FcaBedrock, a formal context creator

ANDREWS, S. <http://orcid.org/0000-0003-2094-7456> and ORPHANIDES, C.

FcaBedrock, a Formal Context Creator

Simon Andrews and Constantinos Orphanides

Conceptual Structures Research Group, Communication and Computing Research Centre, Sheffield Hallam University, Sheffield, UK
s.andrews@shu.ac.uk, corphani@my.shu.ac.uk

Abstract. FcaBedrock employs user-guided automation to convert c.s.v. data sets into Burmeister .cxt and FIMI .dat context files for FCA.

1 Introduction

Data often exists in the form of flat-files of comma separated values. For FCA to be carried out, these data must be converted into Formal Contexts. Many tools exist to carry out analysis of Formal Contexts but few exist that carry out this preparatory task. Elba performs this task for data-base tables to supply ToscanaJ with Formal Contexts [3], but FcaBedrock deals with flat-file data, producing Formal Context files that can be used by a number of tools and programs. Moreover, FcaBedrock has been developed to convert large data sets into Formal Contexts. It has now been made available at *Sourceforge*[1]. FcaBedrock discretizes and Booleanizes data; taking each many-valued attribute and converting it into as many Boolean attributes as it has values and converting continuous values using ranges [4]. Data can be interpreted in many ways leading to inconsistent analysis and problems in measuring the performance of FCA algorithms [1,5]. FcaBedrock solves these problems by documenting data conversions in re-usable, editable, meta-data files called Bedrock files (Figure 1).

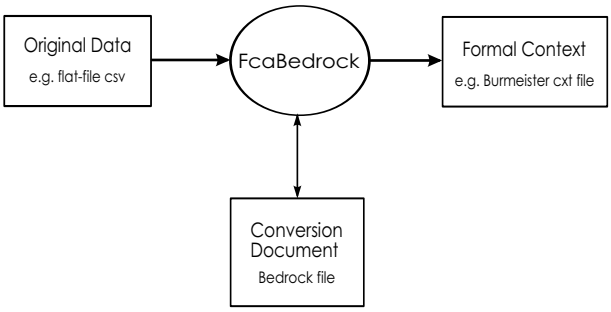


Fig. 1. FcaBedrock Process

[1] http://sourceforge.net/projects/fcabedrock/

M. Croitoru, S. Ferré, and D. Lukose (Eds.): ICCS 2010, LNAI 6208, pp. 181–184, 2010.

2 Operation

Figure 2 shows FcaBedrock. There are fields for the names and types of the original data attributes, whether they are to be converted, their categories and the corresponding category values found in the data file. These can be entered by the user, input via a Bedrock file or auto-detected from a data-file. The names of the categories and the category file values are not always the same, so FcaBedrock uses both; the category values are required for converting data and the category names appear in the Context file. The example shown is the *Mushroom* data set from the UCI Machine Learning Repository [2].

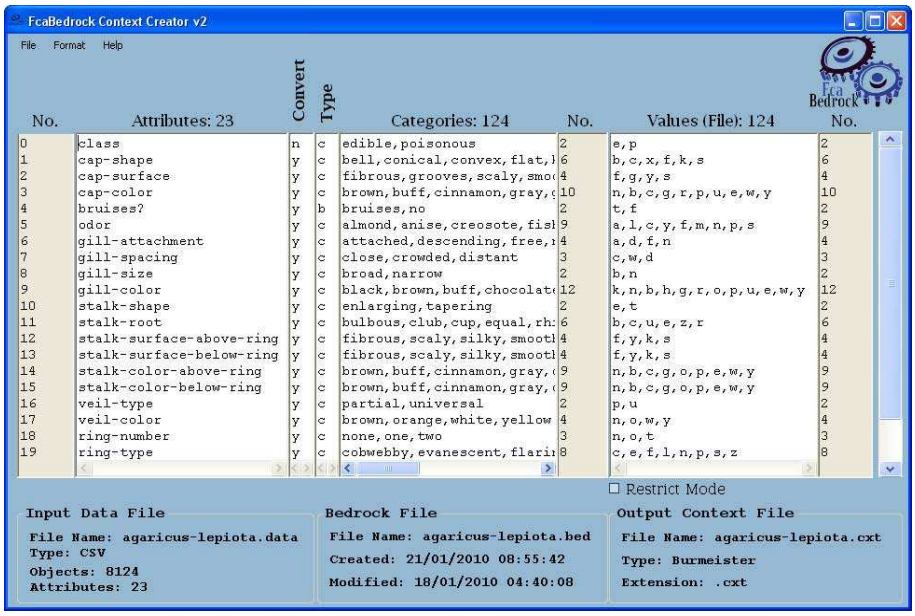


Fig. 2. FcaBedrock

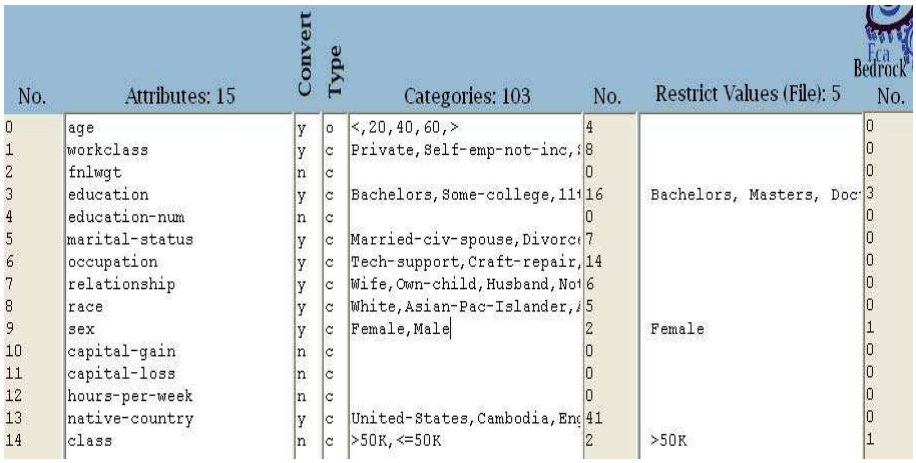
3 Capabilities

The following is a list of some of the capabilities of FcaBedrock (for a more detailed description, see the software documentation at *Sourceforge*). Figure 3 illustrates some of features using the *Adult* data set from UCI [2].

- Data types converted: categorical, continuous and Boolean.
- Input file formats: Many column csv, three-column csv (triples).
- Output file formats: Burmeister (.cxt), FIMI format (.dat)[2].
- Auto-detection of meta-data from data-file.

[2] `http://fimi.cs.helsinki.fi/`

– Interpretation of data:
  - Restrict conversion to user-defined values.
  - Exclude from conversion user-defined values.
  - Freedom over treatment of missing values.



No.	Attributes: 15	Convert	Type	Categories: 103	No.	Restrict Values (File): 5	No.
0	age	y	o	<,20,40,60,>	4		0
1	workclass	y	c	Private,Self-emp-not-inc,S	8		0
2	fnlwgt	n	c		0		0
3	education	y	c	Bachelors,Some-college,11t	16	Bachelors, Masters, Doc	3
4	education-num	n	c		0		0
5	marital-status	y	c	Married-civ-spouse,Divorc	7		0
6	occupation	y	c	Tech-support,Craft-repair,	14		0
7	relationship	y	c	Wife,Own-child,Husband,No	6		0
8	race	y	c	White,Asian-Pac-Islander,A	5		0
9	sex	y	c	Female,Male	2	Female	1
10	capital-gain	n	c		0		0
11	capital-loss	n	c		0		0
12	hours-per-week	n	c		0		0
13	native-country	y	c	United-States,Cambodia,Eng	41		0
14	class	n	c	>50K,<=50K	2	>50K	1

Fig. 3. Creating an *Adult* sub-context using restrict-to values

The following is an example adapted from the UCI *Adult* data set, using a data file called `mini-adult.data` with eight instances and five attributes (*age*, *education*, *employment*, *sex* and *US-citizen*) plus a *salary* class. File 2 shows a corresponding output from FcaBedrock in the cxt format.

```
39, Bachelors, Clerical, Male, Yes, <=50K
50, Bachelors, Managerial, Female, Yes, <=50K
38, HS-grad, Unskilled, Male, Yes, <=50K
53, 11th, Unskilled, Male, Yes, <=50K
28, Bachelors, Professional, Female, Yes, >50K
37, Masters, Managerial, Female, No, <=50K
49, ?, Clerical, Female, No, <=50K
52, HS-grad, Managerial, Male, Yes, >50K
```

File 1. `mini-adult.data`

4 Evaluation

An initial version FcaBedrock was evaluated by a class of final-year Computing undergraduates at Sheffield Hallam University (SHU). Results of this evaluation fed into the development of the version now at *Sourceforge*. This version was then evaluated by successfully converting a number of data sets into Formal Contexts, including the *Mushroom*, *Adult*, *Internet Advertisements*, *Flags* and *Tic-tac-toe* data sets from UCI and several internal student information and supermarket data sets at SHU. The Context files produced were successfully and consistently processed by two Formal Concept generators.

```
B       7                       employment-Unskilled
        age-<30                 sex-Male
8       age-30to<40             sex-Female
15      age-40to<50             US-citizen
        age->=50                .X..X...X...X.X
0       education-Bachelors     ...XX....X...XX
1       education-Masters       .X.....X...XX.X
2       education-11th          ...X..X....XX.X
3       education-HS-grad       X...X.....X..XX
4       employment-Clerical     .X...X...X...X.
5       employment-Managerial   ..X.....X....X.
6       employment-Professional ...X...X.X..X.X
```

File 2. `mini-adult.cxt`, *Burmeister* context file.

There are several other file formats used in FCA, obtainable from a cxt file produced by FcaBedrock using the conversion tool, *FcaStone* [7]. A version of FcaBedrock is being developed that takes RDF-S and OWL as input [6]. This work will form a core part of CUBIST ("Combining and Uniting Business Intelligence with Semantic Technologies"), awarded under the European Union's 7th Framework Programme, 5th ICT call, topic 4.3: Intelligent Information Management; STREP Project No.: FP7 257403.

References

1. Andrews, S.: Data Conversion and Interoperability for FCA. In: CS-TIW 2009, pp. 42–49 (2009), `http://www.kde.cs.uni-kassel.de/ws/cs-tiw2009/proceedings_final_15July.pdf`
2. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
3. Becker, P., Correia, J.H.: The ToscanaJ Suite for Implementing Conceptual Information Systems. In: Ganter, B., Stumme, G., Wille, R. (eds.) Formal Concept Analysis. LNCS (LNAI), vol. 3626, pp. 324–348. Springer, Heidelberg (2005)
4. Ganter, B., 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, Heidelberg (1989)
5. Kuznetsov, S.O., Obiedkov, S.A.: Comparing Performance of Algorithms for Generating Concept Lattices. Journal of Experimental and Theoretical Artificial Intelligence 14, 189–216 (2002)
6. Passin, T.B.: Explorer's Guide to the Semantic Web, Manning, Greenwich, CT (2004)
7. Priss, U.: FcaStone - FCA File Format and Interoperability Software. In: Croitoru, M., Jaschkë, R., Rudolph, S. (eds.) CS-TIW 2008, pp. 33–43 (2008)