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:

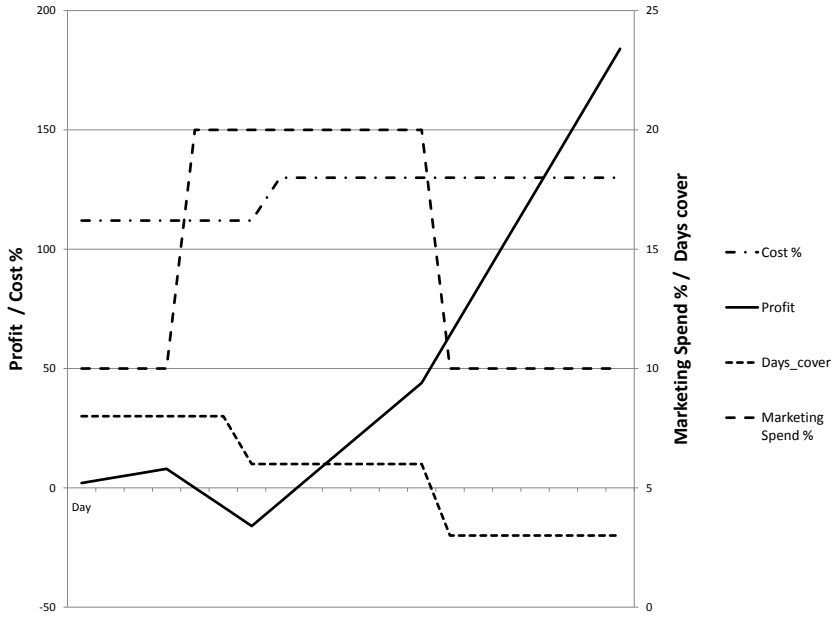

Fig. 1. Example of Input and Output Variables

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)