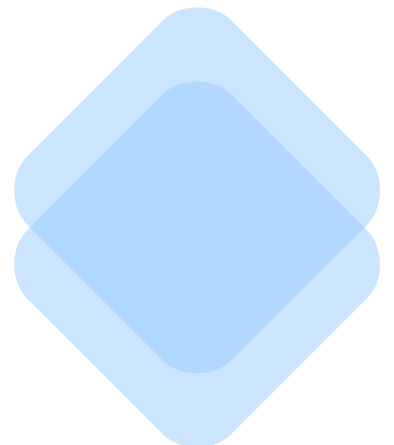


# Contributions to the 11<sup>th</sup> International Conference on Formal Concept Analysis

Dresden, Germany, May 21–24, 2013



Peggy Cellier  
Felix Distel  
Bernhard Ganter (Eds.)



## Preface

Formal concept analysis (FCA) is a mathematical formalism based on order and lattice theory for data analysis. It has found applications in a broad range of neighboring fields including Semantic Web, data mining, knowledge representation, data visualization and software engineering.

ICFCA is a series of annual international conferences that started in 2003 in Darmstadt and has been held in several continents: Europe, Australia, America and Africa. ICFCA has evolved to be the main forum for researchers working on theoretical or applied aspects of formal concept analysis worldwide.

In 2013 the conference returned to Dresden where it was previously held in 2006. This year the selection of contributions was especially competitive. This volume is one of two volumes containing the papers presented at ICFCA 2013. The other volume is published by Springer Verlag as LNAI 7880 in its LNCS series.

In addition to the regular contributions, we have included an extended abstract: Jean-Paul Doignon reviews recent results connecting formal concept analysis and knowledge space theory in his contribution “Identifiability in Knowledge Space Theory: a Survey of Recent Results”.

The high-quality of the program of the conference was ensured by the much-appreciated work of the authors, the Program Committee members, and the Editorial Board members. Finally, we wish to thank the local organization team. They provided support to make ICFCA 2013 proceed smoothly in a pleasant atmosphere.

May 2013

Peggy Cellier  
Felix Distel  
Bernhard Ganter

## Organization

### Executive Committee

### Conference Chair

Bernhard Ganter                      Technische Universität Dresden, Germany

### Local Organizers

Cynthia Vera Glodeanu              Technische Universität Dresden, Germany  
Christian Pech                        Technische Universität Dresden, Germany

### Program and Conference Proceedings

### Program Chairs

Peggy Cellier                         IRISA, INSA Rennes, France  
Felix Distel                          Technische Universität Dresden, Germany

### Editorial Board

Peter Eklund                         University of Wollongong, Australia  
Sebastien Ferré                      Université de Rennes 1, France  
Robert Godin                        Université du Québec à Montréal (UQAM), Canada  
Robert Jäschke                      Universität Kassel, Germany  
Sergei Kuznetsov                    Higher School of Economics, Moscow, Russia  
Leonard Kwuida                    Bern University of Applied Sciences, Switzerland  
Raoul Medina                        LIMOS, Université Clermont-Ferrand 2, France  
Rokia Missaoui                      Université du Québec en Outaouais (UQO), Canada  
Sergei Obiedkov                    Higher School of Economics, Moscow, Russia  
Uta Priss                              Ostfalia University of Applied Sciences, Wolfen-  
büttel, Germany  
  
Sebastian Rudolph                  AIFB, University of Karlsruhe, Germany  
Stefan E. Schmidt                  Technische Universität Dresden, Germany  
Barış Sertkaya                      SAP Research Center, Dresden, Germany  
Gerd Stumme                        University of Kassel, Germany  
Petko Valtchev                      Université du Québec à Montréal, Canada  
Rudolf Wille (Honorary)          Technische Universität Darmstadt, Germany  
Karl Erich Wolff                    University of Applied Sciences, Darmstadt, Ger-  
many

### **Last Year's Chairs**

Florent Domenach	University of Nicosia, Cyprus
Dmitry Ignatov	Higher School of Economics, Moscow, Russia
Jonas Poelmans	Katholieke Universiteit Leuven, Belgium

### **Program Committee**

Simon Andrews	University of Sheffield, United Kingdom
Mike Bain	University of New South Wales, Sydney, Australia
Jaume Baixeries	Polytechnical University of Catalonia, Spain
Radim Bělohávek	Palacky University, Olomouc, Czech Republic
Sadok Ben Yahia	Faculty of Sciences of Tunis, Tunisia
Karell Bertet	L3I, Université de La Rochelle, France
Claudio Carpineto	Fondazione Ugo Bordon, Italy
Stephan Doerfel	KDE Group, University of Kassel, Germany
Vincent Duquenne	ECP6-CNRS, Université Paris 6, France
Alain Gély	Université Paul Verlaine, Metz, France
Marianne Huchard	LIRMM, Université Montpellier, France
Tim B. Kaiser	SAP AG, Germany
Mehdi Kaytoue	INSA Lyon, France
Derrick G. Kourie	University of Pretoria, South Africa
Markus Krötzsch	University of Oxford, United Kingdom
Marzena Kryszkiewicz	Warsaw University of Technology, Poland
Wilfried Lex	Universität Clausthal, Germany
Lotfi Lakhal	Aix-Marseille Université, France
Engelbert Mephu Nguifo	LIMOS, Université de Clermont Ferrand 2, France
Amedeo Napoli	LORIA, Nancy, France
Lhouari Nourine	LIMOS, Université de Clermont Ferrand 2, France
Jan Outrata	Palacky University of Olomouc, Czech Republic
Jean-Marc Petit	LIRIS, INSA Lyon, France
Sandor Radeleczki	University of Miskolc, Hungary
Camille Roth	CNRS/EHESS, Paris, France
Andreja Tepavčević	University of Novi Sad, Serbia
Laszlo Szathmary	University of Debrecen, Hungary

### **External Reviewers**

Daniel Borchmann	Technische Universität Dresden, Germany
Alain Casali	LIF, IUT d'Aix-en-Provence, France
Xavier Dolques	LHYGES, Strasbourg, France
Lankun Guo	Université du Québec, Montréal, Canada
Viet Phanluong	LIF, IUT d'Aix-en-Provence, France

## Table of Contents

### Extended Abstract

Identifiability in Knowledge Space Theory: a survey of recent results . . . . .	1
<i>Jean-Paul Doignon</i>	

### Regular Contributions

Heterogeneous environment on examples . . . . .	5
<i>Lubomír Antoni, Stanislav Krajčí, Ondrej Krídlo and Lenka Pisková</i>	
Attribute Exploration on the Web . . . . .	19
<i>Robert Jäschke and Sebastian Rudolph</i>	
The Detection of Outlying Fire Service’s Reports. The FCA Driven Analytics. . . . .	35
<i>Adam Krasuski and Piotr Wasilewski</i>	
An Approach to Incremental Learning Based on Good Classification Tests . . . . .	51
<i>Xenia Naidenova and Vladimir Parkhomenko</i>	
FCART: A New FCA-based System for Data Analysis and Knowledge Discovery . . . . .	65
<i>Alexey A. Neznanov, Dmitry A. Ilvovsky and Sergei O. Kuznetsov</i>	

# Identifiability in Knowledge Space Theory: a survey of recent results

Jean-Paul Doignon

Université Libre de Bruxelles,  
Bd du Triomphe, c.p. 216,  
B-1050 Bruxelles.  
Belgium.  
`doignon@ulb.ac.be`

**Abstract.** Knowledge Space Theory (KST) links in several ways to Formal Concept Analysis (FCA). Recently, the probabilistic and statistical aspects of KST have been further developed by several authors. We review part of the recent results, and describe some of the open problems. The question of whether the outcomes can be useful in FCA remains to be investigated.

**Keywords:** knowledge space, Basic Local Independence Model, Correct Response Model, model identifiability

In Knowledge Space Theory (KST, see Doignon & Falmagne, 1999; Falmagne & Doignon, 2011), a body of knowledge is represented by a finite set, say  $Q$ , of test items. The knowledge state of a student is identified with the collection of items he masters. Because of dependencies among the items, not any subset of  $Q$  can be a knowledge state; for instance, if  $Q$  is structured by a prerequisite relation, the states should be taken as the ideals of the transitive closure of the prerequisite relation. In general, the collection  $\mathcal{K}$  of all possible *knowledge states* forms a *knowledge structure*  $(Q, \mathcal{K})$ ; it is assumed  $\emptyset, Q \in \mathcal{K}$ . The correctness of the answer provided at a certain time by a student to any item is granted to depend only on his knowledge state, except for careless errors and lucky guesses.

Because variations are routinely observed in such answers, a probabilistic extension of KST was designed. So, assume the knowledge state of a student may vary (around a certain time point of his apprenticeship) in  $\mathcal{K}$  according to a probability distribution  $\pi$  on  $\mathcal{K}$ . Moreover, for any item  $q$  in  $Q$ , let  $\beta_q$  be the probability of a careless error in answering  $q$ , and  $\eta_q$  be the probability of a lucky guess in answering  $q$ . All the numbers  $\pi(K)$  (for  $K$  in  $\mathcal{K}$ ),  $\beta_q$  and  $\eta_q$  (for  $q$  in  $Q$ ) will be considered as parameters with (unknown) latent values. (Of course, the  $\pi(K)$ 's are not independent parameters, because they add up to 1.) The *straight case* obtains when  $\beta_q = \eta_q = 0$ , for any  $q$  in  $Q$ . We now propose two models for the probabilities of correctness of student answers (considered as the observables). Both models are based on the latent knowledge structure together with the various parameters we have just introduced. The second model

is described in Doignon & Falmagne (1999) (see also Falmagne & Doignon, 2011), while the first one is only implicit there.

The first model, the Correct Response Model (CRM), defines the probability  $\tau(q)$  of a correct answer to any isolated item  $q$ . It first conditions the probability of a correct answer to item  $q$  on the state of the student:

$$\tau(q) = \sum_{K \in \mathcal{K}} Pr(q|K) \cdot \pi(K).$$

Then, it specifies each conditional probability  $Pr(q|K)$  by taking into account the careless error probabilities  $\beta_q$  and the lucky guess probabilities  $\eta_q$ :

$$Pr(q|K) = \begin{cases} 1 - \beta_q & \text{if } q \in K, \\ \eta_q & \text{if } q \notin K. \end{cases} \quad (1)$$

A second model, the Basic Local Independence Model (BLIM), defines the probability of a pattern of responses. Here, a pattern is a subset of  $Q$  meant to contain all items to which a student (at a given time) produces a correct answer. Exactly as the CRM, the BLIM conditions the pattern probability on the state of the student. Thus, the probability of a given pattern  $R$  of responses (with  $R \subseteq Q$ ) equals

$$\rho(R) = \sum_{K \in \mathcal{K}} r(R, K) \cdot \pi(K),$$

where  $r(R, K)$  is specified as follows:

$$r(R, K) = \left( \prod_{q \in K \setminus R} \beta_q \right) \left( \prod_{q \in K \cap R} (1 - \beta_q) \right) \left( \prod_{q \in R \setminus K} \eta_q \right) \left( \prod_{q \in Q \setminus (R \cup K)} (1 - \eta_q) \right).$$

Both of our models, the CRM and the BLIM, are instances of probabilistic models. On the basis of a fixed knowledge structure, they predict from any parameter point (that is, any list of values for all parameters) some definite probability values for the observables (in our case, the observables are either individual correct responses, or whole patterns of correct responses). We use the term *predicted distribution* to designate “any distribution of probability values for the observables that are predicted by the model”. The questions we will consider are as follows (the first two are clearly stated in Bamber & van Santen, 2000 for probabilistic models in general).

1. Model testability: is there some distribution of probability values for the observables that the model does not predict?
2. Model identifiability: is each predicted distribution produced from at most one parameter point?
3. Model characterizability: are the predicted distributions susceptible of an effective characterization (without reference to the underlying parameter values)?

Recently Spoto, Stefanutti & Vidotto (2012) have investigated the first two questions for the BLIM, the model for pattern probabilities. Moreover, Stefanutti, Heller, Anselmi & Robusto (2012) have produced additional, nice results about identifiability of the BLIM, especially in its local version: *local identifiability* means identifiability when the model is restricted to some neighborhood of any given parameter point.

On our part, we consider the three types of questions for the CRM, the correct response model, however working mainly in the straight case. First, we are able to characterize testability of the model using a simple criterion (and also to reformulate a variant of it, numerical testability, in a manageable, technical way). Second, about characterizability, we point out unavoidable difficulties in recognizing when it holds. Third, as regards identifiability, we give a tractable equivalent, concluding that identifiability is not often met. On the positive side, we indicate how to modify the parameter domain (consisting of the knowledge state probabilities) in order to restore identifiability while keeping the same prediction range; nevertheless, we show that the construction works well only for the knowledge structures  $(Q, \mathcal{K})$  which are derived from a quasi order on  $Q$  (as it is the case in the presence of a prerequisite relation). As a matter of fact, the construction heavily relies on a theorem of Stanley (1986) for a convex polytope he associates to a partial order.

The results presented during the talk are taken from a manuscript under preparation (Doignon, 2013).



## Bibliography

- Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *J. Math. Psych.*, *44*, 20–40.
- Doignon, J.-P. (2013). A correct response model in knowledge space theory. Manuscript in preparation.
- Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag.
- Falmagne, J.-C., & Doignon, J.-P. (2011). *Learning Spaces*. Berlin: Springer-Verlag.
- Spoto, A., Stefanutti, L., & Vidotto, G. (2012). Considerations about the identification of forward- and backward-graded knowledge structures. Submitted.
- Stanley, R. (1986). Two poset polytopes. *Discrete Comput. Geom.*, *1*, 9–23.
- Stefanutti, L., Heller, J., Anselmi, P., & Robusto, E. (2012). Assessing the local identifiability of probabilistic knowledge structures. *Behavior Research Methods*, *44*, 1197–1211.

# Heterogeneous environment on examples<sup>\*</sup>

Lubomír Antoni, Stanislav Krajčí<sup>\*\*</sup>, Ondrej Krídlo and Lenka Pisková

Institute of Computer Science, University of Pavol Jozef Šafárik, Košice, Slovakia  
lubomir.antoni@student.upjs.sk, stanislav.krajci@upjs.sk,  
ondrej.kridlo@upjs.sk, lenka.piskova@student.upjs.sk

**Abstract.** We propose a running example for heterogeneous approach based on new type of fuzzification that diversifies fuzziness of every object, fuzziness of every attribute and fuzziness of every table value in a formal context. Moreover we suggest another working examples on heterogeneous environment and provide additional utilization and illustration of this new model that allows to use Formal Concept Analysis also for heterogenous data. An interpretation of heterogeneous formal concepts and the resulting concept lattice is included.

**Keywords:** heterogeneous context, longterm preferences, shortterm preferences

## 1 Introduction

Formal concepts consisting of developing countries in supranational groups is one of the earliest example in which has been applied classical Boolean approach of Formal Concept Analysis and appears in [14]. By attribute fuzzification proposed independently by Ben Yahia [11], Bělohávek [3] and Krajčí [16] is possible to think about students and their evaluation in more than two degrees. Such method to process data tables is called one-sided fuzzy approach. Another Krajčí's generalized approach [18], [19] diversifies fuzziness of objects and fuzziness of attribute. Medina, Ojeda-Aciego and Ruiz-Calviño [22] utilize personal preferences to choose suitable journal for a paper submitting and use different adjoint triples to find the best object.

An additional level of generalization based on diversification of every object, every attribute and every table value is proposed in [1]. In this paper we would like to clarify that it has some natural motivation to consider such level of generalization. Also an interpretation of both concept-forming operators and the notions of longterm and shortterm preferences are included in addition to

---

<sup>\*</sup> This work was partially supported by the grant VEGA 1/0832/12, by the Slovak Research and Development Agency under contract APVV-0035-10 "Algorithms, Automata, and Discrete Data Structures" by the Agency of the Slovak Ministry of Education for the Structural Funds of the EU, under project ITMS:26220120007.

<sup>\*\*</sup> Corresponding author at: Institute of Computer Science, Faculty of Science, Pavol Jozef Šafárik University in Košice, Jesenná 5, 040 01 Košice, Slovakia, stanislav.krajci@upjs.sk.

environment introduced in [2]. The cottage example is introduced in more detail here and computed on two larger contexts in compare to [1]. It gives better intuition for the underlying structures. Similarly other applications of heterogeneous Formal Concept Analysis are discussed.

The structure of this paper is as follows. Section 2 recalls the basic notion of heterogeneous approach and details heterogeneous formal context on proposed running cottage example. Section 3 gives appropriate interpretation of heterogeneous concepts. Section 4 describes overview of another working examples of our environment that works also with heterogenous data. Section 5 briefly explains how to construct a heterogeneous formal concept lattice and gives the result for our proposed cottage example. Section 6 concludes the paper and describes future work.

## 2 Heterogeneous formal context

Consider the following situation as a motivation. People (friends, colleagues, classmates) are going to stay at some cottage. In fact, every person can have different requirements and preferences connected with cottage conditions depending on number of days spending at the cottage. One can prefer hot water, other necessarily expect internet connection. Natural requirements based on some actual preferences can be formulated:

- ◇ Eva admits full discomfort on water conditions, partial discomfort on internet/TV and full discomfort on a lake available.
- ◇ Joe accepts half discomfort on water conditions, admits great discomfort on internet/TV and no discomfort on a lake available.
- ◇ Ken allows full discomfort on water conditions, admits discomfort on services and half discomfort on a lake available.

Realize that every person feels full discomfort diverse in general. For instance, full discomfort on water conditions is for Eva connected with absence of hot water even though one arbitrary day at the cottage. Joe is more adaptive and full discomfort is connected with absence of hot water only at second day. Ken is the most adaptive and full discomfort corresponds to two days absence of hot water. So it is natural to inquire which cottage conditions have to be fulfilled to satisfy all people staying at cottage even though different number of days.

In follows we define heterogeneous formal context and formally describe mentioned situation. Let  $A$  and  $B$  be non-empty sets. Let  $\mathcal{P} = ((P_{a,b}, \leq_{P_{a,b}}) : a \in A, b \in B)$  be a system of posets and let  $R$  be a function from  $A \times B$  such that  $R(a,b) \in P_{a,b}$ , for all  $a \in A$  and  $b \in B$ . Let  $\mathcal{C} = ((C_a, \leq_{C_a}) : a \in A)$  and  $\mathcal{D} = ((D_b, \leq_{D_b}) : b \in B)$  be systems of complete lattices. (For simplicity, we will omit the indices of all noticed  $\leq$ , it will be always clear which of one is used.)

Let  $\odot = (\bullet_{a,b} : a \in A, b \in B)$  be a system of operations such that  $\bullet_{a,b}$  is from  $C_a \times D_b$  to  $P_{a,b}$  and it is isotone and left-continuous in both arguments, i. e.

- 1a)  $c_1 \leq c_2$  implies  $c_1 \bullet_{a,b} d \leq c_2 \bullet_{a,b} d$  for all  $c_1, c_2 \in C_a$  and  $d \in D_b$ ,
- 1b)  $d_1 \leq d_2$  implies  $c \bullet_{a,b} d_1 \leq c \bullet_{a,b} d_2$  for all  $c \in C_a$  and  $d_1, d_2 \in D_b$ ,



To go beyond objects, Eva's preferences contain staying in three degrees: not at all, one day (it does not matter which one) or all two days ( $D_{Eva}$ ); Joe in four degrees: not at all, only Saturday, only Sunday or all two days ( $D_{Joe}$ ); Ken again in three degrees ( $D_{Ken}$ ). Beyond attributes, water conditions contain two degrees: hot or cold (there are also cottages having cold water only in facilities) corresponding to  $C_{water}$ ; services include four degrees: internet and television, only internet connection, only television or nothing at all ( $C_{services}$ ); lake availability contains two degrees: yes or no ( $C_{lake}$ ). Finally in case of table values  $\mathcal{P} = (P_{water,Eva}, P_{services,Eva}, P_{lake,Eva}, P_{water,Joe}, P_{services,Joe}, P_{lake,Joe}, P_{water,Ken}, P_{services,Ken}, P_{lake,Ken})$  expresses different scales of degrees for discomfort of every person and every condition. For instance  $P_{services,Eva} = \{0, 1/2, 1\}$  expresses Eva's comfort, partial discomfort, full discomfort, respectively. Further  $P_{services,Ken} = \{0, le, se, 1\}$  correspond to comfort, discomfort on length of staying, discomfort on services and full discomfort, respectively. And this completes description of Figure 1.

Having expressed list of possible values for every person and every condition, further we will consider some concrete **longterm preferences** (diverse perception of discomfort by different conditions in Figure 2) and **shortterm preferences** (actual degree of discomfort that person admits in Figure 3).

**Fig. 2.** Longterm preferences in heterogeneous approach

$\bullet_{water,Eva}$	hot	cold	$\bullet_{services,Eva}$	in+tv	in	tv	no	$\bullet_{lake,Eva}$	yes	no
$\emptyset$	0	0	$\emptyset$	0	0	0	0	$\emptyset$	0	0
1/2	0	1	1/2	0	1/2	1	1	1/2	0	1
Sa+Su	0	1	Sa+Su	0	1	1	1	Sa+Su	0	1

$\bullet_{water,Joe}$	hot	cold	$\bullet_{services,Joe}$	in+tv	in	tv	no	$\bullet_{lake,Joe}$	yes	no
$\emptyset$	0	0	$\emptyset$	0	0	0	0	$\emptyset$	0	0
Sa	0	1/2	Sa	0	1/3	2/3	2/3	Sa	0	1
Su	0	1	Su	0	1/3	1	1	Su	0	1
Sa+Su	0	1	Sa+Su	0	1/3	1	1	Sa+Su	0	1

$\bullet_{water,Ken}$	hot	cold	$\bullet_{services,Ken}$	in+tv	in	tv	no	$\bullet_{lake,Ken}$	yes	no
$\emptyset$	0	0	$\emptyset$	0	0	0	0	$\emptyset$	0	0
1/2	0	0	1/2	0	se	le	1	1/2	0	1/2
Sa+Su	0	1	Sa+Su	0	1	1	1	Sa+Su	0	1

Starting with longterm preferences, notice that every person can have different perception of discomfort depending on cottage conditions and length of stay. In effort to express these longterm preferences, every person has own behavior that expresses  $\odot = (\bullet_{water,Eva}, \bullet_{services,Eva}, \bullet_{lake,Eva}, \bullet_{water,Joe}, \bullet_{services,Joe}, \bullet_{lake,Joe}, \bullet_{water,Ken}, \bullet_{services,Ken}, \bullet_{lake,Ken})$ . That means for instance  $\bullet_{services,Eva}$  is from  $C_{services} \times D_{Eva}$  to  $P_{services,Eva}$ .

Values of isotone and left-continuous operations  $\odot$  (by assumptions of our approach) with respect to the number of days and cottage conditions is known

and for our example included in Figure 2. First of all, notice that higher table values correspond to worse situation (0 as no discomfort, i.e. good situation, 1 as full discomfort, i.e. bad case) that might be in opposite with natural expectation, but this follows from assumptions of our heterogeneous approach.

We describe some remarks and interpretation on these longterm preferences:

- a) Notice that  $c \bullet_{\text{services,Eva}} \emptyset = 0$  for all  $c \in C_{\text{services}}$ , because no staying at the cottage and arbitrary conditions respond to no discomfort.
- b) Notice that  $\text{hot} \bullet_{\text{water,Eva}} d = 0$  for all  $d \in D_{\text{Eva}}$ , because presence of hot water and arbitrary number of days respond to no discomfort.
- c) Notice that  $\text{in} + \text{tv} \bullet_{\text{services,Joe}} d = 0$  for all  $d \in D_{\text{Joe}}$ , because presence of all services and arbitrary number of days respond to no discomfort.
- d) To see monotonicity, staying on Saturday and cold water represent half discomfort for Joe, but two days and cold water lead to big discomfort.
- e) Similarly one day staying and internet only represent half discomfort for Eva, but two days and internet only, or missing internet lead to full discomfort.
- f) Only internet and Saturday represent one third discomfort for Joe, only television and Saturday two third discomfort, only television and Sunday or two days lead to full discomfort.
- g) To see left-continuity, Saturday or Sunday and internet only represent one third discomfort for Joe, but supremum of these days (Saturday+Sunday) and internet only also lead to one third discomfort.

Having known longterm preferences of people, now we would like to express some shortterm preferences corresponding to some actual circumstances or actual sentiment of every person connected with actual staying. So every person appoints degree of discomfort that accepts or admits at the actual situation, i.e.  $R(\text{water, Eva}) \in P(\text{water, Eva})$ ,  $R(\text{services, Eva}) \in P(\text{services, Eva})$ ,  $R(\text{lake, Eva}) \in P(\text{lake, Eva})$ ,  $R(\text{water, Joe}) \in P(\text{water, Joe})$ ,  $R(\text{services, Joe}) \in P(\text{services, Joe})$ ,  $R(\text{lake, Joe}) \in P(\text{lake, Joe})$ ,  $R(\text{water, Ken}) \in P(\text{water, Ken})$ ,  $R(\text{services, Ken}) \in P(\text{services, Ken})$  and  $R(\text{lake, Ken}) \in P(\text{lake, Ken})$ . For example, Eva admits full discomfort on water conditions, half discomfort on services conditions and full discomfort on lake availability. Joes allows half discomfort on water conditions, great discomfort on services conditions and no discomfort on lake availability. Ken admits full discomfort on water conditions, discomfort on services by services conditions and half discomfort on lake availability as it is shown in Figure 3.

**Fig. 3.** Shortterm preferences in heterogeneous approach

	water	services	lake
Eva	1	1/2	1
Joe	1/2	2/3	0
Ken	1	se	1/2

Full discomfort on water conditions for Eva (table value 1) means that by the first table from Figure 2, Eva admits arbitrary number of days and hot or cold water, because all cases from Eva's water table are less or equal to 1 in Figure 2. Partial discomfort on services for Eva (table value 1/2) admits neither presence of all services or maximal one arbitrary day at the cottage and internet only, because these cases from Eva's services table are less or equal to 1/2 in second table of Figure 2. Similarly Ken permit neither all services or only one day at the cottage with internet connection only as you can see from Figure 2 and the eighth table in Figure 3.

Eventually in effort to identify necessary cottage conditions that fulfill all personal requirements we define following mappings  $\nearrow$  and  $\swarrow$ .

Let  $G$  be the set of all functions  $g$  with the domain  $B$  such that  $g(b) \in D_b$ , for all  $b \in B$ . (i. e.  $G = \prod_{b \in B} D_b$ ). Each function  $g$  corresponds to particular person's length of stay (e. g.  $g(\text{Eva}) = 1/2$ ,  $g(\text{Joe}) = \text{Sa}$ ,  $g(\text{Ken}) = 1/2$ ).

And let  $F$  be the set of all functions  $f$  with the domain  $A$  such that  $f(a) \in C_a$ , for all  $a \in A$  (i. e., more formally,  $F = \prod_{a \in A} C_a$ ). Each function  $f$  corresponds to particular cottage conditions (e. g.  $f(\text{water}) = \text{hot}$ ,  $f(\text{services}) = \text{in}$ ,  $f(\text{lake}) = \text{yes}$ ).

Define the following mapping  $\nearrow : G \rightarrow F$ : If  $g \in G$  then  $\nearrow(g) \in F$  is defined by

$$(\nearrow(g))(a) = \sup\{c \in C_a : (\forall b \in B) c \bullet_{a,b} g(b) \leq R(a, b)\}.$$

Mapping  $(\nearrow(g))(a)$  expresses requirement to the worst water or services conditions at the cottage by specific number of staying days that return at most degree of discomfort admitted by people. For instance, if  $g(\text{Eva}) = 1/2$ ,  $g(\text{Joe}) = \text{Sa}$ ,  $g(\text{Ken}) = 1/2$ , then for water we get  $(\nearrow(g))(\text{water}) = \text{cold}$ , that means that one day staying for Eva, staying on Saturday for Joe and one day staying for Ken correspond to the possibility for cold water at the cottage. Another example, if  $g(\text{Eva}) = \text{Sa} + \text{Su}$ ,  $g(\text{Joe}) = \text{Sa}$ ,  $g(\text{Ken}) = \text{Sa} + \text{Su}$ , then for services we get  $(\nearrow(g))(\text{services}) = \text{in} + \text{tv}$ , that admitted only cottage with internet connection and tv as the worst possible cottage in case of Eva's staying on Saturday and Sunday, Joe's staying on Saturday and Ken's staying on Saturday and Sunday.

Symmetrically define the mapping  $\swarrow : F \rightarrow G$ : If  $f \in F$  then  $\swarrow(f) \in G$  is defined as following:

$$(\swarrow(f))(b) = \sup\{d \in D_b : (\forall a \in A) f(a) \bullet_{a,b} d \leq R(a, b)\}.$$

Mapping  $(\swarrow(f))(b)$  expresses natural requirement to maximalize number of days spent at the cottage by specific water and services conditions that return at most degree of discomfort admitted by a person. For instance, for  $f(\text{water}) = \text{hot}$ ,  $f(\text{services}) = \text{in}$ ,  $f(\text{lake}) = \text{yes}$  we get  $(\swarrow(f))(\text{Eva}) = 1/2$ , that means that hot water and internet only correspond to maximal one day staying at the cottage for Eva.

We proved in [2] that the concept-forming mappings defined in this way have worthwhile properties. Here we give some natural interpretation for this theorem written below.

**Theorem 1.** *Let  $f \in F$  and  $g \in G$ . Then the following conditions are equivalent:*

- 1  $f \leq \nearrow(g)$ .
- 2  $g \leq \swarrow(f)$ .
- 3  $f(a) \bullet_{a,b} g(b) \leq R(a, b)$  for all  $a \in A$  and  $b \in B$ .

First part of the theorem can be interpreted as too much superfluous or equal conditions than conditions corresponding to concrete lengths of stay for people. Second part expresses that lengths of stay for people is less or equal than lengths corresponding to concrete cottage conditions. And third part represents that this concrete conditions and lengths of stay certainly satisfy all shortterm preferences.

**Corollary 1.** *Mappings  $\nearrow$  and  $\swarrow$  form a Galois connection.*

*Proof.* It follows from the equivalency of conditions 1 and 2 of the previous theorem.

### 3 Heterogeneous formal concept

We use a Galois connection ( $\nearrow, \swarrow$ ) for the concept lattice construction via classical Ganter-Wille's approach from [14].

By a *concept* we will understand a pair  $\langle g, f \rangle$  from  $G \times F$  such that  $\nearrow(g) = f$  and  $\swarrow(f) = g$ .

**Lemma 1.** *If  $\langle g_1, f_1 \rangle$  and  $\langle g_2, f_2 \rangle$  are concepts then  $g_1 \leq g_2$  iff  $f_1 \geq f_2$ .*

*Proof.* It is a simple consequence of the Corollary 2 (namely, parts 3a and 3b).

This lemma allows to define the following ordering of concepts:  $\langle g_1, f_1 \rangle \leq \langle g_2, f_2 \rangle$  iff  $g_1 \leq g_2$  (or equivalently  $f_1 \geq f_2$ ).

In summary by previous consideration we observed eight concepts in our running cottage example for three people and three cottage conditions shown in Figure 4.

Intents correspond to the worst cottage conditions that fulfill all personal requirements for specific number of days noticed in extent of concept. For example cold water, no services and lake available at the cottage are connected with no staying for Eva, staying on Saturday for Joe and no staying for Ken (second concept). In contrary, hot water, full services and lake available indicates maximal number of days spent for all people (last concept). Similarly one can interpret further concepts. For instance seventh concept shows that one day spend at the cottage by Eva, both days by Joe and one day by Ken requires the worst possible condition with hot water, internet connection and lake available.

Notice that intents do not include possibility of hot water and no services simultaneously. In this case we obtain  $\swarrow(\text{hot, no, yes}) = (\emptyset, \text{Sa}, \emptyset)$  and subsequently  $\nearrow(\emptyset, \text{Sa}, \emptyset) = (\text{cold, no, yes})$ . It can be interpreted as too much superfluous cottage conditions for Joe's stay on Saturday and maybe we can choose cheaper cottage.



**Fig. 4.** Heterogenous formal concepts for 3 people and 3 attributes

extents			intents		
Eva	Joe	Ken	water	services	lake
$\emptyset$	$\emptyset$	$\emptyset$	cold	no	no
$\emptyset$	Sa	$\emptyset$	cold	no	yes
1/2	$\emptyset$	1/2	cold	in	no
1/2	Sa	1/2	cold	in	yes
Sa+Su	$\emptyset$	1/2	cold	in+tv	no
Sa+Su	Sa	Sa+Su	cold	in+tv	yes
1/2	Sa+Su	1/2	hot	in	yes
Sa+Su	Sa+Su	Sa+Su	hot	in+tv	yes

All computations in our cottage example are done for three people, but it is fruitful to consider more complex example as in Figure 1 for six people water conditions, services conditions and lake for swimming available. Also it is possible that two people have the same lattice structures, for instance Eva and Lea have the same water and services lattices. Nevertheless behavior of Eva and Lea by the same condition should be diverse. One day and cold water should correspond to discomfort for Eva, but comfort for Lea or vice versa.

In this sense we make computation of all concepts for cottage example on six people and three cottage conditions and number of concepts was nine. The results are shown in Figure 5.

**Fig. 5.** Heterogenous formal concepts for 6 people and 3 attributes

extents						intents		
Eva	Joe	Ken	Lea	Sue	Tim	water	services	lake
$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	1/2	Su	cold	no	no
$\emptyset$	Sa	$\emptyset$	$\emptyset$	1/2	Su	cold	no	yes
1/2	$\emptyset$	1/2	1/2	1/2	Su	cold	in	no
1/2	Sa	1/2	Sa+Su	1/2	Su	cold	in	yes
Sa+Su	$\emptyset$	1/2	1/2	1/2	Su	cold	in+tv	no
Sa+Su	Sa	Sa+Su	Sa+Su	1/2	Sa+Su	cold	in+tv	yes
1/2	Sa+Su	1/2	Sa+Su	1/2	Su	hot	in	yes
Sa+Su	$\emptyset$	1/2	1/2	Sa+Su	Su	hot	in+tv	no
Sa+Su	Sa+Su	Sa+Su	Sa+Su	Sa+Su	Sa+Su	hot	in+tv	yes

#### 4 Another working examples

Medina, Ojeda-Aciego and Ruiz Calviño in [22] consider situation that we have written a scientific paper and have to decide which journal to choose for submitting. Set of objects consists of particular scientific journal (AMC, CAMWA,

FSS, ...) and set of attributes includes journal properties as impact factor, immediacy index, cited half-life and best position. Furthermore, problem consists in finding a multi-adjoint concept which represent the suitable journal to submit.

We provide the analogous analysis for our heterogeneous approach and propose following situation. People writing a scientific paper have to conclude which attributes of scientific journal is required to satisfy all researchers. Let  $B = \{\text{Ellis, Frank, ...}\}$ , set of objects, consists of people writing a mutual paper with specification of willingness to wait some time period to accepting of an article (till 6 months, till year, till year in case of science is a major job, over year in case of science is a minor job, over year). And let  $A = \{\text{current content, citation, ...}\}$ , set of attributes, includes specific properties of journals. Table values correspond to dissatisfaction with overall process of paper accepting by actual conditions. For example waiting till 6 month and current content means for Ellis no dissatisfaction (notated as table value 0), but waiting over year and uncurrent content full dissatisfaction (notated as table value 1).

Fig. 6. List of possible values for objects and attributes in journal example

		attributes	
		curr.content	citation
objects		<ul style="list-style-type: none"> <li>• no</li> <li>• yes</li> </ul>	<ul style="list-style-type: none"> <li>• slow</li> <li>• aver.</li> <li>• imm.</li> </ul>
Ellis	<ul style="list-style-type: none"> <li>• over year</li> <li>• till year</li> <li>• till 6 mo.</li> </ul>	<ul style="list-style-type: none"> <li>• 1</li> <li>• 0</li> </ul>	<ul style="list-style-type: none"> <li>• 1</li> <li>• 1/2</li> <li>• 0</li> </ul>
Frank	<ul style="list-style-type: none"> <li>• over year</li> <li>• maj. till</li> <li>• min. over</li> <li>• till year</li> </ul>	<ul style="list-style-type: none"> <li>• 1</li> <li>• 1/2</li> <li>• 0</li> </ul>	<ul style="list-style-type: none"> <li>• 1</li> <li>• 2/3</li> <li>• 1/3</li> <li>• 0</li> </ul>

Having expressed complete longterm and shortterm preferences of people working together on a paper, obtained concepts correspond to necessary attributes of journal satisfying all preferences. For instance consider that Ellis requires to publish till 6 months and Frank wishes to publish till one year and research represents his major job. In that case is necessary to submit to uncurrent content journal with medium immediacy index for citation.

Another example is based on a job background. Consider people applying for a job in the same company. The purpose is to specify conditions satisfying all personal requirements and effort to work together. Let  $B = \{\text{Peter, Paul, ...}\}$ , set of objects, consists of people applying for no job, part-time job, job on performance contract or full time job. And let  $A = \{\text{salary, language skills, start date, ...}\}$ , set

of attributes, includes specific job conditions. Salary conditions contain three degrees: high, medium and low; start date includes two degrees: immediately or at a later date; foreign languages requirements contain three degrees: no, one or two foreign language required. Table values express dissatisfaction with conditions connected with type of contract and job properties. Higher value corresponds to more dissatisfaction.

**Fig. 7.** List of possible values for objects and attributes in heterogenous job example

objects \ attributes		salary	start	languages
		$\begin{array}{c} \bullet \text{ low} \\ \bullet \text{ med.} \\ \bullet \text{ high} \end{array}$	$\begin{array}{c} \bullet \text{ late} \\ \bullet \text{ imm.} \end{array}$	$\begin{array}{c} \bullet \text{ two} \\ \bullet \text{ one} \\ \bullet \text{ no} \end{array}$
Peter	$\begin{array}{c} \bullet \text{ full-time} \\ \bullet \emptyset \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 0 \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 1/2 \\ \bullet 0 \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 0 \end{array}$
Paul	$\begin{array}{c} \text{full-time} \\ \text{part} \quad \diamond \quad \text{cont.} \\ \bullet \emptyset \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 1/2 \\ \bullet 0 \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 2/3 \\ \bullet 1/3 \\ \bullet 0 \end{array}$	$\begin{array}{c} \bullet 1 \\ \bullet 1/2 \\ \bullet 0 \end{array}$

Resulting concepts have the following interpretation. By consideration that Peter requires full-time job and Paul claims for job on performance contract, it is for instance necessary to find job with medium salary, immediate start date and most one spoken language.

We do not introduce particular longterm and shortterm preferences for this running examples on journal and job, but we give some motivation about usefulness of this heterogeneous approach in such area.

### 5 Heterogeneous formal concept lattice

The poset of all concepts ordered by  $\leq$  will be called a *heterogeneous concept lattice* and denoted by  $\text{HCL}(A, B, \mathcal{P}, R, \mathcal{C}, \mathcal{D}, \odot, \swarrow, \nearrow, \leq)$ .

The following theorem shows that the word *lattice* in its name corresponds with reality. The proofs of analogous theorems in previous approaches are included in different papers ([13], [14], [18]).

**Theorem 2.** (*The Basic Theorem on Heterogeneous Concept Lattices*)

- 1 A heterogeneous concept lattice  $\text{HCL}(A, B, \mathcal{P}, R, \mathcal{C}, \mathcal{D}, \odot, \swarrow, \nearrow, \leq)$  is a complete lattice in which

$$\bigwedge_{i \in I} \langle g_i, f_i \rangle = \left\langle \bigwedge_{i \in I} g_i, \nearrow \left( \swarrow \left( \bigvee_{i \in I} f_i \right) \right) \right\rangle$$

and

$$\bigvee_{i \in I} \langle g_i, f_i \rangle = \left\langle \bigvee \left( \bigwedge_{i \in I} g_i \right), \bigwedge_{i \in I} f_i \right\rangle.$$

2 For each  $a \in A$ ,  $b \in B$ , let  $P_{a,b}$  have the least element  $0_{P_{a,b}}$  such that  $0_{C_a} \bullet_{a,b} d = c \bullet_{a,b} 0_{D_b} = 0_{P_{a,b}}$ , for all  $c \in C_a$ ,  $d \in D_b$ . Then a complete lattice  $L$  is isomorphic to  $\text{HCL}(A, B, \mathcal{P}, R, \mathcal{C}, \mathcal{D}, \odot, \bigvee, \bigwedge, \leq)$  if and only if there are mappings  $\alpha : \bigcup_{a \in A} (\{a\} \times C_a) \rightarrow L$  and  $\beta : \bigcup_{b \in B} (\{b\} \times D_b) \rightarrow L$  such that:

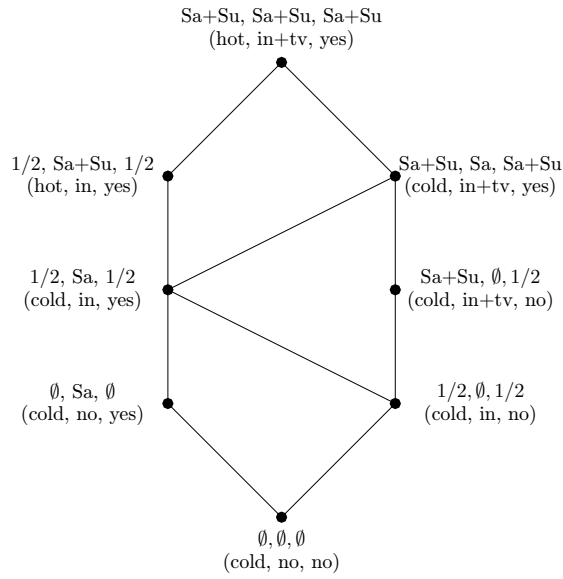
- 1a)  $\alpha$  does not increase in the second argument (for the fixed first one).
- 1b)  $\beta$  does not decrease in the second argument (for the fixed first one).
- 2a)  $\text{Rng}(\alpha)$  is inf-dense in  $L$ ,<sup>1</sup>
- 2b)  $\text{Rng}(\beta)$  is sup-dense in  $L$ .
- 3) For every  $a \in A$ ,  $b \in B$  and  $c \in C_a$ ,  $d \in D_b$

$$\alpha(a, c) \geq \beta(b, d) \quad \text{if and only if} \quad c \bullet_{a,b} d \leq R(a, b).$$

*Proof.* For self-contained proof see [2].

Figure 8 represents the resulting heterogeneous concept lattice for our cottage example with ordered concepts for 3 people and 3 cottage conditions.

**Fig. 8.** Heterogenous formal concept lattice for 3 people and 3 attributes



<sup>1</sup>  $\text{Rng}(\alpha)$  denotes range of mapping  $\alpha$ .

## 6 Conclusions and possible future works

In this paper we introduce some running examples on heterogeneous environment of the Formal Concept Analysis based on cottage, journal or job context. The main idea of heterogeneous approach is to diversify all that can be diversified and it is interesting that process of concept lattice construction still works. Hence, intuitively, it allows to use the Formal Concept Analysis also for tables with data of different types.

Bělohlávek shows how to deal with the problem of generating all concepts of a fuzzy concept lattice in [4]. A fast bottom-up algorithm to compute all concepts of a fuzzy closure operator is presented in [7]. We would like to modify and generalize these algorithms for our heterogeneous approach, too. And in this way we will make assumption of not linearly ordered set of truth degrees. Then it is fruitful to apply it on real-world data.

We would like to put emphasis that there is an similarly called approach working with multi-adjoint concept lattices based on heterogeneous conjunctors. This is done by Medina and Ojeda-Aciego in [21]. The difference is following. Multi-adjoint concept lattices work with different lattices too, but only for sets of attributes and objects. Objects and attributes are evaluated in two different lattices and on heterogeneous conjunctors, finally both different lattices are embedded to new so-called connected lattice and thus resulting concept lattice utilizes the same lattice for objects and attributes.

The next interesting connection is clarifying the relationship of our heterogeneous approach to Bělohlávek & Vychodil's fuzzification working with truth-stressers, so-called hedges (in [9] and [10]). In [19] it is shown that generalized concept lattices cover them in some sense but it seems that this new approach make this relationship more immediate.

In [15] is hedges used as a tool to reduce the size of multi-adjoint concept lattices with heterogeneous conjunctors as unifying of [21] and [10]. Another relationship that seems to be interesting for future work is heterogeneity in multi-adjoint concept multilattices that are more general structures as lattices [26].

## References

1. L. Antoni, S. Krajčí, O. Krídlo, B. Macek, L. Pisková, Relationship between two FCA approaches on heterogeneous formal contexts. In: Laszlo Szathmary, Uta Priss (Eds.): CLA 2012, Universidad de Malaga, pp. 93-102 (2012)
2. L. Antoni, S. Krajčí, O. Krídlo, B. Macek, L. Pisková, On heterogeneous formal contexts. Accepted to Fuzzy sets and systems.
3. R. Bělohlávek, Fuzzy concepts and conceptual structures: induced similarities, JCIS 98, Durham, USA, Vol. I: 179-182 (1998)
4. R. Bělohlávek, Algorithms for fuzzy concept lattices. in Proc. 4th Int. Conf. Recent Adv. Soft Comput., Nottingham, U.K., Dec. 12-13, 200-205 (2002)
5. R. Bělohlávek, Concept Lattices and Order in Fuzzy Logic, Annals of Pure and Applied Logic, 128, 277-298 (2004)
6. R. Bělohlávek, Sup-t-norm and inf-residuum are one type of relational product: Unifying framework and consequences. Fuzzy Sets and Systems 197: 45-58 (2012)

7. R. Bělohlávek, B. De Baets, J. Outrata, V. Vychodil, Computing the Lattice of All Fixpoints of a Fuzzy Closure Operator. *IEEE Transactions on Fuzzy Systems* 18 (3), 546–557 (2010)
8. R. Bělohlávek, V. Sklenář, J. Zacpal, Crispily generated fuzzy concepts, in: B. Ganter and R. Godin (Eds.): *ICFCA 2005, Lecture Notes in Computer Science* 3403, Springer-Verlag, Berlin/Heidelberg: 268-283 (2005)
9. R. Bělohlávek, V. Vychodil, Reducing the size of fuzzy concept lattices by hedges, *FUZZ-IEEE 2005, USA*, 663–668, ISBN: 0-7803-9159-4 (2005)
10. R. Bělohlávek, V. Vychodil, Formal concept analysis and linguistic hedges, *International Journal of General Systems* 41 (5): 503–532 (2012)
11. S. Ben Yahia, A. Jaoua, Discovering knowledge from fuzzy concept lattice. In: Kandel A., Last M., Bunke H.: *Data Mining and Computational Intelligence*, Physica-Verlag, 169–190 (2001)
12. A. Burusco, R. Fuentes-Gonzales, The study of  $L$ -fuzzy concept lattice, *Mathware & Soft Computing* 3, 209–218 (1994)
13. B.A. Davey, H.A. Priestley, *Introduction to Lattices and Order*, Cambridge University Press (2002).
14. B. Ganter, R. Wille, *Formal Concept Analysis, Mathematical Foundation*, Springer Verlag 1999, ISBN 3-540-62771-5 (1999)
15. J. Konecny, J. Medina, M. Ojeda-Aciego: Using intensifying hedges to reduce size of multi-adjoint concept lattices with heterogeneous conjunctors. In: Laszlo Szathmary, Uta Priss (Eds.): *CLA 2012, Universidad de Malaga*, pp. 245–256 (2012).
16. S. Krajčí, Cluster based efficient generation of fuzzy concepts, *Neural Network World* 13,5 521–530 (2003)
17. S. Krajčí, A generalized concept lattice, *Logic Journal of IGPL*, 13 (5): 543–550 (2005)
18. S. Krajčí, The basic theorem on generalized concept lattice, *CLA 2004, Ostrava, proceedings of the 2nd international workshop*, eds. V. Snášel, R. Bělohlávek, ISBN 80-248-0597-9, 25–33 (2004)
19. S. Krajčí, Every Concept Lattice With Hedges Is Isomorphic To Some Generalized Concept Lattice, *CLA 2005, Olomouc*, ISBN 80-248-0863-3, 1–9 (2005)
20. J. Medina, M. Ojeda-Aciego: Multi-adjoint t-concept lattices, *Information Sciences* 180 (5), pp. 712–725 (2010)
21. J. Medina, M. Ojeda-Aciego: On multi-adjoint concept lattices based on heterogeneous conjunctors, *Fuzzy Sets and Systems*, 208 (1), 95–110 (2012)
22. J. Medina, M. Ojeda-Aciego, J. Ruiz-Calviño, Formal concept analysis via multi-adjoint concept lattices. *Fuzzy Sets and Systems*, 160(2), 130–144 (2009)
23. J. Medina, M. Ojeda-Aciego, A. Valverde, P. Vojtáš, Towards biresiduated multi-adjoint logic programming. *Lect. Notes in Artificial Intelligence*, 3040: 608–617 (2004)
24. J. Medina, M. Ojeda-Aciego, P. Vojtáš, Multi-adjoint logic programming with continuous semantics. *Logic programming and Non-Monotonic Reasoning, LP-NMR'01, Lect. Notes in Artificial Intelligence* 2173, 351–364 (2001)
25. J. Medina, M. Ojeda-Aciego, P. Vojtáš, Similarity-based unification: a multi-adjoint approach. *Fuzzy Set and Systems*, 146: 43–62 (2004)
26. J. Medina, J. Ruiz-Calviño, Fuzzy formal concept analysis via multilattices: first prospects and results. In: Laszlo Szathmary, Uta Priss (Eds.): *CLA 2012, Universidad de Malaga*, pp. 69–79, (2012)
27. J. Pócs, Note on generating fuzzy concept lattices via Galois connections, *Information Sciences* 185 (1), pp. 128–136 (2012)

28. J. Pócs, On possible generalization of fuzzy concept lattices, *Information Sciences* 210, pp. 89–98 (2012)
29. S. Pollandt, *Fuzzy-Begriffe – formale Begriffsanalyse unscharfer Daten*, Springer (1997)
30. S. Pollandt, Datenanalyse mit Fuzzy-Begriffen, in: G. Stumme, R. Wille: *Begriffliche Wissensverarbeitung. Methoden und Anwendungen*, Springer, Heidelberg 2000, 72–98 (2000)

# Attribute Exploration on the Web

Robert Jäschke<sup>1</sup> and Sebastian Rudolph<sup>2</sup>

<sup>1</sup> L3S Research Center, Hannover, Germany, [jaeschke@l3s.de](mailto:jaeschke@l3s.de)

<sup>2</sup> AIFB, Karlsruhe Institute of Technology, Germany, [rudolph@kit.edu](mailto:rudolph@kit.edu)

**Abstract** We propose an approach for supporting attribute exploration by web information retrieval, in particular by posing appropriate queries to search engines, crowd sourcing systems, and the linked open data cloud. We discuss underlying general assumptions for this to work and the degree to which these can be taken for granted.

**Keywords:** Formal Concept Analysis, Attribute Exploration, Web Information Retrieval, Linked Open Data

## 1 Introduction

In Formal Concept Analysis [6], objects are described by their attributes. The idea of *attribute exploration* is to determine a minimal set of implicational dependencies between attributes that hold for all objects of a certain domain. To this end, one starts with a partially defined formal context, i.e., a selection of objects and their attributes. From this “sample”, hypothetical implications are computed and presented to an expert who either confirms their universal validity or refutes it by providing an object as counterexample. Normally, obtaining these counterexamples is a laborious task, depending on the domain of the objects. For example, the domain of one of the earliest applications of attribute exploration [18] was lattice theory and confirming a hypothetical implication between properties of lattices meant to find an appropriate proof whereas refuting it meant to provide a specific lattice violating the hypothesis. Attribute exploration has also been used for creating and completing ontologies based on description logics [16,3], or for building access control models [12].

Until now, attribute exploration was mostly driven by an expert who has to check the implications. Typically, each implication is presented to the expert as a question in the form “Is it true that all objects that have the attribute(s)  $l_1, l_2, \dots$  also have the attribute(s)  $r_1, r_2, \dots$ ?” However, the knowledge we are seeking is often already available on the web, e.g., as facts in Wikipedia, or it can be obtained by leveraging massively collaborative Web 2.0 platforms. While we acknowledge the role of the expert and do not want to replace him or her, we aim to better support the expert in employing the knowledge found in the World Wide Web by automatically posing appropriate queries to web search engines in order to retrieve potential counterexamples. Thereby, we assume that the expert is not omniscient but may benefit from external knowledge, at least



for the answer of some questions. Our approach has the potential to speed up the attribute exploration process and, by providing context for all questions, to help the expert to avoid errors due to existing counterexamples unknown to him or her.

In this paper we want to outline the chances and limits of supporting attribute exploration by on-the-fly retrieval of information from the web in various ways. While this endeavor is itself exploratory and only preliminary, we hope that our considerations will pave the way toward exploration methodologies that make intelligent use of the abundance of available web data. The paper is organized as follows: In Section 2 we review related work. After a brief introduction to attribute exploration in Section 3, we describe three specific approaches to tackle attribute exploration using the web in Section 4. We present a first implementation in Section 5 and conclude the paper in Section 6.

## 2 Related Work

Koester’s FooCA system [11], retrieves results from major web search engines and allows users to analyze and visualize them using FCA. Therefore, FooCA uses the title, description, and the URL of web pages to build a formal context. The system allows users to modify queries until they fit their information need. In contrast to our work, attribute exploration is not considered in FooCA and queries are built by the user instead of the system itself. Furthermore, FooCA is considering the web pages themselves as objects while our approach considers objects *within* web pages or draws conclusions about empty result sets.

Rudolph [16] proposed to create description logic [2] knowledge bases by means of attribute exploration coupled with automated reasoning systems. Baader et al. [3] show how an extension of attribute exploration, that is capable of handling partial information, can be employed for completing such knowledge bases. Since description logic is underlying the web ontology language OWL [10] in which DBPedia [1] is represented, their approach could be used to check the completeness of DBPedia and therefore also Wikipedia.

The idea to automatically query web search engines to check or extend a knowledge base has been applied in the area of ontology learning, where Hearst patterns [9] are used to learn relationships between concepts [5].

## 3 Formal Concept Analysis and Attribute Exploration

In the following we briefly introduce the important notions in FCA by means of a small example. Imagine a user that is interested in European politics and therefore investigates political, military, and economic alliances in Europe by considering the membership of European countries in the NATO and the EU and their participation in the Euro and the Schengen Agreement.<sup>3</sup> Collecting

<sup>3</sup> [http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/index\\_en.htm](http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/index_en.htm)

the information for all European countries is a tedious task and since the user is anyway only interested in the implications that hold between the four attributes NATO, EU, Euro, and Schengen, he decides to apply attribute exploration to find these implications.

### 3.1 Formal Contexts and Formal Concepts

Using the notation from [6], we are considering *formal contexts*  $\mathbb{K} := (G, M, I)$  where  $G$  is a set of objects,  $M$  a set of attributes, and  $I$  a binary relation between  $G$  and  $M$ , i.e.,  $I \subseteq G \times M$ . We read  $(g, m) \in I$  as “object  $g$  has attribute  $m$ ”. For  $A \subseteq G$ , let  $A' := \{m \in M \mid \forall g \in A : (g, m) \in I\}$ , and dually, for  $B \subseteq M$ , let  $B' := \{g \in G \mid \forall m \in B : (g, m) \in I\}$ . For an object  $g \in G$  we often write  $g'$  instead of  $\{g\}'$ .

Table 1 shows the formal context with which our user starts. It contains as objects the three countries Czech Republic, Norway, and Germany and as attributes the four alliances/agreements NATO, EU, Euro, and Schengen. The

**Table 1.** A formal context about properties of European countries.

	NATO	EU	Euro	Schengen
Czech Republic	×	×		×
Norway	×			×
Germany	×	×	×	×

information about the membership of the countries is taken from their corresponding pages in Wikipedia. We will use that context as a running example throughout this paper.

### 3.2 Implications Between Attributes

An *implication* between the attributes of a formal context is a pair of subsets  $L, R$  of  $M$  denoted by  $L \rightarrow R$ , in which  $L$  is called the *premise* and  $R$  the *conclusion*. For simplicity, we always assume  $L \cap R = \emptyset$ , i.e., we omit the attributes in the premise from the conclusion. An implication  $L \rightarrow R$  *holds* in a formal context, if each object having all attributes from  $L$  also has all attributes from  $R$ . For instance, the implication  $\{EU\} \rightarrow \{NATO\}$  holds in the context in Table 1.

### 3.3 Attribute Exploration

The goal of attribute exploration is to compute a set of implications between attributes that hold for all objects under consideration. In particular, if it is not feasible to explicitly list all objects of a formal context (e.g., because there are infinitely many of them), attribute exploration supports us in searching and specifying representative objects.

In an interactive process [6], implications between attributes are computed and shown to an expert who checks if they hold. A found implication  $L \rightarrow R$  can either be accepted or refuted by a *counterexample*. A counterexample  $c$  is an object that has all attributes from  $L$  but there exists at least one attribute from  $R$  that  $c$  does not have. Formally, an object  $c$  refutes the implication  $L \rightarrow R$ , if  $L \subseteq c'$  but  $R \not\subseteq c'$  (i.e., there is at least one  $m \in R$  with  $m \notin c'$ ). Assuming a universe of objects, we denote the set of possible counterexamples for an implication  $L \rightarrow R$  (i.e., all those objects refuting the implication  $L \rightarrow R$ ) by  $\mathcal{C}(L \rightarrow R)$ . The algorithm in [6] ensures that a non-redundant and complete set of implications [7] is computed.

Starting the attribute exploration with the context in Table 1 yields the implication  $\emptyset \rightarrow \{\text{NATO}, \text{Schengen}\}$  and thus the question “Is it true that all objects have the attributes NATO and Schengen?” A counterexample would be a country that is not a member of the NATO or does not participate in the Schengen Agreement.

## 4 Web-Based Attribute Exploration

How can we leverage the knowledge available in the web to infer implications between attributes? Returning to our example, in the simplest case the user could gather a list of countries in Europe (e.g., from Wikipedia<sup>4</sup>) and for each country look up the information on the web, or do the same for each of the attributes. On the one hand, this is a very tedious task that involves searching, visiting, and screening many web pages. On the other hand, it is not necessary to build this complete list in order to obtain implications between the attributes. As we have seen in Section 3.3, attribute exploration helps us to find a small set of objects whose attribute logic (i.e., the attribute implications jointly satisfied by them) is universally valid.

We are now investigating how the knowledge available on the web can be leveraged for attribute exploration. While we are focussing on web search engines (Section 4.2) that allow us to query a larger part of the web than any other technology, we want to show the wide range of sources available on the web by investigating three other possible options (Section 4.3): social question answering, crowdsourcing, and the linked open data cloud. We start with an overview on query strategies, since there are a few commonalities between these approaches on an abstract level.

### 4.1 Abstract Query Strategies

In the majority of the cases, the information that we can hope to draw from the web is *factual* (or *assertional*), i.e., it provides information about a singular instance (object) and its properties (attributes). Information about universally

<sup>4</sup> [http://en.wikipedia.org/wiki/List\\_of\\_sovereign\\_states\\_and\\_dependent\\_territories\\_in\\_Europe](http://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_in_Europe)

valid propositions (so-called *terminological* knowledge) is more rare and harder to retrieve.<sup>5</sup> However, implications are of terminological nature, hence it will be hard to draw conclusive evidence from the web for the validity of an implication. On the other hand, information about counterexamples is factual. Thus, the general strategy is to focus on the task of retrieving counterexamples from the web and, applying a sort of closed-world assumption, assume that an implication is valid if no such counterexamples can be found.

In general, information retrieval is performed via queries (be it queries posed toward a web search engine or a database or a human). Now, when considering an implication  $L \rightarrow R$ , which type of factual query helps us to find counterexamples?

From a logical viewpoint, an *instance query* is a formula  $\varphi[x]$  with one free individual variable  $x$ , specified in a fixed logical formalism which is called *query language*. For a given domain (or logically speaking: interpretation) and query  $\varphi[x]$ , the associated *set of answers*  $Ans(\varphi[x])$  is the set of domain elements  $d$  for which  $\varphi[d]$  is true. We will now assume that each attribute  $m \in M$  comes with an instance query  $\varphi_m[x]$  for which the answer set is  $m'$ .

Under these assumptions, a query  $q[x]$  for a counterexample of an implication  $L \rightarrow R$  can be written as

$$q[x] := \bigwedge_{l \in L} \varphi_l[x] \wedge \bigvee_{r \in R} \neg \varphi_r[x] \quad (1)$$

Note that this requires the query language to support conjunction, but also both negation and disjunction. In the case that disjunction is not available, the above query can be split into a set of queries  $Q = \{q_r[x] \mid r \in R\}$  with

$$q_r[x] := \bigwedge_{l \in L} \varphi_l[x] \wedge \neg \varphi_r[x], \quad (2)$$

then the set of counterexamples can be obtained by taking the union over all corresponding answer sets, i.e.,

$$Ans(q[x]) = \bigcup_{q_r \in Q} Ans(q_r[x]). \quad (3)$$

The logical viewpoint presented here helps in setting the stage and formulating a generic counterexample query (set), however, it assumes “perfect querying” which is not realistic in practice, particularly for information retrieval on the web. Web query answers may be unsound (the answer contains instances that do not qualify), incomplete (the answer does not contain instances that would qualify), usually they are both.<sup>6</sup>

<sup>5</sup> With the exception of knowledge bases formulated in expressive ontological languages like OWL.

<sup>6</sup> Note that systems performing retrieval tasks are evaluated via the measures precision and recall. Sound querying would correspond in 100% precision, complete querying in 100% recall, neither of which is normally achieved.

Thus it will normally be necessary to perform some sort of quality assurance (usually via scrutiny by a human) on the answers retrieved for a counterexample query, in order to make sure that they are indeed counterexamples. If the result contains a valid counterexample, we can add it to the context and continue the attribute exploration.

We will investigate this in more detail in the following sections, noting that the overall process is essentially the same in all cases:

- a. Take an implication from attribute exploration and transform it into a query or a set of queries.
- b. Pose the queries to the system and show the user the result.
- c. Let the user decide if the implication holds or not.

The decision whether to employ queries of the form (1) or (2) not only depends on technical restrictions of the query language but also on the degree of human involvement in the querying task. Depending on whether the queries are posed to a large crowd of users or a single user is checking the results for counterexamples and possibly modifying the queries, we must take the capabilities of the involved human into account. While with the form (1), one query per implication is sufficient (and thus only one result set needs to be checked, respectively), the disjunction potentially causes many results to be returned at once and it is not so obvious for each result, which of the attributes from the conclusion it is lacking. This complicates the task of finding counterexamples. Queries of the form (2), on the other hand, are shorter and more simple and therefore their syntax and semantics are easier to understand by the user. Unfortunately, the results of the queries for one implication might overlap if there exist counterexamples that lack several of the attributes.

Now that it is clear that, theoretically, we can formulate web queries that support a user in checking the validity of an implication, the following research questions arise:

- What background knowledge about objects and attributes must we demand from the user? Is it enough to consider objects and attributes as strings or do we need synonyms, regular expressions, or even the corresponding page in Wikipedia? Which query language will we use?
- Is one query sufficient or do we need something like incremental query-refinement?
- Which results should be presented to the user? In which way?

We tackle these questions theoretically in the following two sub-sections and practically in Section 5.

## 4.2 Web Search Engines

Search engines constitute the most common entry point to the web and allow us to search over a corpus of documents that is by far the largest set of documents we can access. Their query languages typically support at least the logical

conjunction and disjunction of query terms. For efficiency reasons,<sup>7</sup> negation is only supported in queries that contain at least one positive term, to retrieve documents that satisfy ⟨list of positive terms⟩ *but not* ⟨list of negative terms⟩. Therefore, implications  $\emptyset \rightarrow R$  with an empty premise must be treated with care, since the corresponding query would entirely consist of negative terms. Except for this case, the query language is therefore suitable to represent queries like the ones in Equations (1) and (2). In principle, any web search engine is useable but for simplicity our examples follow the search syntax of the most popular ones, Bing and Google.<sup>8</sup>

We focus on queries of the form (2), composed only of logical conjunction and negation. Their syntax is more simple and more common to users, since it composes (positive and negative) terms using conjunction only, which is the standard composition operation in most web search engines.<sup>9</sup> Furthermore, the queries are shorter. Both aspects allow the users to easier understand and modify the query. As a drawback, for each implication the users must check as many result sets as there are attributes in the conclusion. The prefix + in front of a term ensures that it is textually contained in the web page (possibly modulo morphological variants), the prefix - ensures that the term is not contained in the web page, a conjunction of terms is achieved by concatenating them by whitespace ( ). The query from Equation (2) can then be written as

$$q_r[x] := +l_1 \_ +l_2 \_ \dots \_ +l_{|L|} \_ -r \quad (4)$$

As already mentioned, special care must be taken when  $L = \emptyset$ , i.e., the premise is empty, since queries containing exclusively negated terms are not supported by web search engines. This can be addressed by specifying the domain  $d$  of the objects (e.g. “Countries in Europe”) and adding it as positive term to the query:

$$q_r[x] := +d \_ +l_1 \_ +l_2 \_ \dots \_ +l_{|L|} \_ -r \quad (5)$$

This is not a strong restriction, since often the objects we are interested in are instances of a common class. Furthermore, many search engines allow us to restrict the web site of the results by adding it to the query with the `site:` prefix. E.g., to restrict the result set to pages from the English Wikipedia, we can add the term `site:en.wikipedia.org` to a query. For our example context from Table 1 both restrictions are actually useful, since we can expect that the objects, i.e., European countries, we are interested in are well described in Wikipedia.

<sup>7</sup> A set of documents that contain the positive terms can be efficiently retrieved using an inverted index, likewise the documents that contain the negative terms can be retrieved. The first set is then filtered by the second.

<sup>8</sup> <http://support.google.com/websearch/bin/answer.py?hl=en&answer=136861>

<sup>9</sup> Note that this default is more and more weakened: when no or few documents could be found, or a term is very general, web search engines occasionally return documents that do not necessarily contain all positive terms. That is also the reason why we prefer to prefix all positive terms with +, which really ensures that they are contained in the document.

In the standard case, the result of a web search query is a set of web documents.<sup>10</sup> Thus, unless the objects of our domain of interest are indeed web documents (as in the setting described by Koester [11]), the retrieved documents will merely serve as an informative resource, hopefully describing objects of the wanted category.

Motivated by the preceding discussion, our approach is based on the following implicit assumptions:

1. Web documents often describe singular objects.
2. For every attribute there is a search term, the presence of which in such a web document can be regarded as an indicator for the described object having the attribute.
3. Likewise, the absence of this search term can be regarded as an indicator for the object *not* having the attribute.

All of these assumptions are arguable and their applicability varies from case to case. E.g., one problem with that approach is that web pages that mention that an object does *not* have a certain attribute are ignored by the corresponding negated term in the query and thus the counterexample can not be found. In Section 5.2, we will see instances of these problems we have to face in reality.

### 4.3 Other Paradigms

Besides web search engines, the number of systems that could possibly be employed to support attribute exploration is abundant: One could post questions to blogging or micro-blogging platforms and hope for answers, or even ask friends on social networks. In this section we first focus on two approaches that are particularly intended for posing queries to humans: *social question answering* and *crowdsourcing*. Then we take a look at an approach with considerably less human involvement on the one hand but a much more formal knowledge representation on the other hand: structured knowledge bases that are part of the so-called *linked open data cloud*.

**Social Question Answering Systems.** As one of the most explicit forms of social search, social question answering systems like Yahoo! Answers<sup>11</sup> or StackOverflow<sup>12</sup> allow users to post questions on the web that can be answered by other users. While Yahoo! Answers is very general and also contains questions like *How do hotels keep their towels so white?*, other systems are focused on certain topics, e.g., StackOverflow on programming (with questions like *How can I draw a flow chart using L<sup>A</sup>T<sub>E</sub>X?*). The systems provide mechanisms to

<sup>10</sup> A notable exception after the recent advent of <http://schema.org/> are cases where the search engine additionally returns data about other entities, as e.g. in <http://www.google.com/?q=Rudolf+Wille>

<sup>11</sup> <http://answers.yahoo.com/>

<sup>12</sup> <http://www.stackoverflow.com/>

rate, comment, and accept answers such that users can more easily find correct answers or discuss alternative solutions.

Leveraging social question answering systems for attribute exploration is straightforward: instead of asking the expert, we could in turn post the question if an implication holds or not to the system and the expert could then conclude from the answers if the implication at hand holds or not. We could even ask the users to answer in a specific format such that we could automatically parse counterexamples and thereby completely automate the process. The rating mechanisms would allow us to judge the quality level of the answers and could support the expert in judging the result. An important drawback of the approach, however, is the high latency of answers (depending on the domain from some minutes to days; some questions are never answered) and the misuse of a social system in an unsocial way. On the other hand, one can retrieve profound answers, if an expert is willing to answer the question.

**Crowdsourcing Systems.** Something similar can be accomplished (and is technically much easier to implement) using crowdsourcing systems like Amazon Mechanical Turk.<sup>13</sup> They allow programmers to create small “human intelligence tasks” that are solved by a crowd of workers that get a small amount of money for solving these tasks (from a few cents to some dollars). Typical examples for such tasks are optical character recognition, information extraction, or image classification. In our scenario, we could again directly ask the workers if an implication holds and if not, which counterexample they can provide. Alternatively, we could break down each implication into questions of the form introduced in the previous section and ask the workers to answer these. The systems provide an application programming interface such that we can program user interfaces where the workers can directly enter counterexamples. In contrast to social question answering systems, crowdsourcing platforms are explicitly intended for this kind of human-machine interaction and therefore better suited as automatic source for attribute exploration. On the other hand, each answer costs money (though we can be lucky that FCA ensures that a minimal number of questions is asked) and the quality of the results often is not very good which requires to distribute each question to several workers and employ voting, or reputation mechanisms [15].

**Linked Open Data Cloud.** An increasing amount of knowledge is published online in the linked open data cloud in knowledge bases like YAGO [17] or DB-Pedia [4] where it is represented in triples of the form *(subject, predicate, object)* that express the fact that the *subject* is in relation *predicate* with the *object*, e.g., *(Germany, is member of, European Union)*. These triple sets – which could be conceived as multi-valued formal contexts – are represented using the Semantic Web standards RDF and OWL [10]. Knowledge bases in these formats can be queried using SPARQL [14] and thus be integrated into the attribute

<sup>13</sup> <http://mturk.amazon.com/>



exploration process in a similar way as described in Section 4.2 by formulating SPARQL queries instead of web search queries. In the sequel, we focus on DBPedia as one of these exemplary knowledge bases. DBPedia contains millions of triples that are automatically extracted from Wikipedia and thus constitutes a valuable source of information for various domains of interest. For each city in Wikipedia, for example, facts like name, country, geo-location, population, etc. are available.

In order to answer the queries posed during the exploration process, in the most simplest case the objects of the formal context should be entities that are represented by Wikipedia pages, preferably from a particular category, like *Countries in Europe*, or *Internet standards*.<sup>14</sup> For a multi-valued context, the attributes should map to properties of DBPedia, e.g., *total population*. The attribute values could then be categories, pages in Wikipedia, or arbitrary data (strings, numbers, dates, etc.). One would then apply conceptual scaling to derive a one-valued context amenable for attribute exploration. One exception to this is the particular case where the attributes directly map to categories of Wikipedia. This situation can be represented by a one-valued context in which the intent of an object is the set of categories associated to this object. During the attribute exploration process, an implication between attributes can be checked by querying the knowledge base with an appropriate SPARQL query. How such a query is built depends on the chosen attributes and scales. Due to space restrictions, we can not present the complete SPARQL queries that would correspond to Equations (1) or (2), but instead we give an example for the context in Table 1.

We require that all objects belong to the category *European countries*<sup>15</sup> and map the attributes to categories of Wikipedia in the following way: *NATO*  $\mapsto$  *Member states of NATO*, and *EU*  $\mapsto$  *Member states of the European Union*. The attribute *Euro* needs special care, since there exists no category for all countries that have the Euro as currency. Instead, we can use the category *Currency* and restrict it to the value *Euro*. Unfortunately, there exists no category for countries of the Schengen area and thus we can not map the attribute *Schengen*. This shows two limitations of our approach we will discuss at the end of this section. As an example, we now consider the query  $q[x] = \neg\varphi_{\text{NATO}}[x] \vee \neg\varphi_{\text{EU}}[x]$  for the implication  $\emptyset \rightarrow \{\text{NATO}, \text{EU}\}$  that comes up during the exploration of the context in Table 1. Using the mappings of attributes to categories presented above, we can map the query to the SPARQL query in Figure 1. The disjunction in our original query is represented by a UNION of two patterns that match countries of Europe that are not in the NATO or not in the EU, respectively. Since the current SPARQL standard does not support negation, we must employ the rather complicated OPTIONAL { ?y ... FILTER (?country=?y) . }

<sup>14</sup> The English Wikipedia contains more than 850,000 hierarchically organized categories (source: extracted category labels in DBPedia, [http://downloads.dbpedia.org/3.8/en/category\\_labels\\_en.nt.bz2](http://downloads.dbpedia.org/3.8/en/category_labels_en.nt.bz2)).

<sup>15</sup> This category has been changed to *Countries in Europe* in Wikipedia, but this change is not available in DBPedia, yet.

```

PREFIX dbc: <http://dbpedia.org/resource/Category:>
PREFIX dcs: <http://purl.org/dc/terms/>
SELECT DISTINCT ?country WHERE {
  {
    ?country dcs:subject dbc:European_countries .
    OPTIONAL {
      ?y dcs:subject dbc:Member_states_of_NATO .
      FILTER (?country = ?y) .
    }
    FILTER (!BOUND(?y))
  }
  UNION
  {
    ?country dcs:subject dbc:European_countries .
    OPTIONAL {
      ?y dcs:subject dbc:Member_states_of_the_European_Union .
      FILTER (?country = ?y) .
    }
    FILTER (!BOUND(?y))
  }
}
ORDER BY (?country)

```

**Figure 1.** A SPARQL query to DBPedia that retrieves Wikipedia pages of the category *European countries* that do either not belong to the category *Member states of the NATO* or *Member states of the European Union*, respectively.

FILTER (!BOUND(?y)) construct.<sup>16</sup> Posing the query to the DBPedia SPARQL Explorer<sup>17</sup> indeed returns a list of European countries that are either not in the NATO or not in the EU, e.g., *Albania*, or *Andorra*. These could now be added as counterexamples to the context and the attribute exploration could continue.

One question quickly comes up with this approach: Why should we build the formal context through attribute exploration when we could as easily create it by a single SPARQL query? On the one hand, we want to show in this paper the wide range of possible sources in the web that can support attribute exploration, and linked open data is an obvious candidate. On the other hand, the goal of attribute exploration could be to find errors in the knowledge bases and check them for completeness. This requires human interaction. Compared to querying web search engines, it is considerably easier to exactly specify the requested or unrequested attributes of an object and one has to deal with fewer ambiguities. On the other hand – as we have seen with the attributes *Euro* and *Schengen* of our example context – only a small fraction of the knowledge available in

<sup>16</sup> The upcoming refinement of the SPARQL standard [8] will allow for a more direct way of expressing negation.

<sup>17</sup> <http://dbpedia.org/snorql/>

Wikipedia (let alone in the web) is accessible using the described method and therefore this approach clearly has its limitations.

## 5 Implementation

In this section we present a prototype that – since it is freely available on the web – allows everybody to test our approach using web search engines and discover its chances and limitations. We first describe the prototype and then present first insights, limitations, and plans for improvement.

### 5.1 Prototype

We developed a web-based prototype<sup>18</sup> in Java that implements the querying strategy described in Section 4.2. On its start page, the application allows the user to upload a formal context in a file in the ConExp [19] XML format CEX or to select one of the predefined example contexts. In addition, the user can specify the domain of the objects and restrict the results to a specific site. The prototype is based on the attribute exploration algorithm available in FCalib.<sup>19</sup> We implemented the **Expert** class of the FCAAPI<sup>20</sup> such that for each implication that the expert shall check, queries are generated and sent to a web search engine (we are using Microsoft Bing,<sup>21</sup> since it provides an API which allows a limited number of free requests per month). The context, the accepted implications, the current implication, the corresponding queries, and the first ten retrieved results for the active query are then shown on a web page to the user who is asked if the result set contains a counterexample (see Figure 2). Each result contains the title and URL of the corresponding web page and a short text snippet from the page that contains the matching query terms. If the user found a counterexample, he or she can add it to the context and the next implication is checked. If no counterexample could be found, the results for the other queries of that implication can be inspected, until either a counterexample can be found or all queries for the implication at hand were inspected. A text input field allows the user to modify the query and retrieve further results from the search engine, if necessary.

### 5.2 Example

Returning to our context in Table 1, the first implication that can be derived is  $\emptyset \rightarrow \{\text{NATO, Schengen}\}$ . A counterexample would be a European country that is not a member of the NATO or does not participate in the Schengen Agreement. Since this is an implication with an empty premise and since all objects we are interested in belong to a common class, namely “Countries in Europe” – a

<sup>18</sup> <http://greymane.l3s.uni-hannover.de:8888/>

<sup>19</sup> <http://code.google.com/p/fcalib/>

<sup>20</sup> <http://code.google.com/p/fcaapi/>

<sup>21</sup> <http://www.bing.com/>

## Web-Based Attribute Exploration

### Formal Context

Countries in Europe	NATO	EU	Euro	Schengen
Czech Republic	x	x		x
Norway	x			x
Germany	x	x	x	x

[change context](#)

### Attribute Exploration

The current implication is: **[]** ⇒ **[Schengen, NATO]**.

You can either  it or  :

	NATO	EU	Euro	Schengen
<input type="text"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Can you find a find a counterexample within the following web search results?

1. **+"Countries in Europe" -"Schengen" site:en.wikipedia.org**
2. [+"Countries in Europe" -"NATO" site:en.wikipedia.org](#)

<input -\"schengen\"="" countries="" europe\"="" in="" site:en.wikipedia.org"="" type="text" value="+\"/>	<input type="button" value="custom search"/>
---	--

1. [Category:Former countries in Europe - Wikipedia, the free encyclopedia](#)  
[http://en.wikipedia.org/wiki/Category:Former\\_countries\\_in\\_Europe](http://en.wikipedia.org/wiki/Category:Former_countries_in_Europe)  
Former countries in Europe after 1815; List of early East Slavic states; List of historic states of Germany; List of historic states of Italy, Grand Duchy of Moscow
2. [Former countries in Europe after 1815 - Wikipedia, the free ...](#)  
[http://en.wikipedia.org/wiki/Former\\_countries\\_in\\_Europe\\_after\\_1815](http://en.wikipedia.org/wiki/Former_countries_in_Europe_after_1815)  
This article gives a detailed listing of all the countries, (including puppet states), that have existed in Europe since the Congress of Vienna in 1815 to the present ...
3. [File:Same sex marriage map Europe detailed.svg - Wikipedia, the ...](#)  
[http://en.wikipedia.org/wiki/File:Same\\_sex\\_marriage\\_map\\_Europe\\_detailed.svg](http://en.wikipedia.org/wiki/File:Same_sex_marriage_map_Europe_detailed.svg)  
Date: 1 August 2007 (2007-08-01) Source: self-made, based on Image:Same sex marriage map Europe.svg. Author: Silje L. Bakke: Other versions: Derivative works of this ...
4. [Category talk:Countries in Europe - Wikipedia, the free encyclopedia](#)  
[http://en.wikipedia.org/wiki/Category\\_talk:Countries\\_in\\_Europe](http://en.wikipedia.org/wiki/Category_talk:Countries_in_Europe)  
This category is clearly broken. When one clicks on category:European countries, they expect to see the list of countries, not a list of further arbitrary subdivisions.

**Figure 2.** A screenshot showing our prototype implementation.

category of the English Wikipedia,<sup>22</sup> we add the domain restriction **+"Countries in Europe"** and the site restriction **+site:en.wikipedia.org** and obtain the two queries

<sup>22</sup> [http://en.wikipedia.org/wiki/Category:Countries\\_in\\_Europe](http://en.wikipedia.org/wiki/Category:Countries_in_Europe)

1. `+"Countries in Europe" -Schengen +site:en.wikipedia.org`
2. `+"Countries in Europe" -NATO +site:en.wikipedia.org`

As can be seen in the screenshot in Figure 2, none of the top four results for the first query is a Wikipedia page about a specific country. The same applies to the following six results and also to the top ten results of the second query (which are actually the same as for the first query). However, as we have seen in Section 4.3, there do exist several countries that would constitute a counterexample.

This raises the question *Why are no countries among the top results?* The two likely reasons for the discovered problem can be found in the way web search engines work: they retrieve pages whose *textual* content *matches* the query and return only the top hits according to a *ranking*, which is typically computed by an algorithm like PageRank [13]. The matching against the text of the pages returns many pages that contain the string *Countries in Europe* but do not belong to the corresponding category, which is a property we can not yet enforce with standard web search engines. In addition, for five of the top ten results the string is contained in the page title, which typically increases their ranking score. A PageRank-like ranking further prefers pages that have a high number of incoming links, which is typical for category and listing pages that constitute a large part of the top results. Hence, for our approach it would be very helpful, if we could enforce to receive only pages of a specific Wikipedia category. As a workaround, we can add terms to our query that we would expect to find on a Wikipedia page that is describing an object from the domain at hand. In our example, where the objects are (European) countries, we can assume that every page that describes a country contains a section about *politics*, *history*, *geography*, etc. This assumption is supported by the results we retrieve for the extended query `+"Countries in Europe" site:en.wikipedia.org -Schengen +politics +history +geography`: The top ten results on Bing contain eight countries: Azerbaijan, Spain, Armenia, Greece, Turkey, United Arab Emirates, Ukraine, Vatican City – with the Ukraine being a valid counterexample. Unfortunately, although at least Spain and Greece are part of the Schengen Area, they are returned among the top results. On the other hand, countries like Romania, that are not part of the Schengen Area, are missing. This brings us to the questions *Why are countries missing that do constitute a counterexample?* and *Why are countries returned that do not constitute a counterexample?* One explanation is the limited validity of our assumptions 2 and 3 from Section 4.2: on the one hand, web pages *do not* always contain search terms corresponding to attributes that the objects they describe *do* have (e.g., *Schengen* is not mentioned on the Wikipedia pages of Spain and Greece), and on the other hand, web pages *do* sometimes mention terms corresponding to attributes that the objects *do not* have (e.g., the term *Schengen* is mentioned on the Wikipedia page of Romania, because it is mentioned that the country wants to join the Schengen Area<sup>23</sup>).

<sup>23</sup> <http://en.wikipedia.org/wiki/Romania>

These examples also show that, even for a human expert, it is often not sufficient to rely on the information about objects that can be found on their web pages. Sometimes it is necessary to investigate further web pages.

## 6 Conclusion

In this paper we have indicated that there is a potentially wide range of options to employ the web for attribute exploration: querying search engines, asking questions on social question answering or crowdsourcing platforms, or retrieving counterexamples from the linked open data cloud. All of these approaches have their limitations, for some of them we provided examples and explanations. In all cases we have to cope with the open world assumption, i.e., in principle we can not assume from the absence of a fact that the fact is not true.

Nevertheless, the approach presented in this paper can ease the attribute exploration process by automatically posing queries to the web that a human would start with to find counterexamples. The approach can also be employed for learning, such that students can interactively investigate the topic of interest and at the same time learn to search the web, understand the underlying mechanism, and learn to judge the quality of the results.

Future extensions of our approach could mitigate some of the limitations:

- The user could provide more information about the objects and the attributes which could be incorporated into the query (like the additional query terms we have added in Section 5.2).
- A combination of the different sources could improve the efficiency of the process. E.g., one could use structured knowledge bases to automatically find counterexamples and only if none could be found query a web search engine. Based on the results, the user could then decide if the implication holds or not or she could forward the question to a social question answering platform or a crowdsourcing service to see if other people know the answer.
- The retrieved information could be subject of further automated analysis. E.g., web pages might be analyzed by deep semantic analysis tools to detect the presence or absence of the wanted information more reliably.

## References

1. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, Berlin/Heidelberg, 2007.
2. Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York, NY, USA, 2003.
3. Franz Baader, Bernhard Ganter, Baris Sertkaya, and Ulrike Sattler. Completing description logic knowledge bases using formal concept analysis. In *Proceedings of the 20th international joint conference on Artificial intelligence*, pages 230–235, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

4. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, 2009.
5. P. Cimiano and S. Staab. Learning by googling. *ACM SIGKDD Explorations Newsletter*, 6(2):24–33, 2004.
6. Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.
7. J.-L. Guigues and V. Duquenne. Familles minimales d’implications informatives résultant d’un tableau de données binaires. *Mathématiques et Sciences Humaines*, 95:5–18, 1986.
8. Steve Harris and Andy Seaborne. SPARQL 1.1 query language. W3C proposed recommendation, W3C, November 2012. <http://www.w3.org/TR/sparql11-query/>.
9. Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.
10. Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From SHIQ and RDF to OWL: the making of a web ontology language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1):7–26, 2003.
11. Bjoern Koester. *FooCA: web information retrieval with formal concept analysis*. Beiträge zur begrifflichen Wissensverarbeitung. Verlag Allgemeine Wissenschaft, Mühlthal, 2006.
12. Sergei Obiedkov, Derrick G. Kourie, and J.H.P. Eloff. Building access control models with attribute exploration. *Computers and Security*, 28(1–2):2–7, 2009.
13. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.
14. Eric Prud’hommeaux and Andy Seaborne. SPARQL query language for RDF. W3C recommendation, W3C, January 2008. <http://www.w3.org/TR/rdf-sparql-query/>.
15. Alexander J. Quinn and Benjamin B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1403–1412, New York, NY, USA, 2011. ACM.
16. Sebastian Rudolph. Exploring relational structures via  $\mathcal{FL}\mathcal{E}$ . In Karl Erich Wolff, Heather D. Pfeiffer, and Harry S. Delugach, editors, *Proceedings of the 12th International Conference on Conceptual Structures (ICCS 2004)*, volume 3127 of *Lecture Notes in Computer Science*, pages 196–212. Springer, 2004.
17. Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM.
18. Rudolf Wille. Restructuring lattice theory: an approach based on hierarchies of concepts. In Ivan Rival, editor, *Ordered sets*, pages 445–470, Dordrecht–Boston, 1982. Reidel.
19. Serhiy A. Yevtushenko. System of data analysis Concept Explorer. In *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, pages 127–134, Russia, 2000. (In Russian).

# The Detection of Outlying Fire Service's Reports

## FCA Driven Analytics

Adam Krasuski<sup>1</sup> and Piotr Wasilewski<sup>2</sup>

<sup>1</sup> Section of Computer Science  
The Main School of Fire Service

ul. Słowackiego 52/54, 01-629 Warsaw, Poland

<sup>2</sup> Institute of Informatics, University of Warsaw  
ul. Banacha 2, 02-097 Warsaw, Poland

krasuski@inf.sgsp.edu.pl, piotr@mimuw.edu.pl

**Abstract.** We present a methodology for improving the detection of outlying Fire Service's reports based on domain knowledge and dialogue with Fire & Rescue domain experts. The outlying report is considered as element which is significantly different from the remaining data. Outliers are defined and searched on the basis of domain knowledge and dialogue with experts. We face the problem of reducing high data dimensionality without losing specificity and real complexity of reported incidents. We solve this problem by introducing a knowledge based generalization level intermediating between analysed data and experts domain knowledge. In the methodology we use the Formal Concept Analysis methods for both generation appropriate categories from data and as tools supporting communication with domain experts. We conducted two experiments in finding two types of outliers in which outliers detection was supported by domain experts.

**Keywords:** outliers detection, formal concept analysis, fire service

## 1 Introduction

Each of approximately 500 Fire and Rescue Unit (JRG) of the State Fire Service of Poland (PSP) conducts around 3 fire and rescue actions on daily basis. There is a report created in the internal computer system of PSP named EWID after every single action. The reports comply to the requirements set by regulations [1]. The data collected in EWID database is divided into two sections – structured (database fields) and unstructured (description in natural language (NL)).

Every day ca. 1 500 reports flow into the Headquarter (HQ) of the State Fire Service of Poland. Due to the number of attributes which commanders have to select during the report submission (about 500), many reports have wrong or omitted information. These errors distort future statistics and impede analyses of the data. There is a special department in HQ which is delegated to run data analysis and check the correctness of the reports. Unfortunately, the large number of the reports which are to be checked, forces use of sampling. Therefore, many reports are saved in EWID with factual errors. There is also another type



of reports which should be intercepted. These are the reports describing very rare occurrence of objects involved in fire, not typical flow of events during the rescue action, unusual combination of threats at the fire ground or weird method used by commanders. This type of incidents due to their peculiarity may (in the future) result in large number of casualties. They should be analysed, discussed and in extreme cases new procedures should be introduced.

In the EWID database, there is no difference between these two types of the reports in question (misspelled and real outlier). From data representation point of view, they contain rare attribute value or attribute values combination. Therefore, the methods focused on detecting them should be generally similar or the same. Whereas successful detecting of this atypical reports may result in improvement of public safety and more reliable analysis of the data.

The EWID till now has collected approximately 7 million incidents. Undoubtedly, EWID is the rich source of information about threats and appropriate but also incorrect methods of their elimination. However, without doubt this database is difficult to process and analyse. The main reason for this is curse of dimensionality (500 attributes) and the necessity of processing of natural language descriptions. The simple methods like filtering, aggregating or statistical analysis do not reflect the phenomena behind the data. Therefore more sophisticated methods are needed in order to discover the knowledge. Recently few works were published, which present the more advanced approach to analysing such data. They used the methods from data mining domain [8, 19, 12], text mining [14] or even granular computing approach [13]. However, in our opinion most promising algorithms of knowledge discovery should interact with domain experts while working. In the data analysing like Fire Service reports an expert who can interpret the semantics of data, find interesting patterns or cases and can set the direction of the research plays a pivotal rôle. The works of Poelmans et. al (see [3, 16, 15, 17]) show that combination of domain experts with tools which can pre-process the information and present it in the way convenient for the experts, may help discovering important information from the structured and unstructured data (police reports).

The Formal Concept Analysis (FCA) is a theory of data analysis which identifies conceptual structures among data sets [5]. The strength of FCA in data analysis is grouping and structuring the information hidden in dataset and its presentation in a perspective convenient for the domain experts. The selected data are presented to experts and they can recognise the interesting pattern or data structure. In the scope of analysing of the Fire Service reports, FCA structures the data, creating at the same time the concepts limited by the attributes and set of objects which possess the same attribute values. For example it may create the concept of incidents which were extinguished by the same equipment set. However, in order to recognise not trivial concepts the interaction with experts is needed.

In this paper we propose a methodology for improving the detection of outlying reports based on domain knowledge and dialogue with domain experts. We analyse reports from the database of Polish Fire Service's reports, with support

of Fire & Rescue domain experts. Presented methods are based on the Formal Concept Analysis (FCA) approach. The rest of this paper is structured as follows. In Section 2 we give the definition of outlying reports which is based on atypical emergencies and F&R methods. In Section 3 we describe the dataset. In Section 4 we present our method of the analysis focused on detecting the atypical emergencies, F&R methods or relation between them. In Section 5 we describe the experiments which we conduct to validate our methodology. The article is concluded with the interpretation of research results and the perspectives for future work.

## 2 The Outlying Reports

The State Fire Service of Poland responses many types of incidents. Main categories include: fires, road incidents, industry disasters, natural calamities, collapses. For every of these main categories we can outline the several levels of subcategories. This taxonomy complies to the regulations set by [1]. However even in the lowest subcategories of the taxonomy, experts can define, according to domain knowledge, particular subclasses of similar events. There are also incidents which cannot be categorized or attached to a particular subclass even fuzzily defined. These cases are labeled by domain experts as unusual incidents, in this paper we will refer to them as **atypical events**. Such events are represented in EWID system by outliers (outlying reports). The main reason for outliers generation is presented in Introduction. In this section we define and categorize the outliers.

For the sake of clarity we need to specify the concepts used in this paper. By an **emergency** we understand the event that poses an immediate risk to health, life, property or environment, requiring urgent intervention of Fire Service, which takes place before rescue unit arrival. By an **emergency scene** we understand location in which emergency occurs, together with all persons, objects or elements involved in that emergency, as it is understood in firefighting theory. We define **fire and rescue (F&R) methods** as the set of all activities undertaken by Fire Service at the emergency scene. We define **F&R action** as all the methods used by the firefighters together with a course of emergent circumstances which take place after fire unit arrival, possibly as a result of application of F&R methods. An **incident** is an event which consists of both emergency and F&R action. A **report** is an information unit stored in the EWID system which describes a singular incident. The **outlying report** is a surprising veridical report which appears to be inconsistent with the subclass it should belong to.

In Table 1 we propose the categorisation of the outliers consisting of forms, kinds and sources. Since reports are computer representations of real phenomena, therefore outlying reports can be generated due to three different reasons: rare report occurrence with respect to other reports stored in the system (connected with reports themselves), atypicality of reported real phenomena according to domain experts knowledge (connected with represented phenomena) and incor-

**Table 1.** The forms, kinds and sources of outliers.

Atypicality form	Atypicality kind	Atypicality source
1. Atypical emergency.	1.1. Very rare occurrence in dataset.	1.1.1. Incorrect report submission.
		1.1.2. Real outlier.
	1.2. Unusual combination or number of elements, threats or objects at the emergency scene.	1.2.1. Incorrect report submission.
		1.2.2. Real outlier.
	1.3. Other circumstances.	1.3.1. Incorrect report submission.
		1.3.2. Real outlier.
2. Atypical F&R method.	2.1. Method does not exist in the firefighting theory (amateurish or innovative methods).	2.1.1. Incorrect report submission.
		2.1.2. Real outlier.
3. Atypical relationship between emergencies and methods.	3.1. Standard emergency & atypical method used.	3.1.1. Incorrect report submission.
		3.1.2. Real outlier.
	3.2. Atypical emergency & standard method used.	3.2.1. Incorrect report submission.
		3.2.2. Real outlier.
	3.3. Atypical emergency & atypical method used.	3.3.1. Incorrect report submission.
		3.3.2. Real outlier.

rectness of report submissions (connected with a representation relation between reports and modeled phenomena).

In Table 1 we pointed that a report can be classified as atypical because of three categories: 1) as emergency itself since it is unusual combination or it contains unusual number of elements, threats or objects at the fire ground from the perspective of fire service domain knowledge or it occurs very rarely in dataset, 2) as containing amateurish or innovative F&R methods from the perspective of fire service domain knowledge, 3) as containing atypical relationship between emergencies (parts of incidents) and F&R methods with specified three kinds of this relationship. In the case of first two categories, in searching for atypicality some standard universal methods (statistical, data mining or machine learning

methods) can be used together with F&R knowledge domain oriented methods. In the case of third category, searching for atypicality is more complex since atypicality here depends on relations. Emergencies or F&R methods are atypical with respect to other F&R methods or emergencies appearing in a given incident. Finding the atypical relationship between emergencies and methods using threats becomes possible. According to the firefighting theory any threat can not be left without reaction. Therefore if in some incident the threat was identified and there is no information about taking the appropriate action, then such incident can be classified as atypical.

### 3 Dataset

Our dataset consists of 291 683 F&R reports. They contain information about the incidents which Fire Service respond, from the years 1992 to 2011. Our set of the reports concerns the incidents which happend in Warsaw City and its surroundings. In this dataset 136 856 reports represent fires, 123 139 local threats and 31 688 false alarms.

Each of the reports consists of an attribute section and a natural language part. The attribute section contains 506 attributes fitted to describe all type of incidents. However depending on category of the incident, the number of non-empty attributes varies from 120 to 180 for the report. Most of the attributes are boolean (True/False) type but there are also numerical values (i.e. fire area, amount of water used).

The natural language (NL) part is an extension to the attribute part. It was designed to store information, which can not be represented in a form of a set of attributes. Unfortunately there is no clear regulation what should be written in the NL part. Therefore, in this part the full spectrum of information, from the detailed information including the time coordinates to the very general and brief descriptions can be found. The simple statistic reveals that NL part contains approximately three sentences that describe the situation at the fire ground, actions undertaken and weather conditions.

In factual aspects the data stored in the EWID contain information about persons, objects involved in the incident and methods used to eliminate the arisen threats.

In our experiments we used subset of this dataset. For the labeling (assigning threats), we selected by domain experts only reports which represent the fire of residential buildings category. The set consists of 31 556 reports. From this set 302 reports were labeled by the experts. We used these reports in our experiments described in Section 5.

### 4 Method

The biggest issue in analysing the data was the large number of dimensions. It leads to higher computational complexity, scalability problems and results in computing difficulties (huge hardware resources are needed). The vast number

of dimensions makes communication with domain experts harder or even impossible due to limited cognitive resources of experts minds (such as attention and working memory). Experts are able to elaborate in a given moment relatively small numbers of attributes. This also decreases usability of conceptual lattices as visual tools supporting communication with domain experts. The dimensionality reduction by a domain experts driven attribute selection does not reflect the real complexity of the incidents. Moreover, in this way we lose the possibility of finding the outliers (for example when the incident has a very rare attribute not considered by the domain experts).

To solve the problem of dimensional complexity in searching dataset for outliers, we decide to add some more abstract (generalization) layer intermediating between analysed data and experts domain knowledge. This layer objective is to reduce the number of dimensions and keeping the specificity of the modeled phenomena at the same time. To construct the generalization layer, we chose threats which can appear at the emergency scene and objects which can suffer from these threats. Further we will refer to this generalization layer as *threats layer*.

The main goal of Fire Services activity at the fire ground is elimination or neutralisation of arisen threats. The specific emergency generate the specific threats. Similar emergency should generate similar threats. If for the similar emergencies (for example from the same category) there exists one with significantly different number of threats or the combinations of threats, then such an emergency can be described as atypical and might be treated as an outlier. Either the emergency is very rare or its internal structure of attributes is unusual. Our approach to searching for atypicality of emergency is based on threats layer. As we pointed out in the Section 2, searching for atypical relationships is more complex however it becomes possible by using threats layer. Since no threat can be left without a proper reaction therefore, if in some incident where a threat was identified and there is no information about taking any appropriate action, then this report is considered as a disruption of the relationship between threats and methods.

This approach allows us to reduce significantly the number of dimensions without losing the information about complexity of the real phenomena. However in our reports database there is no information about threats related to the specific emergency. Our next step was labeling the reports by domain experts with appropriate threats generated by reported emergency. To eliminate this issue, we used the tactic of German Fire Service [2, 6]. After arriving at a fire ground or an emergency scene German commanders have to evaluate and recognise the appearing threats. In order to do this systematically and not to miss any of the threats they have to fulfill the Threats Matrix (in German – Gefahrenmatrix) [6]. The Threats Matrix helps to identify the threats emerging at the scene and the threatened objects. This information plays a pivotal rôle in planning the further action. Having this information, commanders can recognize the primary danger that has to be eliminated at the outset and difficult point

of action. The columns of the matrix represent threats, and the rows represent objects which can be threatened. The Table 4 depicts the Threats Matrix.

**Table 2.** The Threats Matrix used by German commanders. Legend: A1 – Fear, A2 – Toxic smoke, A3 – Radiation, A4 – Fire spreading, C – Chemical substances, E1 – Collapse, E2 – Electricity, E3 – Disease or injury, E4 – Explosion

Threat/object	A1	A2	A3	A4	C	E1	E2	E3	E4
People									
Animals									
Environment	-					-	-	-	
Property	-	-						-	
Rescuers									
Equipment	-	-						-	

In German language, column names are chosen so that they can be easily remembered. In order to help to memorize all threats by commanders, German threats' names were taken to form the following pattern: AAAA-C-EEEE *Angstreaktion, Atemgifte, Atomare Strahlung, Ausbreitung, Chemische Stoffe, Einsturz, Elektrizität, Erkrankung, Explosion*. The sign '-' in table indicates, that this threat in general can not threaten this object. At the background of fulfilled Threats Matrix, German commanders define the Threat Focus and according to them organize their commanding. This tactic method is not used by the Polish Fire Services. In order to apply the Threats Matrix to EWID, we labeled the reports manually. The reports were analysed and labeled by students of the Main School of Fire Service Warsaw (abbreviation from Polish SGSP) that educates the officers of State Fire Service. Among the SGSP students there are also extramural students with commanding experience. From students who agreed to participate in our research, we selected three commanders having at least seven years experience in commanding. They were involved as *experts – practitioners* in labeling real action reports from EWID system.

We created the special system for reports labeling. Labeling process consists of two main phases: *tutorial phase* and *labeling phase*. Tutorial phase was focused on introducing the Threats Matrix and form of EWID incidents reports to experts. It was divided in to three consecutive parts. In the first part, experts were informed about Threats Matrix. In the second part a particular completed and discussed Threats Matrix was presented to experts. In the third part, experts received an exemplary EWID report together with Threats Matrix describing this report. Labeling phase consisted of many evaluating stages. In every evaluating stage experts were provided with one EWID report. On the ground of the information about incident described in the report, they were asked to evaluate threats which appeared during reported incident and to complete its Threats Matrix. Every expert was asked to label at least 100 EWID reports. Every report description was labeled by only one expert. In total we collected 302 labeled

incident descriptions. From this dataset we created the FCA lattices and presented them to the SGSP’s teachers, considered as *experts – theoreticians*. These teachers educate students in Tactic courses. They gave us their remarks how to rebuild the lattice or indicated interesting concepts.

The rest of the paper presents two examples of finding outliers supported by domain knowledge and using FCA methods.

## 5 Experiments

In this section we present the experimental results of the validation of our approach. At this stage of our research, the experiments conducted were mainly focused on evaluation of correctness of our model. Therefore, the obtained results should be interpreted as the preliminary results. We describe the experiments designed to evaluate the efficiency of reports’ detection with two forms of atypicality: emergency and relationship. We conducted the experiments on described dataset with Concept Explorer<sup>3</sup> application version 1.3.

### 5.1 The Detection of Atypicality in Emergencies

As published in [7] there are three fundamental approaches to the problem of outlier detection: unsupervised (clustering), supervised (classification) and semi-supervised. In the current experiment we tried to detect the outlying reports according to the last category – a semi-supervised recognition [4, 11]. This approach needs pre-classified data but it only learns data marked as normal. It is suitable for static or dynamic data as it only learns one class which provides the model of normality. Systems which implement the approach, recognise an exemplar as normal if it lies within the (normality) boundary and as an outlier if it lies outside the boundary.

Taking into consideration the Fire Service’s reports, the approach requires firstly the definition of the standard emergency – model of normality. The concept of the standard emergency is very difficult to define, mainly due to the variety of the emergencies (from fire to local threats). Despite narrowing the scope of emergencies to the lowest category (e.g. fire of residential buildings) we still have the problem with definition of the normality for this category.

We asked the domain experts (SGSP’s instructors) to outline the standard scenario of residential buildings fires category. They were not able to solve this issue. They quoted many aspects of the construction of the buildings, thermal insulation existence, access to the building and equipment of firefighters which differentiate the emergencies. According to them, it was impossible to define the standard emergency (scenario) for such a category.

To eliminate the problem we constructed the formal context from the reports labeled with the threats by the commanders (extramural students). In this context the threats represented the attributes and the incidents – objects. Next, we created lattice from this context. Figure 1 depicts the lattice created.

<sup>3</sup> <http://conexp.sourceforge.net/>

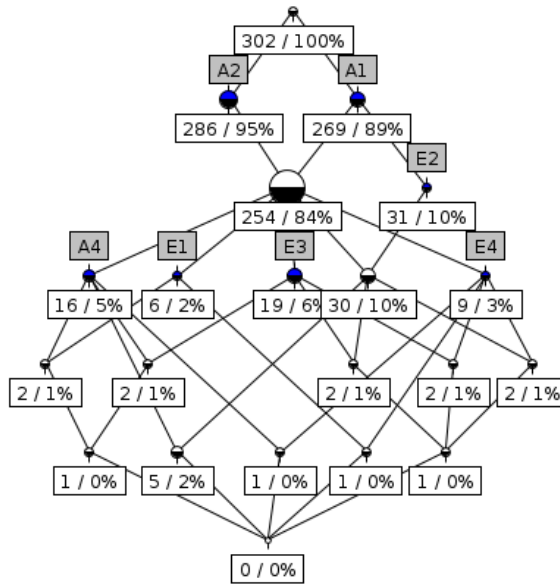


Fig. 1. The line diagrams of threats-incidents lattice.

The obtained lattice is very expressive. In the top-center of the lattice the large node is located. The node defines the concept of emergencies with two threats: A1 – Fear reaction and A2 – Toxic smoke. The node has a large number of own formal objects – 190 (63%). The own objects are not contained within any of its sub-concepts [18]. The features of the node made us believe that the concept related to the node, may define the standard (normal) emergency for the given category. To verify this, we asked the same domain experts (SGSP's instructors) whether these two threats occur mainly at the fire ground of residential buildings. They confirmed. Next, we asked them if the emergency with these two types of threats may be interpreted as the standard emergencies. They agreed to it with one exception. There is another type of threat which should be mentioned while the residential buildings are on fire: A4 – Fire spreading.

To face the problem of discrepancy between the concept lattice and the opinion of the domain experts, we examined the issue in more detail. Finally we realised that these differences are due to perceiving the threats by theoretical (SGSP's instructors) and practical (commanders – extramural students) experts – who labeled the incidents. The practitioners assign the A4 threat only when there is no sufficient team or equipments to extinguish a fire. The theorists state



that the A4 threat always exists if there is a combustible material near the one involved in the fire.

This statement confirmed that we can treat the emergencies with A1, A2 threats as a standard. The large node in the lattice (Figure 1) defines this concept. According to the definition of semi-supervised outlier detection, all the incidents outside these concepts should be treated as outlying reports. However, in our opinion the notion of normality is not crisp but fuzzy. It means that the reports lying closer to the concept of standard emergency are less likely to be outlier than those more distant. To validate our model we analysed in details, with help of the domain experts, all the incidents located far away from the standard emergency concept node. Each of those incidents had assigned more than four threats.

There were 9 emergencies in the lattice which contained more than four threats. According to the experts, 4 out of the 9 incidents were correctly submitted and they weren't real outliers. Rest of them were labeled as potential outliers.

The first report from the set was outlier of category 1.2 (see Table 1). It contained information about a fire of residential building. The owner of this apartment stored the explosive materials inside the apartment. There was an explosion reported during the rescue action. The objects involved in the fire and the scenario allowed us to treat the case as the real outlier. The emergency was so rare and at the same time dangerous, that was chosen for further discussion during courses.

The second report should be also considered as the outlier. However, its atypicality does not satisfy the classification rules presented in Table 1. Its atypicality was caused by the shortcomings of our methodology. The student who labeled the report, assigned too many threats to the common basement fire. This also implies that the methodology is somehow self-controlled.

The next case was also correctly detected as an outlier. However, the atypicality stemmed from improper relationship between the emergency and the methods. There was a overvaluation in equipment. The small fire of residential building involved 6 fire appliances and 27 rescuers. The source of atypicality was difficult to settle. It should have been caused by the incorrect report submission or it was really the wrong F&R method. To clarify this issue, we should have contacted the officer in charge during this incident.

The last two cases were outliers in the category of an atypical emergency, caused by the incorrect report submission. They were wrongly assigned to the category of residential building fires. The first of these two reports was a fire of garden gazebo, the second a small carpenter's workshop located in the residential property. Both of them should be allocated to another category.

The presented experiment demonstrates that there is a potential in detection of outlying reports with utilizing FCA approach. However, in order to evaluate its effectiveness, we can only use the precision measure. In this experiment precision equaled 0.55. At the current level of our research the other measures can not be

calculated. We have not yet calculated how many outlying reports contains the dataset.

## 5.2 The Detection of Atypicality in the Relationship

In the second experiment we concentrated on a recognition of the outlying reports caused by atypical relationship between emergencies and methods used to eliminate them.

In compliance with the firefighting theory, all the threats which occurred at the scene should not be left without proper action. If some of the threats exist and there is no information in the report regarding the action focusing on their elimination, we can suspect that the report is an outlier. In contrary, if there is no information about some threat and there is information in the report about the method used for its elimination, we may consider this report as an outlier. In this experiment the main focus was on first issue: threats without proper F&R methods. For each of incidents labeled by students we extracted information about the used F&R methods. We chose only methods which are associated with residential building fire. That means: extinguishing, evacuation (all types of objects from Threats Matrix) and smoke removal. Figure 2 depicts the concept lattice for this subset.

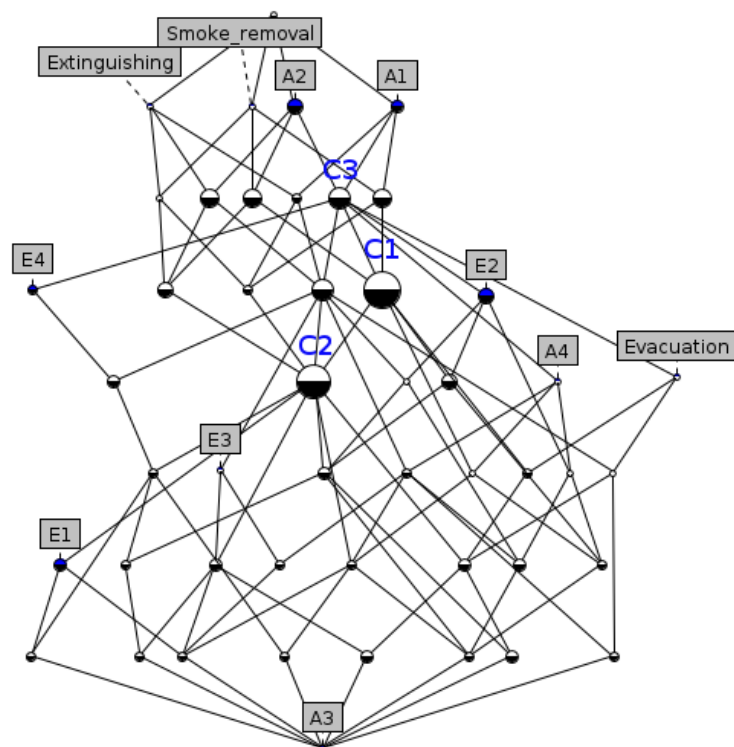
In the lattice there are three large nodes which we took for the further examination. There are respectively: C1 – node with 55 own objects, C2 – node with 40 own object and C3 – node with 17 own objects. They define the formal concepts, which we describe as:

- C1 – emergencies where A1, A2 threats exist and only smoke removal was performed (Figure 3 a)),
- C2 – emergencies where A1, A2 threats exist and extinguishing and smoke removal were performed,
- C3 – emergencies where A1, A2 threats exist and there weren't any rescue activities (Figure 3 b)).

The formal concept C2 represents a proper relationship between emergencies and F&R methods. There was a fire and the firefighters undertook adequate actions (extinguishing and smoke removal). The formal concepts C1 and C3 reveal some peculiar scenarios. There was a fire and only smoke was removed (C1); and there was a fire and there were no activities performed by Fire Service (C3). These both types were considered as outliers. However, the large number of own objects in these concepts (72) indicated that the problem was more systemic.

After deeper investigation it appeared that the problem was related to the definition of the attributes in EWID system. In the system there are three attributes (without natural language part) allocated to store information about extinguishing: *water stream used in the attack*, *water stream used in the defence* and *amount of extinguishing agent used*.

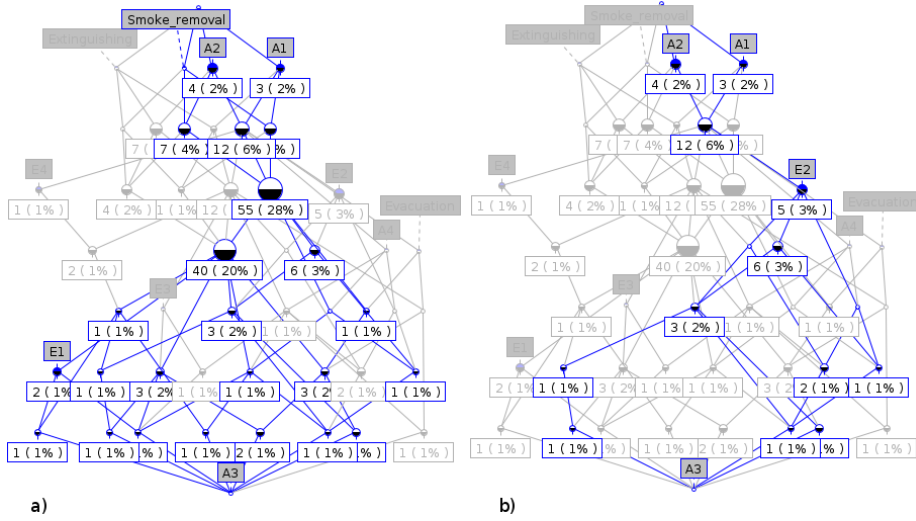
The reports that belonged to the C1 or C3 category were mostly small fires i.e. cooking meals left on an oven unattended. The firefighters extinguished this



**Fig. 2.** The line diagrams of lattice of the threats, F&R methods and incidents.

type of fire using water from a tap or a bucket. According to the commanders opinions, who managed the firefighting actions and submitted the reports, these activities did not meet any of the outlined attributes. They used neither water stream nor water from fire appliances. Therefore, all the attributes were left empty. This selected concepts did not define the outlying reports according to our classification. However, the problem has negative impact on the statistical analysis that why it can be the starting point for the further improving of the EWID system.

To detect the atypical relationship between emergencies and F&R methods we performed some extended analysis. We tried to find the description of firefighting activities in NL section. The method is based on selecting by the experts the set of 17 words which may express the extinguishing activities. Firstly we lemmatized the NL part of the reports. The lemmatization allowed us to recognise the selected words, even if they were in the inflexed form. Then, we created the Document Term Matrix (DTM). The rows of this matrix represented the reports, the columns the set of words which appeared at least once in the NL part. In order to obtain one attribute that express extinguishing activities, we sliced the DTM, selecting only columns which contained words from the experts



**Fig. 3.** The line diagrams of lattice of threats, F&R methods and incidents with selected concepts C1 and C3.

set. Then we run a logical OR on the previously selected columns. We obtained one column which represented the extinguishing activities mentioned in the NL part of the report. Finally, we perform a logical OR of this column with the columns from attribute section which represented the extinguishing. The final column showed the extinguishing actions marked either in the attribute part or NL part. We updated our formal context and created a new lattice. Figure 4 depicts the obtained results.

According to the lattice, there is one large node which represents a concept of most often appearing threats and proper F&R methods of their elimination. There are two nodes left (C4, C5) with 3 own objects where the threats exist and there are no rescue activities. After more detailed analysis done in the cooperation with experts, we came to the conclusions that they were false alarms. That means that they should be categorised as outlying reports caused by the incorrect report submission.

## 6 Conclusions and Further Work

The Incident Data Reporting Systems (in our case EWID) are the vast source of information. They contain description of threats which may appear at the scene and F&R methods to deal with them. In this dataset there are reports which describe very rare or atypical incidents as well as methods which are not in accordance with firefighting theory. Those reports should be detected and analysed to avoid serious accidents in the future.



methods. Moreover, FCA assisted in finding the systemic error in submission of the reports to the EWID system. FCA also revealed differences in perceiving the threats by the practitioners and theoreticians.

One of the most important problems, which has not been addressed yet is the scalability of our approach. The first stage of our method is based on the manual labeling of the incidents by the domain experts. This is not an issue for the German or USA Fire Services where firefighters assign threats to the incidents at the fire ground. However, in the case of the State Fire Service of Poland there is no such a procedure and we recommend assigning AAAA-C-EEEE terms to the rough reports. Before introducing this procedure, the problem is complex and requires multi-label classification and should be solved in the further work – hopefully we can succeed in this field, due to our experience in this domain, including involvement in projects [10] and data mining competitions [9]

Discovering the outliers in the whole dataset at once would be problematic. Analysing 7 million of incidents in one context would be impossible due to system resources and restrictions of FCA-presenting the analysis to the experts. However, the primary use of the solutions is to support the HQ analysts in their daily work. Every day ca. 1 500 reports must be checked against their validity and atypicality. These incidents can be easily divided into three groups: fires, local threats and false alarms. The number of incidents in are as follows: 50% fires, 42% local threats and 8% false alarms. False alarms won't be considered, so what must be comfortably fit on computer display is ca. 750 fires and even less local threat (these two will be treated separately).

The final system detecting outlying reports should not be limited to just one module, e.g. FCA. That means, the system should contain the set of classification, clustering and other algorithms combined in a form of ensemble. The ensemble methods use multiple models to obtain better predictive performance than could be obtained from any of the constituent models<sup>4</sup>.

**Acknowledgements:** The authors would like to thank the students of NP-PS40 extramural study for their participation in this project and sharing of domain knowledge. Moreover, the authors would like to thank the Lecturers from the Section of Firefighting Operations of the SGSP, especially Lt. Col. Aleksander Adamski and Maj. Przemysław Wysoczyński.

The research has been supported by the research project realized within the Homing Plus programme, Edition 3/2011, of the Foundation for Polish Science, co-financed from the European Union Regional Development Fund, and by the grant 2011/01/D/ST6/06981 from the Polish National Science Centre.

## References

1. Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 29 grudnia 1999 r. w sprawie szczegółowych zasad organizacji Krajowego Systemu Ratowniczo-Gaśniczego. Dz.U.99.111.1311 §34 pkt. 5 i 6

<sup>4</sup> [http://en.wikipedia.org/wiki/Ensemble\\_learning](http://en.wikipedia.org/wiki/Ensemble_learning)

2. Bundesamt für Bevölkerungsschutz und Katastrophenhilfe: Feuerwehr-Dienstvorschrift 100 Führung und Leitung im Einsatz : Führungssystem. FwDV 100 Stand: 10. März 1999
3. Elzinga, P., Poelmans, J., Viaene, S., Dedene, G., Morsing, S.: Terrorist threat assessment with formal concept analysis. In: 2010 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 77–82. IEEE (2010)
4. Fawcett, T., Provost, F.: Activity monitoring: Noticing interesting changes in behavior. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 53–62. ACM (1999)
5. Ganter, B., Wille, R.: Formal concept analysis. Springer Berlin (1999)
6. Graeger, A., Cimolino, U., de Vries, H., Stümersen, J.: Einsatz-und Abschnittsleitung: Das Einsatz-Führungs-System (EFS). Ecomed Sicherheit (2009)
7. Hodge, V., Austin, J.: A survey of outlier detection methodologies. *Artificial Intelligence Review* 22(2), 85–126 (2004)
8. Holmes, M., Wang, Y., Ziedins, I.: The application of data mining tools and statistical techniques to identify patterns and changes in fire events (2009)
9. Janusz, A., Nguyen, H., Ślęzak, D., Stawicki, S., Krasuski, A.: JRS'2012 Data Mining Competition: Topical Classification of Biomedical Research Papers. In: Proceedings of the Rough Sets and Current Trends in Computing (RSCTC 2012), Chengdu, China, August 17-20, 2012. Lecture Notes in Computer Science, vol. 7414, pp. 422–431. Springer (2012)
10. Janusz, A., Świeboda, W., Krasuski, A., Nguyen, H.S.: Interactive Document Indexing Method Based on Explicit Semantic Analysis. In: Proceedings of the Joint Rough Sets Symposium (JRS 2012), Chengdu, China, August 17-20, 2012. Lecture Notes in Computer Science, vol. 7415, pp. 156–165. Springer (2012)
11. Japkowicz, N., et al.: A novelty detection approach to classification. In: International Joint Conference on Artificial Intelligence. vol. 14, pp. 518–523. Lawrence Erlbaum Associates LTD (1995)
12. Krasuski, A., Kreński, K., Łazowy, S.: A Method for Estimating the Efficiency of Commanding in the State Fire Service of Poland. *Fire Technology* 48(4), 795–805 (2012)
13. Krasuski, A., Ślęzak, D., Kreński, K., Łazowy, S.: Granular knowledge discovery framework. *New Trends in Databases and Information Systems* pp. 109–118 (2012)
14. Kreński, K., Krasuski, A., Łazowy, S.: Data mining and shallow text analysis for the data of state fire service. In: Proceedings of Concurrency, Specification and Programming - XXth International Workshop, CS&P 2011. pp. 313–321 (2011)
15. Poelmans, J., Elzinga, P., Dedene, G., Viaene, S., Kuznetsov, S.: A concept discovery approach for fighting human trafficking and forced prostitution. *Conceptual Structures for Discovering Knowledge* pp. 201–214 (2011)
16. Poelmans, J., Elzinga, P., Viaene, S., Van Hulle, M., Dedene, G.: Gaining insight in domestic violence with emergent self organizing maps. *Expert systems with applications* 36(9), 11864–11874 (2009)
17. Poelmans, J., Elzinga, P., Neznanov, A., Viaene, S., Kuznetsov, S.O., Ignatov, D.I., Dedene, G.: Concept Relation Discovery and Innovation Enabling Technology (CORDIET). *CoRR* abs/1202.2895 (2012)
18. Willmor, D., Embury, S.M.: An intensional approach to the specification of test cases for database applications. In: Proceeding of the 28th international conference on Software engineering. pp. 102–111. ACM Press, New York (2006)
19. Xiangxin, L.: Rational judging method of fire station layout based on data mining. In: 2nd IEEE International Conference on Emergency Management and Management Sciences (ICEMMS), 2011. pp. 455–458. IEEE (2011)

# An Approach to Incremental Learning Good Classification Tests

<sup>1</sup>Xenia Naidenova, <sup>2</sup>Vladimir Parkhomenko

<sup>1</sup>Military Medical Academy, Saint-Petersburg, Russian Federation

ksennaid@gmail.com

<sup>2</sup>Saint-Petersburg State Polytechnic University, Russian Federation

parhomenko.v@gmail.com

**Abstract.** An algorithm of incremental mining implicative logical rules is proposed. This algorithm is based on constructing good classification tests. The incremental approach to constructing these rules allows revealing the interdependence between two fundamental components of human thinking: pattern recognition and knowledge acquisition.

**Keywords:** Incremental learning, Good classification test, Pattern recognition, Machine learning, Human mental operations

## 1 Introduction

Methods of incremental symbolic machine learning are developing in several directions. The first one is to construct incrementally concept lattice [1-3]. In [3], an incremental algorithm to construct a lattice from a collection of sets is derived, refined, analyzed, and related to a similar previously published algorithm for constructing concept lattices. The second direction is related to incremental mining association rules [4-5]. This direction includes the incremental approach to mining frequent itemsets based on Galois Lattice Theory [6]. Significantly fewer investigations are devoted to incremental mining logical rules. Utgoff proposed three incremental decision tree induction algorithms [7]. Rough set based incremental method is advanced in [8].

This paper is devoted to incremental learning of logical rules in the form of implications based on the concept of good classification test. Incremental learning is considered not only as a model of inductive human reasoning for implicative logical rule generation but also as an essential part of pattern recognition processes.



## 2 The Concept of Good Classification Test

Let  $G = \{1, 2, \dots, N\}$  be the set of objects' indices (objects, for short) and  $M = \{m_1, m_2, \dots, m_j, \dots, m_m\}$  be the set of attributes' values (values, for short). Each object is described by a set of values from  $M$ . The object descriptions are represented by rows of a table  $R$  the columns of which are associated with the attributes taking their values in  $M$ . Let  $R(+)$  and  $G(+)$  be the sets of positive object descriptions and the set of indices of these objects, respectively. Then  $R(-) = R/R(+)$  and  $G(-) = G/G(+)$  are the set of negative object descriptions and the set of indices of these objects, respectively.

The definition of good tests is based on two mapping  $2^G \rightarrow 2^M$ ,  $2^M \rightarrow 2^G$  determined as follows.  $A \subseteq G$ ,  $B \subseteq M$ . Denote by  $B_i$ ,  $B_i \subseteq M$ ,  $i = 1, \dots, N$  the description of object with index  $i$ . We define the relations  $2^S \rightarrow 2^T$ ,  $2^T \rightarrow 2^S$  as follows:  $A' = \text{val}(A) = \{\text{intersection of all } B_i; B_i \subseteq M, i \in G\}$  and  $B' = \text{obj}(B) = \{i: i \in G, B \subseteq B_i\}$ . These mapping are Galois's correspondences [9]. Of course, we have  $\text{obj}(B) = \{\text{intersection of all } \text{obj}(m); \text{obj}(m) \subseteq G, m \in B\}$ . Operations  $\text{val}(A)$ ,  $\text{obj}(B)$  are reasoning operations (derivation operations).

The generalization operations  $\text{generalization\_of}(B) = B'' = \text{val}(\text{obj}(B))$  and  $\text{generalization\_of}(A) = A'' = \text{obj}(\text{val}(A))$  are actually closure operators [9]. A set  $A$  is closed if  $A = \text{obj}(\text{val}(A))$ . A set  $B$  is closed if  $B = \text{val}(\text{obj}(B))$ .

Notice that these generalization operations are also used in FCA [10], [11]. For  $g \in G$  and  $m \in M$ ,  $\{g\}'$  is denoted by  $g'$  and called **object intent**, and  $\{m\}'$  is denoted by  $m'$  and called **value extent**.

**Definition 1. A diagnostic (classification) test** for  $R(+)$  is a pair  $(A, B)$  such that  $B \subseteq M$  ( $A = \text{obj}(B) \neq \emptyset$ ),  $A \subseteq G(+)$  and  $B \not\subseteq \text{val}(g) \ \& \ B \neq \text{val}(g)$ ,  $\forall g, g \in G(-)$ . Equivalently,  $\text{obj}(B) \cap G(-) = \emptyset$ .

In general case, a set  $B$  is not closed for diagnostic test  $(A, B)$ , consequently, diagnostic test is not obligatory a concept of FCA [12].

To say that a collection  $B$  of values is a diagnostic test for the set  $R(k)$  is equivalent to say that it does not cover any object description belonging to the classes different from  $k$ . At the same time, the condition  $\text{obj}(B) \subseteq G(k)$  implies that the following implicative dependency is true: 'if  $B$ , then  $k$ ' and, consequently, a diagnostic test, as a set of values, makes up the left side of an implication.

It is clear that the set of all diagnostic tests for a given set  $R(k)$  (call it ' $DT(k)$ ') is the set of all  $B$  such that the condition  $\text{obj}(B) \subseteq G(k)$  is true. For any pair of diagnostic tests from  $DT(k)$  only one of the following relations is true:  $\text{obj}(B_i) \subseteq \text{obj}(B_j)$ ,  $\text{obj}(B_i) \supseteq \text{obj}(B_j)$ ,  $\text{obj}(B_i) \approx \text{obj}(B_j)$ , where the last relation means that  $\text{obj}(B_i)$  and  $\text{obj}(B_j)$  are incomparable, i.e.  $\text{obj}(B_i) \not\subseteq \text{obj}(B_j)$  and  $\text{obj}(B_j) \not\subseteq \text{obj}(B_i)$ . This consideration leads to the concept of a good diagnostic test: they are maximal elements of partially ordered set  $DT(k)$ .

**Definition 2. A classification test**  $(A, B)$ ,  $B \subseteq M$  ( $A = \text{obj}(B) \neq \emptyset$ ) is **good** for  $R(+)$  if and only if any extension  $A^* = A \cup i$ ,  $i \notin A$ ,  $i \in G(+)$  implies that  $(A^*, \text{val}(A^*))$  is not a test for  $R(+)$ .

**Definition 3.** A good diagnostic test  $(A, B)$ ,  $B \subseteq M$  ( $A = \text{obj}(B) \neq \emptyset$ ) for  $R(+)$  is **irredundant** if any narrowing  $B^* = B \setminus m$ ,  $m \in B$  implies that  $(\text{obj}(B^*), B^*)$  is not a test for  $R(+)$ .

**Definition 4.** A good diagnostic test for  $R(+)$  is **maximally redundant** if any extension of  $B^* = B \cup m$ ,  $m \notin B$ ,  $m \in M$  implies that  $(\text{obj}(B^*), B^*)$  is not a good test for  $R(+)$ .

It is possible to show that good maximally redundant tests (GMRTs) are closed, consequently, they are formal concepts in term of the FCA, but they are not always frequent itemsets [12]. In what follows, we shall consider mining GMRTs.

The first algorithm for inferring all GMRTs for a given class of objects with its theoretical foundation has been proposed in [13] and analyzed in [14]. Then an algorithm ASTRA has been proposed and realized in a program system SIZIF [15]. The algorithms NIAGaRa and DIAGaRa have been described in [16] and [17], respectively. Diagnostic Test Machine (DTM) [18] is a software based on supervised mining good diagnostic tests. The experiments conducted with the publicly available dataset of 8124 mushrooms have showed that the result of the DTM turned out to be better with respect to classification accuracy (97,5%) than the results (95%) informed in [19] for the same set of data.

Any algorithm for mining GMRTs can be used as a part of incremental algorithm solving the same task.

## The Decomposition of Good Test Inferring into Subtasks

To transform good classification test inferring into an incremental process, we introduce two kinds of subtasks [15], [16]: for a given set of positive examples: 1) given a set of values  $B \subseteq M$ ,  $(\text{obj}(B), B)$  is a test, find all  $B^* \subset B$  such that  $(\text{obj}(B^*), B^*)$  is a GMRT; 2) given a non-empty set of values  $X \subseteq M$  such that  $(\text{obj}(X), X)$  is not a test, find all  $Y$ ,  $X \subset Y$ , such that  $(\text{obj}(Y), Y)$  is a GMRT.

**The subtask of the first kind.** We introduce a concept of projection  $\text{proj}(R)[t]$  of a given positive object description  $t$  on a given set  $R(+)$  of positive examples. The  $\text{proj}(R)[t]$  is the set  $Z = \{z: (z \text{ is non empty intersection of } t \text{ and } t') \& (t' \in R(+)) \& ((\text{obj}(z), z) \text{ is a test for } R(+))\}$ .

If  $\text{proj}(R)[t]$  is not empty and contains more than one element, then it is a subtask for inferring all GMRTs that are in  $t$ . If the projection contains one and only one element  $t$ , then  $(\text{obj}(t), t)$  is a GMRT.

**The subtask of the second kind.** We introduce a concept of attributive projection  $\text{proj}(R)[B]$  of a given set  $B$  of values on a given set  $R(+)$  of positive examples. The projection  $\text{proj}(R)[B] = \{t: (t \in R(+)) \& (B \text{ appears in } t)\}$ . Another way to define this projection is:  $\text{proj}(R)[B] = \{t: i \in (\text{obj}(B) \cap G(+))\}$ . If attributive projection is not empty and contains more than one element, then it is a subtask for inferring all GMRTs containing  $B$ . If  $B$  appears in one and only one object description  $t$ , then there is only one GMRT:  $(\text{obj}(t), t)$ .

The following theorem gives the foundation of reducing projections [15], [16].

**Theorem 1.** Let  $m \in M$ ,  $(\text{obj}(X), X)$  be a maximally redundant test for a given set  $R(+)$  of positive objects and  $\text{obj}(m) \subseteq \text{obj}(X)$ . Then  $m$  does not belong to any GMRT for  $R(+)$  different from  $(\text{obj}(X), X)$ .

## 1 An Approach to Incremental Inferring GMRTs

Incremental supervised learning is necessary when a new portion of observations becomes available over time. Suppose that each new object comes with the indication of its class membership. The following actions are necessary with arrival of a new object: 1) checking whether it is possible to perform generalization of some existing rules (tests) for the class to which a new object belongs (a class of positive objects, for certainty); 2) inferring all GMRTs induced by the new object description; 3) checking the validity of rules (tests) for negative objects, and, if it is necessary, modifying the tests that are invalid (test for negative objects is invalid if its intent is included in a new (positive) object description). Thus the following mental acts are performed:

- Pattern recognition and generalization of knowledge (increasing the power of already existing inductive knowledge);
- Increasing knowledge (inferring new knowledge);
- Correcting knowledge (diagnostic reasoning).

The first act modifies already existing tests (rules). The second act is reduced to subtask of the first kind. The third act can be implemented by the following ways. In the first way, we delete invalid tests (rules) and, by the use of subtask of the first kind, we must find new GMRTs generated by negative objects's descriptions that have been covered by invalid tests. In the second way, this act can be reduced to subtasks of the second kind. Then we obtain diagnostic logical assertions in the form:  $X, d \rightarrow$  negative class of objects;  $X, b \rightarrow$  positive class of objects;  $d, b \rightarrow$  false, where  $X, d, b \subset M$ , and  $X$  is object intent of invalid test .

Algorithm DIAGaRa is used for solving both kinds of subtasks. Currently, we realize the first way with deleting invalid tests.

## 2 DIAGaRa: an Algorithm for Inferring GMRTs

The decomposition of inferring GMRTs into subtasks of first and second kinds gives the possibility to construct incremental algorithms. The simplest way to do it consists of the following steps: choose object description (value), form subtask, solve subtask, delete object description (value) after the subtask is over, and check the condition of ending the main task. In this process, already obtained tests are used for pruning the search space.

DIAGaRa is based on using a basic recursive procedure for solving subtask of the first kind. The initial information for finding all the GMRTs contained in a positive example (object) description is the projection of this example on the current set  $R(+)$ . It is essential that the projection is simply a subset of examples (object descriptions)

defined on a certain restricted subset  $B^*$  of values. Let  $A^*$  be the subset of indices of objects from  $R(+)$  which have produced the projection.

Generally, it is useful to introduce the weight  $W(B)$  of any set  $B$  of values in the projection:  $W(B) = \|splus(B)\| = \|\text{obj}(B) \cap A^*\|$  is the number of positive object descriptions of the projection containing  $B$ . Let  $WMIN$  be the minimal permissible value of weight. Currently, we assume that  $WMIN = 1$ .

Let  $STGOOD$  be the partially ordered set of elements  $A \subseteq A^*$  satisfying the condition that  $(A, \text{val}(A))$  is a good test for  $R(+)$ . The basic recursive procedure consists of applying the sequence of the following steps:

**Step 1.** Check whether the intersection of all the elements of projection corresponds to a test and if so, then  $A^*$  is stored in  $STGOOD$  if  $(A^*, \text{val}(A^*))$  is currently a good test; in this case, the subtask is over. Otherwise the next step is performed.

**Step 2.** The generalization operation is performed as follows:  $B' = \text{val}(splus(m))$ ,  $m \in B^*$ ; if  $B'$  is object intent of a test, then  $m$  is deleted from the projection and  $splus(m)$  is stored in  $STGOOD$  if  $splus(m)$  is currently value extent of a good test.

**Step 3.** The value  $m$  is deleted from the projection if  $splus(m) \subseteq s$  for some  $s \in STGOOD$ .

**Step 4.** If at least one value has been deleted from the projection, then the reduction of the projection is necessary. The reduction consists in deleting the elements of projection that do not correspond to tests (as a result of previous eliminating values). If, under reduction, at least one element has been deleted from the projection, then Step 2, Step 3, and Step 4 are repeated.

**Step 5.** Check whether the subtask is over or not. The subtask is over when either the projection is empty or the intersection of all elements of the projection corresponds to a test (see, please, Step 1). If the subtask is not over, then an element of this projection is selected, new subtask is formed, and the basic algorithm runs recursively.

**An Approach for Forming the Set  $STGOOD$ .** The important part of the basic algorithm is how to form the set  $STGOOD$ . Let  $L(S)$  be the set of all subsets of the set  $S$ .  $L(S)$  is the set lattice. The ordering determined in the set lattice coincides with the set-theoretical inclusion. It will be said that subset  $s_1$  is absorbed by subset  $s_2$ , that is  $s_1 \leq s_2$ , if and only if the inclusion relation is hold between them, that is  $s_1 \subseteq s_2$ . Under formation of  $STGOOD$ , a set  $s$  is stored in  $STGOOD$  if and only if it is not absorbed by any element of this set. It is necessary also to delete from  $STGOOD$  all the elements in it that are absorbed by  $s$ . Thus, when the algorithm is over,  $STGOOD$  contains all the collections of objects that correspond to GMRTs and only such collections. Essentially the process of forming  $STGOOD$  is an incremental procedure of finding all maximal elements of a partially ordered set.

The set  $TGOOD$  of all the GMRTs is obtained as follows:  $TGOOD = \{(s, \text{val}(s)), (\forall s) (s \in STGOOD)\}$ .

### 3 INGOT: An Incremental Algorithm for Inferring All GMRTs

The first act is performed by the procedure GENERALIZATION ( $STGOOD, j^*$ ).

The procedure GENERALIZATION ( $STGOOD(+), j^*$ ).

Input:  $j^*$  is the index of new example (object), the set  $STGOOD(+)$  of GMRTs for the class of positive examples, the set  $R(-)$  of negative examples.

Output:  $STGOOD(+)$  modified by the generalization.

Begin

```
( $\forall s$ ) ( $s \in STGOOD(+)$ )
  if to_be_test( $\{s \cup j^*\}, val(\{s \cup j^*\})$ ) = true then
     $s \leftarrow$  generalization ( $s \cup j^*$ );
end
```

The second act is reduced to the subtask of the first kind. The procedure FORMSUBTASK( $j$ ) aims at preparing initial data for inferring all the GMRTs contained in description  $t$  of object with index  $j$ :

The procedure FORMSUBTASK( $j, R(class(j)), G(class(j)), STGOOD(class(j))$ ).

Input:  $j, R(class(j)), R(-) = R/R(class(j)), G(class(j)), STGOOD(class(j))$ . Output:

proj( $R(class(j))[j]$ );

**Begin**

```
proj( $R(class(j))[j]$ )  $\leftarrow$   $\{j\}$ ; nts  $\leftarrow$   $G(class(j))$ ;
( $\forall i$ )  $i \in nts, i \neq j$ 
  if to_be_test( $\{j, i\}, val(\{j, i\})$ ) = true then do
    Begin
    insert  $i$  into proj( $R(class(j))[j]$ );
    end
```

end

Four possible situations can take place when a new object comes to the learning system:

- The knowledge base is empty;
- The knowledge base contains only objects of the positive class to which a new object belongs;
- The knowledge base contains only objects of the negative class;
- The knowledge base contains objects of both the positive and the negative classes.

The second situation conforms to the generalization process taking into account only the similarity relation between examples of the same class. This problem is known in the literature as inductive inference of generalization hypotheses or unsupervised generalization. An algorithm for solving this problem can be found in [20].

Let CONCEPTGENERALIZATION [ $j^*$ ]( $G(+), STGOOD(+)$ ) be the procedure of generalization of positive examples in the absence of negative examples. Next, the procedure INGOT is presented.

The procedure  $INGOT(j^*)$ .

Input:  $j^*$ ,  $class(j^*)$ ,  $t^*$  - description of  $j^*$ -object,  $R$ ,  $G$ ,  $STGOOD = STGOOD(+) \cup STGOOD(-)$ . Output:  $STGOOD$ .

```

begin
 $k \leftarrow class(j^*)$ ;  $G(+) \leftarrow G(k)$ ;  $R(+) \leftarrow R(k)$ ;  $R(-) \leftarrow R/R(+)$ ,  $G(-) \leftarrow G/G(+)$ ;
 $N \leftarrow N + 1$ ;  $j^* \leftarrow N$ , where  $N$  is the number of objects;
 $G(+) \leftarrow j^* \cup G(+)$ ;  $R(+) \leftarrow t^* \cup R(+)$ ;
 $STGOOD(+) \leftarrow STGOOD(k)$ ;
 $STGOOD(-) \leftarrow \cup STGOOD/STGOOD(+)$ ;
if  $N = 1$  then  $STGOOD(+) \leftarrow \{j^*\} \cup STGOOD(+)$ ; else
if  $N \neq 1$  and  $\|G(+)\| = 1$  then
begin
 $STGOOD(+) \leftarrow \{j^*\} \cup STGOOD(+)$ ;
 $(\forall s), s \in STGOOD(-), val(s) \subseteq t^*$ 
 $(\forall j), j \in G(-), s \subseteq val(j)$ 
FORMSUBTASK ( $j, R(-), G(-), STGOOD(-)$ );
DIAGaRa(proj( $R(-)[j]$ ),  $STGOOD(-)$ );
end
end
else if  $N \neq 1$  and  $G(-) = \emptyset$  then
CONCEPTGENERALIZATION [ $j^*$ ]( $G(+), STGOOD(+)$ );
else /*  $N \neq 1$  and  $\|G(+)\| \neq 1$  and  $G(-) \neq \emptyset$  */
begin
if  $STGOOD(+) \neq \emptyset$  then
GENERALIZATION( $STGOOD(+), j^*$ ); end
FORMSUBTASK ( $j^*, R(+), G(+), STGOOD(+)$ );
DIAGaRa(proj( $R(+)[j^*]$ ),  $STGOOD(+)$ );
 $(\forall s), s \in STGOOD(-), val(s) \subseteq t^*$ 
 $(\forall j), j \in G(-), s \subseteq val(j)$ 
FORMSUBTASK ( $j, R(-), G(-), STGOOD(-)$ );
DIAGaRa(proj( $R(-)[j]$ ),  $STGOOD(-)$ );
end
end
end
end

```

The data in Table 1 is for processing by algorithm  $INGOT$  (Example 1) for each object description step by step.

**Table 1.** The Data for Generating GMRTs (Example 1)

Index of example	Outlook	Temperature	Humidity	Windy	Class
------------------	---------	-------------	----------	-------	-------

1	Sunny	Hot	High	No	1
2	Sunny	Hot	High	Yes	1
3	Overcast	Hot	High	No	2
4	Rain	Mild	High	No	2
5	Rain	Cool	Normal	No	2
6	Rain	Cool	Normal	Yes	1
7	Overcast	Cool	Normal	Yes	2
8	Sunny	Mild	High	No	1
9	Sunny	Cool	Normal	No	2
10	Rain	Mild	Normal	No	2
11	Sunny	Mild	Normal	Yes	2
12	Overcast	Mild	High	Yes	2
13	Overcast	Hot	Normal	No	2
14	Rain	Mild	High	Yes	1

**Table 2a.** The Records of Step-by-Step Results of the Procedure INGOT.

$j^*$	Class( $j^*$ )	$STGOOD(1), STGOOD(2)$
{1}	1	$STGOOD(1): \{\{1\}\};$
{2}	1	$STGOOD(1): \{\{1,2\}\};$
{3}	2	$STGOOD(1): \{\{1,2\}\}; STGOOD(2): \{\{3\}\};$
{4}	2	$STGOOD(1): \{\{1,2\}\}; STGOOD(2): \{\{3\}, \{4\}\};$
{5}	2	$STGOOD(1): \{\{1,2\}\}; STGOOD(2): \{\{3\}, \{4,5\}\};$
{6}	1	$STGOOD(1): \{\{1,2\}, \{2,6\}\}; STGOOD(2): \{\{3\}, \{4,5\}\};$
{7}	2	$STGOOD(1): \{\{1,2\}, \{6\}\}; STGOOD(2): \{\{3,7\}, \{4,5\}\};$
{8}	1	$STGOOD(1): \{\{1,2,8\}, \{6\}\}; STGOOD(2): \{\{3,7\}, \{4,5\}\};$
{9}	2	$STGOOD(1): \{\{1,2,8\}, \{6\}\}; STGOOD(2): \{\{3,7\}, \{4,5\}, \{5,9\}\}.$

**Table 2b.** The Records of Step-by-Step Results of the Procedure INGOT (continuation).

$J^*$	Class( $J^*$ )	$STGOOD(1); STGOOD(2)$
{10}	2	$STGOOD(1): \{\{1,2,8\}, \{6\}\};$ $STGOOD(2): \{\{3,7\}, \{4,5,10\}, \{5,9,10\}\};$
{11}	2	$STGOOD(1): \{\{1,2,8\}, \{6\}\};$ $STGOOD(2): \{\{3,7\}, \{4,5,10\}, \{5,9,10\}, \{10,11\}, \{9,11\}\};$
{12}	2	$STGOOD(1): \{\{1,2,8\}, \{6\}\};$ $STGOOD(2): \{\{3,7,12\}, \{\{4,5,10\}, \{5,9,10\}, \{10,11\}, \{9,11\}, \{11,12\}\};$
{13}	2	$STGOOD(1): \{\{1,2,8\}, \{6\}\};$

		<i>STGOOD</i> (2): {3,7,12,13},{4,5,10},{5,9,10,13},{10,11},{9,11},{11,12}
{14}	1	<i>STGOOD</i> (1):{1,2,8}, {6,14};
		<i>STGOOD</i> (2):{3,7,12,13},{4,5,10}, {5,9,10,13},{10,11},{9,11}.

In Tables 2a and 2b, the sets *STGOOD*(1) and *STGOOD*(2) accumulate the sets of objects corresponding to the GMRTs for Class 1 and Class 2, respectively, at each step of the algorithm. Table 3 contains the results of the procedure INGOT.

**Table 3.** The Sets TGOOD(1) and TGOOD(2) Produced by the Procedure INGOT

<i>TGOOD</i> (1)	<i>TGOOD</i> (2)
({1,2,8}, Sunny High)	({4,5,10}, Rain No)
({6,14}, Rain Yes)	({5,9,10,13}, Normal No)
-	({10,11}, Mild Normal)
-	({9,11}, Sunny Normal)
-	({3,7,12,13}, Overcast)

The training set of next example is in Table 4. It contains the description of 25 students (persons) characterized by positive (Class 1) and negative (Class 2) dynamics of intellectual development during a given period of time. The persons are described by factors of the MMPI method modified in Russia by L. Sobchik [21].

**Table 4.** The Training Set of Data for Example 2

	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class
1	4	3	5	3	3	4	3	4	4	2
2	4	4	5	3	3	3	2	4	4	2
3	4	3	4	3	3	3	3	3	4	2
4	3	3	4	3	3	3	3	3	4	2
5	4	3	4	3	3	4	3	3	3	2
6	5	4	5	3	4	2	4	3	3	2
7	5	3	5	4	4	3	4	4	4	2
8	4	3	4	3	3	4	3	3	3	2
1	3	3	5	4	4	4	3	4	4	1
2	2	3	4	3	3	3	3	3	3	1
3	3	3	5	3	3	2	4	4	3	1
4	3	3	4	3	4	4	2	3	5	1
5	3	3	5	4	4	4	3	4	3	1
6	4	2	4	4	4	4	2	3	3	1
7	3	3	3	2	3	4	3	2	5	1
8	3	3	4	3	4	4	3	3	3	1
9	2	4	5	4	3	4	4	4	4	1
10	3	3	5	3	3	2	4	3	4	1
11	3	4	4	3	3	4	2	3	4	1
12	3	3	4	4	2	4	3	3	4	1



13	5	3	5	4	4	4	4	4	4	1
14	3	3	4	3	4	4	2	4	4	1
15	3	3	4	3	3	2	2	3	4	1
16	5	3	4	2	3	3	4	3	3	1
17	3	3	5	4	3	5	4	4	3	1

Incremental learning is partitioned into several stages (Table 5). Stage 1: training set contains 6 first persons of Class 1 and 6 first persons of Class 2. The result of Stage 1 is in Tables 6.

Stage 2 is a pattern recognition stage; the control set contains persons 7 and 8 of Class 2 and persons 7 – 17 of Class 1. All persons of Class 2 and 5 persons (8, 9, 13, 14, 17) of Class 1 have been recognized correctly. Persons 10, 11, 15 of Class 1 have been recognized as persons of Class 2, and persons 7, 12, 16 of Class 1 have been assigned to neither of these classes. Results of Stages 3-7 are given in Tables 7-11. Each table contains only new rules generated in corresponding stage.

**Table 5.** Stages of Incremental Learning

Stage	Training sets		Searching rules for		Rules are in Table
	Class 1	Class 2	Class 1	Class 2	
1	Persons 1-6	Persons 1-6	Yes	Yes	6
2	Pattern recognition				
3	Persons 1-6	Persons 1-8	No	Yes	7
4	Persons 1-6 and 8, 9, 13, 14, and 17	Persons 1-8	Yes	No	8
5	Persons 1-6, and 8-11, 13-15, and 17	Persons 1-8	Yes	No	9
6	Person 1-17	Person 1-8	Yes	No	10
7	Persons 1-17	Persons 1-8	No	Yes	11

**Table 6.** The Result of Stage 1

№ of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
1					4	4				1	{1,4,5,6}
2	3	3	5							1	{1,3,5}
3	2	3	4	3	3	3	3	3	3	1	{2}
1	4			3	3					2	{1,2,3,5}
3				3	3				4	2	{1,2,3,6}
5		4	5	3						2	{2,4}

**Table 7.** The result of Stage 3

№ of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
1	4			3	3					2	{1,2,3,5,8}
2						3			4	2	{2,3,6,7}
3				3	3				4	2	{1,2,3,6}

4	5	5	4	4			2	{4,7}
5		4	5	3				{2,4}

During Stage 4, Rule 4 for Class 2 (see, please, Table 7) is deleted (Rule 4  $\subset$  val(13) for person 13 of Class 1).

During Stage 6, Rule 3 for Class 2 (see, please, Table 7) is deleted (Rule 3  $\subset$  val(11) for person 11 of Class 1).

**Table 8.** The result of Stage 4.

Nº of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
1					4	4				1	{1,4,5,6,8,13,14}
2	2				3					1	{2,9}
3			5		3		4	4		1	{3,9,17}
4				4					3	1	{5,6,17}
5				4		4				1	{1,5,6,9,13}
6	3	3							3	1	{3,5,8,17}
7	3	3						4		1	{1,3,5,14,17}

**Table 9.** The result of Stage 5.

Nº of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
8			4				2			1	{4,6,11}
9	3	3		3	3	2				1	{3,10,15}
10		4			3	4			4	1	{9,11}
11	3					4				1	{1,4,5,8,11,14}
12			5		3		4			1	{3,9,10,17}
13	3	3	5							1	{1,3,5,10,17}

**Table 10.** The result of Stage 6

Nº of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
14			3		2	3				1	{7,16}
15			3	4		3	3		3	3	{2,16}
16		3	3		4					1	{1,5,12,17}

Stage 7: correcting the rules for Class 2. The result is in Table 10.

**Table 11.** Final Rules for Class 2 (Stage 7)

Nº of rule	L	F	K	Hy	Pd	Mf	Pa	Pt	Ma	Class	Persons
1	4			3	3					2	{1,2,3,5,8}
2						3			4	2	{2,3,6,7}
5			4	5	3					2	{2,4}
6			3		3		3		4	2	{1,3,6}

#### 4 The Integrative Inductive-Deductive Model of Reasoning

We considered only supervised learning, but integrative inductive-deductive reasoning includes unsupervised learning too. This mode of learning is involved in reasoning when a new portion of objects (examples) becomes available over time but without indication of their class membership. In this case, a teacher is absent. Only knowledge is available. A new object description can currently be complete or incomplete, i.e. some attribute values can be unknown or not observable. If we deal with completely described object, then the following results of reasoning are possible: 1) class of new object is determined; 2) there are several classes to which new object can belong to (a situation of uncertainty); 3) object is unknown.

In situation with incomplete object description, we can try to infer hypotheses about unknown values of attributes (it is reasoning based on “past experience”); if an object is unknown, we can try to select a set of training examples that are similar to this object in most degree and to infer new rules for describing this set of examples.

Consider some instances of pattern recognition reasoning by using the rules obtained by the procedure INGOT (Table 3).

Example 1. New weather descriptions are complete, for example, <Overcast, Cool, High, No>; <Sunny, Mild, Normal, No>; <Sunny, Mild, High, Yes>. In all these cases, we find the rules, which allow us to recognize the weather class.

Example 2. If weather descriptions are incomplete, then it is possible that neither of the rules is applicable. But we can use the training set of examples to infer possible variants of weather class. Assume that the weather description is: <Rain, Mild, High>. We construct the decision tree as follows: Rain: Class 2 (Observations 4, 5, 10), Class 1 (Observations 6, 14); Mild: Class 2 (Observation 4, 10), Class 1 (Observation 14); High: Class 2 (Observation 4), Class 1 (Observation 14). It is a situation of uncertainty. Consequently, a step of conditional or diagnostic reasoning is needed. We can consider hypothetically some possible values of attribute Windy; then we conclude that “**if** Windy = No, **then** Class 2”; “**if** Windy = Yes, **then** Class 1”. Really, we have obtained the following diagnostic rule: “**If** we observe that (Outlook = Rain) & (Temperature = Mild) & (Humidity = High), then (**if** Windy = No, **then** Class 2; **else** Class 1). Note that, the process of pattern recognition includes some inductive step of reasoning.

Example 3. The weather description is: <Hot, Yes>. The reasoning tree is: Hot: Class 1 (Observations 1, 2), Class 2 (Observations 3, 13); Yes: Class 1 (Observations 2), Class 2 (Observations -). Now we can formulate hypothetically a new forbidden rule: “Hot, Yes → Class 2, **false**” or, in another form, “If we observe that (Temperature = Hot) & (Windy = Yes), then it is never observed Class 2”.

Example 4. The weather description is: <Sunny, Mild, Low, No>. Here we meet a new value of Humidity – “Low”. Assume that the sets of values of Humidity and Temperature are ordered and Low < Normal < High and Mild < Cool < Cold. Assume that the functions of distance on the attribute domains are also defined. Then in the pattern recognition process, it is possible to infer that <Sunny, Mild, Low, No> is nearer to the example of Class 2 <Sunny, Cool, Normal, No> than to the example of

Class 1 <Sunny, Mild, High, No>. A new feature for Class 2 can be formed, namely, <Sunny, Low >.

One of the possible models of deductive plausible human reasoning based on implicative logical rules can be found in [22].

One of the important problems of integrating deductive and inductive reasoning is connected with creating some on-line interactive method for modifying context of reasoning. Failures in reasoning or appearance of new data can require to add new attributes to the context. The task of incremental generating a logical context for email messages classification is considered in [23]. This article presents a method for incremental constructing a logical context by the assignment of new attributes to object descriptions. The existing context plays the role of a dynamic schema to help users to keep consistency in their object descriptions.

## 5 Conclusion

In this paper, the decomposition of inferring good classification tests into subtasks of the first and second kinds is presented. This decomposition allows, in principle, to transform the process of inferring good tests into a “step by step” reasoning process.

We have described some inductive algorithm INGOT for inferring good maximally redundant classification tests. We did not focus on the efficiency of this algorithm; we intend to give more attention to the complexity problems in future contributions.

The development of full on-line integrated deductive and inductive reasoning is of great interest. The main problem in this direction is the development of an on-line interactive model to support users in constructing and modifying the context of deductive-inductive reasoning.

## References

1. Godin, R., Missaoui, R., Alaoui H.: Incremental Concept Formation Algorithm Based on Galois (Concept) Lattices. *Computational Intelligence*, 11(2), 246-267 (1995).
2. Dean van der Merwel, F., Obiedkov, S., & Kouriel, D.: AddIntent: A New Incremental Algorithm for Constructing Concept Lattices. *LNCS*, vol. 2961 (pp. 372-385) (2004).
3. Kourie, D.G, Obiedkov, S., Watsona B.W., & Dean van der Merwe, F.: An incremental algorithm to construct a lattice of set intersections. *Science of Computer Programming* 74, 128-142 (2009).
4. Ravindra Patel, K. Swami & R. Pardasani: Lattice Based Algorithm for Incremental Mining of Association Rules. *International Journal of Theoretical and Applied Computer Sciences*. 1(1), 119-128 (2006).
5. Aaron Ceglar and John F. Roddick: Incremental Association Mining using a Closed-Set Lattice. *Journal of Research and Practice in Information Technology*, 39(1), 35-45 (2007).
6. Valtchev, P., Missaoui, R., Godin, R., Meridji, M.: Generating Frequent Itemsets Incrementally: Two Novel Approaches Based on Galois Lattice Theory. *J. Expt. Theor. Artif. Intell.* 14, 115–142 (2002).

7. Utgoff P. E.: An Improved Algorithm for Incremental Induction of Decision Trees. Proceedings of the Eleventh International Conference of Machine Learning (pp. 318-325) (1994).
8. Z. Zheng, G. Wang, Y. Wu: A Rough Set and Rule Tree Based Incremental Knowledge Acquisition Algorithm. LNAI, 2639 (pp. 122-129).Springer-Verlag (2003).
9. Ore, O.: Galois Connexions. Transactions of American Mathematical Society 55(1), 493-513 (1944).
10. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin/Heidelberg (1999).
11. Kuznetsov, S.O.: Machine Learning on the Basis of Formal Concept Analysis. Automation and Remote Control 62(10), 1543-1564 (2001).
12. Naidenova, X.A.: Good Classification Tests as Formal Concepts. F. Domenach, D.I. Ignatov, and J. Poelmans (Eds): ICFCA 2012, LNAI 7278, pp. 211-226 (2012).
13. Naidenova, X.A., Polegaeva, J.G.: SISIF – the System of knowledge acquisition from experimental facts. Alty, J.L., Mikulich, L.I. (Eds.): Industrial Applications of Artificial Intelligence, (pp. 87–92). Elsevier Science Publishers B.V., Amsterdam (1991).
14. Naidenova, X.A.: Data-Knowledge Transformation. V. Solovyev (Ed.): Text Processing and Cognitive Technologies, Issue 3, (pp. 130-151). Pushchino (1999).
15. Naidenova, X. A., Plaksin, M. V., Shagalov, V. L.: Inductive Inferring All Good Classification Tests. In: Valkman, J. (Ed.): Knowledge-Dialog-Solution, Proceedings of International Conference in Two Volumes, vol. 1 (pp. 79-84). Kiev Institute of Applied Informatics, Jalta, Ukraine (1995).
16. Naidenova, X.A.: An Incremental Learning Algorithm for Inferring Logical Rules from Exemplified in the Framework of the Common Reasoning Process. E. Triantaphyllou and G. Felici (Eds.): Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques, (pp. 89-147). Springer, Heidelberg, Germany (2006).
17. Naidenova, X.A.: DIAGARA: An Incremental Algorithm for Inferring Implicative Rules from Examples. Intern. Journal “Information Theories & Application” 12(2), 171-196 (2005).
18. Naidenova, X.A., Shagalov, V.L.: Diagnostic Test Machine. Auer, M. (Ed.): Proceedings of the ICL 2009 – Interactive Computer Aided Learning Conference, Austria, CD (pp. 505–507). Kassel University Press (2009).
19. Schlimmer, J.S.: Concept Acquisition through Representational Adjustment. Technical Report 87-19. Department of Information and Computer Science, University of California, Irvine (1987).
20. Kuznetsov, S.O.: A Fast Algorithm for Computing All Intersections of Objects in a Finite Semi-lattice. Automatic Documentation and Mathematical Linguistics, 27(5), 11–21 (1993).
21. Sobchik, L.N: Standardized Multi-Factorial Method of Personality Investigation (SMIL (MMPI modified)). Practical Guide. “Rech”, Moscow, Russian Federation (2007).
22. Naidenova, X.A.: Constructing Galois Lattices as a Commonsense Reasoning Process. X. Naidenova and D. Ignatov (Eds.): Diagnostic Test Approaches to Machine Learning and Commonsense Reasoning Systems, (pp. 34-70). IGI Global (2012).
23. Ferré, S., & Ridoux, O.: The use of Associative Concepts in the Incremental Building of a Logical Context. Conceptual Structures: Integration and Interfaces, Proceedings of the 10<sup>th</sup> International Conference on Conceptual Structures (ICCS’02), LNCS 2393 (pp. 299-313). Berlin/Heidelberg, Springer (2002).

# FCART: A New FCA-based System for Data Analysis and Knowledge Discovery

A.A. Neznanov, D.A. Ilvovsky, S.O. Kuznetsov

National Research University Higher School of Economics,  
Pokrovskiy bd., 11, 109028, Moscow, Russia  
ANeznanov@hse.ru, DIlvovsky@hse.ru, SKuznetsov@hse.ru

**Abstract.** We introduce a new software system called Formal Concept Analysis Research Toolbox (FCART). Our goal is to create a universal integrated environment for knowledge and data engineers. FCART is constructed upon an iterative data analysis methodology and provides a built-in set of research tools based on Formal Concept Analysis techniques for working with object-attribute data representations. The provided toolset allows for the fast integration of extensions on several levels: from internal scripts to plugins.

FCART was successfully applied in several data mining and knowledge discovery tasks. Examples of applying the system in medicine and criminal investigations are considered.

**Keywords:** Data Analysis, Formal Concept Analysis, Knowledge Discovery, Software.

## 1 Introduction

We introduce a new software system for information retrieval and knowledge discovery from various data sources (textual data, database queries). The Formal Concept Analysis Research Toolbox (FCART) was designed especially for the analysis of unstructured (textual) data. In case studies we applied FCART for analyzing data in medicine, criminalistics, and trend detection.

The core of the system supports knowledge discovery techniques, including those based on Formal Concept Analysis [1], clustering [2], multimodal clustering [2, 3], pattern structures [4, 5] and other.

## 2 Motivation

Currently, there are several well-known open source FCA-based tools, such as ConExp [6], Conexp-clj [7], Galicia [8], Tockit [9], ToscanaJ [10], FCAStone [11], Lattice Miner [12], OpenFCA [13], Coron [14]. These tools are Java-based, cross-platform, rather easy to use, and they do not need to be installed. However, they cannot completely satisfy the growing demands of the community. There are common drawbacks of these systems which have to be addressed: poor data preprocessing,

extensibility, and scalability, as well as non-universal character of existing software tools and their command line interface (CLI) or “old-fashioned” graphical interface (GUI) in terms of usability. There is a lack of universal integrated environment for knowledge discovery based on FCA, although some attempts were made in the Tockit project [9] (the development of ToscanaJ has been forked into a separate project from Tockit). Therefore, our main task was to create an effective software implementation of full research cycle of data analysis and knowledge discovery.

A new system should provide:

1. Universal integrated environment for knowledge and data engineers.
2. Built-in set of research tools based on FCA and multimodal clustering techniques for working with object-attribute data representation.
3. Additional tools for import/export of data and data preprocessing.
4. Extendibility of the research tools on several levels: from internal scripts to plugins.
5. Generation of rich, visually appealing reports.

### 3 Methodology

#### 3.1 Goal

The goal of developing the software package FCART is to create a universal extensible integrated environment. The methodology of using this software is based on modern methods and algorithms of data analysis, technologies for manipulating big data collections, data visualization, reporting and interactive processing techniques. It allows one to obtain new knowledge from data with full process tracking and reproducibility. Some ideas of data preprocessing were inherited from the CORDIET project [15].

#### 3.2 Fundamentals

##### Analytic artifacts

Some of FCA entities appear to be fundamental to information representation. In FCART we use the term “*analytic artifact*” which denotes the definition of abstract interface, describing the entity of the analytic process.

The basic artifact for FCA-based methods is that of “*formal context*”, i.e., object-attribute representation of a subject domain. Most important artifacts include “*concept lattice*” and “*formal concept*”.

All artifact instances are linked by “origination”. For example, we can generate the concept lattice from the formal context. In this case the formal concept will be an “origin artifact” for the lattice. Another example is lattice and “association rules” – the lattice is the origin of the rules. Any artifact instance is *immutable*. It means that an instance cannot be changed after creation, but can be visualized in various ways.

If we have the predefined set of artifacts in most cases we can use the term “artifact” instead of “artifact instance” without ambiguity. Collection of all artifacts in current analytic cycle forms so-called *analytical session*.

Another interesting option of our system is that of *multicontext* [16] artifact. In the most general case a multicontext can be considered as a network of contexts. In the particular case each context is assigned a specific time point, – then one can use this artifact for describing statements of a dynamic system. An example of applying this artifact is the problem of finding trends. Consider the context where objects are documents and attributes are terms. We find chains of intentionally related concepts (other variants of relations can be used). The terms included in the intersection of these concepts will form the core of the trend, the remaining terms and their number will characterize the stage in the life cycle and the popularity of this trend at a particular time point. Relations between concepts in contexts have suitable visual interpretation: sequence of diagrams with labeled links between related concepts.

### Solvers

All types of artifacts are generated by *solvers*. Each solver requires one or many artifact instances of preassigned types as input and produces one artifact instance of preassigned type as output.

Having predefined types of artifacts and links (assigned by solvers) between immutable artifact instances we can check an integrity of data of particular analytical session. Without explicit user action a session cannot lose any artifact instances and links, and guarantees integrity of a session.

### Visualizers

*Artifact visualizer* is a special solver that generates user-oriented visual representation of input artifact instance. From a technical point of view visualizer produces interactive or non-interactive window with some elements of user interface. Of course, one artifact can have different kinds of visual appearance.

Usually, visualizer is the last in a chain of solvers. But we can get a visual representation of each artifact in a session. For example, lattice browser generates a diagram of a lattice and allows a user to manipulate the diagram, but this browser does not generate new artifacts. We need to distinguish generation of new artifact and drawing of existing artifact for various purposes: working in the batch mode, increasing efficiency of long chains of solvers, benchmarking, etc.

### Reports

*Report* is a final result of research. Every scientific environment must provide a report rich text editor with additional functionality to avoid mistakes while converting and moving multiple results with metadata to an external editor. The main feature of the editor is an automatic insertion of fully decorated artifact representation in the resulting report.



### 3.3 Main principles

1. Iterative process of data analysis using FCA entities and methods.
2. Separation of processes of *data querying* (from various data sources), *data preprocessing* (of locally saved immutable snapshots), *data analyzing* (in interactive visualizers of immutable analytic artifacts), and *results formalizing* (in a report editor).
3. Explicit definition of analytic artifacts and their types. It allows checking integrity of the session data and provides links between artifacts for an end-user.
4. Integrated performance estimation tools.
5. Integrated documentation of software tools and data analysis methods.

## 4 Software properties

### 4.1 Common information

At this moment we introduce the version 0.7 of FCART in the form of local Windows application. We use Microsoft and Embarcadero programming environments and different programming languages (C++, C#, Delphi, Python and other). For scripting we use Delphi Web Script and Python. Native executable (the core of the system) is compatible with Microsoft Windows 2000 and later and has not any additional dependences.

Another line of development is Web-version of system based on Microsoft .NET platform. For now architecture and some key components are ready, but we are going to focus on Web-development after finishing local version 0.9.

### 4.2 Architecture

FCART constructed as multicomponent application. Current version consists of the following components:

- Core component
  - multiple-document user interface of research environment with session manager,
  - snapshot profiles editor (SHPE),
  - snapshot query editor (SHQE),
  - query rules database (RDB),
  - session database (SDB),
  - report builder.
- Local XML-storage for preprocessed data.
- Internal solvers and visualizers.
- Additional plugins and scripts.

### 4.3 Data preprocessing in FCART

#### Obtaining initial artifacts

There are several ways to obtain initial artifacts.

- Load from ready data files of supported formats.
- Generate by plugin or script.
- Query from data snapshots.

Data snapshot (or snapshot) is a data table with structured and text attributes, loaded in the system by accessing external SQL, XML or JSON data sources. Snapshot is described by a *profile*. FCART provides one with a snapshot profile editor (SHPE) and local storage of snapshots with metadata.

#### Constructing binary contexts

Initial formal contexts can be imported from data files in standard format like CXT or CSV. The system has query language for transforming snapshot into formal context. This language describes so-called rules. Main rule types are the following.

- *Simple rule* generates one attribute from atomic fields of a snapshot.
- *Scaling rule* generates several attributes from atomic fields based on nominal or ordinal scale.
- *Text mining rule* generates one attribute from unstructured text fields.
- *Multivalued rule* generates one or many attributes from multivalued field (arrays and sets).
- *Compound rule* merges rules of all types into single rule. This rule uses standard logical operations and brackets to combine elements.

We also implement additional rule types: *Temporal rules* are used for manipulating date and time intervals and *Filters* are used for removing objects from context.

In most cases, it is not necessary to write a query from scratch. One can select some entities in rules DB and automatically generate a query. It is possible because the rule DB is aware of dependencies between rules. Separate queries or full DB of rules can be imported and exported as XML-files.

FCART uses Lucene full text search engine [17] to index the content of unstructured text fields in snapshots. The resulting index is later used to validate quickly whether the text mining or compound rule returns true or false. It is useful for dealing with dynamic data collections, including texts in natural language.

### 4.4 Sessions, solvers and visualizers

#### Session

Multiple-document interface allows one to have each solver in its own window. User can view all artifacts in the *session browser* (independent task pane) in the form of a tree. The main mode of user interaction in FCART is interactive work in various visu-

alizers. Our software manages links between artifacts and guarantees valid state of a working session in case of deleting some objects and restarting the system.

Main solvers in the current version can produce clusters; concept lattices and sublattices; association rules and implications; calculate stability indices, similarity measures for contexts and concepts. All those artifacts can be visualized and inserted into the report. The set of solvers and visualizers can be appended by plugins and scripts.

Any artifact can be exported in several formats. For example, concept lattice can be saved as graph (XGMML) or as picture (EMF, PNG, and JPG). We plan to extend the set of admissible formats on demand of future users.

### **Interactive visualization of concept lattice**

The *concept lattice visualizer* is an example of visualizer. It can be used to browse the collection of objects with binary attributes given as a result of query to snapshot (with structured and text attributes). The user can select and deselect objects and attributes and the lattice diagram is modified according. The user can click on a concept. The screen shows in a separate window names of objects in the extent and names of attributes in the intent. Names of objects and attributes are linked with initial snapshot records and fields. If the user clicks on the name of an object or an attribute, the content of the object or attribute description is shown in a separate window according to snapshot profile.

Fig. 1 demonstrates the result of building sublattice from concept lattice. The Multiple-document interface allows us to inspect several artifacts, so a sublattice will be opened in a new window.

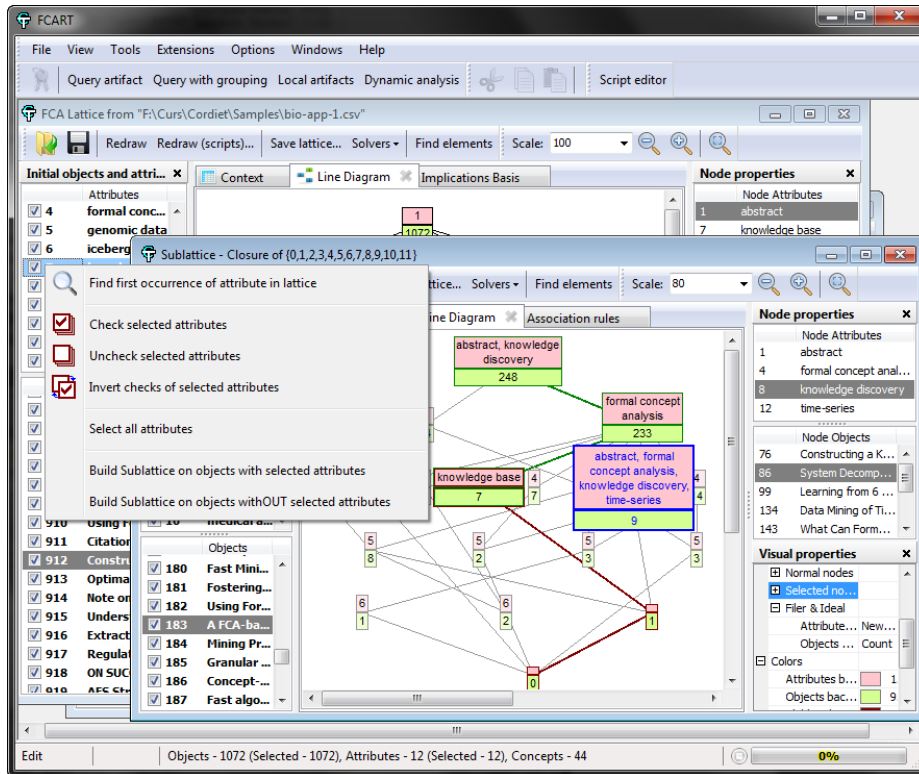


Fig. 1. Concept lattice visualizer

The user can customize settings of lattice browsing in various ways. The user can specify whether nodes corresponding to concepts show numbers of all (or only new) objects and all (or only new) attributes in extent and intent respectively, or names of all (or only new) objects and all (or only new) attributes. Separate settings can be specified for the selected concept, concepts in the order filter, and the remainder of the lattice. The visual appearance can be changed: zooming, coloring, and other tools are available.

Right clicking on the name of an attribute user can choose several options: he can build a sublattice containing only objects with selected attribute; build a sublattice containing only objects without selected attribute; or find the highest concept with selected attribute. Right clicking on the name of object allows the same actions.

### Report generation

FCART supports editing several reports at the same time. A user can add any of valid artifacts from the current session to the report. Source file can be added as text with syntax highlighting (XML or other schemes); snapshot – as a table with profile definition; context – as a table or a bipartite graph; concept lattice – as an XGMML-text or a vector diagram (Fig. 2); and so on. Reports are part of the session and are stored

automatically. The final report can be copied to the clipboard with full content and formatting. Also it can be saved to a file in RTF or HTML format.

Same report engine is used to edit and render documentation: descriptions of solvers, comments to artifacts, and other.

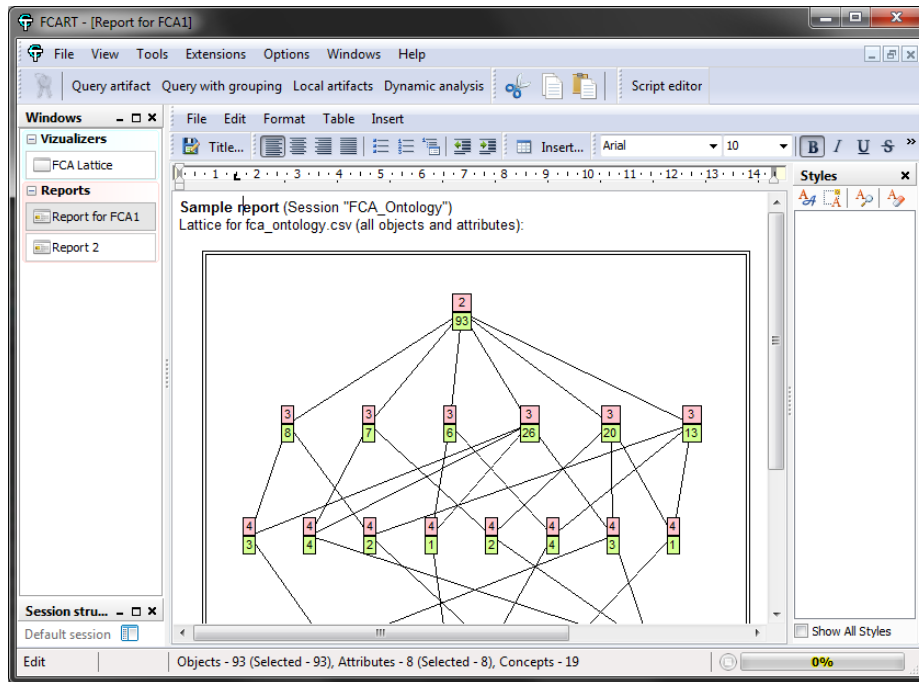


Fig. 2. Report editor with inserted lattice diagram

## 4.5 Extensibility

### Scripts

*Scripts* (macros) are small internal programs, written in Delphi Web Script [18] (already implemented) or Python [19] (implementing now). In the current version following tasks can be automated by scripts:

- Generating artifacts (for example, building contexts on the fly, randomizing and generating of test samples). Of course, the user can generate only artifacts of predefined types.
- Formatting reports.
- Drawing lattice (layouting).
- Calculating similarity measures of artifacts of same types.

The system provides the script editor with syntax highlighting and debugging. The set of possible tasks will be extended in the next versions of FCART.

### Plugins

The system can be extended by special modules called plugins using low-level API. The plugin's API is designed to reach maximum performance. In the current version we tested generators of artifacts.

## 5 Comparison with existing systems

The study of big analysis software like IBM i2 Analyst's Notebook or QSR NVivo shows that this software do not have FCA tools and have a completely different methodology of data analysis as compared to FCA software systems. So we need to compare functionality of the system with well-known tools for building and visualizing FCA artifacts (Table 1). Some criteria for comparison:

- Basic functionality: support for contexts editing, formal concept lattice generation and drawing, sublattice construction, association rules generation and other.
- Performance of basic operations and scalability.
- Rich set of supported formats.
- Data preprocessing capabilities.
- Changing visualization schemes.
- Reporting capabilities.
- Ease of extensibility.

All of the tools mentioned in Table 1 have unique features. For example, Concept Explorer was an important milestone in the development of FCA software tools. It has interesting modes of visualization of a lattice and good default layout. Galicia introduces the generic MultiFCA approach to deal with a set of contexts. ToscanaJ can visualize nested lattices and involves an editor of conceptual schemas on relational databases. FcaStone was primarily intended for file format conversion and other low level operations. Unfortunately, most of useful tools for end-user did not have official updates starting from 2006. Last version of Coron was released in 2010. Only two actively developed projects can be noted: ToscanaJ and Conexp-clj.

**Table 1.** Some accessible FCA software tools

<b>Program title</b>	<b>Authors</b>	<b>Web-site</b>
Concept Explorer (ConExp)	S.A. Evtushenko et al [6]	conexp.sourceforge.net
Galicia	P. Valtchev et al [8]	www.iro.umontreal.ca/~galicia
ToscanaJ (with Siena and Elba)	University of Queensland, Technical University of Darmstadt [10]	toscanaj.sourceforge.net
FcaStone	U. Priss et al [11]	fcastone.sourceforge.net
Lattice Miner	Boumedjout Lahcen [12]	lattice-miner.sourceforge.net
Conexp-clj	TU-Dresden, Daniel Borchman	daniel.kxpq.de/math/conexp-clj
OpenFCA	P. Borza, O. Sabou, et al [13]	code.google.com/p/openfca

(Conflexplore)		
Lattice navigator	M. Radvansky, V. Sklenar	www.fca.radvansky.net
Coron	Szathmary, L., Kaytoue, M., Marcuola, F., Napoli, A. [14]	coron.loria.fr

The common problem for these tools is low limits of size of interactively analyzed artifacts (for example, lattices with more than 8000 concepts can hardly be operated and visualized on modern hardware). This is mainly due to the use Java (or other high-level languages) and cross-platform GUI.

Let's look at an example of scalability. Consider real context (707 objects and 257 attributes) and generate concept lattice (10568 concepts) in different software<sup>1</sup>.

*ConExp* generated concepts, spent 90.0 MB of memory, and *could not* produce layout the lattice (Fig. 3).

*ToscanaJ Siena* generated concepts, spent 203.2 MB of memory, produced layout (Fig. 4), but worked *very slowly* even when viewing initial context.

*FCART* generated concepts, spent 23.5 MB of memory, produced layout (Fig. 5), and provided normal interactive manipulations with context and concepts.

The current version of FCART can construct and manipulate big lattices (more than 16000 concepts), also in interactive mode. After all planned optimizations in version 0.8 we will present deep comparison of implementations of all basic FCA algorithms in the form of compiled components and scripts (the system has built-in tools for benchmarking) on synthetic tests and real data.

FCART is built on top of modern platform, provides powerful preprocessing tools, rich reporting capability, and two levels of extensibility.

---

<sup>1</sup> All tests were conducted on a computer with Intel Core i7-3770 3,4 GHz CPU, 16 GB of RAM, Microsoft Windows 7 Professional x64 (system info is tracked by Process Explorer).

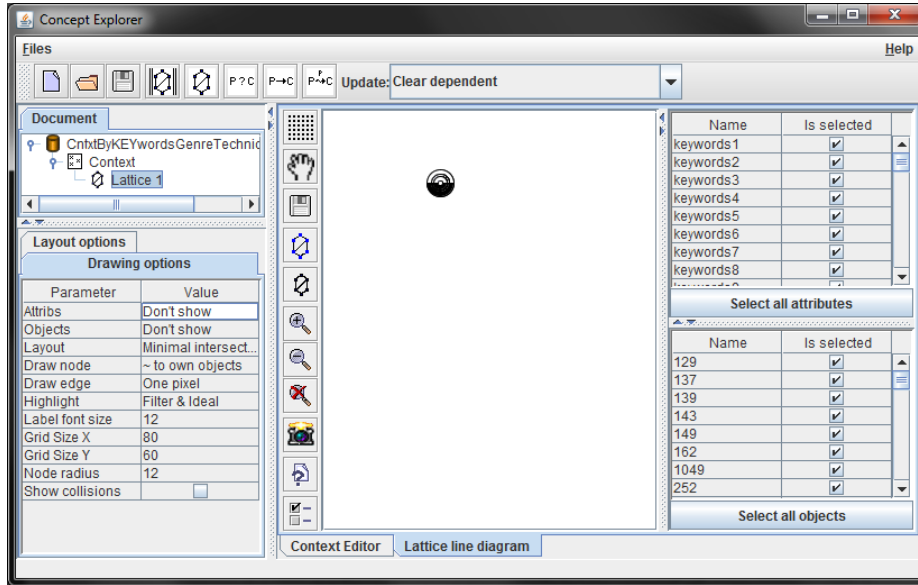


Fig. 3. Sample lattice layout in ConExp

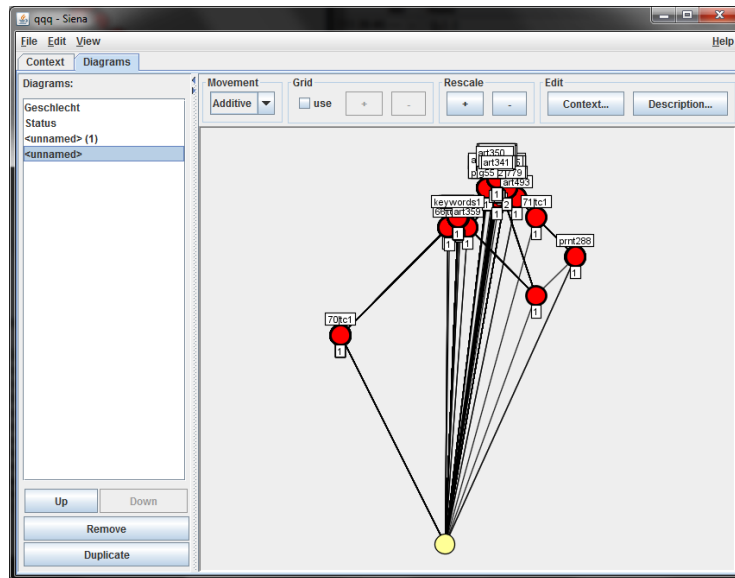


Fig. 4. Sample lattice layout in ToscanaJ Siena



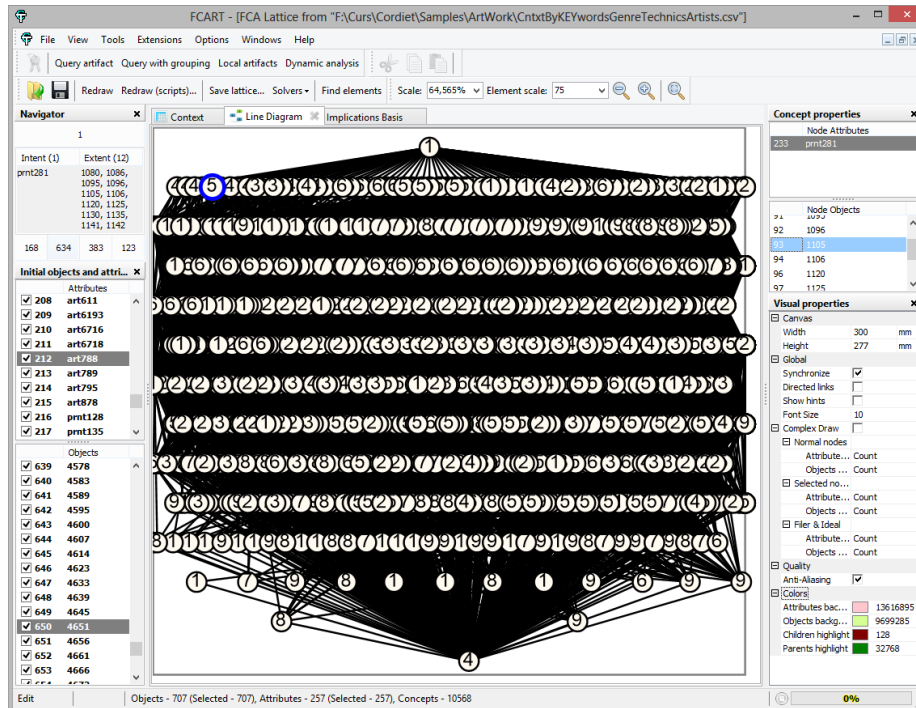


Fig. 5. Sample lattice layout in FCART (simple drawing scheme is used)

## 6 Conclusion and future work

FCART is a powerful environment being in active developing state. The next major release of the local version 0.8 is planned for March 2013. Then this system will be freely available to the FCA community.

We assume to improve methodology, extend the set of solvers, optimize some algorithms, and use proposed system in different knowledge discovery tasks. We already test new solvers based on concept stability [20] and similarity [10]. Biclustering techniques [2, 22] are also being actively tested; we are going to extend our platform to triadic concept analysis and noise-robust triclustering methods [3].

## Acknowledgements

The work of the authors on the project “Mathematical Models, Algorithms, and Software Tools for Intelligent Analysis of Structural and Textual Data” was supported by the Basic Research Program of the National Research University Higher School of Economics.

## References

1. Ganter, B., Wille R. Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
2. Mirkin, B. Mathematical Classification and Clustering, Springer, 1996.
3. Ignatov, D.I., Kuznetsov, S.O., Magizov, R.A., Zhukov, L.E. From Triconcepts to Triclusters. Proc. of 13th International Conference on rough sets, fuzzy sets, data mining and granular computing (RSFDGrC-2011), LNCS/LNAI Volume 6743/2011, Springer (2011), pp. 257-264.
4. Ganter, B., Kuznetsov, S.O. Pattern Structures and Their Projections. Proc. of 9th International Conference on Conceptual Structures (ICCS-2001), 2001, pp. 129-142.
5. Kuznetsov, S.O. Pattern Structures for Analyzing Complex Data. Proc. of 12th International conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Conference (RSFDGrC-2009), 2009, pp. 33-44.
6. Yevtushenko, S.A. System of data analysis "Concept Explorer". (In Russian). Proceedings of the 7th national conference on Artificial Intelligence KII-2000, p. 127-134, Russia, 2000.
7. Conexp-clj (<http://daniel.kxpq.de/math/conexp-clj/>)
8. Valtchev, P., Grosser, D., Roume, C. Mohamed Rouane Hacene. GALICIA: an open platform for lattices, in Using Conceptual Structures: Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03), pp. 241-254, Shaker Verlag, 2003.
9. Tockit: Framework for Conceptual Knowledge Processing (<http://www.tockit.org>)
10. Becker, P., Hereth, J., Stumme, G. ToscanaJ: An Open Source Tool for Qualitative Data Analysis, Proc. Workshop FCAKDD of the 15th European Conference on Artificial Intelligence (ECAI 2002). Lyon, France, 2002.
11. Priss, U. FcaStone - FCA file format conversion and interoperability software, Conceptual Structures Tool Interoperability Workshop (CS-TIW), 2008.
12. Lahcen, B., Kwuida, L. Lattice Miner: A Tool for Concept Lattice Construction and Exploration. In Supplementary Proceeding of ICFCA'10, 2010.
13. Borza, P.V., Sabou, O., Sacarea, C. OpenFCA, an open source formal concept analysis toolbox. Proc. of IEEE International Conference on Automation Quality and Testing Robotics (AQTR), 2010, pp. 1-5.
14. Kaytoute, M., Marcuola, F., Napoli, A., Szathmary, L., Villerd, J. The Coron System. Proc. of the 8th Intl. Conference on Formal Concept Analysis (ICFCA 2010), 2010, pp. 55-58.
15. Poelmans, J., Elzinga, P., Neznanov, A., Viaene, S., Kuznetsov, S., Ignatov, D., Dedene, G. Concept Relation Discovery and Innovation Enabling Technology (CORDIET) // CEUR Workshop proceedings Vol-757, Concept Discovery in Unstructured Data, 2011.
16. Wille, R.: Conceptual structures of multicontexts. In Eklund, P., Ellis, G., Mann, G., eds.: Conceptual Structures: Knowledge Representation as Interlingua. Volume 1115 of LNAI, Springer (1996), pp. 23-29.
17. Apache Lucene (<http://lucene.apache.org>)
18. Grange, E. DelphiWebScript Project (<http://delphitools.info/dwscript>)
19. Python Programming Language – Official Website (<http://www.python.org>)
20. Kuznetsov, S.O. On Stability of a Formal Concept // Annals of Mathematics and Artificial Intelligence, Vol. 49, 2007, pp.101-115.
21. Klimushkin, M.A., Obiedkov, S., Roth C. Approaches to the Selection of Relevant Concepts in the Case of Noisy Data // 8th International Conference on Formal Concept Analysis (ICFCA 2010), 2010, pp. 255-266.

22. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J. Concept-Based Bicustering for Internet Advertisement. Proc. of 12th IEEE International Conference on Data Mining Workshops, 10 December 2012, Brussels, Belgium, pp. 123-130.

## Author Index

Antoni, Lubomír	5
Doignon, Jean-Paul	1
Ilvovsky, Dmitry A.	65
Jäschke, Robert	19
Krajčí, Stanislav	5
Krasuski, Adam	35
Krídlo, Ondřej	5
Kuznetsov, Sergei O.	65
Naidenova, Xenia	51
Neznanov, Alexey A.	65
Parkhomenko, Vladimir	51
Pisková, Lenka	5
Rudolph, Sebastian	19
Wasilewski, Piotr	35