledge may remain a challenge. An alternative view on knowledge is

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

building the Duquenne-Guigues basis (DGB) of implications [8]. An interest of this implication basis is its formulation of pieces of knowledge using a compact and comprehensive formalism. DGB indeed provides a cardinality-minimal set of non-redundant implications.

Through DGB building, expert concern is to discover new knowledge within the implication set. Diverse situations can occur. For instance, if an implication is too obvious for the expert, e.g. because it exclusively describes a domain taxonomy, then it presents a too limited interest to be kept in the implication set. This implication therefore becomes a tacit knowledge, while the others remain explicit. In more complex situations, the implication may contain both obvious knowledge elements (KE) and useful KE. In such cases, the obvious part of the implication may become tacit, when the other part may remain explicit.

The objective of this paper is to observe and discuss different patterns of implications extracted from a dataset on plants used in medical care or consumed as food. With these patterns, we aim to identify obvious KE included in the implications, and how implications can be simplified. The patterns may also provide an opportunity to classify the implications into coherent sets. Section 2 introduces the background and the dataset. Section 3 presents and discusses the preliminary results. Section 4 concludes and draws future work.

2 Background and Dataset

Background FCA elaborates knowledge, including formal concepts or attribute implications, on top of a formal context (FC) $\mathcal{K} = (G, M, I)$ where G is an object set, M is an attribute set and $I \subseteq G \times M$. An implication, denoted by $A \implies B$, is an attribute set pair (A, B) , $A, B \subseteq M$ such that all objects owning the attributes of A (premise) also own the ones of B (conclusion): $\{g | \forall m_a \in A, (g, m_a) \in I\} \subseteq \{g | \forall m_b \in B, (g, m_b) \in I\}$. There are several types of implication bases [1]. Here we consider the Duquenne-Guigues basis (DGB) of implications [8], which is a cardinality minimal set of non-redundant implications, from which all implications can be produced. An implication is held (or supported) by a number of objects, that we call the implication scope (S). The support is the proportion of such supporting objects. Let $Imp = A \implies B$, $S(Imp) = |\{g | \forall m \in A, (g, m) \in I\}|$. $Support(Imp) = S(Imp)/|G|$. For this work, the DGB of implications is built on a FC using Cogui software platform⁴, which includes a Java implementation of LinCbO

The dataset and the taxonomic knowledge To conduct the evaluation, we use an excerpt of the Noctuidae dataset [13], which is itself part of the Knomana dataset [14]. This dataset draws particular attention of experts in the context of One Health initiative [11] for addressing the worrying worldwide invasion of *Spodoptera frugiperda* (Lepidoptera from the Noctuidae family) which was first detected in Africa in 2016 [7] and is continuously spreading. At the end of 2018, *S. frugiperda* was first found in Yunnan Province in China [16]. Furthermore, in

⁴ <http://www.lirmm.fr/cogui/>

2018, it was first recorded in South Asia, namely India [9]. In January 2020, it was trapped in Australia’s special biosecurity zone in the Torres Strait islands of Saibai and Erub, and confirmed on 3 February 2020, and on mainland Australia in Bamaga on 18 February 2020 [15].

Table 1. On the top, excerpt of [13] that describes organisms uses, and, on the bottom, the associated formal context (FC) after the nominal scaling of *Species*, *Genus*, *Family*, *food*, and *medical*. *S*, *G*, and *F* are short notation for Species, Genus, and Family.

<i>Organism</i>		Species	Genus	Family	food	medical
p1	Acorus	Calamus	Acorus	Acoraceae	no	yes
p2	Cychorium	Intybus	Cychorium	Asteraceae	yes	yes
p3	Achillea	Collina	Achillea	Asteraceae	no	no

<i>FC</i>	S_Acorus Calamus	S_Cychorium Intybus	S_Achillea Collina	G_Acorus	G_Cychorium	G_Achillea	F_Acoraceae	F_Asteraceae	food	no-food	medical	no-medical
p1	X			X			X			X	X	
p2		X			X			X	X		X	
p3			X			X		X		X		X

This dataset indicates for each plant organism, out of the 600 in the dataset, its species, its genus, its family, and whether it is consumed as food and used in medical care. There is one-to-one mapping between organisms (objects) and *Species* values (cf. Tab. 1). Species, genera, and families respect a taxonomy which is a 3-level tree structure with this general shape: Species \prec Genus \prec Family; E.g. Species *Acorus calamus* \prec Genus *Acorus* \prec Family *Acoraceae* (see Fig. 1). This taxonomy is a familiar knowledge for the experts. The dataset comprises 600 species from 376 genera and from 98 families.

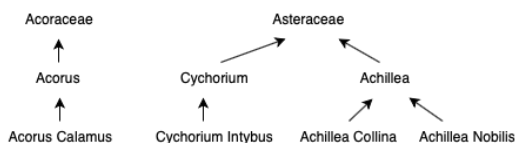


Fig. 1. Example of taxonomy for plants. Family, genus, and species names are respectively presented on the top, in the middle, and at the bottom of the figure. An arrow represents a generalization relation.

Formal context built from the dataset A FC describes a set of objects using a set of Boolean attributes. When the dataset contains a multi-valued attribute, conceptual scaling can be used to obtain Boolean attributes [6]. Various conversion methods are adopted, among which the nominal scaling for categorical attributes, such as the species name (e.g. *Achillea collina* or *Acorus calamus*), where each value is converted into a Boolean attribute [10]. Applied to this

work, the conversion of the dataset as a FC consisted in the nominal scaling of the attributes *Species*, *Genus*, *Family*, *food*, and *medical*. The taxonomy KE is expressed, in the FC (G, M, I) , by the fact that for a plant organism $p \in G$ and a given species s from genus g and family f , with s, g and $f \in M$, if $(p, s) \in I$, then $(p, g) \in I$. Similarly, if $(p, g) \in I$ then $(p, f) \in I$. In addition, the dual of the *food* (i.e. *no-food*) and *medical* (i.e. *no-medical*) attributes are added in the FC to explicit respectively the fact to be not consumed (*no-food*) or not used in medical care (*no-medical*). Note that, in [13], the attribute *medical* is encoded by *Medical_X*, *no-medical* by *Medical_*, *food* by *Food_X*, and *no-food* by *Food_*.

3 Results and Discussion

As noticed in [4], the implications from the DGB are not redundant one with the others. But they may contain redundant attributes in the premise and in the conclusion, due to the fact that pseudo-intents are used instead of minimal generators. Besides, we can expect that implications respect a limited number of patterns, and that some of these patterns have different meanings. A long-term objective of this work is to provide the experts with a minimal set of implications filtered and classified to assist them in the analysis. In this section, we observe patterns in implications of DGB. Then, we discuss how implications may be post-processed and classified making them more appropriate to the domain expert.

3.1 The implications from DGB

Due to the scaling of the attributes *Species*, *Genus*, and *Family*, the FC associated to the dataset has 1078 Boolean attributes, i.e. 600 attributes to inform on the species, 376 on the genus, 98 on the family, and 4 on the medical and food use. For the 600 plant organisms, Table 2 shows that 1168 implications were extracted from this FC. Most of the implications, i.e 1007, are held by one object, and thus are specific to a plant species. Among the 161 remaining ones, 9 implications are supported by more than 9 objects. The maximum scope, i.e. 35, corresponds to the implication informing that none of the 35 species from the Meliaceae Family, present in the dataset, is consumed (cf. ID 1 in Table 3).

Table 2. Number of implications per scope.

Scope	1	2	3	4	5	6	7	8	10	11	16	18	29	35
#Implications	1007	76	37	18	12	3	1	5	3	2	1	1	1	1

These 1168 implications are formulated using 16 patterns, where a pattern corresponds to the pair (Premise, Conclusion) in which each of the declarative sentence is designed using the SGFp schema (Table 3). This schema is the ordered list of presence of the attributes *Species* (S), *Genus* (G), *Family* (F), *food* or *no-food*, *medical* or *no-medical* (p), in the declarative sentence. By grouping *food*,

no-food, *medical* and *no-medical*, our intent is to focus our analysis on the types of KEs, and not the KEs themselves.

Various combinations of S, G, F, and p can be observed in the premise or the conclusion. For instance, pattern 1 informs on the use of all plants from a family. Pattern 2 provides the taxonomic relation of a genus with a family. Some patterns are more extended, such as pattern 5 that states on the genus of plants from a family with a given use.

Table 3. Implication patterns from DGB. The premise and the conclusion are designed using the SGFp schema. KU, KT, and KD are respectively short notations for Knowledge on plant Use, Knowledge on plant Taxonomy, and Knowledge on the Dataset.

ID	Premise	Conclusion	knowledge elements	Example of implication	#implications	Max scope
1	F	p	KU	F.Meliaceae ⇒ no-food	35	35
2	G	F	KT	G.Salvia ⇒ F.Lamiaceae	12	18
3	Fp	p	KU	no-food,F.Annonaceae ⇒ no-medical	10	16
4	G	Fp	KU, KT	G.Trichilia ⇒ no-food,F.Meliaceae	84	10
5	Fp	G	KU, KD	food,medical,F.Rutaceae ⇒ G.Citrus	6	5
6	F	G	KD	F.Piperaceae ⇒ G.Piper	1	5
7	GFp	p	KU, KT	medical,F.Asteraceae,G.Artemisia ⇒ no-food	7	4
8	Fp	Gp	KU, KD	medical,F.Annonaceae ⇒ food,G.Annona	1	3
9	F	Gp	KU, KD	F.Lythraceae ⇒ no-food,no-medical,G.Lythrum	5	2
10	S	GFp	KU, KT	S.ZygophyllumAlbum ⇒ no-food,no-medical,G.Zygophyllum,F.Zygophyllaceae	600	1
11	G	SFp	KU, KT, KD	G.Zygophyllum ⇒ no-food,no-medical,S.ZygophyllumAlbum,F.Zygophyllaceae	280	1
12	Fp	SG	KU, KT, KD	medical,no-food,F.Zingiberaceae ⇒ S.HedychiumCoronarum,G.Hedychium	29	1
13	GFp	S	KU, KT, KD	medical,no-food,G.Cinnamomum,F.Lauraceae ⇒ S.CinnamomumCassia	38	1
14	F	SGp	KU, KT, KD	F.Zygophyllaceae ⇒ no-food,no-medical,S.ZygophyllumAlbum,G.Zygophyllum	42	1
15	Fp	SGp	KU, KT, KD	food,F.Lauraceae ⇒ medical,G.Cinnamomum,S.CinnamomumVerum	3	1
16	GFp	Sp	KU, KT, KD	medical,F.Solanaceae,G.Solanum ⇒ food,S.SolanumLycopersicum	15	1

3.2 Observing patterns and implications

Three types of KEs were identified in the implications. KU type informs on the relationship of a plant, at any taxonomic level, with a use as food or medical care. The second KE type is the Taxonomic relationship type (KT), such as giving the family in the conclusion when the species is indicated in the premise. The third type (KD) corresponds to a KE resulting from the content of the Dataset, which represents a limit in this work. For instance, taxonomic referential web sites list 5 genera from the Piperaceae family⁵. As only one is present in the dataset (i.e. Piper), inferring that “*a plant from the Piperaceae family is of the genus Piper*” is wrong in the real life, but is true in this work as it results from a side effect of the dataset.

Except for patterns 2 and 6, all the implication patterns include KU (Table 3), suggesting at first sight that the latter are useful implications for the expert. But attention should be paid on implications combining KU, KT, or KD, and respected by implications with a scope value of 1, meaning that each one is associated to a single plant organism. Pattern 10 (scope 600) only reports information from the context as these rules only describe an organism by its

⁵ E.g. <http://www.plantsoftheworldonline.org/> lists 5 plant genera.

attributes. Similarly, patterns 11 and 14 indicate that a given genus or family has only one species in the dataset, the other attributes being those of the species, and these patterns could be considered as containing only KD. Most of the patterns include KT. Including the taxonomy was crucial in this work for FC processing in order to discover knowledge at a higher generic level, but corresponds to a redundancy in the implications.

This preliminary analysis gives directions for pattern and implication post-processing and classification. Some may be specific to some characteristics of Knomana, such as the fact that attribute species is an object identifier, and some could be generalized to all datasets.

The post-processing may have different forms for redundant or evident information: either removing it, or simply separating it from the rest, so that it remains written but not distracting. KU, KD, KT can be highlighted in different ways to distinguish them. Highlighting the different reasons under redundancy or evident information (e.g. KT information versus logical redundancy due to the fact that minimal generators are not used) would be useful. We could consider removing redundancy only in premise [5] or only in conclusion, or in both.

Patterns also have to be analyzed. For example, pattern 4 ($G \rightarrow Fp$) could be simplified as $G \rightarrow p$, as F is tacit given G. This would change the pattern classification (initially KU KT), as it is now reduced to KU.

As regard with implications, a post-process could remove KT from the implications, corresponding to a tacit knowledge as experts are familiar with the taxonomy. As the redundancy due to KT may appear in the premise (e.g. *the species and its genus*), in the conclusion, or in both (e.g. *a species implies a family*), the post-process has to consider the implication in its entirety. This approach differs with [5] where authors consider exclusively the left-minimal premises for technical purpose (i.e. a fast computation of attribute closure and a minimal left hand side in the implication). In addition, a filtering could lead to remove pattern 2 implications that only express tacit knowledge.

KUs are the explicit KEs investigated by the expert and thus have to be put forward. KDs present a particular situation in the implications as they result from a lack of knowledge in the dataset. Pattern 6 has to be considered carefully as it contains only KD. Thus, the experts have to be alerted of KD presence to consider this aspect in the dataset analysis.

A classification of implications based on patterns seems to us relevant for presenting them to the expert by groups having a coherent meaning: E.g. implications providing information on the diversity of some plant families in the dataset or pure information on the One Health Approach. Depending of the number of rules in some categories, a classification may have a significant impact for the expert, e.g. discarding or at least separating rules from pattern 10 distinguishes 600 rules that only recall initial data. We also guess that if a post-processing is made, the way it is made has to be notified to the expert and it should be indicated if this is reversible operation.

Finally, literature on association rules also faced the issues of extension of non-redundant rules [17] and redundancy removal introduced by using a taxon-

omy in concept-based rules building [3]. Classifying association rules in a lattice has been addressed in [2] in the context of fault localization, where rules are described by elements of their premises, that can be inspiring in our case, using patterns as an implication description.

4 Conclusion

This paper identifies 3 types of knowledge elements and 16 patterns that constitute the implications from the Duquenne-Guigues basis on a formal context. Each implication needs a specific consideration before being presented to the expert. For instance, a post-process can be conducted to remove tacit knowledge elements from implications, which may drive to delete some of them.

In a future work, we will study how the different patterns can be used to display implications to experts by categories, that may help them to focus on different aspects of the dataset. For instance, the user may focus on knowledge elements related to some plant families in the dataset, or pure knowledge elements on the plant uses at the family level.

This work is a preliminary study to the analysis of the Duquenne-Guigues basis of implications resulting from a Relational Context Family of the Knomana knowledge base. The general objective is to contribute in the decision support process to identify plants that could be used by farmers to control pest. These pesticidal plants will be an alternative to pesticide and antibiotics, considering the One Health approach. Using this approach, one must be aware of the multi-uses of these plants to prevent the intentional effects on the animals, the humans, and their environment.

In a more general perspective, it would be relevant to examine how the forms of post-processing, filtering and classification of patterns and rules can be generalized in order to be able to apply these approaches to other datasets, more particularly when several taxonomic relations are involved.

Acknowledgments. We warmly thank the reviewers for their comments. Part of the discussion has been notably improved thanks to their relevant remarks. This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

References

1. Bertet, K., Demko, C., Viaud, J.F., Guérin, C.: Lattices, closures systems and implication bases: A survey of structural aspects and algorithms. *Theor. Comp. Sci.* **743**, 93–109 (2018)
2. Cellier, P., Ducassé, M., Ferré, S., Ridoux, O.: Dellis: A data mining process for fault localization. In: *Proceedings of the 21st International Conference on Software Engineering & Knowledge Engineering (SEKE)*. pp. 432–437 (2009)
3. Cellier, P., Ferré, S., Ridoux, O., Ducassé, M.: A parameterized algorithm to explore formal contexts with a taxonomy. *Int. J. Found. Comput. Sci.* **19**(2), 319–343 (2008)

4. Cordero, P., Enciso, M., Mora, A., Ojeda-Aciego, M.: Computing minimal generators from implications: a logic-guided approach. In: Szathmary, L., Priss, U. (eds.) Proceedings of The Ninth International Conference on Concept Lattices and Their Applications (CLA). CEUR Workshop Proceedings, vol. 972, pp. 187–198 (2012)
5. Cordero, P., Enciso, M., Mora, A., Ojeda-Aciego, M.: Computing left-minimal direct basis of implications. In: Ojeda-Aciego, M., Outrata, J. (eds.) Proceedings of the Tenth International Conference on Concept Lattices and Their Applications (CLA). CEUR Workshop Proceedings, vol. 1062, pp. 293–298 (2013)
6. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer (1999)
7. Goergen, G., Kumar, P., Sankung, S., Togola, A., Tamò, M.: First report of outbreaks of the fall armyworm *Spodoptera frugiperda* (J. E. Smith) (Lepidoptera, Noctuidae), a new alien invasive pest in West and Central Africa. *PLoS ONE* **11** (2016)
8. Guigues, J.L., Duquenne, V.: Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Math. et Sci. Hum.* **24**(95), 5–18 (1986)
9. Guo, J., He, K., Wang, Z.: Biological characteristics, trend of fall armyworm *Spodoptera frugiperda*, and the strategy for management of the pest. *Chin. J. Appl. Entomol.* **56**(3) (2019)
10. Ignatov, D.I.: Introduction to Formal Concept Analysis and Its Applications in Information Retrieval and Related Fields. *CoRR* **abs/1703.02819** (2017)
11. Kahn, L.H., Kaplan, B., Monath, T.P., Conti, L.A., Yuill, T.M., Chapman, H.J., Carter, C.N., Barrentine, B.: One-health initiative. <http://www.onehealthinitiative.com> (2020), accessed: 2020-04-01
12. Kuznetsov, S.O., Poelmans, J.: Knowledge representation and processing with formal concept analysis. *Wiley Interd. Rev. Data Min. Knowl. Disc.* **3**(3), 200–215 (2013)
13. Martin, P., Gutierrez, A., Marnotte, P., Huchard, M., Keip, P., Mahrach, L., Silvie, P.: Dataset on Noctuidae species used to evaluate the separate concerns in conceptual analysis: Application to a life sciences knowledge base (2021). <https://doi.org/10.18167/DVN1/HTFEST>
14. Martin, P., Silvie, P., Sarter, S.: Knomana - usage des plantes à effet pesticide, antimicrobien, antiparasitaire et antibiotique (patent APP IDDN.FR.001.130024.000.S.P.2019.000.31235) (2019)
15. Queensland-Government: Department of Agriculture and Fisheries. First mainland detection of fall armyworm (2020)
16. Shylesha, A., Jalali, S., Gupta, A., Varshney, R., Venkatesan, T., Shetty, P., Ojha, R., Ganiger, P., Navik, O., Subaharan, K., Bakthavatsalam, N., Ballal, C., Raghavendra, A.: Studies on new invasive pest *Spodoptera frugiperda* (J. E. Smith) (Lepidoptera: Noctuidae) and its natural enemies. *J. Biol. Cont.* **32**(3), 145–151 (2018)
17. Zaki, M.J.: Closed itemset mining and non-redundant association rule mining. In: Liu, L., Özsu, M.T. (eds.) *Encyclopedia of Database Systems*, pp. 365–368. Springer US (2009)

Analyzing the composition of remedies in ancient pharmacopeias with FCA

Agnès Braud¹, Xavier Dolques¹, Pierre Fechter², Nicolas Lachiche¹,
Florence Le Ber¹, and Véronique Pitchon³

- (1) Université de Strasbourg, CNRS, ENGEES, ICube UMR 7357, F67000 Strasbourg
`{agnes.braud,dolques,nicolas.lachiche}@unistra.fr`
`florence.leber@engees.unistra.fr`
- (2) Université de Strasbourg, CNRS, ESBS UMR 7242, F67000 Strasbourg
`p.fechter@unistra.fr`
- (3) Université de Strasbourg, CNRS, Archimede UMR 7044, F67000 Strasbourg
`pitchon@unistra.fr`

Abstract. This paper presents the collaborative work led in an interdisciplinary project on studying remedies in Arabic medieval pharmacopeia. The goal is to find new molecules in plants or combinations of active principles to substitute them for antibiotics which encounter some limits. Formal Concept Analysis is used to discover co-occurrences of ingredients. We describe the difficulties inherent to these data and the results of preliminary analyses.


1 Introduction

This paper presents the preliminary work led in an interdisciplinary project that aims at studying remedies in Arabic medieval pharmacopeia. The project gathers researchers in history, microbiology, botanic and computer science. The general goal is to discover the active ingredients in remedies and to test them experimentally on bacteriae. The extracted knowledge could be used to substitute molecules in plants or combinations of active principles for antibiotics which encounter some limits.

In this work, we focus on medieval remedies prescribed for urinary problems. We present a dataset of remedies extracted from 5 pharmacopeias. This dataset is rather small, but the modeling is not straightforward. Although the building of the database is not finished, some analyses can already be done on the composition of remedies. Formal Concept Analysis (FCA) [3] appears as particularly suitable for this task, and preliminary results we obtained using it seemed very interesting to historians and microbiologists.

Section 2 presents the process for carrying out the project and in particular data collection. Section 3 describes the data, their characteristics and the difficulties resulting from them. In Sect. 4, we present our preliminary analyses and show the benefits of using FCA in this context. Finally, Sect. 5 yields some conclusions.

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Process overview

The whole process of this research work is presented on Fig. 1.

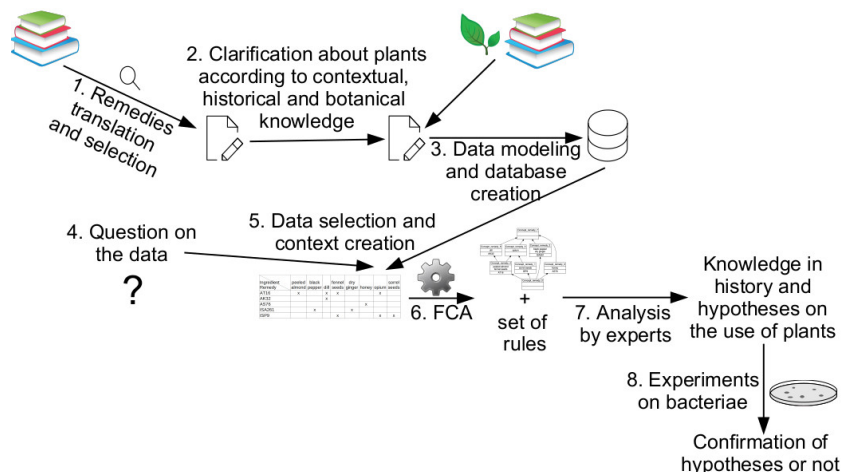


Fig. 1. Overview of the project process

The first step is to select pharmacopeias and remedies in them. This is done by an historian, who reads Arabic, on the basis of the symptoms indicated in the description of the remedy and on her knowledge of Arabic medicine. Let us note that Arabic medicine is based on the theory of Humours that considers that the body is constituted of four humours (water, fire, earth, air) having some qualities (hot, cold, dry, wet). The second step consists in clarifying some plants which are not described clearly enough in the text, with the help of a botanist. The next step is to build a database to store the remedies, whose modeling difficulties will be discussed in the next section. Then comes the formulation of questions that experts have on the data and that can be solved using FCA. The analysis of lattices and rules may bring new insights or questions both in history and on the use of plants in remedies. Some hypotheses on the use of plants stemmed from this analysis will be studied in collaboration with a botanist (for the identification and characterization of therapeutic properties of plants) and a pharmacologist studying drugs based on natural substances (for the identification of the active molecules in plants), and then tested on bacteriae by microbiologists.

In this project, 5 pharmacopeias have been selected:

- two versions of the pharmacopeia of Sābūr ibn Sahl (9th century). Sābūr ibn Sahl was born in Jundishāpūr (Iran) and was chief of the hospital there before moving to Iraq. A version of the pharmacopeia [6] was for the pharmacists of ‘Aḍudī’s hospital in Baghdad (this pharmacopeia will be designated by



Fig. 2. Example of pages of manuscripts - Left: Dioscorides Pedanius of Anazarbos, Kitab Disqūrīdis (Dioscorides’s book), circa 1100–1200, National Library of France. Right: Al-Kindī Folio 98r (photo credit: Aya Sofia, Istanbul, Ms. N° 3603)

ISA), and a shorter one [4] for pharmacists outside of the hospital who were considered less skillful (designated by ISP);

- the pharmacopeia of Al-Kindī (9th century) [8], who studied in Bassora and Baghdad. This pharmacopeia will be designated by AK;
- the pharmacopeia of Al-Tilmīdh (11th-12th centuries) [5], who was chief of ‘Aḍudī’s hospital in Baghdad. This pharmacopeia will be designated by AT;
- the pharmacopeia of Al-Samarqandī [9]. This pharmacopeia will be designated by AS.

The historian has based her choice firstly on the availability of books describing them, then preferably around Baghdad because of her knowledge on this area and also because Arabic pharmacopeias have a very rich content, and on a long period for a wide historical view. A few other criteria are linked to interesting historical questions.

Pharmacopeias are available as translations and/or original manuscripts. Figure 2 shows pages of two manuscripts. Both translations and original manuscripts may contain errors, so in the first case the historian may also have an original manuscript to check, and in the second case there may be errors made by copyists so that several different manuscripts are necessary. Among several hundreds of remedies, 38 prescribed for urinary symptoms were found. Among these, we removed 3 for this work: 1 from ISA which was already in ISP with exactly the same recipe, and 2 universal remedies with more than 50 ingredients (1 in ISP, 1 in ISA).

In the following, remedy identifiers will be prefixed by the acronym of the pharmacopeia in which they appear.

3 Data characteristics

The data that have been collected are rather small, but the collection requires a lot of time and the data are actually complex. The description of a remedy is composed of a list of ingredients (see Table 1 for an example) with quantities,

and a description of the preparation. Let us note that part of the ingredients are present in the recipe to treat the disease, and some others may be present to treat their side effects, but this is not indicated.

Table 1. List of ingredients of a remedy

- Armenian borax	- Ginger	- Peeled sweet almonds
- Kerman cumin	- White pepper	- Rue leaves
- Parsley	- Scammony	- Dates from Hairūn

Ingredients belong to different categories: plant, mushroom, mineral, animal (like honey or musc) and metal. The description of plants is not homogeneous:

- some plant descriptions correspond to different taxonomic rank (e.g. species for Scammony, genus for Euphorbia, several families for Fern);
- some descriptions correspond to a plant, other to a part of a plant, or even a part of a part of... a plant (e.g., Acorn/Acorn shell/Inside peel of acorn shell, or Thymus without part while just a part is probably used). Let us notice that several parts of the same plant can be used in a remedy;
- some ingredients have some transformations associated (e.g., Peeled and grilled bitter almonds);
- some ingredients are a mix of ingredients (e.g., Persian za'atar);
- a few ingredients have an origin (e.g., Kerman cumin) and it is an important information since it usually reflects a high quality;
- some are still to determine (e.g., Idrūmaghmū).

Data modeling requires expert knowledge to determine which information is important to capture; for example, should Acorn and Inside peel of acorn shell be comparable? Moreover, for a work on active principles, having a precise description of plants is important as different species may have different properties. The exact species may be implicit for the author, who is used to a specific one or to the local one. A more precise information can be inferred on the species, either by checking the translation, searching for contextual information in books, or relying on illustrations present in the book. In this last case, the botanist may help as well as with knowledge on the parts of plants traditionally used. At the end, the discussions on data modeling have raised new questions. For example, after a discussion on the problem of representing acorn and the different parts used, the historian has decided to study further in books the use of oak.

All these points make that a relational database is not suitable to store these data. A graph database offers more flexibility to represent them and to adapt the representation to new cases. We have thus chosen this type of database. For the same reasons, building a context for FCA requires some choices among the information.

Currently, the data on which we work are the data issued from the translation of texts. The work to clarify species and parts of plants is in progress, led by

the historian and the botanist, and will take some time, but it is still possible to start some analyses.

4 Preliminary analyses

For our preliminary analyses on the data, we have worked on two questions asked by the microbiologist: "which ingredients appear the most often?" (Q1) and "which ingredients often appear together?" (Q2). As shown in previous works [2], FCA-based approaches are very suitable to solve this type of questions.

We have worked both on the ingredients as they appear in the recipes, and on these ingredients with the ones coming from plants replaced just by the corresponding plant (without part and transformations, such as Oak for Acorn shell or Almond tree for Peeled sweet almonds). The plant can be at the level of species, genus, family... Let us note R the set of remedies, I the set of ingredients without modification and IP the set of ingredients with the ingredients coming from plants replaced by the plant, $contains \subseteq R \times I$ and $containsP \subseteq R \times IP$ two incidence relations describing the composition of remedies. We have thus built one context $CI = (R, I, contains)$ and another context $CIP = (R, IP, containsP)$. Let us recall that we work on 35 remedies concerning urinary symptoms. They are distributed in the 5 pharmacopeias as follows: 3 in AK, 5 in AS, 7 in AT, 6 in ISA and 14 in ISP. Their recipes contain between 2 and 21 ingredients giving 170 different ingredient descriptions in CI, and 144 attributes in CIP. The lattice obtained from CI (\mathcal{L}_{CI}) contains 138 concepts, the one obtained from CIP (\mathcal{L}_{CIP}) 151.

Starting with question Q1 ("which ingredients appear the most often?"), we study the lattices and in particular the concepts with the biggest extents. Table 2 shows some of the concepts obtained in \mathcal{L}_{CI} and the ones in \mathcal{L}_{CIP} corresponding to the same ingredients but considering the plant. The extent is presented so that the different pharmacopeias are separated. Concept ($\{Celery\ seeds\}, \{AK114, AT16, AT20, AT72, ISA137, ISA162, ISP100, ISP16, ISP185, ISP186, ISP78, ISP9\}$) from \mathcal{L}_{CI} reveals that Celery seeds is the ingredient that appears the most often (12 times) in remedies. Then Indian nard appears 9 times, Saffron and Black pepper 7 times. Of course, it is possible to know which ingredients appear most often by computing statistics in the database, but the extension of the concept also shows that Celery seeds appears in 4 out of 5 pharmacopeias. Similarly, lattice \mathcal{L}_{CIP} reveals Celery as appearing the most often in remedies. There is however a little change with Oak appearing with 8 occurrences, while Acorn appeared only 6 times in \mathcal{L}_{CI} . As can be seen in Table 2, this is due to the use of different parts of oak with different transformations (Acorn in 6 remedies, Acorn shell and Grilled inside peel of acorn shell in remedy AT137, Grilled acorn shell in AT80). There are thus 3 concepts introducing oak parts in \mathcal{L}_{CI} . Moreover, among the 6 remedies in the extension of the concept introducing Acorn shell, 2 are from ISA pharmacopeia and none from ISP while they are from the same author. This may be due to the fact that ISP was for pharmacists

outside of the hospital who were considered less skillful than those from the hospital, and acorn shell is toxic and thus should be used carefully.

Table 2. Examples of concepts in \mathcal{L}_{CI} and \mathcal{L}_{CIP}

Concepts of \mathcal{L}_{CI}		Concepts of \mathcal{L}_{CIP}	
Intent	Extent (<i>cardinality</i>)	Intent	Extent (<i>cardinality</i>)
Celery seeds	AK114, AT16, AT20, AT72, ISA137, ISA162, ISP100, ISP16, ISP185, ISP186, ISP78, ISP9 (12)	Celery	AK114, AT16, AT20, AT72, ISA137, ISA162, ISP100, ISP16, ISP185, ISP186, ISP78, ISP9 (12)
Indian nard	AT137, AT72, AT80, ISA137, ISA162, ISA261, ISP16, ISP27, ISP78 (9)	Indian nard	AT137, AT72, AT80, ISA137, ISA162, ISA261, ISP16, ISP27, ISP78 (9)
Saffron	AT16, ISP16, ISP186, ISP27, ISP6, ISP70, ISP9 (7)	Saffron	AT16, ISP16, ISP186, ISP27, ISP6, ISP70, ISP9 (7)
Black pepper	AS78, ISA137, ISA261, ISP47, ISP54, ISP6, ISP70 (7)	Black pepper	AS78, ISA137, ISA261, ISP47, ISP54, ISP6, ISP70 (7)
Acorn	AS124b, ASN39, AT137, AT81 ISA125, ISA231 (6)	Oak	AS124b, ASN39, AT137, AT377, AT80, AT81, ISA125, ISA231 (8)
Cypress, Olibanum bark, Olibanum, Lavender, Cumin, Acorn shell, Grilled inside peel of acorn shell, Indian nard	AT137 (1)		
Cypress, Olibanum bark, Olibanum, Lavender, Grilled acorn shell, Indian nard	AT80 (1)		

For question Q2 ("which ingredients often appear together?"), association rules are a good starting point. We computed those with a minimal support of 3 remedies and a minimal confidence of 80%. For context CI, we obtained 15 rules. Among these rules, 10 were implication rules. Three of these implication rules are given in Table 3.

One appeared with a support of 7: *Celery seeds* \rightarrow *Fennel seeds*, another one with a support of 6: *Opium* \rightarrow *Saffron*. With a support of 4, we had 2

Table 3. Examples of implication rules obtained on context CI

Rule	Support
Celery seeds \rightarrow Fennel seeds	7
Opium \rightarrow Saffron	6
Asarabacca, Indian nard \rightarrow Celery seeds	4

implication rules involving 3 ingredients such as: *Asarabacca*, *Indian nard* \rightarrow *Celery seeds*. Context CIP leads to more rules: 24, among which 20 implication rules. If we consider the first rule (*Celery seeds* \rightarrow *Fennel seeds*), we then explore the subconcepts of the concept introducing Celery seeds to see which ingredients other than Fennel seeds also appear with it. Figure 3 shows an extract of this part of the lattice. This analysis is interesting for microbiologists, in order to make hypotheses on the respective roles of the ingredients, and which ones play similar roles. Finally, experts were really interested by the results obtained with FCA on this first dataset.

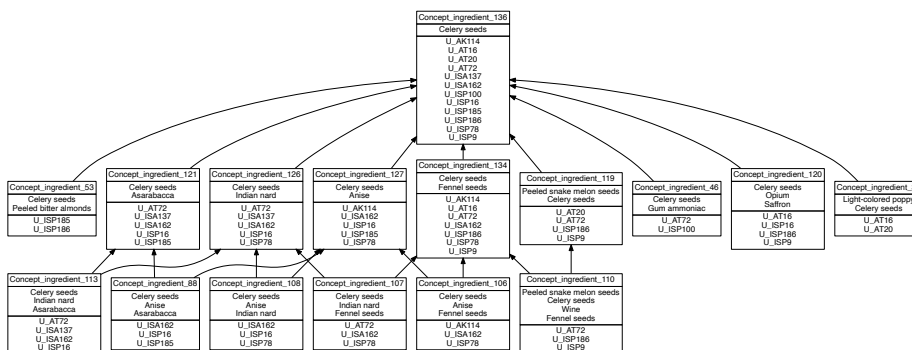


Fig. 3. Extract of lattice \mathcal{L}_{CI} extracted from context CI

The work in [1] is based on the same objective of finding co-occurring ingredients but in a different pharmacopeia and for skin, mouth, or eye infections. It uses community detection in a network of ingredients. FCA is a more direct process that gives a complete view of the combinations of ingredients. Moreover the expert can see the remedies corresponding to each combination and exploit association rules. An FCA-based approach has been used in a similar project focusing on the identification of local plants for addressing sanitary problems in Sub-Saharan Africa [10, 7]. In this project, data represent plants or part of plants, and how, and in which country, they are used to treat animal or plant diseases. They are collected from various contemporary texts and concern contemporary practices, unlike our data for which we thus have more uncertainty.

5 Conclusion

This project highlights several interesting points about the work on such real data. First, the necessity and richness of adopting an interdisciplinary approach. Indeed, interdisciplinarity is required at each step of the process. Data preparation requires historians, botanists and computer scientists. The questioning on data leading to the analyses realized by computer scientists comes from historians and microbiologists. Finally, the results are discussed with historians and microbiologists, who will also perform experimental tests of interesting hypotheses. The questioning evolves all along the process with the discussions and new results. Second point, FCA appears as particularly suitable for expert needs, and even the simple analyses of lattices and rules presented in this paper have revealed useful insights.

The next steps of the project will aim at enriching the database, both by clarifying ingredients and by integrating other data like quantities for the ingredients, symptoms for which remedies are prescribed, preparation of the remedies, remedies for other kind of diseases, and by adding information on plants such as toxicity. More complex analyses will then be possible.

Currently, the amount of data the historian needs to work is rather small, however if the amount of data increases we may have to search for techniques to facilitate the exploration of the results.

References

1. Connelly, E., del Genio, C.I., Harrison, F.: Data mining a medieval medical text reveals patterns in ingredient choice that reflect biological activity against infectious agents **11**(1) (2020)
2. Couceiro, M., Napoli, A.: Elements About Exploratory, Knowledge-Based, Hybrid, and Explainable Knowledge Discovery. In: ICFCA 2019, Frankfurt, Germany. LNAI, vol. 11511, pp. 3–16. Springer (Jun 2019)
3. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag (1999)
4. Kahl, O.: The Small dispensatory: Sābūr ibn Sahl. Brill, Leiden (2003)
5. Kahl, O.: The dispensatory of Ibn at-Tilmīdh, Arabic text, English translation, study and glossary. Brill, Leiden (2007)
6. Kahl, O.: Sābūr ibn Sahl's Dispensatory in the Recension of the 'Aḍudī Hospital. Brill, Leiden; Boston (2009)
7. Keip, P., Gutierrez, A., Huchard, M., Le Ber, F., Sarter, S., Silvie, P., Martin, P.: Effects of Input Data Formalisation in Relational Concept Analysis for a Data Model with a Ternary Relation. In: ICFCA 2019, Frankfurt, Germany. LNCS, vol. 11511, pp. 191–207. Springer (Jun 2019)
8. Levey, M.: The Medical Formulary or Aqrābādhīn of Al-Kindī. Univ. of Pennsylvania Press, Philadelphia (1966)
9. Levey, M., al Khaledy, N.: Chemistry in the Medical Formulary of Al-Samarqandī. Univ. of Pennsylvania Press, Philadelphia (1967)
10. Silvie, P.J., Martin, P., Huchard, M., Keip, P., Gutierrez, A., Sarter, S.: Prototyping a Knowledge-Based System to Identify Botanical Extracts for Plant Health in Sub-Saharan Africa. *Plants* **10**(5) (2021)

Mining the Groceries Database using Triadic Concept Analysis

Pedro H. B. Ruas¹, Rokia Missaoui², Mark A. J. Song¹, Léonard Kwuida³

¹ Pontifical Catholic University of Minas Gerais Belo Horizonte, Brazil
pedrohbruas@gmail.com, song@pucminas.br

² Université du Québec en Outaouais
rokoa.missaoui@uqo.ca

³ Bern University of Applied Sciences, Bern, Switzerland
leonard.kwuida@bfh.ch

Abstract. In this paper we illustrate the potential of Triadic Concept Analysis (TCA) in extracting useful patterns from triadic formal contexts. To that end, we exploit our recent contributions to TCA to analyze the *Groceries* database and identify triadic patterns expressed by concepts and association rules, including implications.


1 Introduction

Since datasets are frequently expressed by ternary and more generally n -ary relations, one can observe a recent and increasing interest to propose new solutions for the analysis and exploration of these multidimensional data, and specially triadic contexts [2, 4, 6, 7, 10]. This is the case of multidimensional social networks, social resource sharing systems such as folksonomies, and security policies. For instance, in the latter application, a 4-ary or quaternary relation indicates that a *user* is authorized to use *resources* with given *privileges* under restricted *conditions*. For example, User 1000 is allowed to access to Files F_1 , F_2 and F_3 with the privilege to *read* F_1 and *update* the last two files in the first three working days of the week.

In this paper, we aim at illustrating the utilization of our software platform using a subset of the well-known dataset named the *Groceries* database [1] of transactions made by customers at a given date for a set of products. Subsets and variants of this dataset have been extensively used by data mining researchers under the name of “market basket analysis” to discover associations between products bought by customers by looking for combinations of items that occur together frequently in customer transactions.

The rest of the paper is organized as follows. In Section 2 we briefly recall the main notions of Triadic Concept Analysis. Section 3 presents our platform, the original dataset and its preprocessing as well as a few examples of the generated patterns which are triadic concepts and association rules, including implications. Finally, Section 4 summarizes the paper and presents the future work in terms of the platform enrichment and validation using synthetic and real data.

RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Triadic Concept Analysis

Triadic concept analysis was originally introduced by Lehmann and Wille [9] as an extension to FCA, to analyze data described by three sets K_1 (objects), K_2 (attributes) and K_3 (conditions) together with a 3-ary relation $Y \subseteq K_1 \times K_2 \times K_3$. $\mathbb{K} := (K_1, K_2, K_3, Y)$ is called a *triadic context*. A triple (a_1, a_2, a_3) in Y means that object a_1 possesses attribute a_2 under condition a_3 . Table 1 shows a row of the context, which partially describes the transactions made by Customer 3782 in K_1 who bought items in K_2 , including *pip fruit* in *January*, *chicken* and *oil* on *February*, and *curd* on *January* and *September* as listed in Table 1. Here the condition set K_3 represents the twelve months (J, F, \dots, D) of the year.

\mathbb{K}	Chicken	Oil	Pip Fruit	Curd
	J F ... S	J F ... S	J F ... S	J F ... S	J F ... S	J F ... S
3782	1	1	1	1 1

Table 1. A row of the triadic context $\mathbb{K} := (K_1, K_2, K_3, Y)$

A *triadic concept* or *closed tri-set* of \mathbb{K} is a triple $c = (A_1, A_2, A_3)$ (also denoted by $A_1 \times A_2 \times A_3$) with $A_1 \subseteq K_1$, $A_2 \subseteq K_2$, $A_3 \subseteq K_3$ and $A_1 \times A_2 \times A_3 \subseteq Y$ is maximal with respect to inclusion in Y . A_1 is called the *extent* of c , A_2 its *intent*, and A_3 its *modus*. We name (A_2, A_3) the *feature* of c . For example, $(\{2512, 4320\}, \{\text{canned beer}\}, \{\text{January, April}\})$ is one of two triadic concepts (see the green box in Figure 2) with the same extent $\{2512, 4320\}$ which indicates that the two identified customers have two common features. The first feature means that both of them bought *canned beer* in *January* and *April* while the second one indicates that they purchased *butter* and *canned beer* in *January*.

Let $\mathcal{L} = (\mathcal{C}, \leq_1)$ be a poset of nodes such that each node represents the set of triadic concepts in \mathcal{C} with the same extent, and the relation \leq_1 is induced by the inclusion on extents. We defined in [10] an adapted version of the *iPred* algorithm to link triadic concepts according to the quasi-order on extents. This allows the Hasse diagram construction of triadic concepts.

A pair (B_2, B_3) is called a triadic feature-based generator [10] (or F-generator) associated with (*i.e.*, compatible with) the feature (A_2, A_3) in a concept (A_1, A_2, A_3) if $A_2 \times A_3 \subseteq (B_2, B_3)^{(1)(1)}$, where the (i) -derivation [9] is defined by:

$$X_i^{(i)} := \{(a_j, a_k) \in K_j \times K_k \mid (a_i, a_j, a_k) \in Y \forall a_i \in X_i\} \quad (1)$$

$$(X_j, X_k)^{(i)} := \{a_i \in K_i \mid (a_i, a_j, a_k) \in Y \text{ for all } (a_j, a_k) \in X_j \times X_k\}. \quad (2)$$

As an illustration, let us consider the portion of the Hasse diagram shown in Figure 2. We can see that each node represents the set of features and F-generators associated with one extent. For example, there are two features and two F-generators attached to the node whose extent is $\{2512, 4320\}$. The dotted

red box attached to the extent $\{2512\}$ contains ten F-generators. One of them is $(\{coffee, pastry\}, \{September\})$ which is simply denoted by $(coffee, pastry - September)$ in Figure 2.

Biedermann [5] defined a *triadic implication* of the form $(A \rightarrow D)_C$ with the meaning that “whenever A occurs under all conditions in C , then D also occurs under the same conditions”. Later on, Ganter and Obiedkov [8] extended Biedermann’s definition by proposing three types of implications. We recall two of them in the following.

A *conditional attribute* implication takes the form: $A \xrightarrow{C} D$, where A and D are subsets of K_2 , and C is a subset of K_3 . It means that A implies D under all conditions in C . In particular, the implication holds for any subset in C . This implication is then linked to Biedermann’s definition of triadic implication as follows [8]: $A \xrightarrow{C} D \iff (A \rightarrow D)_{C_1}$ for all $C_1 \subseteq C$.

In a dual way, an *attributinal condition* implication is an exact association rule of the form $A \xrightarrow{C} D$, where A and D are subsets of K_3 , and C is a subset of K_2 .

Let (B_2, B_3) be a feature-based generator associated with (A_2, A_3) of a triadic concept whose extent is A_1 , i.e., $B_2 \subseteq A_2$ and $B_3 \subseteq A_3$. Then, we define in [10] the *Biedermann conditional attribute implication* (BCAI) as $(B_2 \rightarrow A_2 \setminus B_2)_{B_3}$ with a support equal to $|A_1|/|K_1|$ if $B_2 \subseteq A_2$ and $B_3 \subseteq A_3$, where K_1 stands for the set of objects.

Dually, the *Biedermann attributinal condition implication* (BACI) $(B_3 \rightarrow A_3 \setminus B_3)_{B_2}$ holds with a support equal to $|A_1|/|K_1|$ if $B_3 \subseteq A_3$ and $B_2 \subseteq A_2$. A_2 involved in a BCAI is constrained to be maximal among all the intents of the features associated with (B_2, B_3) and A_3 inside a BACI must be maximal among all the modi of the features associated with (B_2, B_3) .

Given a feature-based generator $g = (B_2, B_3)$ and a feature (A_2, A_3) , both associated with the node whose extent is A_1 such that A_2 is the maximal intent in $((B_2, B_3)^{(1)})^{(1)}$ that contains B_2 . Let the quasi-order $(X_1, X_2, X_3) \lesssim_1 (A_1, A_2, A_3)$ holds between the current concept $c = (A_1, A_2, A_3)$ and $c_1 = (X_1, X_2, X_3)$ found in the lower cover of the node that contains c . Then, we claim the following statement:

If $A_2 \subset X_2$ and $X_3 \subseteq A_3$ and $B_2 \subset X_2$ and $B_3 \subseteq X_3$, then the following *Biedermann conditional attribute association rule* *BCAAR* holds: $(B_2 \rightarrow X_2 \setminus A_2)_{B_3}$ with a support equal to $|X_1|/|K_1|$ and a confidence of $|X_1|/|A_1|$, where K_1 stands for the set of objects.

In a dual way, the *Biedermann attributinal condition association rule* *BACAR* can be defined.

3 A Case Study

3.1 Platform

Our development platform for TCA is implemented mostly with Python and has the following modules [10] as presented in Figure 1:

1. Call of *Data-Peeler* procedure [6] to get triadic concepts
2. Computation of the precedence links by adapting the *iPred* algorithm [3] to the triadic setting
3. Calculation of two types of triadic generators, including the feature-based generators
4. Computation of association rules, including implications
5. Adaptation of stability and separation indices to the triadic framework.

In this paper, we will mainly illustrate our work on Module 2 and parts of Modules 3 and 4.

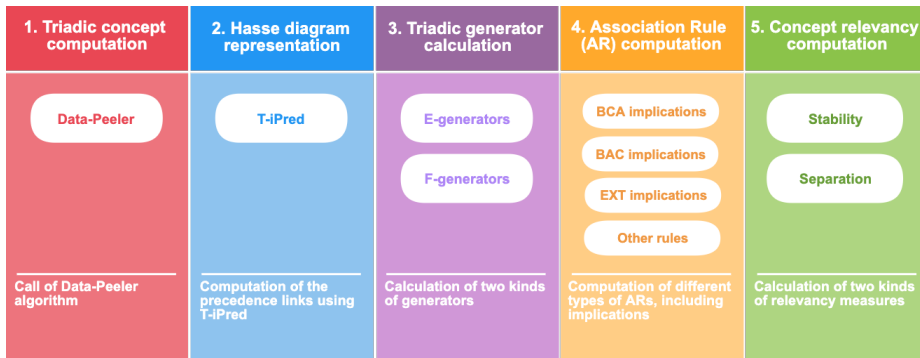


Fig. 1. The architecture modules of our proposed solution.

3.2 Data set

The *Groceries* database [1] contains 38765 transactions, 3898 customers, 167 products (items) and 728 distinct transaction dates. From this database, we extracted and analyzed many samples of different sizes w.r.t. to the number of customers, the number of items and variants of transaction dates (day and month, month only, day of the week, ..). The subset we are analyzing in this paper contains 8121 transactions made by 1000 customers who bought 30 items during a given month (rather than a specific date) between 2014 and 2015. To select items, we took the 23 frequently bought products (after removing the first top 5 ones) and added seven other items which are turkey, chicken, chocolate, meat, ham, ice cream, and napkins. The identified customers are those who bought at least one of the 30 selected products. A portion of the input data after the preprocessing step (but before a conversion into a triadic context) can be seen in Table 2.

The reason we used such a sample is due to the fact that it allows us to illustrate the meaning of the generated patterns that are either triadic concepts, implications or association rules with a confidence lower than 1. Due to the

sparsity of the initial dataset and many large subsets, the set of implications was frequently empty or small and many generated association rules were trivial with a very low support.

The sample is then converted into a triadic context where the value 1 in a cell indicates that a given customer bought an item at least once during a given month. For example, Customer 3782 purchased a pip fruit on January as shown in Table 1.

Customer number	Item	Month
2512	butter	January
2512	coffee	September
2512	bottled_water	September
2512	domestic_eggs	April
2512	root_vegetables	January
2512	fruit/vegetable_juice	September
2512	newspapers	September
2512	bottled_water	April
2512	ham	April
2512	domestic_eggs	January
2512	canned_beer	January
2512	canned_beer	April
2512	pastry	September

Table 2. Input data.

3.3 Patterns

In the following we show a part of the output produced for the subset of the *Groceries* database by first displaying a part of the Hasse diagram and then a set of implications and association rules.

Figure 2 shows a portion of the Hasse diagram of triadic concepts where the node in the green box represents the extent of two triadic concepts as a set of two customers who share two features found inside the corresponding blue box. The first feature tells us that Customers 2512 and 4320 bought the item *canned beer* on *April* and *January* while the second feature indicates that they both purchased *butter* and *canned beer* on *January*. Since the set of F-generators is equal to the set of features associated with the extent $\{2512, 4320\}$, no implication can be generated from the node labeled with this extent. If we look at the upper covers of this current node, we observe three groups of customers: those who bought *canned beer* in *January* (winter season), those who purchased *canned beer* in *April* (spring) and those who bought *butter* in *January*.

Statistics	Groceries dataset
Nb. of links	11124
Nb. of distinct extents	3128
Nb. of triadic concepts	3912
Nb. of minimal F-generators	5124
Nb. of implications (sup >0)	2164
Nb. of association rules (conf. <1)	7290
Execution time in seconds per step	
T-iPred	2.524
F-generator computation	95.721
Implication computation	0.053
Association rule computation (conf. <1)	16.281
Total time	114.579

Table 3. Statistics and execution times in seconds for the sample ($1000 \times 30 \times 12$) of the *Groceries* dataset.

4 Conclusion and Further Development

In this paper we illustrated the application of algorithms for Triadic Concept Analysis to a subset of the *Groceries* database as a triadic context to discover concepts and association rules, including implications. The merits of TCA lie in the fact that patterns are in a compact form and under many perspectives in the sense that for a given extent (set of objects) of a triadic concept, we obtain different views expressed by distinct features, and for a given generator and compatible features, we may get more than one association rule or implication.

Our current work is to complete the development of our software solution to make it an open source for everyone and continue our investigation of new kinds of association rules [10] on other datasets.

Finally, in order to help researchers validate their present and future contributions in TCA and more generally in Polyadic Concept Analysis [4, 6, 11], the FCA research community members need to join their forces and share their best experiences in collecting, exchanging and using clean and coherent multidimensional datasets. Software tools to generate synthetic data of different sizes and density levels are also welcome to evaluate the performance and scalability of the designed algorithms.

Acknowledgment

This study was financed in part by the NSERC (Natural Sciences and Engineering Research Council of Canada) and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001.

References

1. Groceries dataset for market basket analysis. <https://www.kaggle.com/heeraldedhia/groceries-dataset>
2. Ananias, K.H., Missaoui, R., B. Ruas, P.H., Zarate, L.E., Song, M.A.: Triadic concept approximation. *Information Sciences* (2021)
3. Baixeries, J., Szathmary, L., Valtchev, P., Godin, R.: Yet a Faster Algorithm for Building the Hasse Diagram of a Concept Lattice. In: *ICFCA'09*. pp. 162–177 (2009)
4. Bazin, A.: On implication bases in n-lattices. *Discrete Applied Mathematics* 273, 21–29 (2020), *advances in Formal Concept Analysis: Traces of CLA 2016*
5. Biedermann, K.: How triadic diagrams represent conceptual structures. In: *ICCS*. pp. 304–317 (1997)
6. Cerf, L., Besson, J., Robardet, C., Boulicaut, J.: Data peeler: Constraint-based closed pattern mining in n-ary relations. In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*. pp. 37–48. SIAM (2008)
7. Felde, M., Stumme, G.: Triadic exploration and exploration with multiple experts. *CoRR* abs/2102.02654 (2021)
8. Ganter, B., Obiedkov, S.: Implications in triadic formal contexts. In: Wolff, K.E., Pfeiffer, H.D., Delugach, H.S. (eds.) *Conceptual Structures at Work*. pp. 186–195. Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
9. Lehmann, F., Wille, R.: A triadic approach to formal concept analysis. In: *ICCS*. pp. 32–43 (1995)
10. Missaoui, R., Ruas, P.H., Kwuida, L., Song, M.A.: Pattern discovery in triadic contexts. In: *ICCS*. pp. 117–131. Springer (2020)
11. Voutsadakis, G.: Polyadic concept analysis. *Order* 19(3), 295–304 (2002)

Leveraging Formal Concept Analysis to Improve n -fold validation in Multilabel Classification

Francisco J. Valverde-Albacete¹ * and Carmen Peláez-Moreno¹

Depto. Teoría de Señal y Comunicaciones, Univ. Carlos III de Madrid, Madrid, Spain,
fva@tsc.uc3m.es carmen@tsc.uc3m.es

Abstract. In this application note we revisit previous work on the exploratory analysis of Multilabel Classification (MLC) tasks in Machine Learning. We combine Information Theory and Formal Concept Analysis (FCA) to formalize the intuition that inference of classifiers in MLC tasks can only proceed when the training and testing data concerning the labels define the same Concept Lattice. We instantiate our procedure on the `emotions` dataset, but the procedure is independent of the dataset being explored. An R language interactive notebook carrying out the procedure is available upon request.

Keywords: Multilabel Classification · Entropy-based assessment · Data preprocessing · n -fold validation · Formal Context Analysis.

1 Introduction and Motivation

1.1 The Multilabel Classification Task

MLC is a relatively recently-formalized task in Machine Learning [1] with applications in *text categorization*, *gene expression analysis*, etc. A recent tutorial explains the progress in methods and concerns [2], while a more up-to-date exposition with special emphasis on software tools is [3]. We describe here a variant of the MLC task setting to accommodate feature transformations, as depicted in Fig. 1.




Fig. 1: Basic scheme for multi-label classification

The following way of solving the problem is based on the theory of Statistical Learning [4]: consider a binary multivariate source \bar{Y} , emitting binary label vectors or *labelsets* from a label space $\mathcal{Y} \equiv 2^{m_Y}$, by virtue of the isomorphism between sets and their characteristic vectors¹. Suppose that the labelsets are *hidden* and we can only access the result of an *observation mechanism* of the labelsets in terms of *visible instance*

* Corresponding author.

¹ We assign to each of the labels a certain “meaning” but this is out of this mathematical model. RealDataFCA'2021: Analyzing Real Data with Formal Concept Analysis, June 29, 2021, Strasbourg, France

 Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

or *observation* in a feature space $\mathcal{X} \equiv \mathbb{R}^{m_X}$. The *multi-label classification problem* is to tag any (feature) vector $\vec{x} \in \mathcal{X}$, with a labelset $\vec{y} \in \mathcal{Y}$ [3].

The usual way to solve the problem in the simpler, *supervised setting* is:

1. Model the *source* of labelsets as random label vectors $\bar{Y} \sim P_{\bar{Y}}$ and their *observations* as the feature vectors $\bar{X} \sim P_{\bar{X}}$ over their respective spaces.
2. Collect a (*task*) *dataset* with s samples, $\mathcal{D} = \{(\vec{x}^i, \vec{y}^i)\}_{i=1}^s$ of labelsets and their observed features to induce a classifier from observations to labelsets $h: \mathcal{X} \rightarrow \mathcal{Y}$.
3. Choose the *classifier type* and induction scheme for it.
4. In order to assess the classifiers, choose an adequate measure of performance, and implement (any of a number of) schemes of iterated *re-sampling* of the data into a set of *training examples* $\mathcal{D}_T = \{(\vec{x}^j, \vec{y}^j)\}_{j=1}^{s_T}$ and a set of *test examples* $\mathcal{D}_E = \{(\vec{x}^k, \vec{y}^k)\}_{k=1}^{s_E}$ so that the training data are used to induce the classifiers and the testing data to test them, and *validate* these results, e.g. using *k-fold validation* [4].

Example 1 (emotions[5]). A traditional dataset to test new ideas in MLC, `emotions` captures $m_X = 72$ features and the emotions they produce—as $m_Y = 6$ labels like `amazed-surprised`, `angry-aggressive`, etc.—for a total of $s = 593$ song tracks. As expected, distinct labels have distinct expressions in samples.

Note that only 27 of the $2^6 = 64$ possible labelsets occur, and as many as 4 of them are *hapaxes*—they occur only once—while at least one of the labelsets appears 81 times. This wildly imbalanced behaviour is typical of MLC datasets [6]. \square

1.2 An FCA-based Interpretation of MLC

Note that in the standard approach above FCA is all but absent. However, it is easy to see that FCA and its variants are relevant in modelling this ML task.

First, with the labelsets $\{\vec{y}^i\}_{i=1}^s$ of the task dataset we can build a formal context using the set of labels L as attributes, with $|L| = m_Y$, and taking for each sample—considered as a formal object $i \in G$ with $|G| = s$ —its labelset as a row of the incidence matrix $I_i = \vec{y}^i$, whence $\mathbb{D}_L = (G, L, I)$ is the *binary formal context of samples and their labels*. This context captures the information in the stochastic source \bar{Y} .

Secondly, for observation vectors $\{\vec{x}^i\}_{i=1}^s$ their context $\mathbb{D}_F = (G, F, R)$ —with F the set of features—will not be binary, in general, but many-valued $R_i = \vec{x}^i$. This context captures the information in the observations \bar{X} ².

With the previous modelling, relevant notions in MLC correspond to relevant notions in FCA. For instance, *labelsets are object intents*, and they can be found through the polars $\vec{y}_i = \{i\}^\uparrow$. In this way, **we can reason about the sampling of the stochastic variables \bar{Y} and \bar{X} —the dataset—in terms of the contexts above, and vice-versa.**

For instance, each labelset present in the dataset—that is, each object-extent—may appear with different samples (formal objects). Recall that the concept-forming function $\bar{\gamma}$ induces a partition $\ker \bar{\gamma}$ on G by equality of labelsets: $(i_1, i_2) \in \ker \bar{\gamma} \iff \{i_1\}^\uparrow = \{i_2\}^\uparrow = \vec{y}_{i_1}$. *This partition is crucial for defining the entropy of the label source* (see § 2). In the other direction, we expect the sampling to be good enough $m_Y \ll s$ so it is

² Note that Fig. 1 suggests another context formed by the *transformed observations* captured by the random vector \bar{Z} , but this will not be dealt with here.

safe to suppose that no two labels are predicated of the same set of objects. Therefore we expect the partition on labels induced by $\bar{\mu}$ to be the identity $\ker \bar{\mu} = \iota_L$, which holds on standard testing datasets (see e.g. those in [3]).

1.3 The Problem and the FCA-induced Solution

The following problem and its theoretical solution were proposed in [7] as an example of the above-suggested operation: The MLC classifier induction and assessment procedures demand that we generate train and test resamplings of the original data [8, 4]. This amounts to splitting the original context \mathbb{D}_L into two subposed subcontexts of training \mathbb{D}_T and testing \mathbb{D}_E data: $\mathbb{D} = \mathbb{D}_T/\mathbb{D}_E$. Note that

1. Since the samples are supposed to be independent and identically distributed, the order of these contexts in the subposition—indeed the row order of I —is irrelevant.
2. The resampling of the labelset context is tied to that of the observations: we decide on the labelset context and this carries over to the observation context.

Since the data are a formal context we know that an important part of the information contained in it comes from the concept lattice, hence:

Hypothesis 1. *1. (FCA intuition) A necessary condition for the resampling of the data \mathcal{D} into training part \mathcal{D}_T and testing part \mathcal{D}_E to be meaningful for the MLC task, is that the concept lattices of all of these be isomorphic:*

$$\mathfrak{B}(\mathbb{D}) \cong \mathfrak{B}(\mathbb{D}_T) \cong \mathfrak{B}(\mathbb{D}_E)$$

2. *(ML intuition) Not only the labelsets but also their occurrence frequencies as quantified by the cardinalities of the blocks of $\ker \gamma$ are important.*

The last consideration follows because if we only retained the meet- and join-irreducibles to obtain these concept lattices, then the labelsets of reducible attributes would be lost so the relative importance of the samples—both labels and observations—would change, impacting the induction scheme of the classifiers.

The above proposition suggests that the analogue of *stratified sampling* in MLC is a procedure in which the stratification must proceed on a block-by-block basis with respect to $\ker \bar{\gamma}$. However this comes at a price, when there are hapaxes in the data. If we choose, for instance, to maintain 80% of the data for training and 20% for testing, regardless of these proportions, stratified sampling will force us to include all hapaxes with the following deleterious consequences:

- The relative frequency of the hapaxes will be distorted wrt to other labelsets.
- We will be using some data (the hapaxes) both for training and testing, which is known to obtain too optimistic performance results in whichever measure of it.

Furthermore, if we use, e.g. k -fold validation we have to repeat this procedure and ensure that the resamplings are somehow different. A usual procedure is to distribute the original dataset into k blocks in order to aggregate $k - 1$ of them into the training dataset \mathbb{D}_T and use the leftover as the testing dataset \mathbb{D}_E . It is common to use $k = 5$ or $k = 10$. This can only compound the previous problem, therefore *FCA allows us to spot possible problems with the classifier induction and validation schemes using resampling.*

In this paper we will use side-by-side FCA-based and Information Theory-based devices to support the claim that they improve the understanding of the task when used in synergy. For that purpose we first review an Exploratory Data Analysis (EDA) technique in Sec. 2 to help in interpreting the experimental results in Sec. 3.

2 Methods: Source Multivariate Entropy Triangles

The Source Multivariate Entropy Triangle (SMET, [9]) is a visual tool based on Information Theory to explore the informational content of multivariate sources. It is better suited to discrete sources—be they binary or multiclass—and optimally suited for analyzing the set of label assignments of MLC tasks.

Due to space limitations, we will only describe the graphic approach to the SMET. We refer the reader to [9] for its theoretical bases. In brief, the information content of a discrete multivariate source can be depicted in the simplex of Fig. 2, where the vertices describe extreme behaviours of the labels. If a label appears in the lowest vertex, is not stochastic. If it appears in the right vertex, its information can be predicted from that of the rest of the labels, whereas if it appears in the left vertex, it has a lot of private information. The opposite sides to these vertices represent their opposite behaviours. In general, the behaviour of labels is a mix of these three types (see Fig. 3).

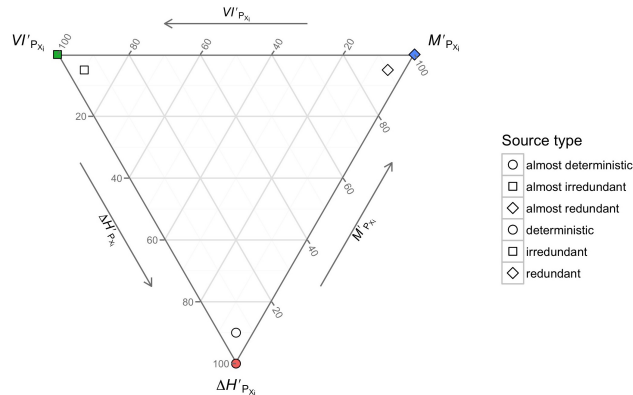


Fig. 2: General Source Multivariate Entropy Triangle from [9]

To assess the effectiveness of stratified sampling schemes for MLC, this paper proposes to use the visualization of the information balance of a formal context of labels \mathbb{D}_L using the SMET.

3 Results: Example Data Analysis for an MLC Dataset

SW resources. The following analysis is carried out on the Example 1 emotions dataset [5], as pre-processed and presented by the *mldr* R package [6]. The interactive R notebook embodying the analysis is available from the authors upon request.

Basic EDA of the labels. Since we are only considering the set of labels \bar{Y} , we extracted the histogram of the labelsets $\{\vec{y}_j, n_j\}_{j \in J}$ from the dataset and considered a set of minimal frequencies of occurrence $n_T \in \{0, 1, 4, 9, 16, 25\}$ acting as thresholds based on it. The case $n_T = 0$ actually represents the original dataset and Fig. 3 shows the information balance of the six labels of `emotions` as well as the average balance for them all. We see that most labels are rather random, with `relaxing-calm` completely so. No label is completely specified by the rest of them, nor is any totally independent. This in essence means that the dataset is truly multilabel.

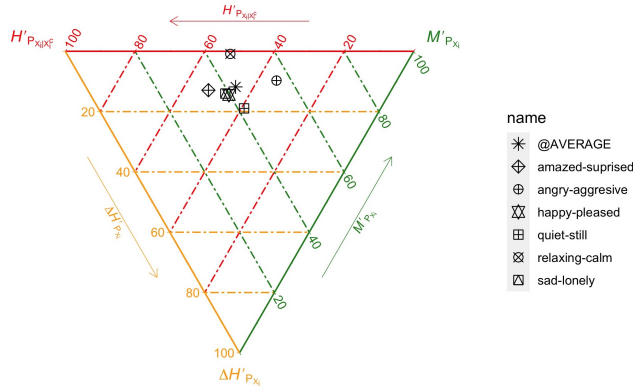


Fig. 3: SMET for the original `emotions` dataset, e.g. with $n_T = 0$

Disposing of hapaxes to improve stratified sampling. Previous analyses of the histogram of labelsets made us realize that this dataset is not adequate for resampling due to hapaxes and in general low-counts of many labelsets [7]. This applies to most MLC datasets used at present [3].

To dispose of hapaxes without disposing of samples we must re-assign each to a more frequent labelset. The rationale for this decision is because we consider hapaxes errors in label codification, and assume that the “real” labelset is the closest non-hapax in Hamming distance³. However, this re-assignment changes the histogram of labelsets what results in an impoverishment of the information balance of the labels and the dataset in general.

To explore this trade-off, at each threshold n_T , a labelset \vec{y} was considered a *generalized hapax* if $n_{\vec{y}} < n_T$. For each threshold n_T we calculated the Hamming distance between each generalized hapax \vec{y}_{n_T} and the non-hapaxes, and found the set of those closest to it. Then we re-assigned \vec{y}_{n_T} to one of them uniformly at random (allowing for repetitions)⁴.

³ Recall that the Hamming distance between two sequences of bits of identical length is the number of positions in which they differ.

⁴ Note that an alternative strategy would have been a scheme considering the original frequencies in the histogram, to simulate a rich-get-richer phenomenon. But such a procedure would decrease the source entropy more than the one we have chosen.

This reassignment defined a new dataset whose information balance was represented by the SMET, as suggested in Section 2, whence Fig. 4 ensued. What we can

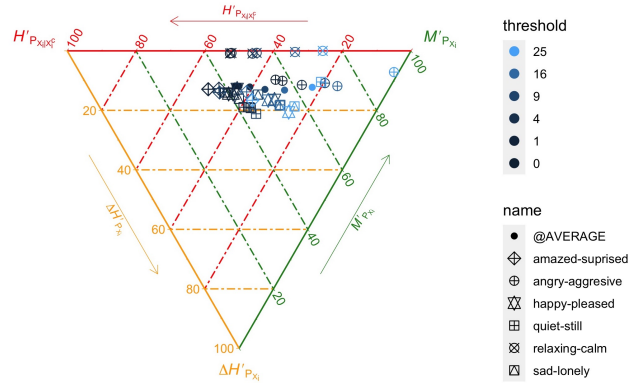


Fig. 4: SMET for emotions in several thresholds.

see is a general tendency to the increment of the total correlation as the thresholds increase. But this entails that the individual distinctiveness of each label is diminished. See, for instance, the case for *angry-aggressive* that can actually be predicted from the other labels when $n = 25$, confirming that too aggressive a threshold will substantively change the relative information content of the labels in the dataset.

Choosing the adequate threshold. Note that a threshold of n is needed to request an $(n + 1)$ -fold cross-validation of any magnitude about the dataset, since all labelset will have at least $(n + 1)$ representatives for the stratified sampling requested by the cross validation procedure. Is it possible to balance the identical sampling property on train and test, yet avoid too much loss of information content?

Figure 5 depicts a choice of thresholds typically used in validation—1, 4 and 9, corresponding to 2-, 5- and 10-fold validation—for three differently behaving labels—*angry-aggressive*, *quiet-still* and *relaxing-calm*—and the average of the dataset, both for the ensembles of training and testing folds.

- As applied to the estimation of the entropies, the $(n + 1)$ -fold validation yields the same result in train and test, the sought-for result.
- We can see the general drift towards increased correlation in all labels, but much more in, say, *angry-aggressive* than in *quiet-still*.
- For this particular dataset, a threshold of $n_T = 4$ with 5-fold validation seems to be a good compromise for attaining statistical validity vs. dataset fidelity.

FCA confirmation. To strengthen the validity of the last two conclusions, we calculated the number of concepts of all of the train and test label contexts using the `fcaR` package [10]. This is possible since such datasets are just binary tables. After creating the contexts, we clarified and obtained the lists of concepts, then we compared the cardinality of the training and test concept lattices both for the unsplit dataset—after reassigning the generalized hapaxes, when needed—and the $(n + 1)$ -cross validated versions. The results are shown in Fig. 6.

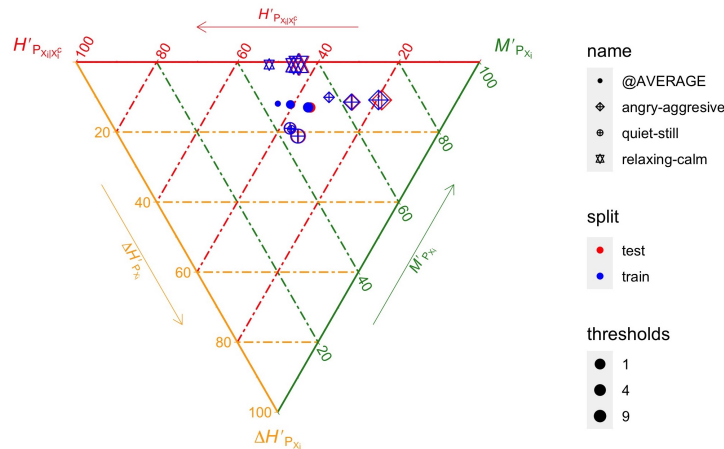


Fig. 5: SMET for emotions with cross-validated entropies (some labels not shown for clarity).

As expected, for $n_T = 0$ the difference in number of concepts between the non-sampled and sampled versions of the dataset make it non-adequate for appropriate sampling⁵. The training and test splits had the same number of concepts for every other threshold. For $n_T \in \{1, 4, 16\}$, the number of concepts was constant among folds, but due to the randomness inherent in sampling for $n_T \in \{9, 25\}$ one of the folds was different (not shown explicitly).

4 Summary and Discussion

We have tested and strengthened a data hypothesis put forward in previous work [7]. This hypothesis demands that the training and testing splits for solving a MLC task have the same concept lattice associated with their multilabel contexts, and also similar entropies in the induced partition of the sample set.

We first explored the influence of carrying out different degrees of smoothing of the labelset frequencies and found a balance between it and the increase in total correlation of the labels, leading to a loss of their individual information content.

We also checked that the recipe for defining an appropriate threshold n to carry out an $(n+1)$ -fold cross validation actually works by comparing the informational balances of training and testing splits of the dataset as random estimates of their actual values.

Further work should start applying this rationale to solving of the MLC task. Preliminary work on using both FCA- and information theoretic-approaches suggest that present-day solutions are far from satisfactory in this respect. In the future, we plan to use more FCA-induced techniques from [7] to improve on this.

⁵ It is a fluke of the dataset that both the training and test subcontexts have the same number of concepts as some of the hapaxes are singletons

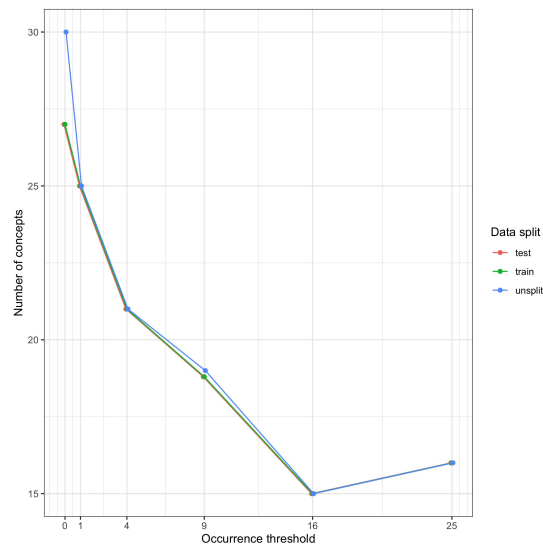


Fig. 6: Number of concepts vs. threshold for different n and splits of the dataset

References

- [1] Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. *Pattern Recognition* **37** (2004) 1757–1771
- [2] Gibaja, E., Ventura, S.: A Tutorial on Multilabel Learning. *ACM Computing Surveys* **47** (2015) 1–38
- [3] Herrera, F., Charte, F., Rivera, A.J., del Jesus, M.J.: *Multilabel Classification. Problem Analysis, Metrics and Techniques*. Springer (2016)
- [4] Murphy, K.P.: *Machine Learning. A Probabilistic Perspective*. MIT Press (2012)
- [5] Wiczorkowska, A., Synak, P., Raś, Z.W.: Multi-label classification of emotions in music. In Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds.: *Intelligent Information Processing and Web Mining*, Springer Berlin Heidelberg (2006) 307–315
- [6] Charte, F., Charte, F.D.: Working with multilabel datasets in R: The mldr package. *The R journal* **7** (2015) 149–162
- [7] Valverde Albacete, F.J., Pelaez-Moreno, C., Cabrera, I.P., Cordero, P., Ojeda-Aciego, M.: Exploratory Data Analysis of Multi-Label Classification Tasks with Formal Context Analysis. In Trnečka, M., Valverde Albacete, F.J., eds.: *Concept Lattices and their Applications CLA*, Tallinn University of Technology (2020) 171–183
- [8] Sechidis, K., Tsoumakas, G., Vlahavas, I.P.: On the Stratification of Multi-label Data. In: *Lecture Notes in Computer Science*. Volume 6913., Berlin, Heidelberg, Springer Berlin Heidelberg (2011) 145–158
- [9] Valverde-Albacete, F.J., Peláez-Moreno, C.: The evaluation of data sources using multivariate entropy tools. *Expert Systems with Applications* **78** (2017) 145–157
- [10] Lopez Rodriguez, D., Mora, A., Dominguez, J., Villalon, A.: *fcaR: Formal Concept Analysis*. (2020) R package version 1.0.7.