About Universality and Flexibility of FCA-based Software Tools

A.A. Neznanov, A.A. Parinov

National Research University Higher School of Economics,
20 Myasnitskaya Ulitsa, Moscow, 101000, Russia
ANeznanov@hse.ru, AParinov@hse.ru

Abstract. There is a big gap between variety of applications of Formal Concept Analysis (FCA) methods and general-purpose software implementations. We discuss history of FCA-based software tools, which is relatively short, main problems in advancing of such tools, and development ideas. We present Formal Concept Analysis Research Toolbox (FCART) as an integrated environment for knowledge and data engineers with a set of research tools based on Formal Concept Analysis. FCART helps us to illustrate methodological and technological problems of FCA approach to data analysis and knowledge extraction.

Keywords: Formal Concept Analysis, Knowledge Extraction, Data Mining, Software.

1 Introduction

Formal Concept Analysis (FCA) [1] is a mature group of mathematical models and a well foundation for methods of data mining and knowledge extraction. There is a huge amount of publications about many aspects of FCA using in different application fields.

Why are there no universal and flexible FCA based tools have implemented in data mining software? We can assume that last ten years is an enough time to implement such tools not only in the most popular big analytic software but also in separate instruments are tightly integrated with other data access tools. Is there a problem in methodology of current "FCA data mining" approach?

Most popular FCA tools are small utilities purposed for final drawing of relatively small (tens to hundred concepts) formal concept lattice and calculate its properties. The other tools are very experimental and separately implement one or several mathematical models.

Can we have exchangeable set of tools for iterative building a formal context from raw data, for working with big lattices, for deep analysis of properties of interesting concepts in good user interface?

2 Problems and solutions

The FCA community started a discussion of the universal FCA workflow several years ago. There were a number of approaches to extending FCA to richer descriptions and at the same time to extending scalability of FCA-based methods proposed:

1. Pattern structures [2, 3, 4, 5] for dealing with complex input objects (such objects may have, for example, graph representation).
2. Various scaling techniques [6, 7, 8] for dealing with many-valued contexts.
3. Similarity measures on concepts [9, 10].
4. Various concepts indices for important concepts selection [11, 12, 13].
5. Alpha lattices [14] (and other coarser variants of concept lattice).
6. Relational concept analysis [15].
7. Attribute exploration [16].
8. Approaches for fragmentary lattice visualization (from iceberg to embedded lattices).
9. "Fuzzy FCA" approach in form of biclustering methods [17] and other techniques.

Near the middle of the last decade there were very successful implementations of transforming a small context into a small line diagram and calculate implications and association rules. The most well-known open source projects are ConExp [18], Conexp-clj [19], Galicia [20], Tockit [21], ToscanaJ [22], FCAStone [23], Lattice Miner [24], OpenFCA [25], Coron [26], Cubist [27]. These tools have many advantages. However, they suffer from the lack of rich data preprocessing, the abilities to communicate with various data sources, session management, reproducibility of computational experiments. It prevents researchers from using these programs for analyzing complex big data without different additional third party tools.

The goal of our efforts is an integration of ideas, methods, and algorithms, which are mentioned all above, in one environment. It is not only a programing task. At first, it is a challenge facing methodologists of the FCA community. The most significant problems:

1. Sense of FCA approaches in AI tasks.
2. Performance of basic FCA algorithms.
3. Logical complexity of raw data preprocessing task.
4. Scaling of many-valued contexts.
5. Interactive work with FCA artifacts.

Now we want to discuss some aspects of integration problems. In a previous article [28], we have described the stages of development of a software system for information retrieval and knowledge extraction from various data sources (textual data, structured databases, etc.). Formal Concept Analysis Research Toolbox (FCART) was designed especially for the analysis of semi structured and unstructured (including textual) data. In this article, we describe FCART as an extensible toolset for checking different integration ideas. As an example of current development activities, the main FCA workflow will be demonstrated.

3 Methodology

FCART was originated from DOD-DMS platform which creation was inspired by the CORDIET methodology (abbreviation of Concept Relation Discovery and Innovation Enabling Technology) [29] developed by J. Poelmans at K.U. Leuven and P. Elzinga at the Amsterdam-Amstelland police. The methodology allows analyst to obtain new knowledge from data in an iterative ontology-driven process. In FCART we concentrated on main FCA workflow and tried to support high level of extensibility. FCART is based on several basic principles that aim to solve main integration problems:

1. Iterative process of data analysis using adjustable data queries and interactive *analytic artifacts* (such as concept lattice, clusters, etc.).
2. Separation between processes of *data querying* (from various data sources), *data preprocessing* (of locally saved immutable snapshots), *data analysis* (in interactive visualizers of immutable analytic artifacts), and *results presentation* (in report editor).
3. Extendibility at three levels: customizing settings of data access components, query builders, solvers and visualizers; writing scripts (macros); developing components (add-ins).
4. Explicit definition of all analytic artifacts and their types. It provides consistent parameters handling, links between artifacts for end-user and allows one to check the integrity of session.
5. Realization of integrated performance estimation tools and system log.
6. Integrated documentation of software tools and methods of data analysis.

The core of the system supports knowledge discovery techniques, based on Formal Concept Analysis, clustering, multimodal clustering, pattern structures and the others. From the analyst point of view, basic *FCA workflow* in FCART has four stages. On each stage, a user has the ability to import/export every artifact or add it to a report.

1. Filling Local Data Storage (LDS) of FCART from various external SQL, XML or JSON-like data sources (querying external source described by External Data Query Description – EDQD). EDQD can be produced by some External Data Browser.
2. Loading a data snapshot from local storage into current analytic session (snapshot is described by Snapshot Profile). Data snapshot is a data table with annotated structured and text attributes, loaded in the system by accessing LDS.
3. Transforming the snapshot to a binary context (transformation described by Scaling Query).
4. Building and visualizing formal concept lattice and other artifacts based on the binary context in a scope of analytic session.

FCART has been already successfully applied to analyzing data in medicine, criminalistics, and trend detection.

4 Technology

We use Microsoft and Embarcadero programming environments and different programming languages (C++, C#, Delphi, Python and other). Native executable of FCART client (the core of the system) is compatible with Microsoft Windows 2000 and later and has no additional dependences. Scripting is an important feature of FCART. Scripts can do generating and transforming artifacts, drawing, and building reports. For scripting we use Delphi Web Script and Python languages. Integrated script editor with access to *artifacts API* allows quick implementation of experimental algorithms for generating, transforming, and visualizing artifacts.

Current version of FCART consists of the following components.

- Core component includes
  - multiple-document user interface of research environment with session manager and extensions manager,
  - snapshot profiles editor (SHPE),
  - snapshot query editor (SHQE),
  - query rules database (RDB),
  - session database (SDB),
  - main part of report builder.
- Local Data Storage (LDS) for preprocessed data.
- Internal solvers and visualizers of artifacts.
- Additional plugins, scripts and report templates.

FCART Local Data Storage (LDS) plays important role in effectiveness of whole data analysis process because all data from external data storages, session data and intermediate analytic artifacts saves in LDS. All interaction between user and external data storages is carried out through the LDS. There are many data storages, which contain petabytes of collected data. For analyzing such Big Data an analyst cannot use any of the software tools mentioned above. FCART provides different scenarios for working with local and external data storages. In the previous release of FCART an analyst could work with quite small external data storage because all data from external storage is converted into JSON files and is saved into LDS. In the current release, we have improved strategies of working with external data. Now analyst can choose between loading all data from external data storage to LDS and accessing external data by chunks using paging mechanism. FCART-analyst should specify the way of accessing external data at the property page of the generator.

All interaction between a client program and LDS goes through the web-service. The client constructs http-request to the web-service. The http-request to the web service is constructed from two parts: prefix part and command part. Prefix part contains domain name and local path (e.g. http://zeus.hse.ru/lds/). The command part describes what LDS has to do and represents some function of web-service API. Using web-service commands FCART client can query data from external data storages in uniform and efficient way.

5 Interactive part of main FCA workflow

The most interesting part for analyst is an interactive work with artifacts in multiple-document user interface of FCART client.

We will illustrate our vision with real examples of querying data and working with big lattices. FCART supports interactive browsing of concept lattices with more than 20000 concepts (see Fig. 1) with fine adjustment of drawing. User may feel some interaction delays but system have special instruments to visualize and navigate big lattices: scaling of lattice element sizes, parents-children navigator, filter-ideal selection, separation of focused concepts from other concepts. We have tested several drawing techniques for visualizing fragments of big lattices and prepared a collection of additional drawing scripts for rendering focused concept neighborhood, selected concepts, "important" concepts, and filter-ideal with different sorting algorithms for lattice levels.

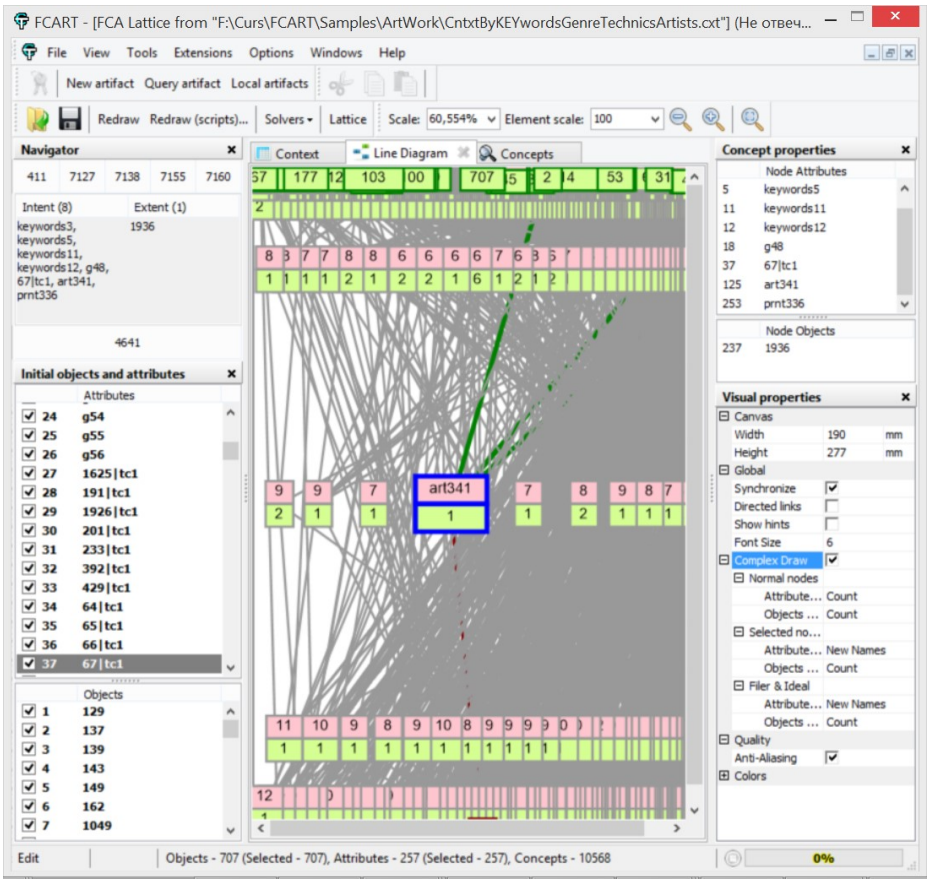

Fig. 1. Lattice browser with focused concept separation

User also can build linked sublattices and calculate standard association rules, implication basis, etc. All artifacts can be commented, annotated and appended to one of reports in a current session.

FCART also supports working with concepts in a form of table with sorting and grouping abilities (see Fig. 2). Concept properties in the last column of this table (indices [30], similarity measures, etc.) can be calculated by scripts.

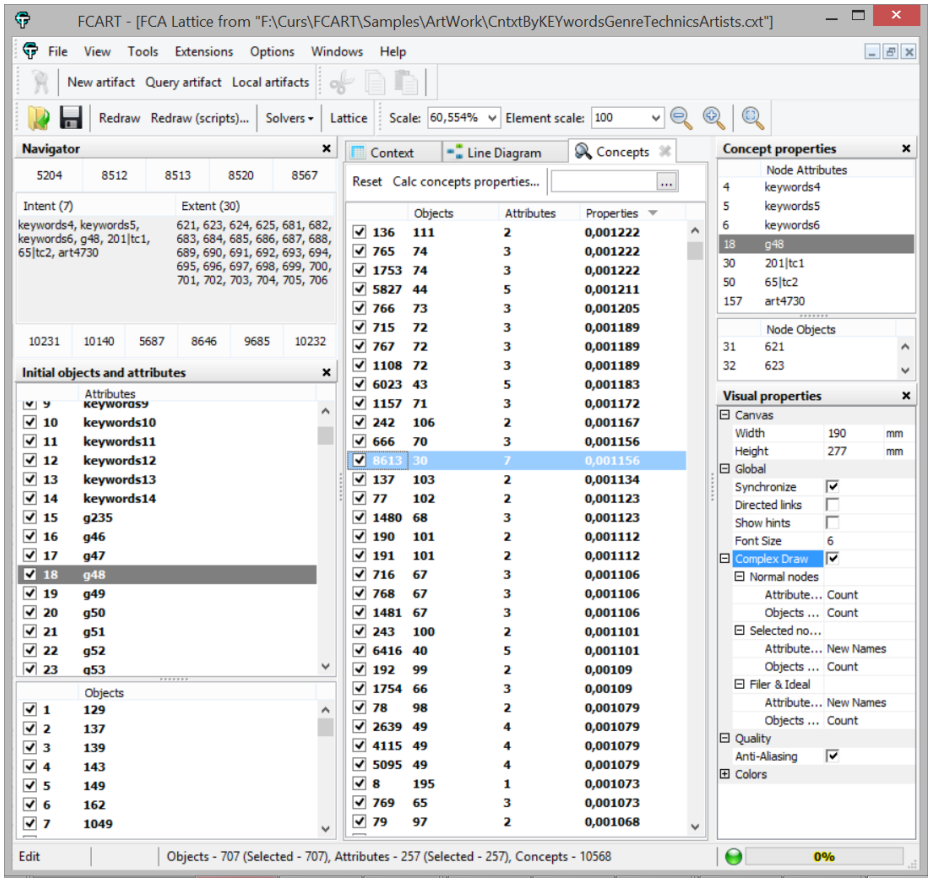

Fig. 2. List of all formal concepts, sorted by value of concept stability index

6 Conclusion and future work

In this paper, we have discussed important questions for all researchers in FCA field about implementation of essential software tools. We will try to suggest solutions for some of those problems in our system. The version 0.9.4 of FCART client (http://ami.hse.ru/issa/Proj_FCART) and version 0.2 of LDS Web-service (http://zeus2.hse.ru) have been introduced this spring.

We have tested preprocessing in the current release with Amazon movie reviews dataset (7.8 million of documents – it is big enough for check standard FCA limitations) [31]. This dataset was loaded to FCART LDS and transformed into many-valued contexts, binary contexts, lattices and other artifacts. The main goal of the current release is to develop architecture, which can work with really big datasets. For now, FCART is being tested on other collections of CSV, SQL, XML, and JSON data with unstructured text fields.

The release of the version 1.0 of FCART is planned for August 2014. We will discuss this version at the seminar and touch on flexibility, extensibility, and overall performance issues. We are opened for ideas to improve methodology and various aspects of implementation.

Acknowledgements

This work was carried out by the authors within the project "Mathematical Models, Algorithms, and Software Tools for Intelligent Analysis of Structural and Textual Data" supported by the Basic Research Program of the National Research University Higher School of Economics.

References

1. Ganter, B., Wille R. Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
2. Ganter, B., Kuznetsov, S.O. Pattern Structures and Their Projections. Proc. 9th International Conference on Conceptual Structures (ICCS-2001), 2001, pp. 129-142.
3. Kuznetsov, S.O. Pattern Structures for Analyzing Complex Data // Proc. of 12th International conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC-2009), 2009, pp. 33-44.
4. Kuznetsov, S.O. Fitting Pattern Structures to Knowledge Discovery in Big Data. Proc. of International Conference on Formal Concept Analysis (ICFCA'13), 2013, pp. 254-266.
5. Buzmakov, A., Egho, E., Jay, N., Kuznetsov, S.O., Napoli, F., Chedy Raïssi: On Projections of Sequential Pattern Structures (with an Application on Care Trajectories) // Proc. of conference on Concept Lattices and their Applications (CLA-2013), 2013: pp. 199-208.
6. Ganter, B., Stahl, J., Wille, R. Conceptual Measurement and many-valued contexts. In Gaul, W., Schader, M. Classification as a Tool of Research. Elsevier, 1986, pp. 169-176.
7. Prediger, S. Logical Scaling in Formal Concept Analysis // Proc. of ICCS'97, LNAI 1257, 1997, pp. 332-341.
8. Stumme, G. Hierarchies of Conceptual Scales // Proc.Workshop on Knowledge Acquisition, Modeling and Management KAW'99, 2, 1999, pp 78-95.
9. Ceja , J.M.O., Arenas, A.G. Concept similarity measures the understanding between two agents. Natural Language Processing and Information Systems, Lecture Notes in Computer Science, Volume 3136, 2004, pp. 182-194.
10. Alqadah, F., Bhatnagar, R. Similarity measures in formal concept analysis. Annals of Mathematics and Artificial Intelligence, Volume 61, Issue 3, 2011, pp. 245-256.
11. Kuznetsov, S.O. On stability of a formal concept. Annals of Mathematics and Artificial Intelligence, Vol. 49, 2007, pp. 101-115.

12. Ilvovsky, D.A., Klimushkin, M.A. FCA-based Search for Duplicate Objects in Ontologies, Proceedings of the Workshop Formal Concept Analysis Meets Information Retrieval, Vol. 977. CEUR Workshop Proceeding, 2013.
13. Belohlávek, R., Trnecka, M. Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications // Proc. 23th International Joint Conference on Artificial Intelligence, 2013, pp. 1233-1239.
14. Soldano, H., Ventos, V., Champesme, M., Forge, D. Incremental construction of Alpha lattices and association rules // Proc. 14th International Conference, 2010, pp. 351-360.
15. Hacene, M.R., Huchard, M., Napoli, A., Valtchev, P. Relational concept analysis: mining concept lattices from multi-relational data. Ann. Math. Artif. Intell. 67(1), 2013, pp. 81-108.
16. Ganter, B. Attribute exploration with background knowledge. Theoretical Computer Science, 217(2), 1999, pp. 215-233.
17. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J. Concept-Based Biclustering for Internet Advertisement // ICDM Workshop, IEEE Computer Society, 2012, pp. 123-130.
18. Yevtushenko, S.A. System of data analysis "Concept Explorer" (In Russian) // Proceedings of the 7th national conference on Artificial Intelligence KII-2000, pp. 127-134, Russia, 2000.
19. Conexp-clj (http://daniel.kxpq.de/math/conexp-clj/)
20. Valtchev, P., Grosser, D., Roume, C. Hacene, M.R. GALICIA: an open platform for lattices, in Using Conceptual Structures // Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03), pp. 241-254, Shaker Verlag, 2003.
21. Tockit: Framework for Conceptual Knowledge Processing (http://www.tockit.org)
22. Becker, P., Hereth, J., Stumme, G. ToscanaJ: An Open Source Tool for Qualitative Data Analysis // Proc. Workshop FCAKDD of the 15th European Conference on Artificial Intelligence (ECAI-2002). Lyon, France, 2002.
23. Priss, U. FcaStone - FCA file format conversion and interoperability software, Conceptual Structures Tool Interoperability Workshop (CS-TIW), 2008.
24. Lahcen, B., Kwuida, L. Lattice Miner: A Tool for Concept Lattice Construction and Exploration // Suplementary Proceeding of International Conference on Formal Concept Analysis (ICFCA'10), 2010.
25. Borza, P.V., Sabou, O., Sacarea, C. OpenFCA, an open source formal concept analysis toolbox // Proc. of IEEE International Conference on Automation Quality and Testing Robotics (AQTR), 2010, pp. 1-5.
26. Szathmary, L. The Coron Data Mining Platform (http://coron.loria.fr)
27. Cubist Project (http://www.cubist-project.eu)
28. Neznanov A.A., Ilvovsky D.A., Kuznetsov S.O. FCART: A New FCA-based System for Data Analysis and Knowledge Discovery // Contributions to the 11th International Conference on Formal Concept Analysis, 2013. pp. 31-44.
29. Poelmans, J., Elzinga, P, Neznanov, A.A., Viaene S., Kuznetsov, S.O., Ignatov, D., Dedene, G. Concept Relation Discovery and Innovation Enabling Technology (CORDIET) // CEUR Workshop proceedings Vol. 757, Concept Discovery in Unstructured Data, 2011.
30. Neznanov, A., Ilvovsky, D., Parinov, A. Advancing FCA Workflow in FCART System for Knowledge Discovery in Quantitative Data // 2nd International Conference on Information Technology and Quantitative Management (ITQM-2014), Procedia Computer Science, 31, 2014, pp. 201-210.
31. Web data: Amazon movie reviews (http://snap.stanford.edu/data/web-Movies.html)