An Implementation for Fault Tolerance and Experimental Results

Frithjof Dau, SAP AG

Abstract. Even for small or medium sized contexts, the corresponding concept lattice soon becomes too large to be nicely displayed and understood, which renders a naive application of FCA for visual analytics challenging. Concept lattices depict all information of the underlying formal context. In certain settings, one can consider the context to be noisy or incomplete, and moderate and meaningful changes of the context such that the corresponding lattices significantly shrink in size and complexity are in those settings permissible. In this paper, such an approach is taken. The formal definitions for the approach are given, and using an implementation, several examples are provided which show the usefulness of the approach.

1   Introduction

CUBIST[1] is an EU funded research project with an approach that leverages BI to a new level of precise, meaningful and user-friendly analytics of data by following a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and FCA-based Visual Analytics. A main goal of CUBIST is the provision of novel Visual Analytics based on meaningful diagrammatic representations. The Visual Analytics part of CUBIST is complementing traditional BI-means by utilizing Formal Concept Analysis (FCA) for analyzing the data in a triple store: whereas traditional BI focuses on the visualization and analysis of *numerical, quantitative* data ("show me the numbers"). FCA is a well-known theory of data analysis which allows objects to be conceptually clustered with respect to a given set of attributes and then visualize the (lattice-ordered) set of clusters, e.g. by means of Hasse-diagrams, thus is best suited for visualization and analysis of *structural dependencies* in the data.

The starting point of FCA is a so-called *formal context (O,A,I)* consisting of a set O of formal objects, a set A of formal attributes, and an incidence-relation $M \subseteq O \times A$ between the formal objects and attributes. The corresponding concept lattice preserves all the information of the formal context, and it might grow exponential in size compared to O and A (one might have $2^{\max\{|O|,|A|\}}$ many formal concepts). Hence even for data sets of moderate size, the corresponding lattices become very soon too complex to be visualized and thus incomprehensible. This is a well-known problem of FCA and has already been addressed with different approaches. *Iceberg-lattices,* introduced by Gerd Stumme, show only the topmost part of a concept lattice, thus reducing the

---

number of concepts shown to the user [8,9]. It is not only the number of concepts which often render the visualization of lattices complicated: Even with good layout algorithms, the Hasse-diagram of a lattice often contains a significant number of crossing edges, which is the main problem for the visual comprehension of graphs. Cassio Melo et al. present in [4,5] different metrics which allow for each node in a concept lattice to choose a uniquely given upper neighbor. By doing so, one can transform a lattice into a tree (with loss of information), which in turn can be better displayed compared to lattices. Finally, Boulicaut et al consider in [2,7] "noisy" formal contexts and directly change the incidence relation *I* in order to reduce the number of formal concepts. For CUBIST, in line with the approach of Boulicaut et al, Simon Andrews has provided in [1] an example with real-data of one the CUBIST use cases which clearly shows the advantage of this approach.

In this paper, we follow up the notion of considering contexts to be noisy and to extend the incidence relation in order to simplify the derived concept lattices. In the next section, we provide the mathematical definition on how we change the incidence relation. As described in the following section, this approach has been implemented into a small FCA-tool in order to show its effectiveness. A thorough, though artificial example is given in the next section, before we turn to examples based on real data out of the CUBIST project.

The authors want to stress that apart from providing the mathematical definitions, this paper focuses on the implementation and the experimental results. A thorough mathematical investigation is out of scope of this paper and will we provided in a different paper which is currently in preparation.

2   Formal Definition

Let a formal context *(O,A,I)* be given. Our basic idea is to understand the context to be noisy or incomplete, and we are aiming at adding crosses to *I* such that the derived lattice becomes smaller (hence, easier to read and understand). For an object $o \in O$ and an attribute $a \in A$, our starting point is the well-known equation

$$oIa \quad \Leftrightarrow \quad o^{II} \subseteq a^{I} \quad \Leftrightarrow \quad a^{II} \subseteq o^{I} \quad \Leftrightarrow \quad o^{I} \rightarrow a \qquad (1)$$

That is, we do *not* have *oIa* if some objects and/or some attributes violate (1). We define the sets of those objects and attributes as follows:

$$Diff_{Obj}(o,a) := \{ \, x \in O \mid x \in o^{II} \text{ and } x \notin a^{I} \, \} \text{ and}$$
$$Diff_{Att}(o,a) := \{ \, y \in A \mid y \in a^{II} \text{ and } y \notin o^{I} \, \}$$

The straight-forward approach taken in this paper, is, roughly speaking, as follows: the bigger $Diff_{Obj}(o,a)$ or $Diff_{Att}(o,a)$, the less we like to add an additional cross between *o* and *a*. In order to do so, it is possible to take only $Diff_{Obj}(o,a)$ into account (this approach is reasonable in contexts having a small amount of attributes, but a large amount of objects, which is standard in traditional BI settings), or only $Diff_{Att}(o,a)$, or both. Moreover, we can either relate $Diff_{Obj}(o,a)$ and $Diff_{Att}(o,a)$ to the

overall sets of objects and attributes, i.e. taking a global approach, or relate them only to the concepts generated by o and a, i.e. taking a local approach. This gives rise to six different ways to measure "an approximate incidence measure" between o and a as follows:

$$G_{Obj}(o,a) \quad := \quad 1-(\qquad |Diff_{Obj}(o,a)|/|O| \qquad\qquad\qquad)$$
$$G_{Att}(o,a) \quad := \quad 1-(\qquad\qquad\qquad\qquad |Diff_{Att}(o,a)|/|A| \quad)$$
$$G_{Obj,Att}(o,a) \quad := \quad 1-1/2 \,(\quad |Diff_{Obj}(o,a)|/|O| \quad + \quad |Diff_{Att}(o,a)|/|A| \quad)$$
$$L_{Obj}(o,a) \quad := \quad 1-(\quad |Diff_{Obj}(o,a)|/|\,o^{II}\,| \qquad\qquad\qquad)$$
$$L_{Att}(o,a) \quad := \quad 1-(\qquad\qquad\qquad |Diff_{Att}(o,a)|/|\,a^{II}\,|\,)$$
$$L_{Obj,Att}(o,a) \quad := \quad 1-1/2 \,(\quad |Diff_{Obj}(o,a)|/|\,o^{II}\,| \quad + \quad |Diff_{Att}(o,a)|/|\,a^{II}\,|\,)$$

Fig 1: Six approaches to measure the incidence between objects and attributes

For any of these incidence measures $S \in \{\, G_{Obj},\, G_{Att},\, G_{Obj,Att},\, L_{Obj},\, L_{Att},\, L_{Obj,Att}\,\}$ we obviously have for all o,a:

$$0 \le S(o,a) \le 1 \quad \text{and} \quad S(o,a) = 1 \Leftrightarrow oIa \qquad (2)$$

So, for a given threshold t with $0 \le t \le 1$ we can naturally set

$$J_{S,t} := \{\, (o,a) \mid S(o,a) \ge t\,\} \qquad (3)$$

Due to (2), we obviously have $s \le t \Rightarrow J_{S,t} \subseteq J_{S,s}$ and $J_{S,1} = I$
.Finally, the local measure for objects corresponds to the confidence of the according association rule, which is a nice rationale for that measure:

$$conf(o^I \to a) = \frac{|(o^I \cup (a))^I|}{|o^{II}|} = 1 - \frac{|o^{II}| - |(o^I \cup (a))^I|}{|o^{II}|} = 1 - \frac{Diff_{Obj}(o,a)}{|o^{II}|} = L_{Obj}(o,a)$$

We will now exemplify (3) for $S := G_{Obj}$ (and $L_{Obj}(o,a)$. which is for this example the same relation) with the following simple context:

	a1	a2	a3
o1		X	X
o2			X
o3	X	X	
o4	X	X	
o5	X	X	
o6	X	X	
o7	X	X	
o8	X	X	
o9	X	X	
o10	X		X

Please note the following:
- For o_{10}, we have $o_{10}{}^I \to a_3$ and $o_{10}{}^I \to a_1$, but not $o_{10}{}^I \to a_2$. This implication is violated by exactly one object, namely o_{10} itself. If we consider a threshold of 0.9 (that is, as we have 10 objects in total, we set a cross between o and a iff the condition $o^I \to a$ is violated by at most one object), then we should add a cross between o_{10} and a_2. Note that the high amount of objects which have a_1 as attribute (i.e. o_3 - o_9) is not relevant.
- For o_1, we have $o_1{}^I \to a_3$ and $o_1{}^I \to a_2$, but not $o_1{}^I \to a_1$. This implication is violated by exactly two objects, namely o_1 itself, and o_2. If we consider a

threshold of 0.8 (that is, now we set a cross between o and a iff the condition $o^I \rightarrow a$ is violated by two or less objects), then we should add a cross between o_1 and a_1

- A similar consideration applies of course to o_2.

The approximate incidence measure $S:=G_{Obj}$ and the concept lattices for the incidence relations $I=J_{S,1}$, $J_{S,0.9}$ and $J_{S,0.8}$ is depicted in the next figure.

cxt	a1	a2	a3
o1	0.80	1.00	1.00
o2	0.80	1.00	1.00
o3	1.00	1.00	0.30
o4	1.00	1.00	0.30
o5	1.00	1.00	0.30
o6	1.00	1.00	0.30
o7	1.00	1.00	0.30
o8	1.00	1.00	0.30
o9	1.00	1.00	0.30
o10	1.00	0.90	1.00

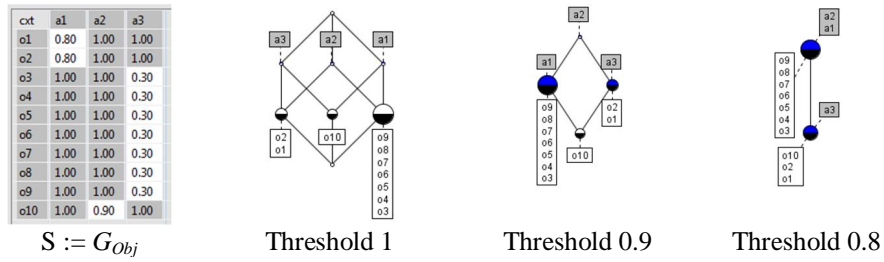$S := G_{Obj}$ Threshold 1   Threshold 0.9   Threshold 0.8

Fig 2: A simple example for G_{OBJ}

Generally, adding crosses to an incidence relation can decrease or increase the number of formal concepts, and one can hardly make statements on how the concept lattices change. For the herein defined extensions of the incidence relations, the situation is different. We start with a simple result which does not generally apply to extensions of incidence relations.

Lemma: Let (O,A,I) be a formal context, let $S \in \{ G_{Obj}, G_{Att}, G_{Obj,Att}, L_{Obj}, L_{Att}, L_{Obj,Att} \}$ be an approximate incidence measure, let t with $0 \leq t \leq 1$ be a threshold, and let $J := J_{S,t}$. Then, for all $o_1, o_2 \in O$; we have $o_1^I \subseteq o_2^I \Rightarrow o_1^J \subseteq o_2^J$. Similarly, for all $a_1, a_2 \in A$ we have $a_1^I \subseteq a_2^I \Rightarrow a_1^J \subseteq a_2^J$.

Proof: We only show the lemma for objects (i.e. the first implication), the proof for attributes is done analogously.

Let $a \in A$ be an attribute . As we have $o_1^I \subseteq o_2^I$ and hence $o_2^{II} \subseteq o_1^{II}$, we obtain $Diff_{Obj}(o_2,a) \subseteq Diff_{Obj}(o_1,a)$ and $Diff_{Att}(o_2,a) \subseteq Diff_{Att}(o_1,a)$; thus $S(o_1,a) \leq S(o_2,a)$. Particularly, we get $o_1 J a \Leftrightarrow S(o_1,a) \geq t \Rightarrow S(o_2,a) \geq t \Leftrightarrow o_2 J a$.   Q.e.d.

As stated in the introduction, a thorough investigation of the approximate incidence relations will be the subject of a different paper. Here we only provide a remark that even though we can prove statements about the relationship between I and $J_{S,t}$ which do not hold for I and *any* extension $J \supseteq I$, the dependencies between I and $J_{S,t}$ have to be investigated further. For example, the author conjectured that we have $P^J = P^{IIJ}$ for all sets of objects $P \subseteq O$, but the context in Fig 3, together with the approximate incidence measure $GObj(o,a)$, shows that this equation does not generally hold. To see this, consider a threshold $t=0.8$, $P:=\{o_0,o_1,o_2\}$: we then have $P^{IIJ} = \{o_1, o_2, o_3, o_4\}$ and $P^J = \{o_1, o_2, o_3, o_4, o_6, o_7, o_8\}$

cxt	a1	a2	a3	a4	a5	a7	a8
o0	.	X	.	.	.	.	.
o1	X	X	X	X	.	.	X
o2	X	X	X	.	.	.	.
o4	.	.	.	.	.	.	X
o5	X	X	X	X	.	X	.
o6	X	X	.	.	.	.	.
o7	.	.	.	.	X	.	.

cxt	a0	a1	a2	a3	a4	a5	a6	a7	a8
o0	1.00	0.85	1.00	0.85	0.85	0.85	1.00	0.85	0.85
o1	0.85	1.00	1.00	1.00	1.00	0.85	1.00	0.85	1.00
o2	0.57	1.00	1.00	1.00	0.85	0.57	0.71	0.71	0.71
o4	0.71	0.85	0.85	0.85	0.85	0.71	0.85	0.71	1.00
o5	0.85	1.00	1.00	1.00	1.00	0.85	0.85	1.00	0.85
o6	0.42	1.00	1.00	0.85	0.71	0.42	0.57	0.57	0.57
o7	1.00	0.85	0.85	0.85	0.85	1.00	0.85	0.85	0.85

Fig 3: Counterexample for $P^J = P^{IIJ}$

3 Implementation

We have extended the tool presented in [3] in order to implement the approach presented in this paper. The tool, initially developed to transform SPARQL-queries into formal contexts, can now load formal contexts (independently from SPARQL) and transform them according to (3). A partial screenshot is provided below.

Fomal Context

Context Computations — Approx. Incidence Relations — Load/Save

[Transitive Closure] [Reflexive Closure] [global similarity: objects only ▼] ‹ [ ] › 90% [Load] [Save As] [Save All As]

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90
O3	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.90	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.80	1.00	0.60	0.90	0.90	0.90	0.90	0.80	0.80
O7	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O8	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O9	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70
O10	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70

Fig 4: screenshot of a tool which implements approximate incidence relations

The tool allows to load a formal context (as a Burmeister file) and computes all six approximate incidence measures. Either the original context or one of the measures can be selected for display, as the next screenshot shows.
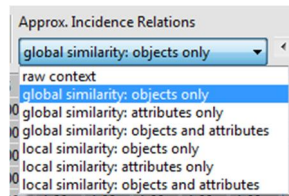
Approx. Incidence Relations

[global similarity: objects only ▼] ‹

raw context
global similarity: objects only
global similarity: attributes only
global similarity: objects and attributes
local similarity: objects only
local similarity: attributes only
local similarity: objects and attributes

Fig 5: The different metrics as the appear in the tool

The slider allows to adjust the threshold t. A cell for an object o and an attribute a is marked grey iff $S(o,a) \geq 1$ (where S stands for the selected measure). One can now either store only the single derived context with the incidence relation $J_{S,t}$, or a set of derived contexts, where all thresholds 0.9, 0.91, …, 0.99 (these thresholds are manually chosen and so far hardcoded) and all six measures are used (thus, 60 Burmeister files are stored).

4   An thorough, artificial example

In this section, we exemplify all measures and different thresholds with an artificial example. Let us consider the following formal context and its concept lattice:
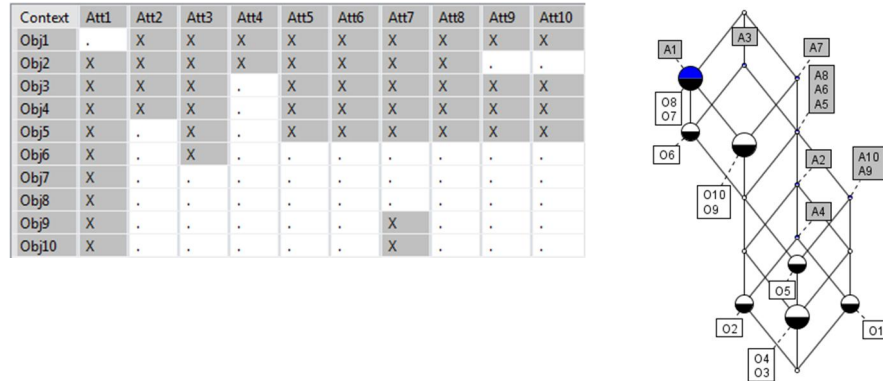
Context	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Att8	Att9	Att10
Obj1	.	X	X	X	X	X	X	X	X	X
Obj2	X	X	X	X	X	X	X	X	.	.
Obj3	X	X	X	.	X	X	X	X	X	X
Obj4	X	X	X	.	X	X	X	X	X	X
Obj5	X	.	X	.	X	X	X	X	X	X
Obj6	X	.	X	.	.	.	.	.	.	.
Obj7	X	.	.	.	.	.	.	.	.	.
Obj8	X	.	.	.	.	.	.	.	.	.
Obj9	X	.	.	.	.	.	X	.	.	.
Obj10	X	.	.	.	.	.	X	.	.	.

Fig 6: Example context and lattice for approximate incidence relations

For this context, we obtain the following values for the six different measures:

Global — Objects

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90
O3	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.90	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.80	1.00	0.60	0.90	0.90	0.90	0.90	0.80	0.80
O7	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O8	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O9	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70
O10	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70

local — Objects

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
O3	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.66	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.60	1.00	0.20	0.80	0.80	0.80	0.80	0.60	0.60
O7	1.00	0.33	0.55	0.11	0.44	0.44	0.66	0.44	0.33	0.33
O8	1.00	0.33	0.55	0.11	0.44	0.44	0.66	0.44	0.33	0.33
O9	1.00	0.50	0.66	0.16	0.66	0.66	1.00	0.66	0.50	0.50
O10	1.00	0.50	0.66	0.16	0.66	0.66	1.00	0.66	0.50	0.50

Global — Attributes

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.80
O3	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.90	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.50	1.00	0.40	0.60	0.60	0.90	0.60	0.40	0.40
O7	1.00	0.40	0.90	0.30	0.50	0.50	0.90	0.50	0.30	0.30
O8	1.00	0.40	0.90	0.30	0.50	0.50	0.90	0.50	0.30	0.30
O9	1.00	0.50	0.90	0.40	0.60	0.60	1.00	0.60	0.40	0.40
O10	1.00	0.50	0.90	0.40	0.60	0.60	1.00	0.60	0.40	0.40

local — Attributes

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.71	0.71
O3	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.83	1.00	0.71	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.16	1.00	0.14	0.20	0.20	0.00	0.20	0.14	0.14
O7	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
O8	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
O9	1.00	0.16	0.00	0.14	0.20	0.20	1.00	0.20	0.14	0.14
O10	1.00	0.16	0.00	0.14	0.20	0.20	1.00	0.20	0.14	0.14

Global — objs and atts

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.85
O3	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.90	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.65	1.00	0.50	0.75	0.75	0.90	0.75	0.60	0.60
O7	1.00	0.40	0.75	0.25	0.50	0.50	0.80	0.50	0.35	0.35
O8	1.00	0.40	0.75	0.25	0.50	0.50	0.80	0.50	0.35	0.35
O9	1.00	0.60	0.85	0.45	0.70	0.70	1.00	0.70	0.55	0.55
O10	1.00	0.60	0.85	0.45	0.70	0.70	1.00	0.70	0.55	0.55

local — objs and atts

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.35	0.35
O3	1.00	1.00	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.75	1.00	0.35	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.38	1.00	0.17	0.50	0.50	0.40	0.50	0.37	0.37
O7	1.00	0.16	0.27	0.05	0.22	0.22	0.33	0.22	0.16	0.16
O8	1.00	0.16	0.27	0.05	0.22	0.22	0.33	0.22	0.16	0.16
O9	1.00	0.33	0.33	0.15	0.43	0.43	1.00	0.43	0.32	0.32
O10	1.00	0.33	0.33	0.15	0.43	0.43	1.00	0.43	0.32	0.32

Fig 7: All six approximate incidence measures for the example

For the thresholds *0.9* (upper half of table) and *0.8* (lower half of table), the next figure provides for each measure S the corresponding concept lattice.
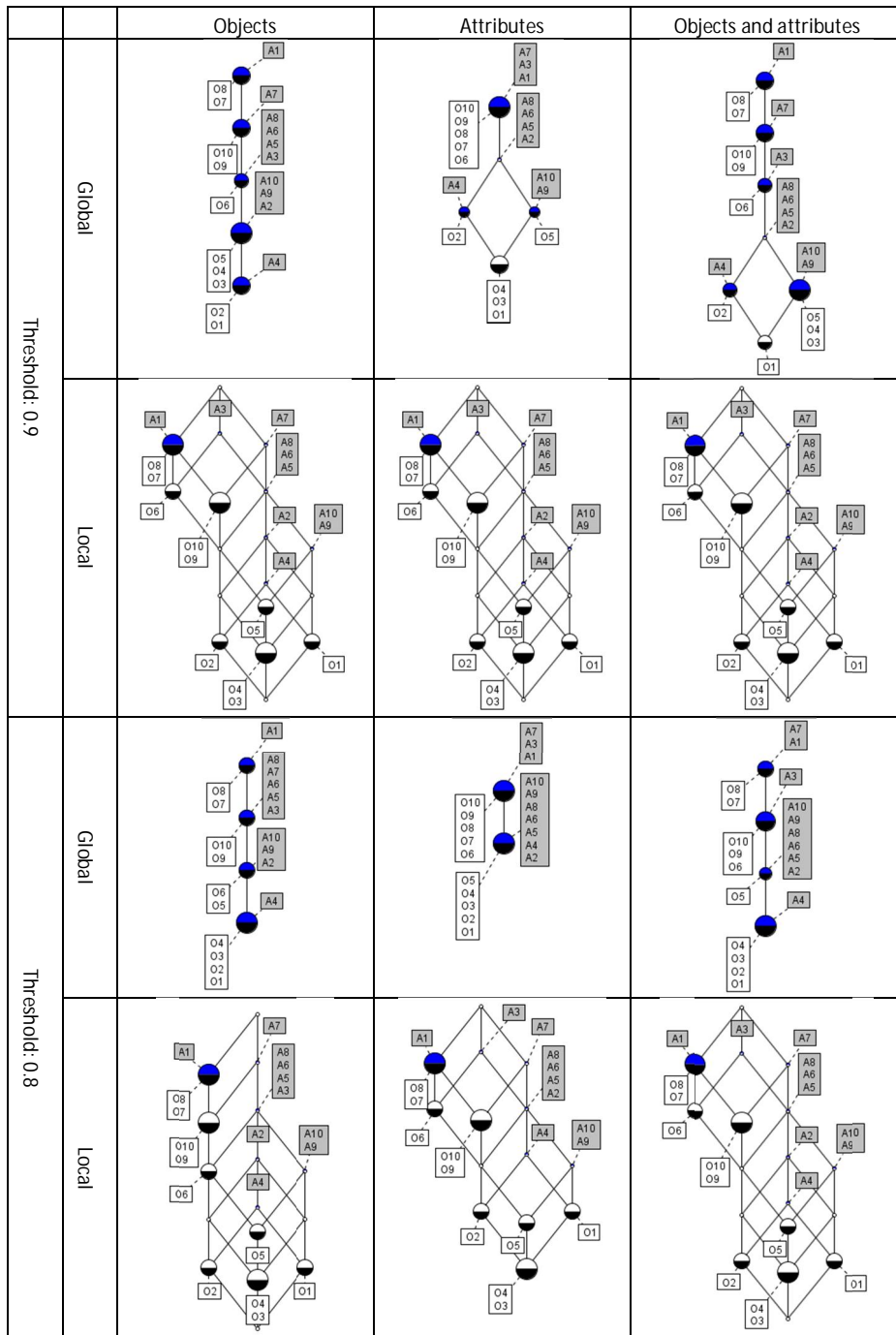
Fig 8: Concept lattices for thresholds 0.8 and 0.9 and all six measures

5   Heriot-Watt University Example

The example of this section is based on the Heriot-Watt University use case in the research project CUBIST, in particular, it is built upon a real data set, namely - embryo gene expression data. Gene expression information describes whether or not a gene is expressed (active) in a location. There is a fixed set of levels of expressiveness: A gene can in some location be *detected*, or more one can more precisely state that it is *weakly*, *moderate* or *strongly detected*. Moreover, it can be stated that a gene is *not detected*, that it is *possible* for the gene to be expressed, or that it is *not examined*. Based on public available data, we have generated a context where the objects are (names of) mouse genes (we have a total of 6613 formal objects in the context), the attributes are the seven different levels of expressiveness, and a cross is set between a gene and a level if that gene is detected in some location with that level of expressiveness. The resulting lattice, having 81 formal concepts, is depicted below.
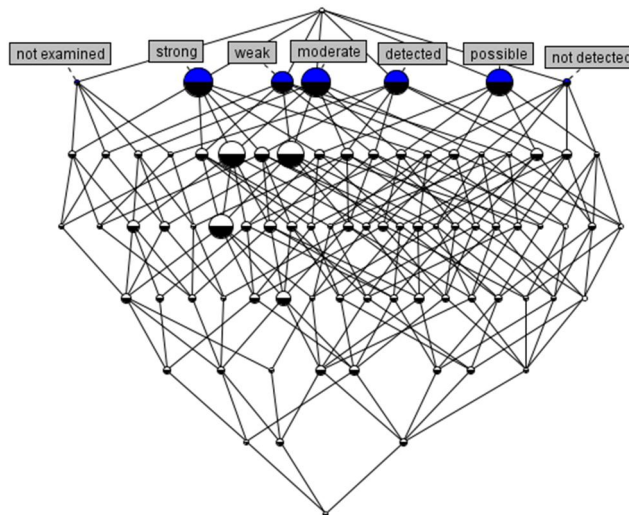
Fig 9: Genes and Levels of Expressiveness without approximation

The lattice clearly shows a well-known challenge for FCA. Most of possible combinations of attributes are indeed instantiated objects in the context, often by few only, (the clarified context still contains 77 objects), which leads to an explosion of the number of concepts. Amongst the concepts, we have some formal concepts with large number of objects, but most concepts are rather small and render the lattice clutterd and hard to read. As we have many objects and only few attributes, it is sensitive to apply the global object-based measurement G_{Obj} to the context. With a (quite high) threshold of 97 already, the concept lattice is reduced significantly, as shown in Fig 10.
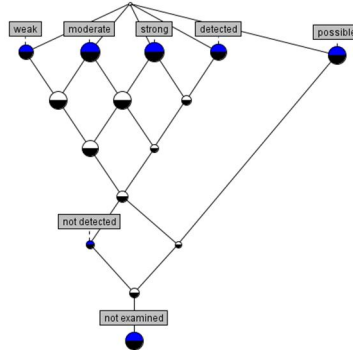
Fig 10: Genes and Levels of Expressiveness with approximation, threshold 0.97

6   Summary and next steps

In this paper, we have introduced a fault tolerance approach for FCA and shown promising experimental results.

Of course, first of all, this approach has to be thoroughly investigated from a formal point of view. It has to be proven that for each approximate incidence measure, the number of formal concepts decreases when the threshold is decreased. More specifically, it has to be investigated how for a given measure and two thresholds $t_1 \geq t_2$ how the concept lattice for t_1 can be mapped onto the concept lattice for t_2. It has to be scrutinized how the different measures relate to each other. And it has to be investigated how the approach taken in this paper relates to other research, most importantly the work of Boulicaut et al.

From a conceptual point of view, it has to be worked out on how derived concept lattices are to be understood, particularly how the relationship of derived lattices and association rules of the orgin lattice is.

Next, from an implementational point of view, we have to elaborate on the algorithm which computes the measures, i.e. scrutinizing its complexity and possibly improving its performance (this is e.g. important for big data sets and real-time applications).

Finally, from an user interface point of view, it would be desirable to have a lattice visualization with a slider to adjust the threshold for a given measure such that adjusting the threshold with the slider leads to animated transitions between the derived lattices. Here the mappings which have been mentioned in the section about the formal investigations are likely to be helpful, and further mathematical investigations might be needed.

In the long run, assuming that the steps mentioned above have been undertaken, the approach taken in this paper is envisioned to bring FCA closer to business intelligence applications and visual analytics, where one needs easy-to-understand and in-

teractive visualizations for large amounts of data, and where a certain and controlled loss of information for the sake of simplified diagrams is permissible.

7   References

1. Andrews, S. and McLeod, K.: Gene Co-Expression in Mouse Embryo Tissues. In: Dau, F. (ed.) 1st CUBIST Workshop, at ICCS 2011, Derby, UK. CEUR Workshop Proceedings, Vol. 753, pp. 1-10. ISSN: 1613-0073
2. Besson, J., Robardet, C., Boulicaut, J.F.: Mining a New Fault-Tolerant Pattern Type as an Alternative to Formal Concept Discovery. Scharfe et al. ICCS. LNAI 4065. Springer (2006)
3. Dau, F.: Towards Scalingless Generation of Formal Contexts from an Ontology in a Triple Store In: Dau, F and Andrews, S: Proceedings of the second CUBIST workshop 2012. KULeuven press, 2012
4. Melo, C.A., Aufaure, M.-A., Le Grand, B. and Bezerianos, A.: Extracting and Visualizing Tree-like Structures from Concept Lattices In: 15th International Conference on Information Visualization (IV 2011). London, UK, 2011.
5. Melo, C.A., Aufaure, M.-A., Bezerianos, A. and Le Grand, B.: Cubix: A Visual Analytics Tool for Formal Concept Analysis. In: 23ième Conférence Francophone Sur l'IHM (IHM 2011) – Demo. Sophia-Antipolis, France, 2011.
6. Pensa, R. G., Boulicaut, J-F.: Towards Fault-Tolerant Formal Concept Analysis. In: Banidini, S., Manzoni, S. (eds.) AI*IA 2005, LNAI 3673, pp. 212{223, Springer-Verlag, Berlin Heidelberg (2005
7. Pensa, G.R:, Boulicaut, J.F.: Towards fault-tolerant formal concept analysis. In: Congress of the Italian Association for Artificial Intelligence AI* IA, 2005
8. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data & Knowledge Engineering, Volume 42, Issue 2, August 2002, Pages 189–222
9. Stumme, G., Taouil, R., Bastide, Y.: Conceptual clustering with iceberg concept lattices. In: Proc. of GI-Fachgruppentreffen Maschinelles Lernen'01, Universität Dortmund, 2001