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 (	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

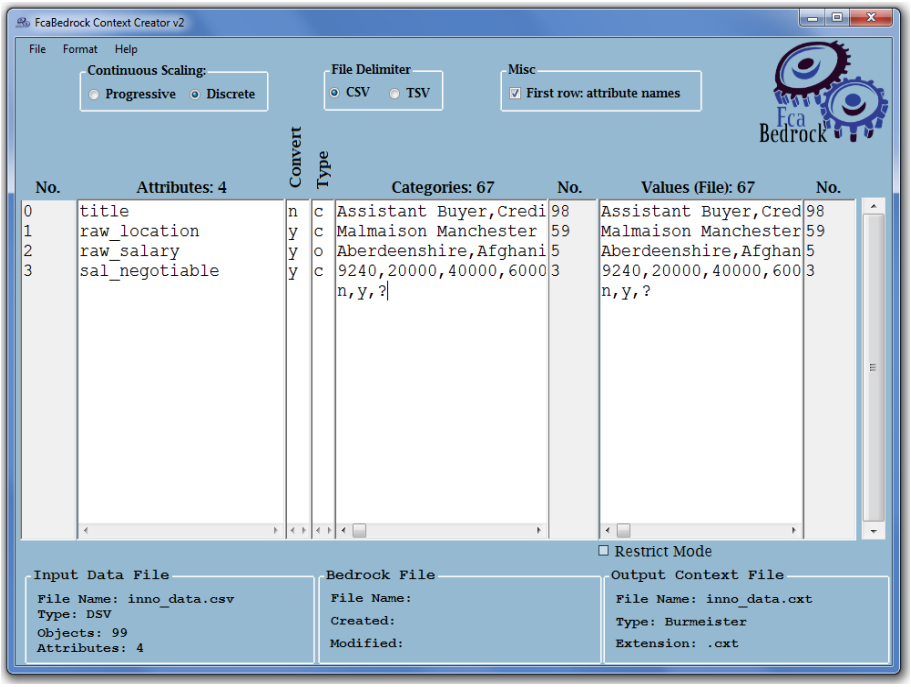

Fig. 3. Autodetecting the metadata in FcaBedrock.

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

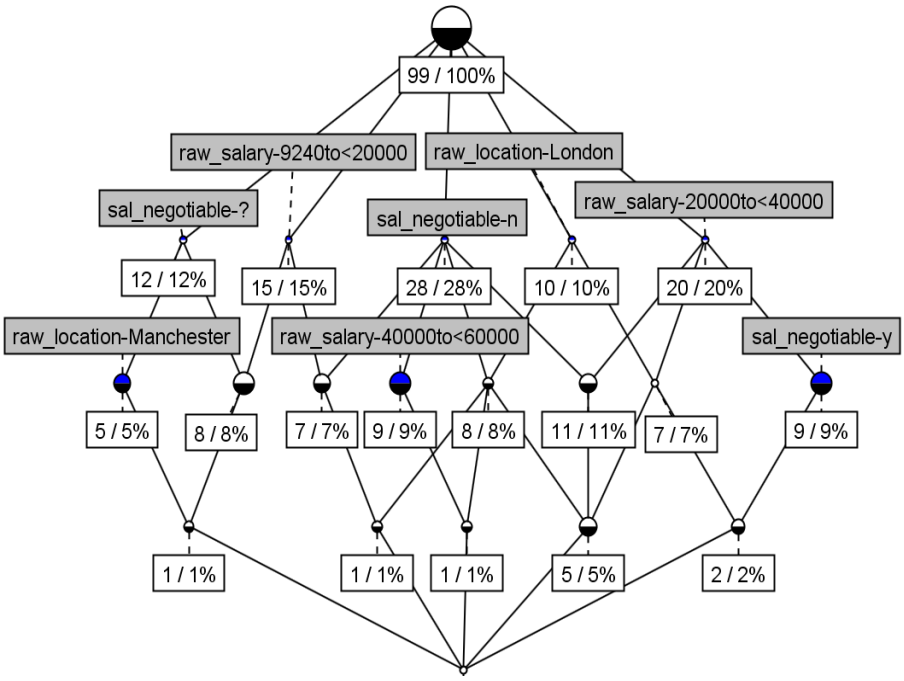

Fig. 4. Visualising the resulting formal context in ConExp.

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".

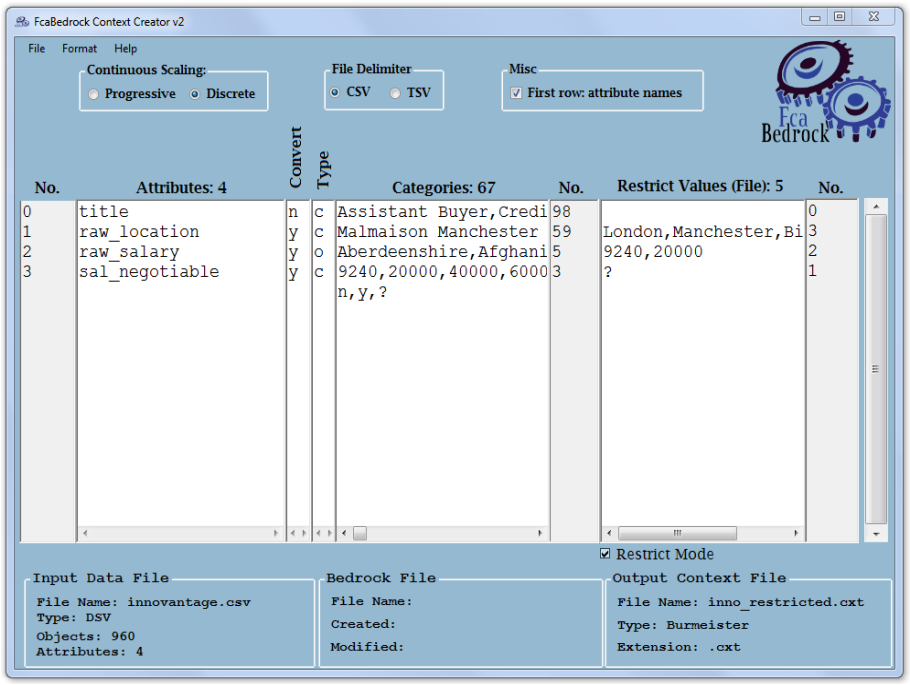

Fig. 5. Using FcaBedrock's restriction capabilities to focus the analysis on specific attributes and attribute values.

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.