
# The 1st CUBIST Workshop

## CUBIST-WS-11

Frithjof Dau (Ed.)

Frithjof Dau (Ed.):

# The 1st CUBIST Workshop

## CUBIST-WS-11

# Preface

This volume contains the papers accepted to the first CUBIST workshop. The workshop has been held in conjunction with the 19th International Conference on Conceptual Structures (ICCS 2011), which was held on 25 - 29 July 2011 at the, University of Derby, United Kingdom.

CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) is a research project funded by the European Commission under the 7th Framework Programme of ICT, topic 4.3: Intelligent Information Management, which has started in October 2010. CUBIST follows a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and Visual Analytics. CUBIST aims to

- persist the federated data in a semantic Data Warehouse; a hybrid approach based on a BI enabled triple store,
- and provide novel ways of applying visual analytics in which meaningful diagrammatic representations based on Formal Concept Analysis will be used for depicting the data, navigating through the data and for visually querying the data.

A project like CUBIST requires expertise from a variety of research fields, which can hardly be provided by only one research organization, thus it requires the collaboration of several partners. Moreover, in order to show the practical benefits of the findings in CUBIST, a prototype will be implemented, and the resulting technology stack will be demonstrated in three use cases from the fields of market intelligence, computational biology and control centre operations. Following this approach, the first CUBIST workshop aimed at providing a forum for both research and practice for CUBIST-related research topics and technologies in order to facilitate interdisciplinary discussions. Naturally, the majority of submissions originated from some partners in the CUBIST project, but I am glad to announce that we had submissions from outside the CUBIST consortium as well, which indicates that the project and the workshop are on the right track.

I want to express my appreciation to all authors of submitted papers on the one hand and to the program committee members for their work and valuable comments on the other hands.

Dresden, Germany, June 2011

Frithjof Dau

# CUBIST-WS-11 Organization

## Chair

Frithjof Dau (SAP AG, Germany)

## Program Committee

Alexander Mikhailian (Space Applications Services NV, Belgium)

Kenneth Mcleod (Heriot-Watt University, UK)

Simon Polovina (Sheffield Hallam University, UK)

Simon Andrews (Sheffield Hallam University, UK)

Constantinos Orphanides (Sheffield Hallam University, UK)

Anastasia Bezerianos (Centrale Recherche S.A. (CRSA) – Laboratoire MAS, France)

Marie-Aude Aufare (Centrale Recherche S.A. (CRSA) – Laboratoire MAS, France)

Cassio Melo (Centrale Recherche S.A. (CRSA) – Laboratoire MAS, France)

# Table of Contents

# Gene Co-Expression in Mouse Embryo Tissues

Simon Andrews[1] and Kenneth M$^{c}$Leod[2]

[1] Conceptual Structures Research Group
Communication and Computing Research Centre
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, Sheffield, UK
s.andrews@shu.ac.uk
[2] School of Mathematical and Computer Sciences,
Heriot-Watt University, Edinburgh, UK
kenneth.mcleod@hw.ac.uk

**Abstract.** This paper develops some existing ideas in FCA to provide an analysis of a large data set of mouse embryo gene expressions. It develops new techniques for managing complexity and visualisation in FCA to identify and approximate large groups of co-expressed genes. This work has been carried out as part the European CUBIST Project: http://www.cubist-project.eu/

## 1 Introduction

Formal Concept Analysis (FCA) has already proved useful in the study of gene co-expression. FCA is attractive in the field because formal concepts are natural representations of maximal groups of co-expressed genes. In [5] FCA was used to extract groups of genes with similar expressions profiles from data of the fungus *Laccaria bicolor* and in [4] human SAGE data provides the example from which clusters of concepts with similar properties are visualised. In both approaches the complexity, in terms of the large number of formal concepts present in the raw data, is managed by specifying a concept's minimum size (the well known idea of minimum support in FCA and frequent itemset mining). In [4], tools were developed to query the set of extracted concepts according to various criteria (e.g., presence of a keyword in a gene description) and then to cluster concepts according to similarity, in terms of the attributes (samples) and objects (genes above a threshold of expression) in them. They called these clusters, *quasi-synexpression-groups* (QSGs). By contrast, in [5], ranges of a measure of gene concentration were used as attributes and the genes as objects. Individual concepts that satisfied a specified minimum size were then examined by, for example, plotting the actual measures of concentration of genes together in a line plot.

In this paper we develop some of these ideas and use some freely available, open-source, tools to apply them to a set of mouse-embryo gene expression data. We employ the idea of minimum support to focus on 'large' co-expressions and use a similar notion to that of QSGs in identifying clusters of similar large co-expressions, giving rise to larger *approximate* co-expressions by using a similar

notion to that of FCA 'fault tolerance' [7]. We show that the technique of clustering co-expressions can be straightforward and is a simple way of approximating and visualising a large amount of gene expression data. We demonstrate that FCA can act as a tool for knowledge discovery and can be used to identify possible 'gaps' in knowledge; data that may be missing, erroneous or inconsistent and thus where further investigation or experimentation may be required.

## 2 The mouse-embryo gene expressions data set: EMAGE

A *gene* is a unit of instructions that provides directions for one essential task, i.e., the creation of a protein. Gene expression information describes whether or not a gene is expressed (active) in a location. Broadly speaking there are two types of gene expression information: those that focus on where the gene is expressed, and those whose primary concern is the strength of expression. This work concentrates on the former category, and in particular a technology called *in situ* hybridisation gene expression.

Information on gene expression is often given in relation to a tissue in a particular model organism. Here the model organism is the mouse. This organism is studied from conception until adulthood. The time window is split into 28 Theiler Stages (TS). Each stage has its own anatomy, and corresponding anatomy ontology called EMAP [3].

Gene expression information allows biologists to discover relationships between genes, in particular when genes are active in the same location. This *co-expression* information provides insights into the ways in which relationships between genes affect the development of a tissue.

The result of an *in situ* experiment is documented as an image displaying an area of a mouse (from a particular Theiler Stage) in which some subsections of the mouse are highly coloured. Areas of colour indicate that the gene is expressed in that location. Additionally, the image provides some indication of the level (strength) of expression: the more intense the colour, the stronger the expression.

Results are analysed manually under a microscope. A human expert determines in which tissues the gene is expressed, and at what level of expression. As volume information is not the main focus of the experiment, its description uses vague natural language terms such as strong, moderate, weak or present. For example, the gene *Bmp4* is strongly expressed in the future brain from Theiler Stage 15.

Completed *in situ* gene expression experiments are published online. One of the main resources in this field is EMAGE [8]. EMAGE documents the result of an experiment using a series of *textual annotations*. Each annotation is a triple: gene - tissue - level of expression. The entire collection of annotations is used as the data set for this work.

For the sake of brevity, both genes and tissues will be referred to by short names or identifiers rather than their full name. For example, the gene "bone morphogenetic protein 4" will be referred to as "*Bmp4*". Likewise, the tissue

"mouse.embryo.skeleton.cranium.viscerocranium.orbito-sphenoid from TS 23" will be known by its unique EMAP identifier "EMAP:8385".

## 3 An approach using freely available FCA tools

The approach was to convert the EMAGE data into a formal context, mine the context for concepts satisfying a specified minimum size and then approximate the results using FCA 'fault tolerance'. To do this, three tools that are open source and freely available at *Sourceforge* were used:

- **FcaBedrock** [1] to convert the EMAGE data into a formal context by converting *(tissue, level, gene)* triples into *(tissue-level, gene)* pairs.
- **In-Close** [2] to mine to context for concepts satisfying a specified size and produce a corresponding 'reduced' context.
- **Concept Explorer** (ConExp) [10] to visualise the 'large' concepts and apply 'fault tolerance' to produce even larger, 'approximate' concepts.

In addition to their main tasks, In-Close was used to sort the formal context to allow easy identification of clusters of similar concepts (a simple way of finding QSG-type groupings [4]) and ConExp was used as a context editor to extract these clusters and to provide a simple manual method of producing the larger approximate concepts.

### 3.1 Converting and concept-mining the raw EMAGE data

The EMAGE data set was obtained in the form of csv triples. FcaBedrock was used to automatically convert the data set into a formal context in the standard Burmeister .cxt format. The context contained 6838 attributes (tissue-levels) and 4627 objects (genes). In-Close was used to mine the context generating 208,377 concepts. By a process of trial and error, a minimum size of concept of 14 tissue-levels and 18 genes was determined that produced a reduced context that was possible to visualise in ConExp (Figure 1). Note that the process of visualising the reduced context shows concepts additional to those satisfying the minimum size because where two concepts that satisfy the minimum size 'overlap' in the context grid (share relations), smaller concepts will exist.

### 3.2 Identification of co-expression clusters

There are two large concepts at the bottom of the lattice in Figure 1 giving a suggestion of two distinct clusters of concepts. A clear visualisation of the two groups is shown in the reduced context produced by In-Close (Figure 2). Because In-Close, as part of its processing, sorts context rows to reduce the difference between them, patterns that would otherwise be difficult to detect become clear. It is apparent that there are two disjoint clusters of concepts, i.e., two disjoint clusters of gene co-expression.

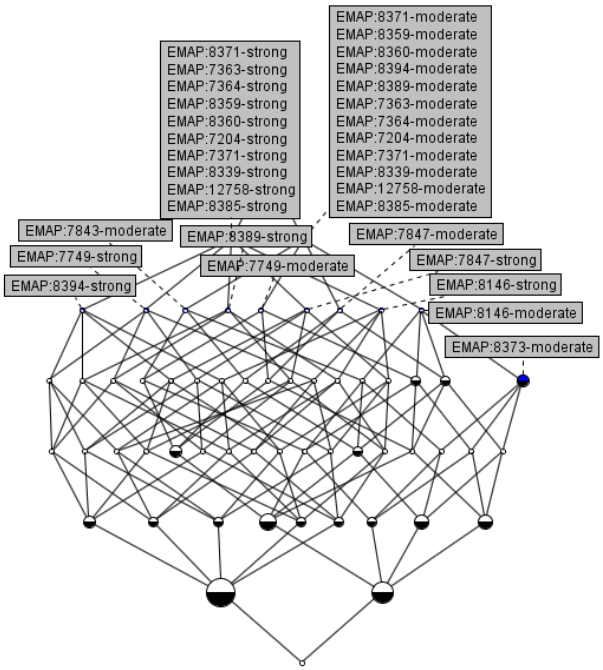



**Fig. 1.** Concept lattice produced from EMAGE gene expression data (for clarity, only the tissue-levels are displayed)

### 3.3 Using fault tolerance to produce large approximate concepts

Figures 3 and 4 show the concept clusters as separate context grids. They now appear as dense grids of crosses with only a few crosses 'missing'. The notion of fault tolerance in FCA [7] is that a certain amount of missing information can be tolerated as being errors of omission, or that at least a sensible approximation is possible by adding a limited number of relations to 'complete' a concept. In Figure 3, for example, there is only one cross missing from the column of EMAP:8394-strong. A fault tolerance level of one gene would add that cross and the ones missing for EMAP:7749-strong, EMAP:8389-strong and EMAP:7847-strong. A fault tolerance level of two would also complete the column for EMAP:8146-strong. It is perhaps equally legitimate to apply fault tolerance to missing attributes, thus a fault tolerance level of three would supply all the missing crosses in both grids. Such an approximation results in the lattice in Figure 5.

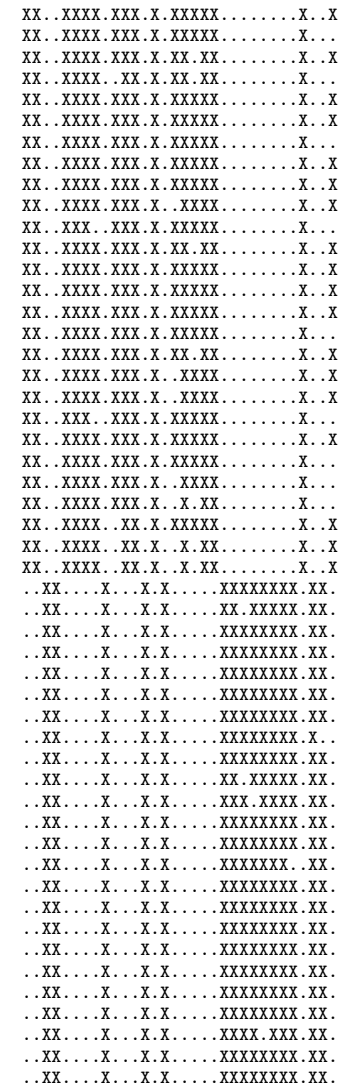



**Fig. 2.** Two distinct clusters in the reduced context

| Gene Co-Exp 1 | EMAP:8385-strong | EMAP:12758-strong | EMAP:8339-strong | EMAP:7371-strong | EMAP:7204-strong | EMAP:8360-strong | EMAP:8359-strong | EMAP:8146-strong | EMAP:7847-strong | EMAP:8389-strong | EMAP:7364-strong | EMAP:7363-strong | EMAP:7749-strong | EMAP:8371-strong | EMAP:8394-strong |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mapk8ip2 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Tgfbi | × | × | × | × | × | × | × | | × | × | × | × | × | × | × |
| Zcchc6 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Brpf3 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Caly | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Bcl9l | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Zc3h18 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Dnajc18 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | |
| H2-T22 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Colec12 | × | × | × | × | × | × | × | | × | × | × | × | × | × | × |
| Wwp2 | × | × | × | × | × | × | × | × | | × | × | × | × | × | × |
| Emp3 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Tcam1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Papss2 | × | × | × | × | × | × | × | × | × | × | × | × | | × | × |
| Ubxn10 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Cytl1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| BC024814 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Plekhb1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Apitd1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Cebpz | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| 1110017D15Rik | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Copb1 | × | × | × | × | × | × | × | × | × | | × | × | × | × | × |
| Unc5cl | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Haus4 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |

**Fig. 3.** Cross-table for cluster 1

| Gene Co-Exp 2 | EMAP:8385-moderate | EMAP:12758-moderate | EMAP:8339-moderate | EMAP:7371-moderate | EMAP:7204-moderate | EMAP:8146-moderate | EMAP:7847-moderate | EMAP:7364-moderate | EMAP:7363-moderate | EMAP:8389-moderate | EMAP:7749-moderate | EMAP:8394-moderate | EMAP:7843-moderate | EMAP:8360-moderate | EMAP:8359-moderate | EMAP:8371-moderate | EMAP:8373-moderate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Sult1c2 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Klk7 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | |
| Ing4 | × | × | × | × | × | × | × | × | × | × | × | × | | × | × | × | × |
| Mir300 | × | × | × | × | × | × | | × | × | × | × | × | | × | × | × | |
| Plcg2 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| U2af1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Fzr1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | |
| Gm22 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Gm10232 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Tnfsf4 | × | × | × | × | × | × | × | × | × | × | | × | × | × | × | × | × |
| Acot6 | × | × | × | × | × | | × | × | × | × | × | × | × | × | × | × | |
| Cit | × | × | × | × | × | × | × | × | × | × | × | × | | × | × | × | × |
| Pop4 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Atp6v0a2 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Ebi3 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Fcgrt | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | |
| Mvk | × | × | × | × | × | × | × | × | × | × | × | × | | × | × | × | × |
| Mdfi | × | × | × | × | × | × | × | × | × | × | | × | × | × | × | × | × |
| Adrm1 | × | × | × | × | × | × | × | × | × | × | | × | × | × | × | × | × |
| Clca1 | × | × | × | × | × | | × | × | × | × | × | × | × | × | × | × | |
| Tnfaip1 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × |
| Arhgap27 | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | × | |
| Tmc5 | × | × | × | × | × | × | × | × | × | × | | × | × | × | × | × | |
| Cops7b | × | × | × | × | × | × | × | × | × | × | | × | | × | × | × | |
| Itpr3 | × | × | × | × | × | × | | × | × | × | × | × | × | × | × | × | × |
| Tsga10ip | × | × | × | × | × | × | | × | × | × | | × | | × | × | × | × |
| Upk2 | × | × | × | × | × | × | | × | × | × | | × | | × | × | × | × |

**Fig. 4.** Cross-table for cluster 2

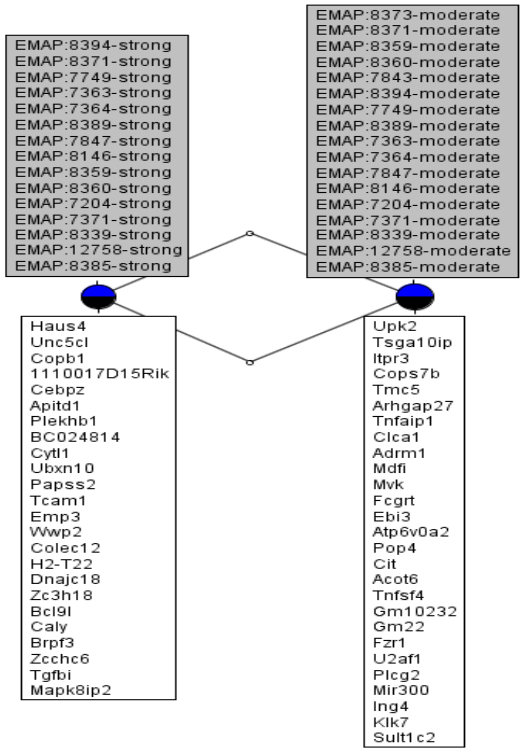



**Fig. 5.** Original lattice with 'fault tolerance' applied

## 4 Analysis of the gene co-expression results

When analysing the output of the FCA process, the first task was to convert the EMAP identifiers back into tissue names to determine which locations were flagged. This revealed that with the exception of EMAP:8146 and EMAP:7749, which are both cartilages, all the tissues are bones.

Intuitively all bones will have a similar expression profile, as they are essentially very similar structures. The list of tissues obtained through FCA demonstrates this as it covers the majority of the mouse including the limbs, body, and head. However, interesting gaps remain: for example, in the list there are no bones from the tail. Why not? Answering this question requires further study.

Looking at the output of the above processes, a reader may ask why this expression profile is found in only one of the twenty-eight Theiler Stages? The answer to this question is that TS 23 has the most experimental data; there are many experiments performed on TS 23 that are not repeated on other stages. As such, this pattern of expression may, or may not, be realised in other stages. Until the requisite experiments have been performed it is impossible to tell.

In a similar vein, many of the "missing" crosses from the cross table are a consequence of EMAGE having no experimental result discussing the gene - tissue pair. As such, it is unknown at what level the gene is expressed in the tissue, or if it is expressed at all. Accordingly, the process has revealed future experiments to perform.

Additionally, observe that EMAGE is only one of a number of resources that serve the current domain. Some of these resources provide proprietary *in situ* gene expression information that is not available to EMAGE, whilst others publish the results of different types of gene expression experiment. By reviewing just one extra resource, GXD [9], it is possible to add a cross missing from the initial lattice: *Acot6* - EMAP:8146 - moderately expressed. This leads to the conclusion that if the data from the other resources were integrated with the data from EMAGE it may be possible to add further crosses. Doing so may produce a "better" cross table, and thus a "truer" lattice. Unfortunately, there are significant difficulties in integrating such data [6], and this has been left as future work.

Future work may also investigate whether or not FCA can help resolve inconsistent information. Unfortunately, due to the nature of biology, a small number of textual annotations are inconsistent, i.e., they suggest different levels of expression for the same gene in the same tissue. Perhaps the process documented in this paper can help identify the most likely level. Furthermore, it might be possible to suggest the probable level of expression when EMAGE contains no data.

## 5 Conclusion

This paper explored FCA within a biological use case. In particular it demonstrated how FCA can be used to analyse *in situ* gene expression data for the developmental mouse.

Analysis was based on large concepts (14 by 18), leaving smaller concepts to be considered as future work. Additionally, further research will be required to understand the full significance of the cross tables documented in this paper.

The list of tissues contained within the cross tables is comprised of a wide selection of bones covering the vast majority of the mouse's skeleton. Yet certain anatomical structures are missing, e.g., the tail. Why are the absent structures not present? What unique features of tail bones prevent them being included in the cross tables?

A further biological question arises in that all the expression levels in each group are the same, i.e., there is a group of genes expressed strongly and a group expressed moderately. There is no reason from an FCA point of view why this should be the case. There may be a biological explanation, perhaps either to do with the nature of the experiments or the nature of the mouse embryo.

From an FCA perspective there are a number outstanding questions too. The appropriateness, and reliability, of fault tolerance needs to be investigated. Additionally, within the context of CUBIST, there is a requirement to improve

the user friendliness of FCA to the extent that a biologist is able to perform the analysis independently of an expert.

Manifestly, the work documented here is at an early stage. Nevertheless, this paper demonstrates there is significant potential that can be exploited for the benefit of both the biological and FCA communities.

**Acknowledgement** This work is part of the CUBIST project ("Combining and Uniting Business Intelligence with Semantic Technologies"), funded by the European Commission's 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.

## References

1. Andrews, S., Orphanides, C.: FcaBedrock, a Formal Context Creator. In: Croitoru, M., Ferre, S. and Lukose, D. (eds.): Proceedings of ICCS 2010, Kuching, Malaysia. LNAI 6208, Springer-Verlag (2010)
2. Andrews, S.: In-Close2, a High Performance Formal Concept Miner. In: Hill, R., Andrews, S., Polovina, S., Akhgar, B. (eds.): Proceedings of ICCS 2011, Derby, UK. Springer-Verlag (in press)
3. Baldock, R., Davidson, D.: Anatomy ontologies for bioinformatics: principles and practise, chap. The Edinburgh Mouse Atlas, pp. 249–265. Springer Verlag (2008)
4. Blachona, S., Pensab, R. G., Bessonb, J., Robardetb, C., Boulicautb, J-F., Gandrillona, O.: Clustering Formal Concepts to Discover Biologically Relevant Knowledge from Gene Expression Data. *In Silico* Biology 7, pp. 467–483 467, IOS Press (2007)
5. Kaytoue-Uberall, M., Duplessis, S., Napoli, A.: Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes Mehdi Kaytoue-Uberall. In Le Thi, H. A., Bouvry, P., Pham Dinh, T. (eds.): Proceedings of MCO 2008, CCIS 14, pp. 445–455, Springer-Verlag, Berlin Heidelberg (2008)
6. M$^c$Leod, K., Ferguson, G., Burger, A.: Argudas: arguing with gene expression information. In: Paschke, A., Burger, A., Splendiani, A., Marshall, M.S., Romano, P. (eds.) Proceedings of the 3rd International Workshop on Semantic Web Applications and Tools for the Life Sciences (December 2010)
7. 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)
8. Richardson, L., Venkataraman, S., Stevenson, P., Yang, Y., Burton, N., Rao, J., Fisher, M., Baldock, R.A., Davidson, D.R., Christiansen, J.H.: EMAGE mouse embryo spatial gene expression database: 2010 update. Nucleic Acids Research 38, Database issue, D703–D709 (2010)
9. Smith, C.M., Finger, J.H., Hayamizu, T.F., M$^c$Cright, I.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., Ringwald, M.: The mouse gene expression database (GXD) : 2007 update. Nucleic Acids Research 35, D618–D623 (2006)
10. 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)

# An Extension of ToscanaJ for FCA-based Data Analysis over Triple Stores

Frithjof Dau and Barış Sertkaya

SAP Research Center Dresden, Germany
(frithjof.dau|baris.sertkaya)@sap.com

**Abstract.** Classical Business Intelligence (BI) solutions provide different means like OLAP, data mining or case based reasoning to explore structured data. The data is usually stored in dedicated repositories like data warehouses and then explored by standard BI means, which are usually based on mathematical statistics and provide a quantitative analysis of the data. In this paper, we complement this approach in two respects. First of all, FCA, as a qualitative approach for data analysis, is used for analyzing the data. Second, the analyzed data is not stored in a data warehouse, but in a triple store instead. To make this possible, the existing FCA-tool ToscanaJ has been extended in order to act on triple stores. The approach in this paper is exemplified on a dataset of documents crawled from the SAP community network.

## 1 Introduction

Business Intelligence (BI) solutions provide different means like OLAP, data mining or case based reasoning to explore data. In the standard BI approach, this data is usually extracted from transactional databases, transformed into a unified schema which is tailored to BI needs, and stored in a dedicated repository like a data warehouse. The standard BI means which are used to explore this data are focusing on attributes which can be numerically measured, thus they provide a *quantitative analysis* of the data (aka "number crunching") based on mathematical statistics. To some extent, though arguably oversimplified, one can understand BI as acting on lists or tables filled with numbers.

Compared to number crunching, Formal Concept Analysis (FCA) [5] provides a complementing approach. The starting point of FCA is a *dyadic* formal context of of formal objects formal attributes and a incidence relation which (only) describes whether attributes apply to objects or not. As FCA builds meaningful clusters based on the objects and attributes, it is a means for *qualitative analysis* of the data, complementing the classical quantitative analysis. Moreover, the clusters (aka formal concepts) are from a hierarchy (a complete lattice) which can be naturally visualized, thus FCA qualifies as a *visual analytics* tool.

A general overview over the benefits of FCA in information science is provided by Priss in [11]. Relevant for this paper are the relationships between FCA and both Business Intelligence (BI) and Semantic Technologies (ST).

With respect to BI, FCA can be for example considered as a data mining technology, particularly for mining association rules [15, 10]. More relevant to this paper is the approach to explore data in relational databases with FCA. As described in the next section, a method called "conceptual scaling" allows transforming columns in a database, being filled with arbitrary values, into formal contexts. Such scales can be compared to dimensions in BI applications. The exploration of data in databases with FCA is for example described in [6, 8, 16, 14]. A number of tools for FCA have been developed. Most important for this paper is Toscana [17, 13, 12] , developed in C, and its Java-based successor ToscanaJ [1][1]. Moreover, it should be mentioned that FCA has been used for exploring data warehouses as well [7].

FCA targets a formalization of the philosophical understanding of concepts with their extensions and intensions, thus FCA indeed is a semantic technology. Though it does not belong to the core of Semantic Web technologies, FCA provides decent means to define and analyze concept hierarchies, so it comes as no surprise that FCA has been used in the realm of querying, browsing and visualization of ontologies (e.g. OWLFCAViewTab for Protege and OntoViz), ontology alignment (e.g. FCA-Merge and OntEx), ontology engineering (e.g. relational exploration or role exploration) and ontology learning (e.g., Text2Onto).[2] In this paper, we exemplify the benefits of FCA for (semantically enabled) BI by analyzing data in a triple store with FCA methods. In order to do so, the existing ToscanaJ tool has been modified such that it can retrieve data from triple stores instead of relational databases. A short introduction into ToscanaJ and its modifications are provided in Sec. 3. An often named benefit of ST compared to relational databases are the ST capabilities to better deal with unstructured data like text-documents. FCA has already been employed to create concept hierarchies out of the content of text documents, e.g. [2] targets the creation of taxonomies out of text corpora, and [3] uses FCA to analyze a corpus of html-pages about rental offers for flats and houses.

In this paper, we apply FCA on a dataset of documents crawled from the SAP community network[3] (SCN), but do not target to investigate the contents of the documents, but utilize meta-data of the documents (which have been created in the crawling process) for FCA-purposes. This use case is described in Sec. 4.

## 2 FCA and Conceptual Scaling

FCA per se can only deal with *binary* attributes. For real data, the situation is usually different: Attributes assign specific values (which might be strings, numbers, etc) to data. For example, RDF-triples $(s, p, o)$ are exactly of this form: The attribute $p$ - from now on we will use the RDF-term "property" instead- assigns the value $o$ to the entity $s$. In FCA, a process called "conceptual scaling"

[1] `http://toscanaj.sourceforge.net`

[2] As this field is only weakly related to this paper, no references are given.

[3] `http://www.sdn.sap.com/irj/scn/index`

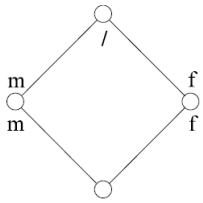
is used to deal with this issue. As ToscanaJ heavily uses scales, we recapitulate their core notions.

Let a specific property be given with a set of possible values. A *conceptual scale* is a specific context with the values of the property as formal objects. The choice of the formal attributes of the scale is a question of the design of the scale: The formal attributes are meaningful attributes to describe the values; they might be different entities or they might even be the values of the property again. To exemplify conceptual scaling, we reuse a toy example from [18], which is the following table provided on the right with two many-valued properties "sex" and "age". Note that empty cells are possible as well.

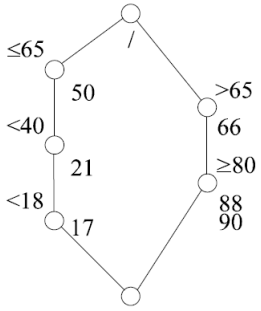
| | sex | age |
|---|---|---|
| Adam | m | 21 |
| Betty | f | 50 |
| Chris | | 66 |
| Dora | f | 88 |
| Eva | f | 17 |
| Fred | m | |
| George | m | 90 |
| Harry | m | 50 |

Next, two conceptual scales for the properties "sex" and "age" and their line diagrams are provided.

| $S_1$ | m | f |
|---|---|---|
| m | × | |
| f | | × |




| $S_2$ | $< 18$ | $< 40$ | $\leq 65$ | $> 65$ | $\geq 80$ |
|---|---|---|---|---|---|
| 17 | × | × | × | | |
| 21 | | × | × | | |
| 50 | | | × | | |
| 66 | | | | × | |
| 88 | | | | × | × |
| 90 | | | | × | × |




With the conceptual scales, the initial many-valued context can be transformed into a standard context, so that the corresponding concept lattice can be displayed. In [5], a number of standard scales is listed, like nominal scales, ordinal scales, and interordinal scales. An introduction into these scales is anyhow beyond the scope of this paper.

For conceptual scales, the following two points should be noted:

1. There is usually no standard or even necessary interpretation of an attribute: It has to be decided by the field expert which scale is appropriate. As discussed in [4], it can be argued that this is indeed not a drawback, but an advantage of FCA. On the other hand, particularly for repositories where the data schema is not stable but continuously changed and improved (which is

particularly the case for semantic repositories) it might not always be feasible to create all conceptual scales beforehand. For example, a nominal scale (like for the property "sex") could be created on the fly.
2. Conceptual scales do not depend on the real data, but only on the properties (and their values, of course) used in the data set. As one can see in the example, a realized context is derived from the scales and the real data in a later step after the scales have been created.

Both points are important for ToscanaJ, which is discussed in the next section.

## 3 ToscanaJ

There is a variety of software for FCA available. Most of them support the creation of contexts from scratch and the subsequent computation and display of the corresponding concept lattices. Contrasting this approach, Elba and ToscanaJ are a suite of mature FCA-tools which allow to query and navigate through data *in databases*. They are intended to be a *Conceptual Information System (CIS)*. CISs are "systems that store, process, and present information using concept-oriented representations supporting tasks like data analysis, information retrieval, or theory building in a human-centered way." Here, a CIS is an FCA-based system used to analyze data stored in one table of an RDBMS.

Similar to other BI-systems, in CIS we have to distinguish between a design phase and a run-time-phase (aka usage phase), with appropriate roles attached to the phases. In the design phase, a CIS engineer (being an expert for the CIS) together with a domain expert who has limited knowledge of a CIS) develops the CIS schema, i.e. those structures which will be later on used to access the system. This schema consists of manually created conceptual scales. Developing the scales is done with a CIS editor (Elba) and usually a highly iterative process. In the run-time phase, a CIS browser (ToscanaJ) allows a user to explore and analyse the real data in the database with the CIS schema. The diagram in Fig. reffig:ElbaToscanaJWorkflow depicts this overall workflow.

From the author's point of view, ToscanaJ is that tool which allows best (if only) to apply a BI-like approach to analyze data with FCA-methods, thus –to some extent– it comes closest to the envisioned CUBIST-system (though the envisioned CUBIST-system will provide different functionalities than ToscanaJ). Anyhow, the original Ebla/ToscanaJ-suite has been developed to analyze data in a *relational table*, i.e. a table in a RDBMS or an excel-file. We have extended the suite in order to be able to access data in a *triple store*. The reasons for doing so are twofold:

- On the one hand, with the extended version of ToscanaJ we have a first FCA-system which is capable to apply FCA-based visual analytics on top of a triple store.
- On the other hand, analyzing the pros and cons of ToscanaJ w.r.t. the envisioned CUBIST-system will help to develop a CUBIST prototype which will overcome the shortcomings (with respect to BI-applications) of existing FCA-tools.

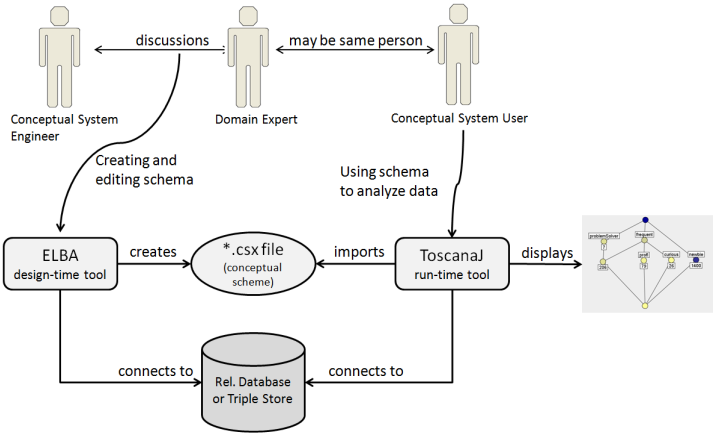



**Fig. 1.** Elba and ToscanaJ workflow

This extended version of the suite uses the Sesame framework[4] for accessing a triple store and and querying the RDF data therein. It provides two ways of connecting to a triple store over Sesame. One of them is over HTTP via Apache Tomcat[5], the other one is over the SAIL API[6]. Tomcat is an open source software implementation of the Java Servlet and JavaServer Pages technologies by the Apache Software Foundation. The SAIL API (Storage And Inference Layer) is a low level system API for RDF stores and inferencers. It is used for abstracting from the storage details, allowing various types of storage and inference to be used.

In a triple store we do not directly have the notions of tables and columns like in databases. As table information we use the type information in the triple store: we treat the objects of triples with the predicate `rdf:type` as tables. As column information, we use the predicates relating the subjects of the selected type to any object. More precisely, in order to detect the columns we get those subjects of the selected type and retrieve all distinct predicates that relate these subjects to an object.

The Elba/ToscanaJ-suite provides different kinds of conceptual scales. We have extended three of them –namely nominal scales, attribute scales and context tables– in order to act on triple stores.

**Nominal scales** are the simplest type of scales one can automatically create in Elba. They are used for properties like `gender` or `country` with mutually exclusive values. The formal attributes of the nominal scale are selected values of that property. As each object is assigned at most one of these values, the attribute concepts form an anti-chain, thus the scale cannot reveal any insight into attribute dependencies.

[4] see http://www.openrdf.org/doc/sesame2/system
[5] see http://tomcat.apache.org/
[6] see http://www.openrdf.org/doc/sesame2/system/ch05.html

**Attributes scales** offer an attribute-centered view, being close to "classical" formal contexts and allowing to create complex scales in an intuitive manner. Here each attribute is a property-value pair, which is manually selected from the triple store. Moreover, the CIS engineer can choose between a) "use only combinations existing in the database" and b) "use all possible combination". For option a) the diagram will only consist of concepts that could be derived from the data in the triple store, thus the diagram will reveal insights into dependencies between property-value pairs. If b) is chosen, a diagram of a Boolean lattice of all listed property-value pairs will be created independently of whether there exists objects in the triple store for each property-value combination or not.

**Context table scales** offer the most freedom to the CIS engineer. In context tables, arbitrary labels act as formal attributes. In contrast to the last two types of scales, no property-value pairs are chosen as attributes, thus now it has explicitly to be specified which objects of the data set fulfil the formal attributes. This is done by entering SPARQL expressions, which act as formal objects, and by entering the incidence relation as well, i.e. the relation which here relates the formal objects (SPARQL expressions) to the attributes (labels).

## 4 Use case

In order to evaluate our approach, we have used a dataset crawled from the SAP Community Network (SCN). SCN contains a number of forums for SAP users and experts to share knowledge. Our dataset is taken from the forum *Service-Oriented Architecture (SOA)*, which contains 2600 threads and 10076 messages. The dataset is annotated by the crawler using ontologies from the NEPOMUK project. The used ontologies and short descriptions[7]. are provided below.

- NEPOMUK Information Element Ontology (NIE): The NIE Framework is an attempt to provide unified vocabulary for describing native resources available on the desktop.
- NEPOMUK file ontology (NFO): The NFO intends to provide vocabulary to express information extracted from various sources. They include files, pieces of software and remote hosts.
- NEPOMUK Message Ontology (NMO): The NMO extends the NIE framework into the domain of messages. Kinds of messages covered by NMO include Emails and instant messages.
- NEPOMUK Contact Ontology (NCO): The NCO describes contact information, common in many places on the desktop.

From these ontologies, our dataset uses the following classes as types:

- nie#DataObject: A unit of data that is created, annotated and processed on the user desktop. It represents a native structure the user works with. This may be a file, a set of files or a part of a file.
- nfo#RemoteDataObject: A file data object stored at a remote location.

[7] taken from the project website `http://www.semanticdesktop.org/ontologies`

– nie#InformationElement: A unit of content the user works with. This is a superclass for all interpretations of a DataObject.
– nco#Contact: A Contact. A piece of data that can provide means to identify or communicate with an entity.
– nmo#Message: A message. Could be an email, instant messanging message, SMS message etc.

For analyzing experience levels of the users of the SOA forum, we used the *Contact* type above and created a scale based on the number of posts, number of questions, number of resolved questions information provided in the data. We have named users that have less than 50 posts as *newbie*, users that have more than 400 posts as *frequent*, users that have more than 1000 posts as *profi*. Note that by definition, every profi is a frequent user as well, but no one can be both newbie and frequent (or profi) user. Moreover, we label users that have asked more than 200 questions as *curious* and people that have resolved more than 300 questions as *problem solver*. Note that this scale uses different measures (number of posts, number of questions, numbers of answers). In Fig. 2 it is shown how the context table in Elba is used to create the appropriate scale, and in Fig. 3 we see in Elbe the corresponding lattice. Note how SPARQL-queries are utilized to describe set of objects.

| | newbie | frequent | profi | curious | problemSolver |
|---|---|---|---|---|---|
| ?s <http://forums.sdn.s... | X | | | | |
| ?s <http://forums.sdn.s... | | X | | X | X |
| ?s <http://forums.sdn.s... | | X | X | | |
| ?s <http://forums.sdn.s... | | | | X | |
| ?s <http://forums.sdn.s... | | | | | X |

**Fig. 2.** Designing in Elba a context table for contacts based on number of posts, questions, resolved questions

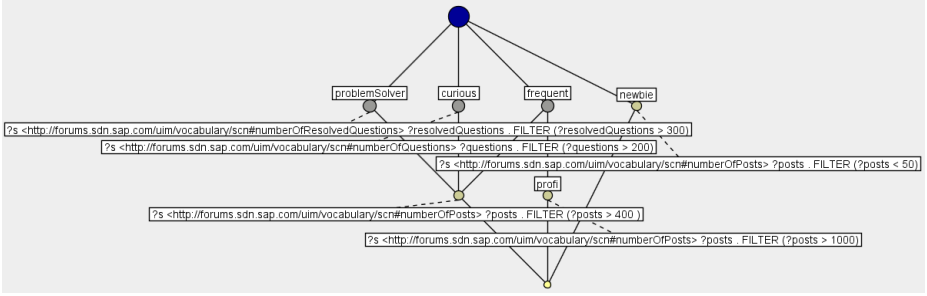



**Fig. 3.** Diagram of the context in Fig. 2

Next, for analyzing experience levels based on the *number of points* information we created another scale. This time, as labels we took contributor types that are officially defined by SCN as *bronze*, *silver*, *gold* and *platinium* contributors, which have more than 250, 500, 1500 and 2500 points respectively. Of course, there might be persons who do not have any of these "medals" at all. We have two possibilities to understand the medals, as depicted in Fig.4.

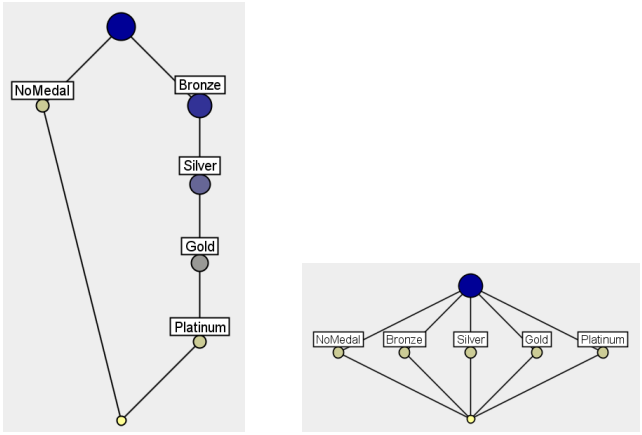



**Fig. 4.** Different scales for the status of contributors

The two approaches for modelling the scales shall be exemplified with gold owners. A gold owner has at least 1500 points. Either we consider a gold owner to be a silver owner and bronze owner as well (as he has more than 500 points to reach silver status and more than 250 points to reach bronze status), or we consider him to be not a silver or bronze owner (that is, a gold owner has *between* 1500 and 2499 points). Thus the two scales in Fig. 4 capture the different notions (aka meanings or semantics) of the status.

Besides these three scales, more scales have been created in Elba, which are not described here due to space limitations. We show next how the scales are utilized in ToscanaJ. First, in Fig. 5 and Fig. 6, we show how the diagrams we have created in Elba now appear with actual numbers in ToscanaJ.

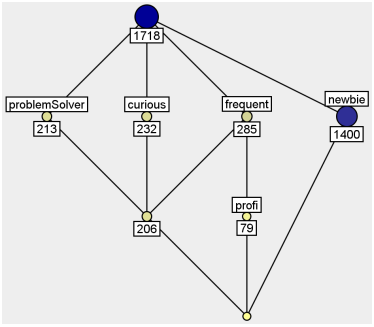



**Fig. 5.** Diagram of the scale based on number of posts, questions, resolved questions

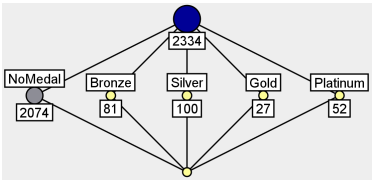



**Fig. 6.** Diagram of the scale based on number of points

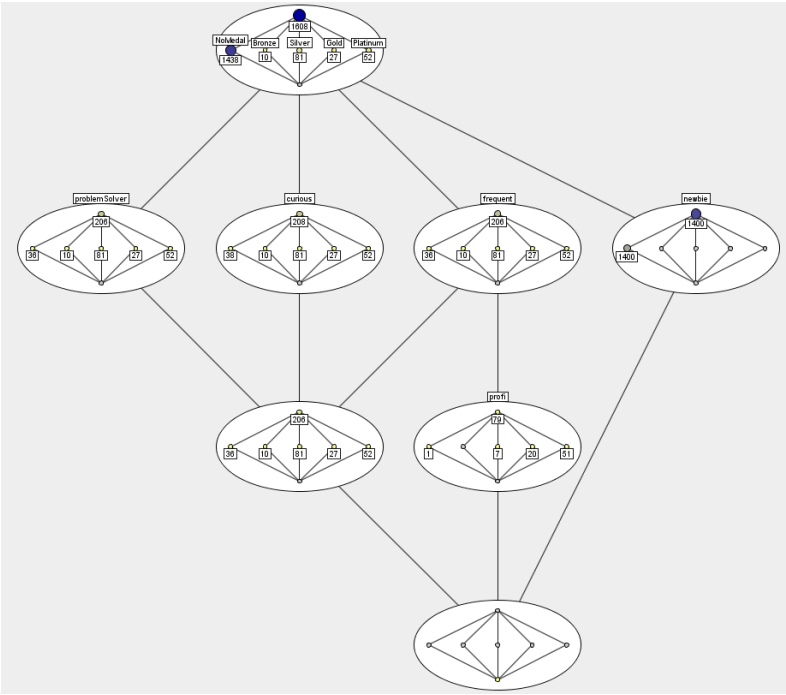



**Fig. 7.** Nested Diagram of two scales

The above displayed concept lattices are separately informative about the properties of forum users, i.e., the first one about experience levels based on number of posts, questions and resolved questions, and the second one about number of points. One of the most powerful techniques of FCA is to "combine" such lattices to give a combined view of several lattices together, which is called a *nested line diagram*. In its simplest form, a nested line diagram is a concept lattice whose concepts are themselves also concept lattices. Nested line diagrams allow the user to select a concept and zoom into it to see the lattice nested in that concept.

Figure 7 shows the nested line diagram of the diagrams in the Figures 5 and 6. Note that the outer diagram is actually the one in Figure 4:experienceLevels. The inner diagrams are the diagram in Fig. 6. Figure 8 shows an excerpt of the nested diagram that corresponds to the node *profi contributor*, and Figure 9 shows the inner diagram of this node. Note that the number of users corresponding to different levels of experience in this diagram differs from that of diagram in Figure 5. The reason is that, now we zoomed into the node gold contributor so the information in the inner diagram is restricted to the profi contributors only. For instance, as seen in this diagram that we have a high number of gold and platinum contributors (but no bronze contributors) amongst profi users, which is

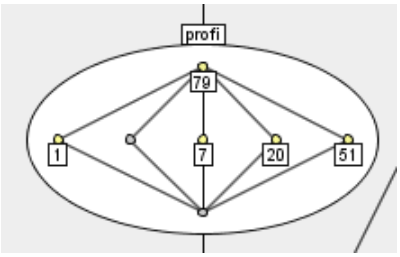



**Fig. 8.** Detail of Figure 7

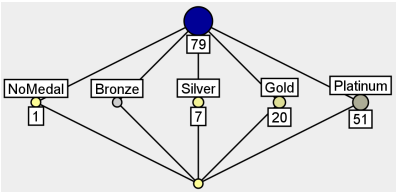



**Fig. 9.** Zooming into the node in Figure 8

quite natural. On the other hand, the one profi user with no medal is a surprising outlier which might deserve a dedicated investigation. In ToscanaJ, and thus in our extension of it to triple stores, one can nest an arbitrary number of diagrams and can browse nested diagrams easily by zooming in and out.

## 5 Conclusion and Further Research

This paper has shown how FCA can be applied data analysis methodology for data in triple stores, and the existing ToscanaJ suite has been extended in order to act not only on relational databases, but on triple stores as well.

As scales in the ToscanaJ workflow are manually crafted in the design phase of a CIS, this workflow is feasible for stable schemata. For ST, this is usually not the case: here the paradigm of agile schema development is prevalent. We plan to implement automatic or at least semi-automatic generation of scales based both on the schema information and the actual data in the triple store.

As stated in the introduction, FCA should be understood to *complement* existing BI approaches. ToscanaJ utilizes scales, thus lattices, for *all* kind of data, even for numerical attributes (which can be modelled as ordinal scales) and nominal attributes. The lattice structure of the scales do not reveal any structural insights, only the distribution of objects amongst the lattice nodes is interesting. Thus it can be argued that for these types of attributes, the standard visualization of BI like pie or bar charts are more appropriate. For this reason, another future research direction of CUBIST is the development of hybrid solutions, combining "classical" BI with FCA. This covers combinations of scales and their diagrams with BI diagrams for numerical data, like pie charts or sun-burst diagrams, and compared to nesting of scales, different approaches for using simultaneously several scales. In the long run, CUBIST intends to combine visualizations for both quantitative and qualitative analytics.

Disclaimer: Parts of this work have been carried out in the CUBIST project, which is funded by the European Commission under the 7th Framework Programme of ICT, topic 4.3: Intelligent Information Management.

## References


1. P. Becker, J. Hereth, and G. Stumme. ToscanaJ: An open source tool for qualitative data analysis. In V. Duquenne, B. Ganter, M. Liquiere, E. M. Nguifo, and G. Stumme, editors, *Advances in Formal Concept Analysis for Knowledge Discovery in Databases, (FCAKDD 2002)*, 2002.
2. P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24(1):305–339, 2005.
3. R. Cole and P. Eklund. Browsing semi-structured web texts using formal concept analysis. In *Proceedings of the 9th International Conference on Conceptual Structures (ICCS 2001)*, pages 319–332, 2001.
4. F. Dau and J. Klinger. From formal concept analysis to contextual logic. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*, pages 81–100. Springer-Verlag, 2005.
5. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, Berlin, Germany, 1999.
6. J. Hereth. Formale begriffsanalyse und data warehousing. Masters thesis, TU Darmstadt, Germany, 2000.
7. J. Hereth. Relational scaling and databases. In U. Priss, D. Corbett, and G. Angelova, editors, *Conceptual Structures: Integration and Interfaces, Proceedings of the 10th International Conference on Conceptual Structures, (ICCS 2002)*, volume 2393 of *Lecture Notes in Computer Science*, pages 62–76. Springer-Verlag, 2002.
8. J. Hereth, G. Stumme, R. Wille, and U. Wille. Conceptual knowledge discovery - a human-centered approach. *Journal of Applied Artificial Intelligence (AAI)*, 17(3):281–301, 2003.
9. W. H. Inmon, D. Strauss, and G. Neushloss. *DW 2.0: The Architecture for the Next Generation of Data Warehousing*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.
10. L. Lakhal and G. Stumme. Efficient mining of association rules based on formal concept analysis. In B. Ganter, G. Stumme, and R. Wille, editors, *Formal Concept Analysis: Foundations and Applications*, volume 3626 of *Lecture Notes in Artificial Intelligence*, pages 180–195. Springer-Verlag, 2005.
11. U. Priss. Formal concept analysis in information science. *Annual Review of Information Science and Technology*, 40, 2005.
12. M. Roth-Hintz, M. Mieth, T. Wetter, S. Strahringer, B. Groh, and R. Wille. Investigating snomed by formal concept analysis. In *Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2000.
13. P. Scheich, M. Skorsky, F. Vogt, C. Wachter, and R. Wille. Conceptual data systems. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 72–84. Springer, Berlin-Heidelberg, 1993.
14. G. Stumme. Conceptual on-line analytical processing. In K. Tanaka, S. Ghandeharizadeh, and Y. Kambayashi, editors, *Information Organization and Databases*, chapter 14. Kluwer, Boston-Dordrecht-London, 2000.
15. G. Stumme. Efficient data mining based on formal concept analysis. In A. Hameurlain, R. Cicchetti, and R. Traunm uller, editors, *Database and Expert Systems Applications. Proceedings of DEXA 2002*, volume 2453 of *Lecture Notes in Computer Science*, pages 534–546. Springer-Verlag, 2002.

16. G. Stumme, R. Wille, and U. Wille. Conceptual knowledge discovery in databases using formal concept analysis methods. In J. M. Zytkow and M. Quafofou, editors, *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)*, volume 1510 of *Lecture Notes in Artificial Intelligence*, pages 450–458. Springer-Verlag, 1998.
17. F. Vogt and R. Wille. Toscana — a graphical tool for analyzing and exploring data. In R. Tamassia and I. G. Tollis, editors, *Graph Drawing*, pages 226–233. Springer-Verlag, 1995.
18. K. E. Wolff. A first course in formal concept analysis. In F. Faulbaum, editor, *Proceedings of Advances in Statistical Software 4*, pages 429–438, 1993.

# Parent Selection Criterion for Extracting Trees from Concept Lattices

Cassio Melo[1], Bénédicte Le-Grand[2], Anastasia Bezerianos[1], Marie-Aude Aufaure[1],

[1] École Centrale Paris – MAS Laboratoire,
69121 Chatenay-Malabry, France
[2] Laboratoire d'Informatique 6 – LIP6,
69121 Paris, France

{Cassio.Melo, Anastasia.Bezerianos, Marie-Aude.Aufaure}@ecp.fr
Benedicte.Le-grand@lip6.fr

**Abstract.** Traditional software in Formal Concept Analysis makes little use of visualization techniques, producing poorly readable concept lattice representations when the number of concepts exceeds a few dozens. This is problematic as the number of concepts in such lattices grows significantly with the size of the data and the number of its dimensions. In this work we propose several methods to enhance the readability of concept lattices firstly though colouring and distortion techniques, and secondly by extracting and visualizing trees derived from concept lattice structures. These contributions represent an important step in the visual analysis of conceptual structures, as domain experts may visually explore larger datasets that traditional visualizations of concept lattice cannot represent effectively.

**Keywords:** Concept Lattices, Formal Concept Analysis, Tree Extraction.

## 1 Introduction

The vast amount of data generated over the last decades has brought new challenges to the analytics science. Visual data analysis and knowledge representation employ methods such as Formal Concept Analysis (FCA) in order to identify groupings of patterns from the analysis process [14]. FCA provides an intuitive understanding of generalization and specialization relationships among objects and their attributes in a structure known as a concept lattice. A concept lattice is traditionally represented by a Hasse diagram illustrating the groupings of objects described by common attributes. A Hasse diagram is a graph where concepts appear as vertices on the plane connected by line segments or curves. The layout of the partially ordered set may be seen as a layered diagram [2]. Lattices visualization becomes a problem as the number of clusters grows significantly with the number of objects and attributes. Interpreting the lattice through a direct visualization of the line diagram rapidly becomes impossible and more synthetic representations are needed.

In this work we propose alternatives to the traditional lattice representation, firstly by enhancing the readability of concept lattices though colouring and distortion

techniques; secondly by extracting and visualizing trees derived from the lattices structure. The tree extraction from the original lattice has some unique advantages: it eliminates all edges crossing and the resulting hierarchy is also easier to interpret and to represent. Moreover, this representation still provides an overview of the dataset, highlighting significant properties of the lattice. In order to extract trees from lattices, we define a set of parent concept selection criteria, including the stability and support indexes [1,4] provided by FCA literature, confidence index as well as topological features of the lattice.

The paper is organized as follows. Section 2 provides background on lattice representations; Section 3 proposes a set of criteria for transforming concept lattices into trees; Section 4 discusses colouring and distortion techniques for enhancing interpretations of lattices. Section 5 presents instantiations of the suggested criteria and visualizations in the biology domain, followed by a discussion in section 6. Section 7 finally concludes and presents perspectives for future work.

## 2. Visual Representation of Concept Lattices

As mentioned above, FCA analysis produces lattices, usually represented as layered directed acyclic graph graphs, named Hasse diagrams, that illustrate the groupings of objects described by common attributes. Hasse diagrams display the partially ordered sets (posets) between concepts in a hierarchical fashion, where each concept may have several parent concepts as illustrated in figure 1. The partial order among concepts of the lattice is materialized through the generalization and specialization relationships: for instance the concept representing the set of *flying birds*, containing *Finch* and *Eagle* objects, is more specific than the one which contains all *birds* – flying or not-, and thus contains a smaller number of objects (the first concept has an extra one, the *ostrich*). This partial order provides different levels of abstraction and native navigation links from a given concept.

As mentioned earlier, such diagrams are usually layered graphs, where concept vertices are assigned to horizontal layers according of the number of common attributes, and are ordered within each layer to reduce edge crossings. FCA lattices in particular suffer from considerable edge crossings, especially if the number of concepts exceeds a few dozen as is the case in more real word applications [13], which leads to reduced graph readability and aesthetics [3].

To reduce the complexity of lattices, simplified diagrams can be produced by displaying only concepts with a sufficient support [4]. Visualisations can also be restricted to portions of the data [5], and concept number reduction is possible by incorporating conditions into the data mining process [6]. Finally, conceptual measures can be applied to identify the most relevant concepts and filter outliers [7].

To deal specifically with the visual complexity of Hasse diagrams, several approaches allow users to dynamically explore and reveal specific parts of the diagram, using visual query languages [8-10]. However these techniques do not provide a clear view of the entire lattice.

Other FCA visualization approaches map the distances between concepts to visual variables, in order to highlight patterns. For example in [11] similar concepts are

represented as similarly coloured pixels placed in the 2D space along a Peano-Hilbert curve, so that similar concepts are placed close to each other. Nevertheless in these representations detailed relationships between concepts are lost. Finally, systems often provide users with hybrid/combined lattice visualization, e.g. showing both a general Hasse diagram and a tag cloud for representing the neighbours of a specific concept (for a review see [12]).

Our approach consists in representing lattices not as Hasse diagrams, but as trees. We use different criteria to extract trees from lattices, and visualize the resulting trees. Trees are inherently simpler hierarchical structures than Hasse diagrams and due to their applicability in many domains, there is a plethora of tree representations.

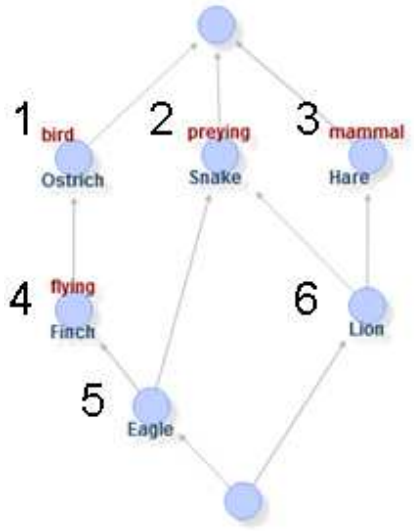



**Figure 1**. An example of animal's concept lattice.

## 3. Tree Extraction from Concept Lattices

Trees are a common and easily understandable visual representation. We consider them as a visualization alternative to large cluttered concept lattices, which preserves all lattice entities and some of its structure. In order for a tree visualization to be an effective alternative to a lattice, the extraction of the tree from the lattice needs to preserve the most essential features of the original structure.

The present approach consists in extracting a tree from a concept lattice by choosing one single parent concept for each concept of the lattice. We start from the most specific concepts i.e. the parent concepts of the lower bound of the lattice, at the bottom of the Hasse diagram and select a single parent concept for each of them, and reproduce this recursively. Choosing a single parent concept at each step leads to an information loss. Our goal is to minimize this loss by selecting parents using the most

relevant criteria according to the kind of analysis performed by the analyst. Before proceeding, we briefly recall the FCA terminology [14]. Given a (formal) context K = (G,M, I), where G is called a set of objects or extent, M is called a set of attributes or intent, and the binary relation $I \subseteq G \times M$ specifies which objects have which attributes, the *derivation* operators $(\cdot)'$ are defined for $A \subseteq G$ and $B \subseteq M$:

$$A' = \{m \in M \mid \forall g \in A : g\mathrm{I}m\};$$
$$B' = \{g \in G \mid \forall m \in B : g\mathrm{I}m\}.$$

In the following sections we consider various strategies for selecting parent concepts, including the *stability* and *support* indexes from FCA literature, *confidence*, as well as topological features of the lattice.

### 3.1. Parent Selection based on the highest Stability or Support

The stability index measures the proportion of subsets of *objects* of a given concept whose derivation is equal to the *intent* of this concept [1]. In other words, the *stability* indicates the probability of preserving a concept *intent* while removing some objects of its *extent*. We recall the definition of stability:

**Definition 1**. Let *K = (G,M,I) be a formal context and (A,B) be a formal concept of K. Card is a cardinality function. The stability index of (A,B) is defined as:*

$$\sigma(A,B) = \frac{Card(\{C \subseteq A \mid C\prime = B\})}{2^{Card(A)}} \quad (1)$$

Using the lattice in figure 1 as an example, we calculate the *stability* for concepts 2 and 4 in order to select a parent for concept 5 (0.25 and 0.5 respectively); we keep the one with highest *stability*, in this case we therefore remove the edge between concepts 2 and 5. The idea behind the choice of the parent concept with the highest *stability* is that we expect to keep parent concept's meaning even if some of the objects or attributes are removed. Another measure which can be used for assigning to each concept a unique upper neighbor is the notion of *'support'* [4]:

**Definition 2**. Let $B \subset M$. *The support count of the attribute set B in K is:*

$$\varphi(B) = \frac{Card(B\prime)}{Card(G)} \quad (2)$$

The use of support as parent selection criteria may lead to trees containing concepts that have fewer specialization levels since in general, generic concepts have higher support values than their most specific counterparts [4]. Concept *stability* and *support* measures have been widely used in FCA and their combination has been promising [1] in reducting the lattice.

### 3.2. Parent Selection Based on Shared Attributes and Objects

This approach relies on clustering parent and child concepts which share most of their attributes or objects. Parent and child having a great number of attributes in common are supposed to be grouped together following the principle of similarity clustering and local predictability [15]. Its definition is:

**Definition 3**. Let *Parent Concept (A,B) be such that* $A \subset G$ *and* $B \subset M$. *Let Child Concept (C,D) be* $C \subset G$ *and* $D \subset M$. *The shared attribute index of an edge E (C,D)→(A,B):*

$$\phi(E) = \frac{Card(B \cap D)}{Card(M)} \quad (3)$$

In the same animal's context illustrated by the lattice in figure 1, we have potential parent concepts 2 and 4 sharing the same number of objects with concept 5, but concept 4 has more attributes in common with 5, so it should be chosen as the unique parent of concept 5.

### 3.3. Parent Selection Based on Confidence

The *confidence* value of a concept estimates how likely an object which has an attribute set A, also has an attribute set C [14]. In other words, it tries to measure how strong the *implication* of the parent attributes in the child objects is. For instance, considering the lattice in figure 1, what is the probability of a given object that is {*Bird, Flying*} to be also {*Bird, Flying, Preying*}? The following paragraph formalizes its definition.

**Definition 4**. Let *Parent Concept (A,B) be such that* $A \subset G$ *and* $B \subset M$. *Let Child Concept (C,D) be* $C \subset G$ *and* $D \subset M$. *The confidence of an edge E (C,D)→(A,B):*

$$\delta(E) = \frac{Card(C)}{Card(A)} \quad (4)$$

An advantage of this method is its consistency with the interpretation of concept lattices. Taking our animals context as example, there is a 50% probability that an animal that is a *flying bird* is also a *flying* and *preying bird*. By contrast, an animal that is *preying* has only 33% of chance to be also a *flying bird*.

## 4. Using extraction criteria to enhance Lattice and Tree Interpretation through Drawing, Sizing and Shaping

Common graph drawing techniques include the assignment of different colours, shapes and sizes to nodes and edges, according to different dimensions or properties. This approach is underused in traditional lattice visualizations, where the main visual

variable used is node/link colour to reflect user selections or node size to indicate the immediate presence of an extent or intent as displayed in ConExp[1].

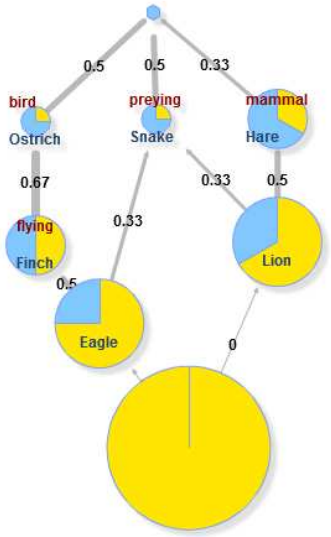



**Figure 2**. Animal lattice with nodes as pie charts sized by stability, and edge thickness by confidence. Pie charts indicate the ratio intent/extent of the concept.

In our work we use these as well as other visual variables in a Hasse diagram to represent possible tree extraction criteria. This provides several benefits to lattice and extracted tree understanding. First, it enables users to rapidly associate the dimension/criteria in question (e.g. *stability*, *support* in Figure 2) with concepts, thus justifying the choices made during the tree extraction process. Second, visualizing different extraction criteria using various visual variables, allows users to compare these criteria in order to choose the one that better fits their needs. Third, irrespective of the tree extraction process, matching visual attributes to concept attributes establishes a benchmark/comparison among concepts, making it possible to compare at a glance different concepts, even if they do not have a link in common, as well as gain insights on the whole lattice itself. Finally, prominent features of the lattice like specialization and generalization can be better understood: for instance the power of implications of different concepts can be rendered by edge thickness. The concept node itself can be a visual metaphor for the intent and extent. In the example of figure 2, a pie chart replaces the traditional box representation to depict the proportion of objects (blue) and attributes (yellow). In this way users can be guided in understanding and choosing criteria for extracting trees to simplify the lattice representation.

[1] *ConceptExplorer*. http://conexp.sourceforge.net/

## 5. A Qualitative Analysis of the Proposed Parent Selection Criteria

In this section we discuss a case study of a concept lattice to qualitatively examine the nature of the trees resulting from different criteria. The techniques for lattice transformation and drawing were implemented in a visual analytics tool called CUBIST Analytics and applied to a dataset containing 8 animals and 9 attributes which produced a lattice with 19 concepts (figure 3). Each of the measures proposed revealed particular aspects on the analysis of a lattice, illustrated in table 2.

Table 1 a) shows the tree generated with stability as parent selection criterion. In practice, it resulted in a tree with very stable concepts more likely to retain their subsequent children. For instance, the concept {lives in land} was the preferred parent of the concept that holds our notion for amphibians: {lives on land, lives in water} because it is more stable than its counterparts.

The measure of shared objects was the criterion that generated the tree in table 1 b). Parent concepts sharing most objects with child concept were the preferred candidates. As an example, the concept {lives on land} shares more objects with {lives on land, needs chlorophyll} than concept {needs chlorophyll} does, therefore it was the chosen parent in this case.

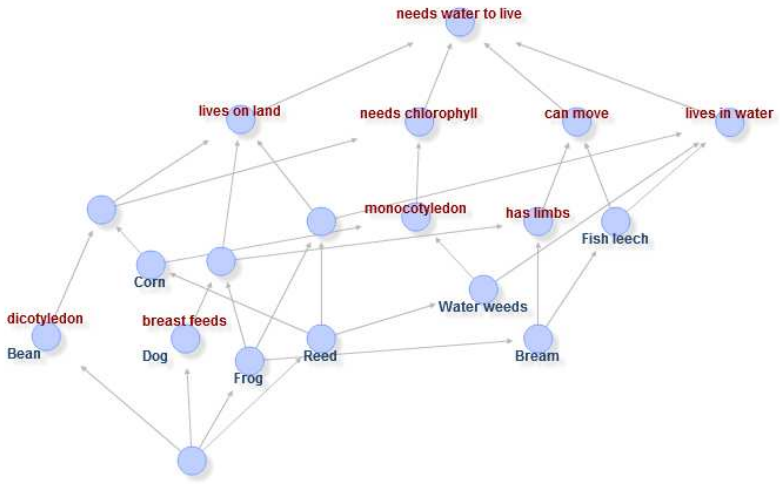



**Figure 3**. Concept lattice of the biology domain.

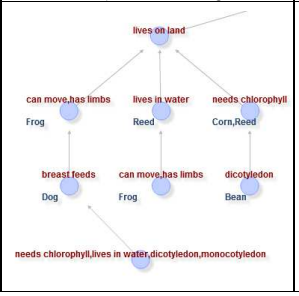
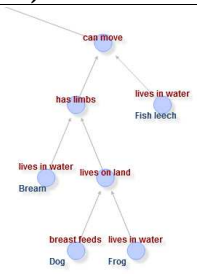
Table 1 c) the tree was generated from confidence criterion, therefore children nodes are associated with the parent with which the relationship of confidence is the highest among the candidates. As a result, the relation {can move, has limbs} has a stronger implication in {lives on land} than {lives on land} has for {can move, has limbs}, for example.

## 6. Discussion

Some may argue that due to the tree construction, the present approach breaks the original lattice meaning, and therefore subsequent mathematical models based on this structure. It is noteworthy to observe however, that only the links in the lattice graph structure are removed and the lattice structure remains semantically valid, since there is no need to take out the attributes or objects that concepts have in common with their parents.

The choice of parent selection criteria for tree transformation corresponds to a classification problem to some extent. Deciding if a Lion is more "mammal" than it is "preying" it's not always straightforward, hence we rely on the measures that attempt to keep the context semantics when looking at the entire concept lattice. For instance, if we have more objects described by mammal which are "closer" to Lion than other concepts, then it may reasonable to be chosen as its parent. As general recommendations, one should use the criteria that best fits to their analysis task (table 2).

**Table 1**. Examples of trees generated from the lattice in figure 3 for each of the proposed measure.

| | a) Stability | b) Shared attributes | c) Confidence |
|---|---|---|---|
| Example |   |   |   |

In addition to the tree-extraction strategies, the use of colours, size, shaping and thickness for both nodes and edges in the original lattice to represent the criteria metrics (such as stability, support, specialization or implication) can enhance the interpretation of a concept lattice, and aid users in their choice and interpretation of the created trees.

The labelling strategy for identifying concepts should be taken into account as well. Merely placing attributes and objects names on concepts may be cumbersome for large lattice analysis (used in most FCA visualizations). In this case, it is recommended to represent the concept's intent and extent with visual metaphors like the pie chart shown in figure 2.

**Table 2**. General guidelines on the usage of the proposed metrics.

| Criteria | Description | Rationale | Suitable for |
|---|---|---|---|
| *Stability* | It measures how likely a concept is to change if some of their attributes or objects are removed. | Stable concepts are less impacted by noise and usually represent strong correlation with real world entities (e.g.: a concept that encapsulates our notion of "mammal"). | Observing real world analogies |
| *Support* | It measures the frequency of the concept itemset. | Frequent concepts are usually generic concepts since they aggregate a larger number of objects than the specialized ones. | Frequent pattern analysis |
| *Shared objects / attributes* | It represents the degree of similarity between parent and child nodes. | Concepts that share most attributes or objects should be linked together because they are similar. | Similarity analysis |
| *Confidence* | It measures how strong the implication is between a parent concept in a child concept. | Implication is one of the desired interpretation of a concept lattice. | Confidence analysis |

## Conclusions and Future Work

Traditional software in FCA makes little use of visualization techniques, producing poorly readable lattice graphs when the number of concepts exceeds a few dozens. In this work we have presented a transformation approach to extract trees from concept lattices, attempting to minimize both semantic and conceptual loss in favour of readability and interpretation. We have also presented ways to visually show the extraction criteria in the original lattice. This is an important step in the visual analysis of conceptual structures, as the resulting tree structures are visually easier to understand than cluttered lattice graphs. Domain experts can thus visually explore larger datasets that traditional visualizations of concept lattice cannot represent effectively. Each of the tree construction measures proposed in our work provides particular insights valuable to different analysis tasks, identified in our paper as recommendations.

In the future we plan to combine two or more criteria for parent selection with other lattice reduction techniques (e.g. icebergs lattices [4]). We also plan to conduct user experiments to understand when users want to have full lattice views vs. tree views, which metrics for creating trees are of most interest to them and under which circumstances, and assess if our visual indications allow users to understand the extraction tree process.

**Acknowledgments.** This work is partly funded by the CUBIST project ("Combining and Uniting Business Intelligence with Semantic Technologies"), funded by the European Commission's 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.'

## References

1. Kuznetsov, S.O.: Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. Nauchn. Tekh. Inf., Ser.2 (Automat. Document. Math. Linguist.) 12 (1990) 21–29
2. Di Battista, G.; Tamassia, R. (1988), "Algorithms for plane representation of acyclic digraphs", Theoretical Computer Science 61: 175–178.
3. C. Ware, H. Purchase, L. Colpoys, and M. McGill. Cognitive measurements of graph aesthetics. Information Visualization, 1(2):103–110, 2002.
4. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. Computing iceberg concept lattices with Titanic. In Data & Knowledge Engineering, Volume 42, Issue 2, pp. 189-222, 2002.
5. Ducrou, J., Eklund, P., and Wilson, T. An Intelligent User Interface for Browsing and Searching MPEG-7 Images Using Concept Lattices. In S. Ben Yahiaet al. (Eds.): CLA 2006, LNAI 4923, pp. 1–21, Springer-Verlag Berlin Heidelberg 2008.
6. Zaki, M.J., Hsiao, C-J. Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. In IEEE Transactions on Knowledge and Data Mining, Vol. 17, No. 4, IEE Computer Soc., 2005.
7. Le Grand, B., Soto, M., Aufaure, M.-A. (2009) "Conceptual and Spatial Footprints for Complex systems Analysis: Application to the Semantic Web", in 20th International Conference on Database and Expert Systems Applications 2009, pp.114-127.
8. Blau, H., Immerman, N.,and Jensen, D.. A Visual Language for Querying and Updating Graphs. University of Massachusetts Amherst, Computer Science Department Tech: Report 2002-037. 2002.
9. Cruz, I. F., Mendelzon, A. O., and Wood, P. T.. A Graphical Query Language Supporting Recursion. In Proc. of the Association for Computing Machinery Special Interest Group on Management of Data, pages 323–330. ACM Press, May 1987.
10. Consens, M., and Mendelzon, A. Hy+: a Hygraph-based query and visualization system. SIGMOD Record, 22(2):511–516, 1993.
11. Michel Soto, Benedicte Le Grand, Marie-Aude Aufaure, "Spatial Visualisation of Conceptual Data," International Conference Information Visualisation, pp. 57-61, 2009.
12. Eklund, Peter, Villerd, Jean. A Survey of Hybrid Representations of Concept Lattices in Conceptual Knowledge Processing Formal Concept Analysis. Lecture Notes in Computer Science 2010, Springer Berlin/Heidelberg, pp. 296- 311
13. C. Roth, S. Obiedkov, D. G. Kourie. "Towards Concise Representation for Taxonomies of Epistemic Communities", CLA 4th Intl Conf on Concept Lattices and their Applications. 2006.
14. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin (1999)
15. Hannan, T., Pogel, A.: Spring-based lattice drawing highlighting conceptual similarity. In: Proceedings of the International Conference on Formal Concept Analysis, ICFCA 2006, Berlin. LNCS, vol. 3974, pp. 264–279. Springer, Heidelberg (2006)

# Applying conceptual analysis to space data

Alexander Mikhailian[1], Saliha Klaï[1], Christian Muller[2], Bernard Fontaine[1], Didier Moreau[2], and Martin Ursík[1]

[1] Space Applications Services
325 Leuvensesteenweg
1930 Zaventem, Belgium
cubist@spaceapplications.com,
http://www.spaceapplications.com
[2] Belgian User Support and Operation Centre (B.USOC)
3 Avenue Circulaire
1180 Brussels, Belgium
{christian.muller,didier.moreau}@busoc.be,
http://www.busoc.be

**Abstract.** This paper describes the Space Control Centres use case of the CUBIST FP7 project[3]. It introduces the concept of telemetry and describes the format and the contents of the housekeeping telemetry of SOLAR, one of the payloads flying on board of the International Space Station (ISS). Further on, it discusses several approaches to conceptual analysis of space telemetry. It also gives an overview of conversion options into a form suitable for FCA and concludes with an outlook of the expected outcomes of the conceptual analysis of telemetry.

**Keywords:** conceptual analysis, formal concept analysis, telemetry, telecommand, mission control centre

## 1 Introduction

This paper describes the Space Control Centres use case of the CUBIST project. The readers are not expected to understand the terminology used in the space industry. However, the following three terms are essential:

**Telemetry** is a common name for technologies that allow remote reporting of information. Telemetry is usually characterized by the rate and the set of parameters.

**Payload** is the useful cargo of a spacecraft, such as a set of scientific instruments and sensors packed in a single assembly.

**Operator** is the person who is constantly monitoring the payload operations. Operators usually work in shifts and are located in space mission control centres.

[3] http://cubist-project.eu

The use case covers the processing of the telemetry produced by the SOLAR payload. SOLAR is a set of devices used for Sun observation and mounted on the Columbus module, which is a part of the International Space Stations. SOLAR is operated from the Belgian User Support and Operations Centre (B.USOC) by the operators working for Space Applications Services.

Next to the scientific data, SOLAR generates tens of gigabytes per year of so called "housekeeping telemetry", which contains an overview of the "health" of the SOLAR devices.

### 1.1 Belgian User Support and Operation Centre (B.USOC)

In 1998, ESA's Manned Space Program board decided to adopt a decentralized infrastructure for the support of European payloads on-board the International Space Station (ISS). This concept was based on operating multiple User Support and Operation Centres (USOCs), distributed throughout Europe. Each of the USOCs controls one or several payloads. For example, the Biolab payload is operated from the Microgravity User Support Centre (MUSC) in Cologne, Germany. EuTEF and EDR (European Drawer Rack) payloads are operated from the Erasmus USOC in Noordwijk, The Netherlands. SOLAR payload is operated by the Belgian USOC (B.USOC) located in Brussels.

While USOCs are responsible for the payload operations, the Columbus Control Centre (Col-CC) at Oberpfaffenhofen, Germany, has the responsibility of the European Columbus module on board of ISS. Together with the USOCs, the Col-CC coordinates all the European Columbus operations.

### 1.2 Space Applications Services

Space Applications Services is an independent Belgian space technology company, founded in 1987, whose aim is to develop innovative systems, solutions and products for the aerospace markets as well as related industries. In addition to developing mission control systems, the company also installs, sets up, operates and manages these systems.

Space Applications Services has operations teams at B.USOC and Erasmus USOC. Operations have been ongoing continuously (mainly 24/7) since the launch of the Columbus payload in February 2008. Missions and experiments include Frank De Winne's OdISSea mission, BOP, PromISS-4, SOLAR, PCDF, LES-2, EPO-3, EuTEF, EDR, etc.

### 1.3 SOLAR payload

SOLAR (see Fig. 1) is an integrated platform accommodating three instruments complementing each other to allow measurements of the solar irradiance throughout virtually the whole electromagnetic spectrum. SOLAR is also one of the first external payloads of the ISS.

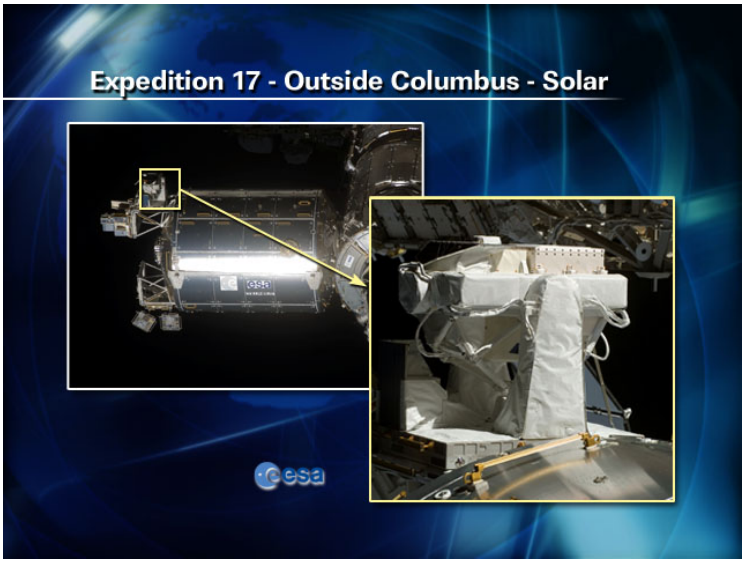



**Fig. 1.** SOLAR mounted outside ISS (Source by ESA)

SOLAR was launched together with the European Columbus laboratory in February 2008 and has been operating since that moment. Since the beginning, B.USOC supports the SOLAR operations on a 24/7 basis.

The SOLAR instruments are mounted on a Coarse Pointing Device (CPD) and make use of the CPD Common Control and Power Distribution Unit (CU) to get power, to collect, packet and dispatch to ground the instruments generated telemetry data and to receive the ground issued telecommands, ISS data and timing synchronization. The CPD accommodates the instruments and provides tracking capability thanks to a two-axis rotating platform and a Sun sensor. The first axis is used to compensate the ISS orbital motion (de-rotation function), while the second axis is used to correct the orbital plane drift and seasonal Sun apparent motion (indexation function). The selected reference position for SOLAR allows the following useful pointing ranges for CPD:

- around Y axis: $\pm 40°$ for de-rotation from sunrise to sunset
- around X axis: $\pm 24°$ for elevation toward outboard and inboard

Three instruments are deployed on the payload:

**SOVIM** (Solar Variable and Irradiance Monitor) measures irradiance in the near-ultraviolet, visible and thermal regions of the spectrum (200 nanometers - 100 micrometers).

**SOLSPEC** (SOLar SPECtral Irradiance Measurements) covers the 180 nanometer - 3000 nanometer range with high spectral resolution.

**SOLACES** (SOLar Auto-Calibrating Extreme UV/UV Spectrometers) measures the EUV/UV spectral regime (17 nanometers - 220 nanometers) with moderate spectral resolution.

The primary objective of the SOLAR mission on Columbus[1] is the quasi-continuous measurement of the solar irradiance variability with highest possible accuracy. For this reason the total spectral range is recorded simultaneously by the three sets of instruments: SOVIM , SOLSPEC and SOLACES. SOLAR operates on Columbus since February 2008. As of today, the SOLACES and SOLSPEC instruments are still achieving their scientific objective. SOVIM stopped operating in October 2008 following an electrical malfunction. The 2008-2010 period was marked by an exceptional minimum of the solar spectrum [2] making the SOLAR data the best available reference for solar intensity at the low solar activity.

These data have both scientific and societal importance as the mechanisms of the solar variations are far from being completely understood and as the Sun is the main energy input to the climate system. A historical cold climate period in the late seventeenth century has been simultaneous with a low solar activity known as the Maunder minimum. This importance justifies thus any mean used to study the health of the monitoring instruments and the quality of their data.

## 2 CUBIST Space data pack

The data used in CUBIST is derived from the housekeeping telemetry stream of SOLAR. Next to the science data, housekeeping telemetry is the biggest in size, amounting to many tens of gigabytes per year [3].

### 2.1 Overview of the Space data pack

SOLAR has been operational for more than three years, already, sending one telemetry packet every second or so. Over a year, this represents approximately $3 \cdot 10^7$ packets. Each telemetry packet contains 343 parameters. 44 parameters do not change at all or very rarely. Among the others, 135 have binary readings, such as ON and OFF. Others have readings that span between 3 and $2 \cdot 10^6$ distinct values.

The data released to CUBIST consortium partners covers 30 days between September 26, 2008 and October 25, 2008.[4] This period has been selected because a major event occurred on the 25th of October, at 04.28 AM. The DC/DC converter powering SOVIM, one of the three instruments used by SOLAR, broke down.

As CUBIST unrolls, additional data sets may be considered for conceptual analysis in the project.

[4] All times are given in GMT.

### 2.2 Overview of parameters

In order to properly monitor the health of SOLAR payload, the B.USOC operators have access to a part of the SOLAR telemetry, the housekeeping data. This telemetry is organized into packets that are sent by the payload to the control centre. It contains readings and variables, such as temperatures, voltage and current readings, operational states and reports from all different aspects of the payload. We provide below an overview of the different parameters sent downstream in the housekeeping telemetry.

**Temperatures** Although the SOLAR thermal control keeps the platform and the instruments within their operational boundaries, many temperature sensors allow the operators to closely monitor the temperatures. These temperature readings are of the type `float` and are available for the Control Unit, the Power Boards (PB1 and the slave boards PB2 and PB3), the motors, the three instruments and the Sun Sensor. The temperature limits are defined in the mission database and, although thermal control should insure these will not be crossed, they will be flagged when the limit is near to be reached. Besides these hard limits which could damage the hardware, soft operational limits are also defined for the scientific measurements. For example, the SOLACES instrument can only perform science measurements within the temperature range of $17 - 20°C$. These limits are currently not set within the mission database, but are common knowledge of the operator.

**Power supply** Housekeeping of the different power boards of SOLAR is also available. The main power board (PB1) powers the CU and the slave boards (PB2) and (PB3). The later power the motors and thus allow the tracking of the platform and the instruments. For each of the boards, the status of the board and its outlets, temperatures, DC/DC converters, voltages and current are available. This housekeeping data is closely checked during power-on activities and continuously monitored during operations. The status parameters are defined as strings (ON/OFF) and readings are of the `float` type. For some, an operating range is defined.

**Instrument housekeeping** For each of the instruments, a set of housekeeping data and their range with respect to the operations, has been defined by the scientists. Besides those, the SOLAR CU also provides the status (ON/OFF) of the instruments, temperature readings and communication status (OK,NOK) in the telemetry. Additionally, for the Sun Sensor, the sun presence reading is also included in the downstream data.

**Pointing Device telemetry** To support the actual SOLAR operations, and to execute science measurements, an additional set of parameters is made available to the operator. These parameters are related to the status of the platform, the movements and the sun observation.

Based on ancillary data of the ISS, containing the station's attitude, SOLAR software calculates the start time of the sun observation, the duration, the indexation and de-rotation angles for both axes, next midnight and noon, and the observation counter. These integer parameters will then trigger the actual Sun tracking of the platform when they are within the platform mechanical and structural range. Supporting the actual movement, the status of the platform such as the `Zero_Proc_Flag` and the `On_Target_flag` indicate whether the pointing device is calibrated and motor controlled, allowing proper Sun tracking.

Another important parameter is the `SOLARMode`. This string parameter indicates whether the operator can define software settings (SCM mode) or update the Software (SMM Mode), whether SOLAR can perform Sun Tracking (PM Mode and submodes) or in case of an anomalous situation (SBM).

**System variables** SOLAR allows the operators to change some system variables and to declare them to the housekeeping in the so-called User Selected Data area (USD). A total of 12 USD are available and each have a specified type (integer, float, string). For the on-going operations, around 6 are used.

### 2.3 Differences from the original telemetry

Several steps [4] have been taken to reduce the size of the data set, in order to make it easier to run FCA analysis on it:

- The period from the start of September 26, 2008 to midnight of October 25, 2008 has been extracted.
- A list of 208 parameters that are relevant for the operations has been established. Other parameters have been filtered out.
- Nominal limits for parameters of type float have been adjusted by an operator to reflect realistic operational limits. All the float parameters have been replaced by the textual labels `NOMINAL`, `WARNING` and `DANGER`, signifying that the value is within operational limits, in the warning zone or is dangerously abnormal.
- In order to reduce the cardinality of the data, readings of the moving axes of the SOLAR payload have been averaged to 1°.

## 3 Further work

Today, in space control centre operations, much time is spent on transferring, discussing, reviewing, and copying information between the operations partners. Also the search and replay of operational data in order to be able to correctly analyse the on-board situation is time consuming. This problem is especially emphasised by the fact that many of the data stores are distributed and have different user interfaces. The current operations show an increasing need for a system providing the operator with adequate and fast information and analytics, especially during anomalous situations. A system federating and managing the broad amount and variety of available operations data, from Console logs to House Keeping Path Telemetry where the operator would benefit the accessibility to other related data in a transparent way. In case of an anomalous situation the operator would have the possibility to immediately identify the possible failure and to propose a way forward or work around, mitigating the science loss.

This section lists use case scenarios that are studies by CUBIST. Each anomaly discussed below has its roots in the real operations. Depending on the anomaly, a particular research strategy for CUBIST has been identified.

### 3.1 Telecommand analysis

SOLAR Science measurements are often performed through so-called command schedules. A command schedule is a dedicated pre-programmed time-ordered sequence of time tagged commands to be sent to the SOLAR instruments or the CPD system On various occasions the B.USOC operator encountered a unexpected event during a run of a command schedule, from a sudden stop of the script to the anomalous behaviour of the instruments. The analysis of the failure that follows is often restricted to that particular command schedule. The

CUBIST project could allow the operator to trace back the command itself inside the script that was sent and presumably generated the error and even back track in the archive whether this particular command has been sent before, independent of the command schedule. With a user-friendly form, the operator can check previous events and the reaction of the payload on it. This will benefit the failure analysis in finding the actual cause of the anomaly, rather than the environmental circumstances.

### 3.2 Telemetry mining

It may happen that the payload manifests an unforeseen thermal situation. That is, a situation when the temperature of one or several sensors changes in an unusual way, albeit within the nominal limits. The operator is then charged with finding similar occurrences inside the telemetry archive and with the determination of typical thermal and power profiles. Currently, the search in the telemetry archive is usually done as a real-time replay of the telemetry archive. A more intelligent solution that employs the results of the concept analysis may be implemented, so that an automated agent finds occurrences of similar situations by taking into account the telemetry parameters.

### 3.3 Parameter correlation

Since the start of the SOLAR operations SOLAR experienced on a regular basis a reset of the internal Analogue Interface Board, causing the platform to go in the anomalous Stand-by Mode and halting all on-going science measurements. Resuming science can only be done by power cycling the payload. The occurrence of the anomaly seems rather random, but due to the amount of SOLAR parameters this has not been confirmed. The CUBIST project will correlate the occurrences of these failures with other parameters and might reveal a pattern with SOLAR parameters or even with factors outside the payload. This would allow the operator to act proactively and avoid a hard stop of the on-going science measurements. In the case of operational instruments, frequently a recovery procedure may not be practiced after the commissioning period and presents a risk if it is improvised. So an orderly power cycle, fully documented, will always present less risk than a general shutdown proceeding from an undetermined state.

### 3.4 Forensics analysis

A few months after the launch of the SOLAR payload, SOVIM, one of its three scientific instruments died because of an electric failure in a DC/DC converter. It is still unknown whether this failure could have been predicted given the previous telemetry stream. The objective of the CUBIST system would be to find patterns of failure in the flow of telemetry parameters with the aim to transpose these to the prediction of future failures.

## 4 Legal notice and disclaimer

Parts of this work have been carried out in CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) project. CUBIST is funded by the European Commission under the 7th Framework Programme of ICT, topic 4.3: Intelligent Information Management.



## References

1. Schmidtke, G., Fröhlich, C. and Thuillier, G.: ISS-SOLAR: Total (TSI) and spectral (SSI) irradiance measurements, Advances in Space Research, 37, 2006, pp 255-264
2. Thuillier,G., Bolsée, D., Schmidtke, G., Schmutz, W., Shapiro A. and Nikutowski. B.: The Absolute Solar Irradiance Spectrum at Solar Minimum Activity Measured by the SOLSPEC and SOL-ACES Spectrometers from 17 to 3000 nm Placed on Board the International Space Station, Bremen COSPAR presentation, 2010, to be published in Advances in Space Research
3. Mikhailian, A., Wislez, J.-M., Fontaine, B., Klai, S.: CUBIST Requirements Document (Space Control Centres) Issue 2.9.1, 8 April 2011
4. Ganter, B., Meschke, C.: A Formal Concept Analysis Approach to Rough Data Tables, In Proceedings of 12th International Conference, RSFDGrC 2009 Delhi, India, December 2009, 117–126

# Exploring the Applicability of Formal Concept Analysis on Market Intelligence Data

Constantinos Orphanides

Conceptual Structures Research Group
Communication and Computing Research Centre
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, Sheffield, UK
c.orphanides@shu.ac.uk

**Abstract.** This paper examines and identifies issues associated with the applicability of FCA on sample data provided by a CUBIST use-case partner. The paper explains the various steps related to the transformation of these data to formal contexts, such as preprocessing, cleansing and simplification, as well as preprocessing and limitation issues, by using two FCA tools currently being developed in CUBIST, FcaBedrock and InClose. The paper demonstrates what is achievable to date, using the above-mentioned tools and what issues need to be considered to achieve more meaningful and intuitive FCA analyses. The paper concludes by suggesting and explaining techniques and features that should be implemented in later iterations of these tools, to deal with the identified barriers. This work has been carried out as a part of the European CUBIST FP7 Project: http://www.cubist-project.eu

## 1 Introduction

It has been shown that a variety of datasets can be converted into formal contexts [8,2] by a process of discretising and Booleanising the data [10]. However, depending on the nature of the dataset, manual or automated means of preprocessing have to be deployed first, in order for FCA to be succesfully carried out. Although the open-source and freely available FCA tools currently being developed in CUBIST, FcaBedrock [3,6] and InClose [1,9], are configured to cater for most preprocessing and data cleansing issues [3,4], further issues might arise: types of attribute that have not been catered for or considered so far, such as free-text data and data inconsistencies.

This paper attempts to identify such issues, by conducting FCA on a dataset provided by Innovantage, a CUBIST use-case partner, providing market and competitive intelligence in the United Kingdom. The paper concludes on further work and explains what techniques will be deployed, in later iterations of the tools, to cater for the issues identified while analysing the specific dataset.

## 2 Dataset Description

The specific dataset consists of job vacancies advertised on the United Kingdom's leading job boards, as well as employers' own websites, tracked in real-time using Innovantage's proprietary software. The dataset is in XML format and has been extracted from a MySQL RDBMS. The dataset comprises of 900 jobs accompanied by their details:

- Title: The job's title.
- Description: A brief description outlining the requirements of the job.
- Date Found: The exact time of when the job was tracked.
- URL: The website where the job was found at.
- Raw Location: The location of the employer.
- Raw Salary: The advertised salary, sometimes also including information about bonuses and benefits.

An example of a job entry is shown below (File 1).

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<jobs>
   <job>
      <title>Data Centre Developer</title>
      <description>Data centers Developer opportunity based in Amsterdam
      on a 6 months rolling contract. This is a very senior position and
      requires the candidate to have at least 7 years experience and
      have extensive knowledge and expertise in the construction
      of a data centers facilities including electrical systems,
      cooling plants etc and familiar with EU regulations and
      best practices.</description>
      <date_found>2011-01-12 17:01:58.0</date_found>
      <url>http://www.itjobspost.com/JobSeeker/</url>
      <raw_location>Amsterdam, Other Countries, UK</raw_location>
      <raw_salary>400-600 Per Day</raw_salary>
   </job>
</jobs>
```

**File 1** `innovantage_sample.xml`, *XML* file.

## 3 Data Conversion Process

### 3.1 Preprocessing

Some issues surfaced during preprocessing, mainly due to the XML file failing to render properly because of illegal, non-UTF-8 characters contained in the data, possibly related to the automated process in which jobs are tracked and recorded. This was worked-around by parsing the XML file using a custom-made algorithm and removing all illegal characters and symbols, without loss of information and without affecting the quality of the data. The XML file was

| title | description | date_found | url | raw_location | raw_salary |
|---|---|---|---|---|---|
| Software Engineer | Software EngineerWe have a | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Birmingham, Coventry, Tar | 35, 000 |
| Linux Systems Administrator - Ubuntu, Apache, MYSQL | Linux Systems Administrator | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | London | 45k - 50k pa + Bonus & benefits |
| Assistant Buyer | Assistant Buyer reporting to F | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Basingstoke | 18k - 23k pa + Pension |
| Credit Controller | The individuals will be part o | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Liverpool | 16000 - 18000 |
| Lead Firmware Engineer - Embedded, C/ C++, Linux, J2EE | Our client is recognised as th | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Poole | 45k pa |
| Bid Manager - Contract Cleaning - Soft FM | Bid ManagerContract Cleanin | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | London | 4k - 45k pa + benefits |
| Implementation Consultant for leading Buy Side Vendor | Key words -Implementation ( | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | London, Uk, Europe | 70,000 - 90,000 |
| Business Analyst - Solvency 2 | Due to an expanding Program | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | London | Neg |
| Area / Branch Manager | Area / Branch ManagerDomic | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Tyne & Wear, Newcastle, S | 28k - 30k pa |
| SAP HCM Technical Architect | We are looking for a SAP HR T | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | England | 650 - 750 p/day + Expences |
| Principal Developer C#. NET (Application Architect) | Principal Developer C#.NET (A | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Basingstoke, Reading, Alde | 45k - 70k pa |
| Business Development Manager (Cloud / AMS) | Business Development Mana | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | London | Salary plus benefits |
| Human Resource Administrator | We are currently recruiting fc | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Manchester | Negotiable |
| Gas Engineer | Benefits- Pension Scheme- C | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | North Allerton And North Y | Negotiable |
| Senior Estimator/ Bid Manager | Bid Manager/Senior Estimato | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | North West England, Lanca | Negotiable + package |
| Site Manager | Benefits- Bonus discretionary | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Cambridgeshire | 30,000 to 40,000 |
| Experienced Admin Assistant | YOU WILL NOT BE CONSIDERE | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Shrewsbury | Negotiable |
| Electricial - Clerk of Works | Our client is looking for a Ele | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Leicestershire | 18 to 22 |
| Senior Quantity Surveyor | This major civil engineering c | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Cambridgeshire | 28,000 to 35,000 |
| General Manager - Water Bottling Plant | Benefits- tax free salary- accc | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Afghanistan | Negotiable |
| Area Sales Manager - Building Products - Kent | JOB TITLE: Area Sales Manage | 2011-01-12 17:00:46.0 | http://www.jobsite.co.uk | Kent, Essex, London | 30k basic, 35k ote |

**Fig. 1.** XML-to-CSV transformation of the dataset.

then loaded in MS Excel and converted to CSV (Figure 1), as FcaBedrock does not currently support XML as input.

Another issue that affected the analysis was the fact that free-text attributes, as their name implies, are inconsistent, mostly due to the fact that the recorded data originate from various sources. Taking the 'Raw Location' attribute as an example, a job's location can be recorded as "Manchester" for one job, "Manchester, United Kingdom" for another job and "Greater Manchester" for another job. The same problem applied for the 'Raw Salary' attribute, as some employers use ranges (e.g "15000-20000"), some use finite values and also include currencies (e.g. "18000 GBP") and others also include additional information (e.g. "25000 per annum, negotiable"). For this type of free-text attribute to be succesfully and meaningfully converted, some kind of Semantic Extract Transform Load (SETL) or Natural Language Processing (NLP) process would be required first, in order to identify values. As such, the data had to be manually modified; 100 jobs were randomly selected from the lot and had the above-mentioned attributes reconfigured for consistency. In addition, some attributes were excluded from the analysis; in particular, 'Description' was excluded as it is not an attribute, but rather the title (or descriptive annotation) of the object, although it could be useful as part of the case-study in terms of adding meaning to an analysis. The

'Date Found' attribute was also excluded, as all of the jobs in the dataset were tracked on the same date, thus adding no specific value to the analysis. The 'URL' attribute was excluded, as URLs are unique for each job posted (thus considered free-text data). In terms of reconfiguring attributes, the 'Raw Location' attribute was configured to hold only city names and the 'Raw Salary' attribute was configured to be purely numeric. An extra attribute was created to hold additional information originally contained in the 'Raw Salary' attribute, such as whether the salary is negotiable or not. This resulted in four attributes remaining: 'Title', 'Raw Location', 'Raw Salary' and 'Negotiable Salary' (the new attribute that resulted during preprocessing). A screenshot showing how the dataset looks after preprocessing is shown at Figure 2 below.

| title | raw_location | raw_salary | sal_negotiable |
|---|---|---|---|
| Assistant Buyer | Aberdeenshire | 18000 | n |
| Credit Controller | Afghanistan | 16000 | n |
| Lead Firmware Engineer - | Antrim | 45000 | n |
| Bid Manager - Contract Cle | Barrow-in-Furness | 45000 | n |
| Business Analyst - Solvenc | Basingstoke | Neg | y |
| Business Development Ma | Bedfordshire | Salary plus benefits | y |
| Human Resource Administ | Birmingham | Negotiable | y |
| Site Manager | Birmingham | 30000 | y |
| Experienced Admin Assist | Brighton | Negotiable | y |
| Electricial - Clerk of Works | Brighton | 22000 | y |
| Senior Quantity Surveyor | Bristol | 28000 | y |
| General Manager - Water I | Bristol | Negotiable | y |
| Administrator | Bromborough | 14874 | n |
| Customer Service Advisor | Cambridge | 14000 | n |
| Solidworks Design Engine | Cambridge | NEG. | n |
| Assistant Accountant | Cambridgeshire | 22000 | n |
| Product/ Configuration En | Cambridgeshire | 26000 | n |
| Web Tester - Berkshire - 1 | Cheltenham | 39600 | n |
| Project Manager - URGENT | Colchester | 45000 | n |

**Fig. 2.** Final version of the dataset, after preprocessing.

### 3.2 Transforming the Dataset into a Formal Context

The dataset was loaded in FcaBedrock and the 'Title' attribute was excluded from the analysis, using the attribute exclusion feature. The metadata autodetection feature of FcaBedrock was used to avoid entering metadata manually (Figure 3).

Converting the dataset with FcaBedrock resulted in a formal context with 67 formal attributes. Feeding the formal context in InClose resulted in 110 formal concepts; although not quite a large amount, the concepts had to be reduced to an amount where the concept lattice would be readable and manageable. Over a trial-and-error process, using InClose and the well-known idea

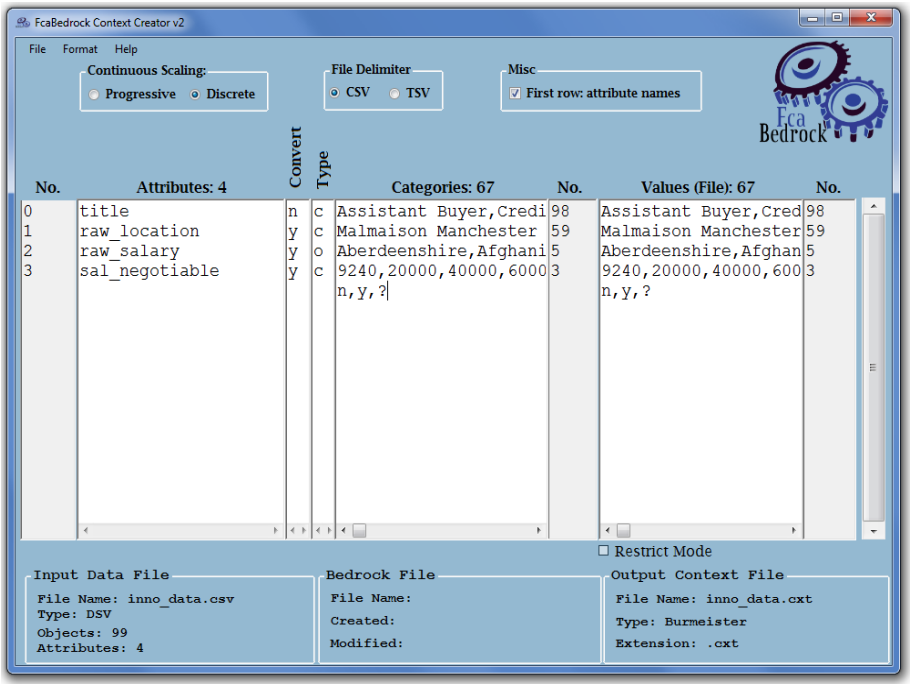



**Fig. 3.** Autodetecting the metadata in FcaBedrock.

of minimum-support (a semi-automated form of lattice 'iceberging' [12]), the minimum-support for the intent was set to 2 and the minimum-support for the extent was set to 5. This resulted in 9 concepts. When visualised in ConExp [13,5], however, 20 concepts are displayed. This is because where the large concepts 'overlap', other concepts are found during a *second pass* of concept mining, with no minimum support, when producing the concept lattice [4]. In this way, possibly significant concepts, that would not have satisfied the initial minimum-support are retained and a complete hierarchy is maintained in the resulting concept lattice (Figure 4).

## 4 Analysis

Even with a small amount of objects and attributes, interesting information can be extracted from the lattice. For example, all non-negotiable salaries are the ones that fall in the £40000-60000 range, while the negotiable salaries fall in the £20000-40000 range. For salaries where it is undefined (or unknown) if they are negotiable, there seems to be no distinct indication as to why this is the case. As for salaries in the £9240-20000 range, they all fall under the 'sal_negotiable-n' and 'sal_negotiable-?' attributes, with a 50-50 ratio. The overall conclusion

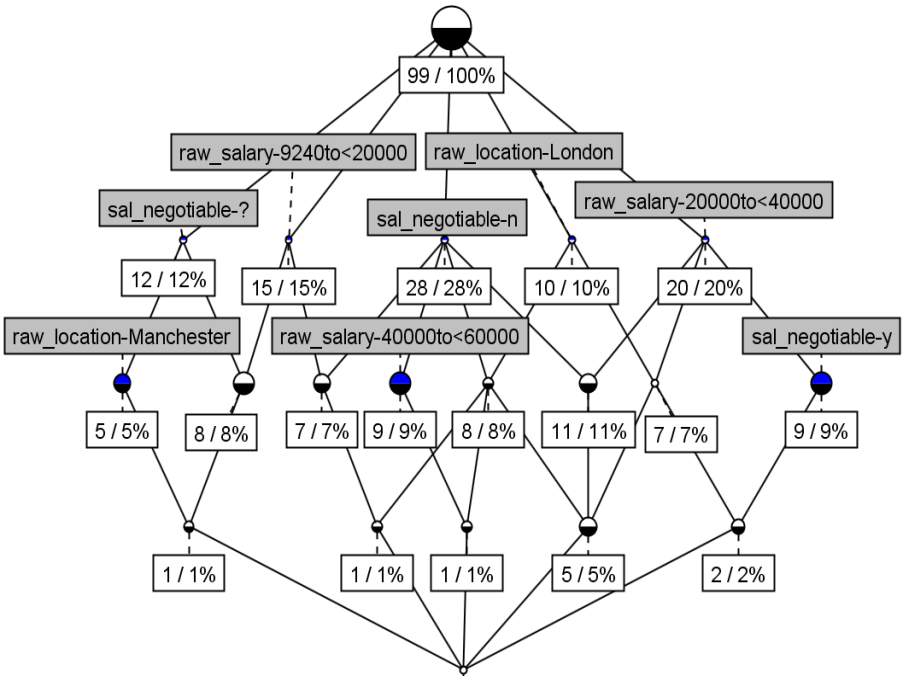



**Fig. 4.** Visualising the resulting formal context in ConExp.

indicates that for high-end salaries negotiation is not an option, while negotiation is possible for mid-end salaries. Interesting is the fact that low-end salary jobs tend not to specify whether their salaries are negotiable or not, though when they do they are not negotiable. Why is that the case? Questions of these nature require further investigation.

While the insights provided by the resulting lattice might not be groundbreaking, a close collaboration of the FCA analyst with the domain expert would help in refining the requirements, to produce meaningful business questions that would be more suitable for analysis of such data. For example, the domain expert might want to investigate why employers tend to not specify or negotiate jobs with low-end salaries. Could it have something to do with their geographic location or the domain of the job? Such kind of analysis is perfectly feasible in FCA, by restricting the context to specific attributes (and attribute values) of interest. Figure 5 shows how this can be done in FcaBedrock, where the location was restricted to London, Manchester and Birmingham, the raw salary was restricted to low-end only and the negotiable salary attribute was set to unknown. As such, the business question has been redefined to "display jobs in London, Manchester or Birmingham with low-end salaries, where salary negotiation is unknown or unspecified".

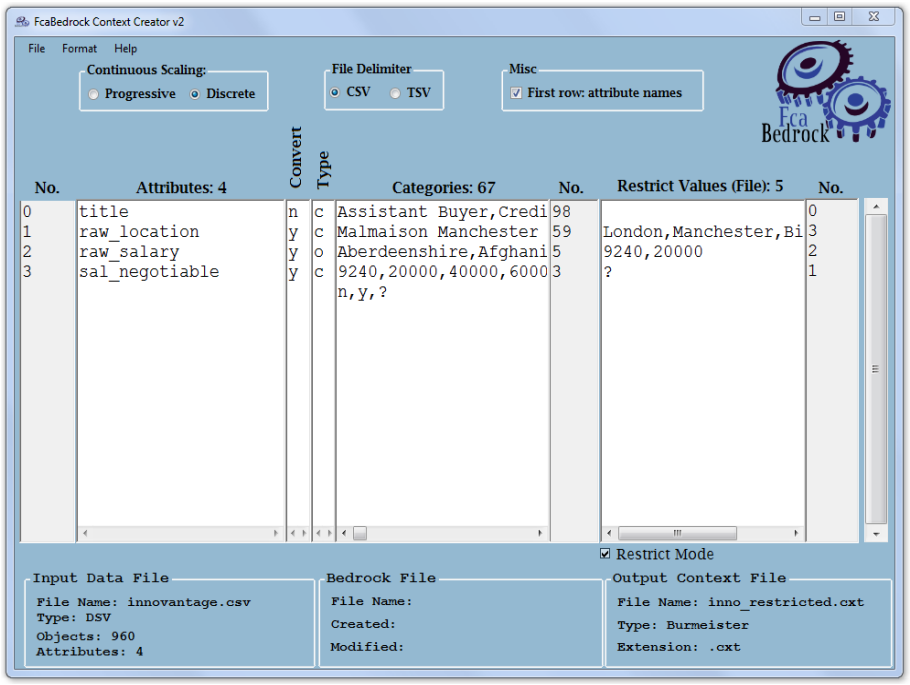



**Fig. 5.** Using FcaBedrock's restriction capabilities to focus the analysis on specific attributes and attribute values.

## 5 Further Work

It is evident that as new data sources and data types are introduced, more preprocessing issues arise. With regards to the tools used in this analysis (FcaBedrock and InClose), further development is currently in process and various issues, mentioned below, are already being considered.

In terms of data sources, XML should be added as a default data source, to avoid XML-to-CSV transformations. Pulling data directly from an RDBMS source would be quite useful as well, by selecting specific database tables, or even specific columns from each table, to use in the analysis. Manipulation of RDF data are of high importance as well, given the fact that CUBIST revolves around semantic technologies.

Free-text data have proven to be not suitable for FCA, unless some kind of Semantic ETL or NLP process, in order to identify values, is deployed first. Use of thesauri, such as the approach described in [11], to tokenize free-text data into categories could prove useful as well, although whether these kind of processes will be manual, semi-automated or automated remain research questions which require further study.

Another feature that would be particularly useful would be to embed additional functionality in the autodetection features of FcaBedrock, particularly for selecting appropriate scales and intervals for continuous attributes. Understanding the true nature of a continuous attribute at the moment, using FcaBedrock, is only feasible when datasets include documentation, such as the ones in the UCI Machine Learning Repository [7], or by manually investigating the data. As such, suggesting ranges and scales, using the same 'guided automation' approach that FcaBedrock uses [3] would make analyzing such attributes more meaningful and insightful.

## 6 Conclusion

The paper has explored the application of FCA within a market intelligence scenario, using real-life data from a CUBIST use-case partner, deploying freely-available and open-source FCA tools, currently being developed in CUBIST, for the analysis. Several preprocessing issues have been identified and suggestions, techniques and features have been proposed for further work.

Although the work presented in this paper is still at an early stage, it demonstrates how the market data and FCA communities can benefit from each other. The market data community has provided new challenges that FCA has to consider, mostly in terms of usability and user-friendliness. Within the context of CUBIST, we envisage that the market data analysts will be able to conduct FCA analysis on their data, without collaborating with FCA experts.

**Acknowledgement** This work is part of the CUBIST project ("Combining and Uniting Business Intelligence with Semantic Technologies"), funded by the European Commission's 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management. More information on the project can be found at `http://www.cubist-project.eu`

## References


1. Andrews, S.: *In-Close, A Fast Algorithm for Computing Formal Concepts*. In: Rudolph, S., Dau, F., Kuznetsov, S.O. (eds.) ICCS'09, `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-483/` (2009)
2. Andrews, S.: *Data Conversion and Interoperability for FCA*. In: CS-TIW 2009, pp. 42-49, `http://www.kde.cs.uni-kassel.de/ws/cs-tiw2009/proceedings_final_15July.pdf` (2009)
3. Andrews, S. and Orphanides, C.: *FcaBedrock, a Formal Context Creator*. In: Croitoru, M., Ferre, S. and Lukose, D. (eds.) ICCS 2010, LNAI 6208. Springer-Verlag, Berlin/Heidelberg (2010)
4. Andrews, S. and Orphanides, C.: *Analysis of Large Data Sets using Formal Concept Lattices*. In: Kryszkiewicz, M. and Obiedkov, S. (eds.). Proceedings of the 7th International Conference on Concept Lattices and Their Applications (CLA) 2010, ISBN 978-84614-4027-6. Seville: University of Seville. pp. 104-115 (2010)
5. ConExp (Concept Explorer). Available at `http://sourceforge.net/projects/conexp`
6. FcaBedrock Formal Context Creator. Available at `http://sourceforge.net/projects/fcabedrock`
7. Frank, A. and Asuncion, A.: *UCI Machine Learning Repository* [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science (2010)
8. Ganter, B. and Wille, R.: *Conceptual Scaling*. In: Roberts, F. (ed.) Applications of Combinatorics and Graph Theory to the Biological and Social Sciences. IMA, vol. 17, pp. 139-168, Springer, Berlin-Heidelberg-New York (1989)
9. InClose Formal Concept Miner. Available at `http://sourceforge.net/projects/inclose`
10. Kaytoue-Uberall, M., Duplesssis, S. and Napoli, A.: *Using Formal Concept Analysis for the Extraction of Groups of Co-expressed Genes*. In: Le Thi, H.A., Bouvry, P., Pham Dinh, T. (eds.) MCO 2008. CCIS vol. 14, pp. 439-449. Springer-Verlag, Berlin/Heidelberg (2008)
11. Poelmans, J., Elzinga, P., Viaene, S. and Dedene, G.: *Formal Concept Analysis in Knowledge Discovery*. In: Croitoru, M., Ferre, S. and Lukose, D. (eds.) ICCS 2010, LNAI 6208. Springer-Verlag, Berlin/Heidelberg (2010)
12. Stumme, G., Taouil, R., Bastide, Y. and Lakhal, L.: *Conceptual Clustering with Iceberg Concept Lattices*. In: Proceedings of GI-Fachgruppentreffen Maschinelles Lernen'01, Universitat Dortmund, vol. 763. (2001)
13. 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.

# Generic Tools for Data Analysis and Visualisation

Uta Priss

Edinburgh Napier University, School of Computing,
www.upriss.org.uk

**Abstract.** This position paper discusses challenges that need to be overcome in order to build generic tools for data analysis and visualisation. Its intention is to stimulate discussion among the CUBIST workshop participants, not to present results.

In the Star Trek television programs, data analysis is usually accomplished by speaking to the computer and asking it to analyse some problem. The computer then provides a succinct, coherent and relevant analysis of the data which allows the Star Trek crew to make their decisions. In reality, although humanity has made some progress with implementing Star Trek technology[1], achieving automated computerised data analysis is probably at least as difficult as implementing automated natural language processing because the computer would need to understand the problem within its full context. A more achievable but still difficult goal would be to build data analysis software that collaborates with human users during the data preprocessing and modelling stages and then automates the rest of the analysis. In recent years, great advances have been made with respect to the availability of large toolkits for data analytical methods, visualisation, storage and retrieval. Thus, although the general problem is difficult (or impossible), many of the building blocks for achieving somewhat more modest solutions are available, even using low cost or free, open source tools.

With respect to using Formal Concept Analysis (Ganter & Wille, 1999) as a tool for data analysis, drawing from our own experience in past projects (Priss & Old, 2010), the most labour-intensive part of using FCA is usually the preprocessing stage during which one builds the formal contexts from the raw data or during which one decides how to select smaller data sets if the original data is too large to be visualised in a single lattice. In our experience each new data set provides new challenges. One usually has to write scripts or use other computational means to preprocess the data. It is not always possible to reuse methods (or scripts) from one project directly for the next one. In some cases (Endres et al 2010), a custom application has to be purpose-built for the data. Ideally, there should be methods and tools which speed-up the data preprocessing stage. It would be nice if it was possible to apply FCA quickly to any new data set that one encounters in order to explore the data. So, with respect to FCA, the general problem of building a generic data analysis tool can be scaled down to the problem of building a generic data preprocessing tool which makes it easier to apply FCA to any

[1] For example, we finally have Star Trek's PADDs in the form of modern tablet computers, such as Apple's iPad, and there is a list of further "Star Trek Technologies that Actually Came True" available at http://electronics.howstuffworks.com/10-star-trek-technologies.htm.

given data set. Additionally, it would be desirable if FCA tools could be more easily integrated with existing tools for data preprocessing, mining, extraction, modelling and so on to allow for a combination of methods.

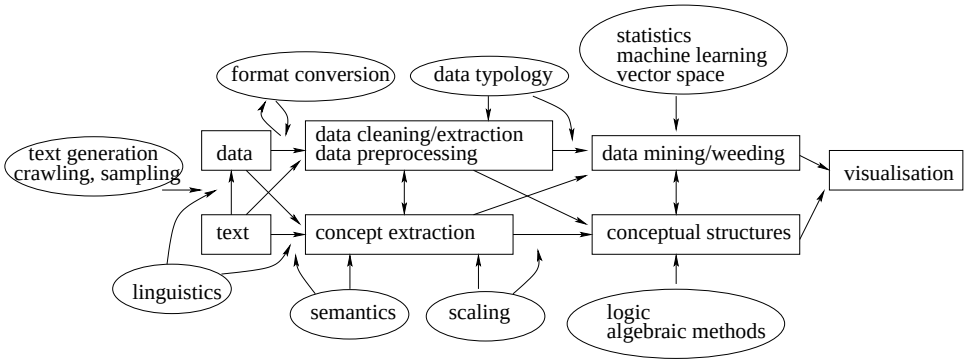



**Fig. 1.** Components of a generic tool for data analysis and visualisation

The purpose of this position paper is to provide an overview of the challenges that need to be overcome in order to build such generic tools for data analysis (both FCA-based ones and tools where FCA is just one component among many others). Figure 1 provides an overview of the steps, tasks and tools that we envisage as building blocks for generic data analysis tools. In the case of FCA, concept extraction, conceptual structures and visualisation are meant to refer to the corresponding FCA techniques. In the general case, conceptual structures could be represented using ontologies or other formal methods and visualisation could be any commonly used method (graph- or network-based, statistical plots or charts, 3-d visualisation, and so on). Many parts of Figure 1 are well established and supported by many existing tools (for example, tools for linguistic parsing and stemming, statistics, data mining and machine learning). In some cases, tools cover a variety of areas. For example, data mining tools tend to provide methods for data preprocessing and often have plugins for some linguistic processing. A tool such as NLTK[2] provides linguistic methods as well as some machine learning methods. A tool such as Sage[3] combines a wide range of mathematical methods (including statistics, network modelling and mathematical plotting) in a single toolbox.

But we believe that there are also some aspects of Figure 1 that are still missing and pose challenges, in particular:

**Text generation, crawling, sampling:** In modern applications text is not just extracted from existing databases and document collections but also generated on demand from web-based or other continually updated sources. This area is probably most difficult to automate because it heavily depends on the structures of the textual data. Semi-

[2] http://www.nltk.org/
[3] http://www.sagemath.org/

structured text (such as XML, linked data) is easier to process than unstructured text. It is easier to look for specific patterns than to discover previously unknown structures.

**Text/data:** In this paper, "data" denotes text that is slightly more structured (using mark-up, database or spreadsheet tables, linked data, etc) whereas text can be in any format or medium. The main challenges with respect to the data itself are the amount (for example, processing all of Wikipedia would require giga- or terabytes of space) and internationalisation. For example, although Unicode is an international coding standard, in our experience, it can still pose problems because not all software supports it perfectly. Unexpected effects can occur. For example, printing a mixture of characters from languages that write right to left and those that write left to right can confuse printed output. Some major languages (such as Chinese) are often written in non-Unicode encodings.

**Format conversion:** Many tools for format conversion exist[4] but not all formats tend to be supported and errors may be introduced in the conversion process. For example, even though Weka is a very popular data mining toolkit[5] and data is commonly stored in spreadsheet or csv formats, importing such formats into Weka's internal format can introduce errors because, for example, leading zeros in string data are automatically deleted.

**Data cleaning:** Data cleaning is often discussed in a database context, probably because databases provide rules for consistency and integrity checks. In a more general context, some form of conceptual modelling is required in order to determine what constitutes an error.

**Data typology, scaling:** The idea for data typology or scaling is that once a datatype or conceptual type is established it should predict the kind of analyses that are suitable for the data. Commercial tools often provide "Wizards" that help users with modelling decisions, but this is less supported in free tools or tools that have more general functionality.

**Data weeding:** We see a difference between mining and weeding (Priss & Old, 2011) in that mining explores all of the data simultaneously whereas weeding allows for a careful (concept-guided) selection of subsets of the data.

**Selection of programming language:** A promising programming language for toolkits of mathematical, mining and machine learning software is currently Python. Although Python is a scripting language, more complex algorithms and treatment of large data sources can be accomplished by writing relevant routines in C or C++ which are accessed by Python scripts. Unfortunately, because Python does not have a standard graphical component, different visualisation software requires different additional graphical software which can make tools difficult to install. The main software for graphical, GUI applications is probably Java. Scripting is easier with Python than Java. Both Python and Java are cross-platform but Python is probably more suited for Unix than PCs.

[4] For FCA these are tools such as FcaStone (http://fcastone.sourceforge.net/), FcaBedrock for reading csv files (http://sourceforge.net/projects/fcabedrock/) and ToscanaJ for database connections (http://tockit.sourceforge.net/).

[5] http://www.cs.waikato.ac.nz/ml/weka/

**Cross-disciplinary approaches:** Toolkits often combine software at a syntactic level but not necessarily in a semantically consistent manner. For example, an analysis of using FCA and Sage (Priss, 2010) demonstrates that although Sage allows to combine a variety of tools with FCA software, coding is still required to model the data appropriately for each tool. Also different tools in a single toolkit can be of varying quality. Approaches from other disciplines are sometimes missing.

**Testing, validation, evaluation:** Methods and standards need to be established that allow for comparison and evaluation of data analysis methods across a variety of disciplines. In particular, it would be interesting to discuss the methods and techniques provided by FCA that are not already available through more traditional methods. Evaluation and testing must involve both the domain experts and the tool builders.

**Usability and learning curve:** Tools must be usable, well documented and fairly easy to learn in order to attract sufficient users.

In summary, modern toolkits (such as Sage and NLTK) are a good starting point for building generic data analysis and visualisation tools. But there are numerous challenges that need to be overcome in order to truly integrate a large variety of methods in a manner that renders them widely applicable. So far FCA software does not appear to be integrated into any of the existing toolkits, but integration of FCA and Sage, for example, can be accomplished (Priss, 2010).

## References


1. Endres, Dominik M.; Foldiak, Peter; Priss, Uta (2010). *An Application of Formal Concept Analysis to Semantic Neural Decoding.* Annals of Mathematics and Artificial Intelligence, 57, 3, Springer-Verlag, 2010, p. 233-248.
2. Priss, Uta; Old, L. John (2010). *Concept Neighbourhoods in Lexical Databases.* In: Kwuida; Sertkaya (eds.), Proceedings of the 8th International Conference on Formal Concept Analysis, ICFCA'10, Springer Verlag, LNCS 5986, p. 283-295.
3. Priss, Uta (2010). *Combining FCA Software and Sage.* In: Kryszkiewicz; Obiedkov (eds.), Proceedings of the 7th International Conference on Concept Lattices and Their Applications (CLA'10), 2010, p. 302-312
4. Priss, Uta; Old, L. John (2011). *Data Weeding Techniques Applied to Roget's Thesaurus.* In: Knowledge Processing in Practice. Springer Verlag, LNAI 6581, p. 150-163.
5. Ganter, Bernhard, & Wille, Rudolf (1999). *Formal Concept Analysis. Mathematical Foundations.* Berlin-Heidelberg-New York: Springer.

# Evaluation of an Approach for Teaching Formal Concept Analysis

Martin Watmough

Conceptual Structures Research Group, Communication and Computing Research Centre / Department of Computing, Sheffield Hallam University, UK
martin.watmough@ciber.com

**Abstract.** This paper describes the evaluation of coursework set for final year degree students designed to teach Formal Concept Analysis (FCA).
The usefulness of this approach is discussed with respect to its application in future iterations of the coursework. The source data was the result of a simulation between competing student teams undertaken on a mainstream ERP system provided by the business software vendor SAP A.G. and using the ERPsim software provided by ERPsim Lab at HEC Montreal. The simulation generated data on which Business Intelligence (BI) is typically based and is representative of business activity. The data generated by the simulation exercise was not specifically for FCA, thus it provides a meaningful test of FCA in BI.

**Keywords:** Formal Concept Analysis, FCA, Enterprise Resource Planning, ERP, Business Intelligence, ERPSim

## 1 Introduction

This paper describes the evaluation of coursework set for final year degree students designed to teach Formal Concept Analysis (FCA). The assessment applied a set of FCA tools and conventional Business Intelligence (BI) using graphical or statistical methods in Microsoft Office Software. There were to distinct objectives to this activity, firstly to fulfil the students learning objectives and secondly to support an action research project about the application of FCA within Enterprise Resource Planning (ERP) systems. The fulfilment of the learning activity was assessed in two ways, firstly as a comparison between the use of conventional BI analysis and FCA tools and secondly as a comparison of FCA against established theory. Topics for the action research project are highlighted in the conclusions and further work sections, however, this is not the primary focus of this paper.

FCA is mathematical theory of data analysis using formal contents and concept lattices [10], [14], [3] and has the potential to compliment and advance current forms of analysis.

The rationale for selecting this research is due to the demands being placed on BI systems to improve and the difficulties in identifying semantic data. A simple definition is "semantics = data + behaviour" [7]. This suggests that if the semantic content can be identified it may be possible to understand or determine behaviour.

The coursework is described in more detail later, however the principle is to introduce frameworks and techniques for representing and reasoning with knowledge for smart applications [12]. The principle of the coursework is to compare how analysis using tools such as Microsoft Excel compares to a FCA tool set using data generated through the realistic use of an ERP system. The students entered into the analysis with a practical knowledge of the processes that generated the data set but having performed no analysis or reflection on the impact of decisions made during the simulation.

The need for analysis and decision making within enterprises is not new but competition and complexity do combine to make the task vast and difficult to execute efficiently or accurately. Business Intelligence (BI) is frequently used to support analysis and decision making and can be traced back at least as far as 1958 [6], however, it remains a field that is subject to much ongoing research and development. Gartner [5] predicts that business units will control at least 40 per cent of the total budget for BI, a reason cited for this is that a significant percentage of companies regularly fail to makie insightful decisions about their business and markets. This implies that tools must be suitable for non technical users while encompassing the reliability and flexibility for application in modern environments.

ERP systems are essentially transactional systems that support a vast array of business functions within the majority of organisations that exist today. They are designed to be explicit and accurate in terms of control and data but often lack the analysis tools and communication methods to support all of an organisation's functions. This is where complimentary tools have a role to play.

ERP systems support integration and control across various functional areas of a company, therefore supporting the achievement of the company's plans [9]. This makes them an excellent source of raw data in a relatively well defined format and structure, however the volume and granularity of the data make analysis inefficient or inadequate without the application of BI tools.

CUBIST [4] argues that the complexity of BI tools is the biggest barrier to successful analysis, particularly because they do not work with the meaning of data (semantics) and are not capable of effectively handling unstructured and structured data.

In this specific example the source data was the result of a game between competing student teams undertaken on a mainstream ERP system provided by the business software vendor SAP A.G. [11] and using the ERPsim software provided by ERPsim Lab at HEC Montreal [8]. The simulation generated data on which Business Intelligence was performed. The data generated by the simulation exercise represents typical business activity and is not specifically for FCA, thus it provides a meaningful test of FCA in BI from ERP data.

ERPsim is based on SAP ECC 6.0 which is an ERP system capable of supporting in this example logistics and financial activities for a number of competing companies. All sales, procurement, master data, inventory, marketing and financial transactions are captured real time in additional to a limited number of reports to show sales, inventory, balance sheet and profit and loss. These are transaction based reports and offer no analysis without the application of further tools.

As an ERP system is effectively a relational database with data held in joined tables it is possible to extract data that contributed towards a goal via a query. Therefore a query using the table relationships was able to extract all the transactional data available that contributed towards the outcome. For example all sales transactions within the time period could be found via the connection from billing through the outbound shipments to the sales orders. Correspondingly individual sales order profit based on the materials cost price could also be extracted.

The chart in figure 1 provides an example of the input and output variables plotted to highlight the relationships that can exist in the simulation game. On the right hand side cumulative profit and percentage profit above cost per sale are shown. Cost is indexed at 100%, therefore 105 equates to 5% profit over cost. On the left hand side days inventory cover and the percentage of sales price attributable to marketing spend is shown. In summary the data could represent a number of relationships including:

- Increasing cumulative profit has an inverse relationship with decreasing days of inventory cover (how many days the stock will last given the sales forecast). [Holding less stock will result in more profit]
- Increasing profit has a direct relationship with increasing marketing spend. [Spending on marketing leads to more sales, therefore more profit]
- Increasing profit has a direct relationship with increasing profit per sale. [More profitable individual sales leads to higher overall profit]

## 2 Method

The primary problem is how to analyse data and identify semantic data or relationships from a generic transactional data set. The coursework addressed three of the learning outcomes from the course [12]:

1. Describe the notion of representing and reasoning with knowledge for smart applications.
2. Draw on one or more frameworks and techniques for representing and reasoning with knowledge for smart applications.
3. Identify the practical use of software tools for developing smart applications.

The scenario presented to the students was:

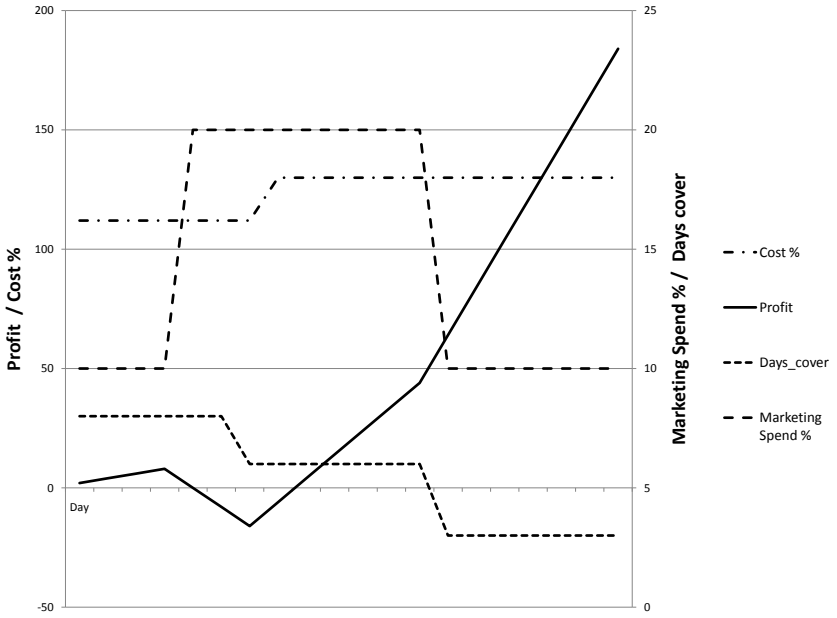



**Fig. 1.** Example of Input and Output Variables

*You are performing the role of a business analyst who has been tasked with analysing the performance of your ERP Water Company by understanding how a) your decision-making and that of others has impacted the organisation and b) identifying rules that could be used to help this decision making in the future. You are also evaluating the method of analysis in order to refine the approach employed for future iterations of this process. It is therefore less the intention to learn ERP; rather through this experience you will explore business intelligence and the role that Formal Concept Analysis (FCA) might play in this context.*

The coursework had three main sections consisting of conventional BI, FCA and Evaluation / Conclusions. The BI analysis would use MS Access and Excel in order to familiarise the students with the data using tools that would already be familiar and offer graphical analysis techniques that are common and taught at a school level of mathematics. Secondly FCA tools are applied based on essentially the same data with any calculated values added to support the analysis. This is expected to be an iterative process in order to produce the best results possible but the core section of the data extract should be stable and reusable. The final section is an evaluation of the two approaches and conclusions.

It is acknowledged that the goal of both the BI and FCA approach is to identify potentially the same relationships, this is deliberate in order to encourage an understanding of the data using tools and techniques in applications such as

| Technique | % Occurrences |
|---|---|
| Line chart (2 variables) | 100 |
| Graph on graph comparisons | 69 |
| Cumulative and actual data charts | 62 |
| Detailed Focus with annotation. | 46 |
| Line chart (3 variables) | 23 |
| Pie chart | 23 |
| Data table | 15 |
| Pivot table | 15 |
| Summary table (annotated) | 8 |
| Use of trend lines | 0 |

**Table 1.** Methods Applied under BI

Excel that will be familiar and well supported with documentation and guides. An understanding of the data and relationships was deemed necessary given the students had no prior knowledge of ERP systems or an understanding of the processes in operation.

The tools set consists of five key software packages: MS Access as a mechanism for extracting data from the ERPSim SAP system and creating the initial data file (CSV) for analysis, MS Excel and FCA tools including: FcaBedrock [2], In-close [1] and Concept Explorer [15].

The method selected was generally an experimental and iterative approach in order to extract and analyse key data, gradually refining the method to explore the anticipated relationships and evaluate the capabilities of the tool set. The aim was to supply a consistent set of data to the FCA tool set making it a repeatable process.

## 3 Student Results

The basic analysis methods applied across all the course work are shown in Table 1 and 2 with a percentage occurrences. It is noted that the marking of the coursework did take into account more than the range of techniques applied.

A minority of students also attempted to identify rules that explicitly stated relationships and could be reused in future iterations of the simulation.

The average mark achieved was 57 % with a standard deviation of 15.3.

Tables 3 and 4 contains a summary list of points made within the Evaluations and Conclusions section of the coursework for BI and FCA respectively.

## 4 Discussion

The initial reaction of the students was one of confusion in how to tackle the coursework, this is reflected to a degree in Tables 1 and 2, these show that less

| Technique | % Occurrences |
|---|---|
| Analysis over 2 data ranges | 69 |
| Percentages of occurrences | 54 |
| Identification of Relationships | 54 |
| Analysis by product profitability | 38 |
| Use of Ranges | 38 |
| Analysis over 3 data ranges | 38 |
| Analysis by Profit by quarter | 23 |
| Performance measures / KPIs | 23 |
| Graph on Graph Comparison | 8 |

**Table 2.** Methods Applied under FCA

| Pros | Cons |
|---|---|
| Good compatibility with data sources / MS Access etc | Data can be manipulated / changed manually at the interpretation or error of the user |
| Easy to learn | Have to drive the analysis and discover trends, no automation |
| Can manipulate data and combine with charts/diagrams | Required manual input to compare multiple charts etc |
| Hands on, easy to manipulate data. | Difficult to represent hierarchies in the data |
| Graphical options give quick visual descriptions of any rules/trends | Tools do not replace expert knowledge |
| Handles different data types, formulas | Data can be misunderstood |
| Reliable software | |
| Widely available | |
| Reuse / Refresh of charts etc | |

**Table 3.** Pros and Cons for BI

complex forms of analysis were prevalent in all work, for example line charts with two variables, but relatively few progressed onto considering more complex selections such as line charts with three variables. A little trial and error coupled with confidence could have eliminated most problems, this could also be supported better with guided examples attempting the coursework.

The marking of the coursework produced a normal distribution of marks with an unexpected enthusiasm for FCA although this was tempered by the difficulties in using the tool set. It is not surprising that they experienced difficulties given the difference between the development effort behind the FCA tool set and BI tools from providers such as Microsoft. It could be surmised that the students understood the advantages of analysing large and relatively unstructured data without expert knowledge or time consuming analysis. It would have been nice to see the students experimenting more with the data and discovering or at least looking for less obvious relationships.

| Pros | Cons |
|---|---|
| Good for analysing small data sets | Difficult to refine data, particularly large data sets |
| Data can be refined in FCA | Involved manual manipulation of data source |
| Good for displaying large amounts of data | Difficult to identify anomalies in the data and to correct. |
| Lattice covers all possible aspects (with Concept Explorer) | Many different formats, applications time consuming |
| Relationships are highlighted visually | Difficult to pin point trends/rules in concept form (for this example) |
| A level of interaction with the data | Any data must be calculated for going into FCA and was therefore reliant on other tools to structure the data, i.e. Excel |
| Analysis of relationships between unconnected data categories. | Comparing multiple lattices etc. is not supported directly. |
| Good for viewing hierarchies | Lack of statistics or alternative graphical analysis or drill down to raw data |
| | Data has to be consolidated to a large extent (to much) before the lattice is readable. |
| | Difficult to reuse not integrated with source data. |

**Table 4.** Pros and Cons for FCA

A consistent criticism of the FCA tool set, see table 4, was the difficulty level involved in data preparation and use of the tools. It would have been nice to eliminate some of the repetitive tasks required by the exercise as the students struggled to grasp and achieve a reusable data extraction mechanism, therefore consuming time that could have been spent more productively on the analysis. A problem that is not uncommon in real life applications.

The presentations produced for assessment made it relatively easy to mark however it was sometimes difficult to understand what was trying to be communicated especially where annotations or additional notes were not present or of low quality. The graphical nature of the presentation medium did form a good basis for presenting the analysis and forced a summary rather than lengthy descriptions of the process and mechanisms involved.

It was clear from the conclusions in Tables 3 and 4 that an appreciation about the difficulties involved in delivering BI was achieved even from this relatively small data set.

The students really failed to identify data or relationships outside of the key parameters, this is partially due to the data available as it was only a partial extract of ERP systems. Even so there are many factors that could have been offered for consideration even if they could not directly be included in the anal-

ysis. Examples of this could include the team structure or the decision making of certain individuals being categorically better or worse in outcome to others.

Graph on graph comparisons featured highly in the BI analysis, essentially this included graphical comparisons that were either overlaid or annotated to illustrate an event or relationship. Considering the frequency of this type of analysis when it came down to the FCA tools set it was hardly applied, even though the concept lattices are primarily a visual tool. The reasons for this were not clear and possibly related to the difficulty experienced in using the tool set. This feature was not supported in the tool set but it was easily possible to capture and present images side by side within the presentation.

Discrete values proved much easier to understand than ranges, in order for ranges to be understood manual input is required in order to create meaningful sub ranges. Progressive scaling was applied but the definition of the discrete values was not appropriate to take advantage of this. With this in mind a bi-ordinal scale would be more useful when representing such values but this will require a different approach when extracting the data or within FCA.

As soon as the analysis required calculations to be performed it started to face many of the challenges also faced by BI. Firstly there may be differences in the calculations between analysts, regions or indeed of interpretation. Secondly, calculated figures and performance measures can lack scale. The analysis was more successful when focus was given to a specific attribute, this was achieved by restricting the data being analysed. The down side of this was that it was a manual process with relatively long iterations even though the source data set did not alter. This limits the scope of data available and potentially the results obtained which could be a significant disadvantage.

It was clearly difficult to analyse the lattices unless a specific feature was chosen as the focus for the analysis, primarily due to their size and complexity. A possible side effect of focussing would be the accidental exclusion of data that could highlight unknown or unexpected relationships which should have been a major benefit for this type of analysis. The whole problem of visualising and exploring or "concept exploration" as termed by Stumme [10] is proving to limit the usefulness of this approach graphically at this time but alternative methods of applying the results may be possible that either solve this issue or do not require graphical representation.

The analysis was limited as it only included attributes that could be attributed to a strategic goal within the ERP system. Making the link within relational database is relatively straight forwards however a far greater challenge would be including data from sources with less well defined relationships. This maybe possible using tentative links such as times and dates but further work is required. This could be achieved within the data extract query as applied in MS Access for this approach.

## 5 Review of Learning Outcomes

Learning objective 1 - *Describe the notion of representing and reasoning with knowledge for smart applications.* This was visible in the coursework by the use of techniques such as performance measures / key performance indicators (KPI) within the data extraction on graphical interrogation of the outcome.

Learning objective 2 - *Draw on one or more frameworks and techniques for representing and reasoning with knowledge for smart applications.* This was visible in the coursework by the application of the tools and presentation of the analysis in the form of the coursework. The range of techniques applied further demonstrated the depth of analysis. There are a wide range techniques available and a reasonable range have been applied but only the minority of students have applied them.

Learning objective 3 - *Identify the practical use of software tools for developing smart applications.* This was visible in the coursework clearly by the conclusions where the ability to interact with the analysis and discover relationships was a clear advantage for FCA tools.

An emergent learning outcome was with regard to a developed appreciation of how the application of relatively simple analysis can highlight major flaws in the decision making processes employed during the game therefore resulting in poor performance. A number of teams indicated this and identified where mistakes had been made due to a lack of analysis or assumptions based on incomplete knowledge.

## 6 Conclusion

The learning outcomes have been achieved with all students appreciating the value and difficulties associated with analysing ERP data. The results did reflect a reasonable range of marks being awarded with all students able to perform both BI and FCA over the data set provided.

The difficulty involved in data preparation had a significant impact on the analysis performed, particularly with respect to the application of more complex analysis techniques and semantic discovery. This was the main factor that detracted from the learning outcomes.

The coursework would benefit from more focus on the analysis and less effort required for the preparation of data. It is expected that significant manual input will still be required in terms of defining any calculations and manipulation of graphical outputs.

A structured criteria for the analysis techniques expected could lead to an improvement in the marks awarded. This could include a pre-configured solution containing the basic forms of analysis, therefore forcing the use of more advanced analysis methods as a minimum criteria for the coursework. This could be achievable by reducing in the amount of data preparation activity required, however this must not place a constraint on the experimental aspect of this coursework and the ability to perform an open analysis.

There is a continued value in applying two methods of analysis, the BI approach is already familiar to the audience and clearly help understanding of the data set. Applying purely an FCA approach would be very challenging at this point in time.

As part of the action research aspect a number of factors should be changed for the next deployment of this coursework in order to permit the students to progress towards more advanced use of FCA, this is detailed in Further Work.

It was clear from the conclusions that the notion of applying BI and FCA was understood and the value it has in real life applications. The value of good analysis and the ability to evaluation unknown relationships was imparted. Equally the potential for error, misunderstanding and potential lack of uptake because of the complexity was clear and echoed the comments from Gartner in the introduction with respect to what how analysis will be controlled by business units and not technical experts [5].

### 6.1 Further Work

The further work section will contribute towards the action research agenda and includes ideas or approaches to be included in the next iteration of the coursework.

A solution to reduce the amount of data preparation is required in order to support a focus of more advanced FCA. The first area for consideration is providing a starting point that already supports the simpler forms of analysis.

More advanced forms of analysis should be directly supported, this includes utilising qualitative data, better visualisations such as lattice on lattice comparisons and concept clustering with iceberg lattices [13] and different scaling methods such as bi-ordinal. This is likely to impact the choice of tools selected.

Utilising a solution that integrates directly to the data instead of the restricted data set contained in the MS Access extract is definitely a requirement in order to support the forms of analysis highlighted above while providing a mechanism for an iterative and experimental approach to finding relationships within the data.

# Bibliography


[1] Andrews, S. [2010], 'In-close'.
**URL:** *http://sourceforge.net/projects/inclose/* (accessed 2009-08-11)

[2] Andrews, S. and Orphanides, C. [2010], Fcabedrock, a formal context creator, *in* F. S. Croitoru, M. and D. Lukose, eds, '18th International Conference on Conceptual Structures (ICCS).', Springer, pp. 181–184.

[3] Andrews, S., Orphanides, C. and Polovina, S. [2011], Visualising computational intelligence through converting data into formal concepts, *in* 'To appear in: Next Generation Data Technologies for Collective Computational Intelligence, Bessis, N., Xhafa, S. (eds.), Studies in Computational Intelligence book series, Springer.', Springer.

[4] CUBIST [2010], 'Cubist'.
**URL:** *http://www.cubist-project.eu/* (accessed 2010-12-11)

[5] Gartner [2009], 'Gartner reveals five business intelligence predictions for 2009 and beyond'.
**URL:** *http://www.gartner.com/it/page.jsp?id=856714* (accessed 2010-12-15)

[6] Luhn, H. [1958], 'A business intelligence system', *IBM Journal of Research and Development* .

[7] McComb, D. [2004], *Semantics in Business Systems: The Savvy Manager's Guide*, San Francisco, US, Elsevier.

[8] Montreal, H. [2011], 'Erpsim lab'.
**URL:** *http://erpsim.hec.ca/*

[9] Portousal, V. and Dunderam, D. [2006], 'Business processes: Operation solutions for sap implementations', *Idea Group Inc* .

[10] Priss, U. [2007], 'Formal concept analysis in information science', *Annual Review of Information Science and Technology* .

[11] SAP [2011].
**URL:** *http://www.sap.com/*

[12] Sheffield-Hallam-University [2009], 'Smart applications'.

[13] Stumme, G., Taouil, R., Bastide, Y. and Lakhal, L. [2002], 'Conceptual clustering with iceberg concept lattices', *Data & Knowledge Engineering* **42**.

[14] Wormuth, B. and Becker, P. [2004], Introduction to formal concept analysis, *in* '2nd International Conference of Formal Concept Analysis', Springer.

[15] Yevtushenko, S. [2010], 'Concept explorer'.
**URL:** *http://sourceforge.net/projects/conexp/* (accessed 2010-09-15)