A System for Knowledge Discovery in Big Dynamical Text Collections

Sergei O. Kuznetsov, Alexey A. Neznanov, Jonas Poelmans

National Research University Higher School of Economics,
Myasnitskaya 20, 101000, Moscow, Russian Federation
SKuznetsov@hse.ru, ANeznanov@hse.ru,
Jonas.Poelmans@econ.kuleuven.be

Abstract. Software system Cordiet-FCA is presented, which is designed for knowledge discovery in big dynamic data collections, including texts in natural language. Cordiet-FCA allows one to compose ontology-controlled queries and outputs concept lattice, implication bases, association rules, and other useful concept-based artifacts. Efficient algorithms for data preprocessing, text processing, and visualization of results are discussed. Examples of applying the system to problems of medical diagnostics, criminal investigations are considered.

Keywords: Formal Concept Analysis, Data Mining, Natural Language Processing, Software Tool, Visualization

1 Introduction

In this paper we introduce the software system Cordiet-FCA for data mining and knowledge discovery based on the Cordiet-DMS (Data Mining System) platform and used primarily the Formal Concept Analysis (FCA) [1] as theoretical basis. FCA emerged in the 1980's from attempts to restructure lattice theory in order to promote better communication between lattice theorists and potential users of lattice theory. Since its early years, Formal Concept Analysis has developed into a research field in its own right with a thriving theoretical community and a rapidly expanding range of applications in information and knowledge processing including visualization, data analysis (mining) and knowledge management.

The system was designed especially for unstructured data analysis. In case studies we applied Cordiet-FCA to the analysis of publications on FCA. The real-life datasets include criminal data (for example, chat conversations of pedophiles) and, in nearest future, medical and emergency rescue data.

2 Methodology

Software package Cordiet-DMS is a universal extendible software platform intended to build data mining and knowledge discovery tools for various application fields. This platform inspired by CORDIET methodology (abbreviation of Concept Relation Discovery and Innovation Enabling Technology) [2], developed by J. Poelmans in K.U. Leuven and P. Elzinga in Amsterdam-Amstelland police. The methodology allows one to obtain new knowledge from the data in iterative ontology-controlled process. The package is based on modern methods and algorithms of data analysis, technologies for manipulating big data collections, data visualization, reporting, and interactive processing techniques. There are four base principles:

1. Iterative process of data analysis using ontology-controlled queries and interactive artifacts (such as concept lattice, etc.).
2. Separation of processes of *data querying* (from various data sources) and *data analyzing* (of locally saved immutable snapshots).
3. Dividing data processing into four stages: access to external data sources and loading data to local storage; access to the local storage and generating snapshots; access to one or many snapshots and building basic analysis artifacts; access to the artifact and analyzing derivative artifacts.
4. Expendability on three levels: customizing settings of data access components, query builders, solvers and visualizers; writing scripts (macros); developing components (add-ins).

3 Current software properties

At this moment we introduce the version 0.9 of Cordiet-FCA in form of local Windows application. This version uses local XML-storage and integrated research environment with snapshot profiles editor, query builder, ontology editor, and a set of solvers and visualizers. The main solvers can produce concept lattice, sublattices, association rules, and implications, calculate stability indexes, similarity measures for contexts and concepts, etc.

We use Microsoft and Embarcadero programing environments and different programing languages (C++, C#, Delphi, Python and others). For scripting we use Delphi Web Script [3]. Also we are developing a distributed version based on Web-services.

3.1 Text processing

Cordiet-FCA has a query language for transforming data snapshot into basic analysis artifacts. The main artifact for FCA methods is a *formal context*.

The language describes so called *rules* and consists from four main rules types:

- *Simple rule* generates one attribute from structured fields of snapshot.
- *Scaling rule* generates several attributes from structured fields based on nominal or ordinal scale.

– *Text mining rule* generates one attribute from unstructured text fields.
– *Multivalued rule* generates one or many attributes from multivalued field (array).

Also we have *temporal rules* (for manipulating with date and time) and *compound rules* (for merging all types of rules into one). As usual we don't need to write a query from scratch. We can select some entities in the ontology editor and automatically generate a query. Text mining rule can use terms (set of synonymous) and term-clusters (set of terms) from ontology entities.

Cordiet-FCA uses Lucene [4] to index the content of the unstructured text fields in the snapshots using the description of the term attributes in the ontology editor. The resulting index is later used to quickly validate whether the text mining or compound rule return true or not. In fig. 1 we show how system visualizes the profile-controlled description of snapshots records (Report Viewer) and query builder (the list of rules and Rule Editor).

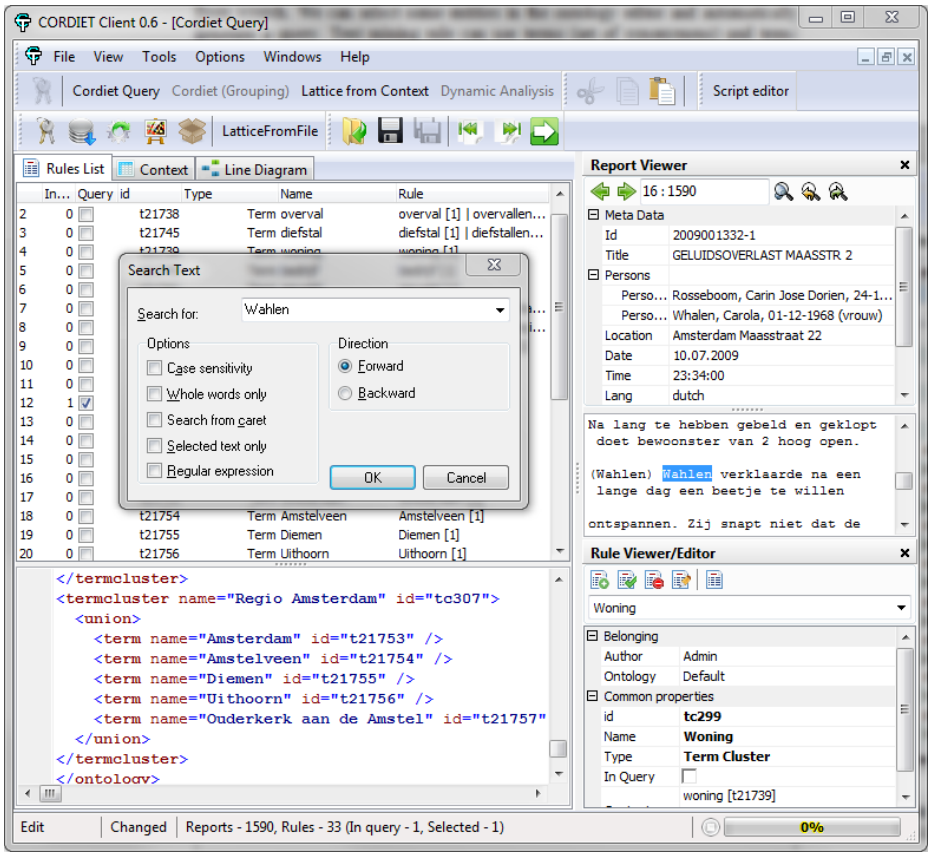

Fig. 1. Opened base of police reports and query builder

3.2 Concept lattice browser

The main mode of user interaction in Cordiet-FCA is interactive work in the *concept lattice browser*. The lattice can be used to browse the collection of objects with binary attributes given as a result of query to snapshot (with structured and text attributes). The user can select and deselect objects and attributes and the lattice diagram is modified accordingly. The user can click on a concept. The screen shows in a separate window the names of the objects in the extent and the names of the attributes in the intent. Names of objects and attributes are linked with initial snapshot records and fields. If the user clicks on the name of an object, the content of the object is shown in a separate window according to snapshot profile. If the user clicks on the name of an attribute, its content is also shown in a separate window.

Fig. 2 demonstrates the browser (building sublattice). The multidocument interface allows us to inspect several lattices and moreover the system remembers all links between derivative artifacts.

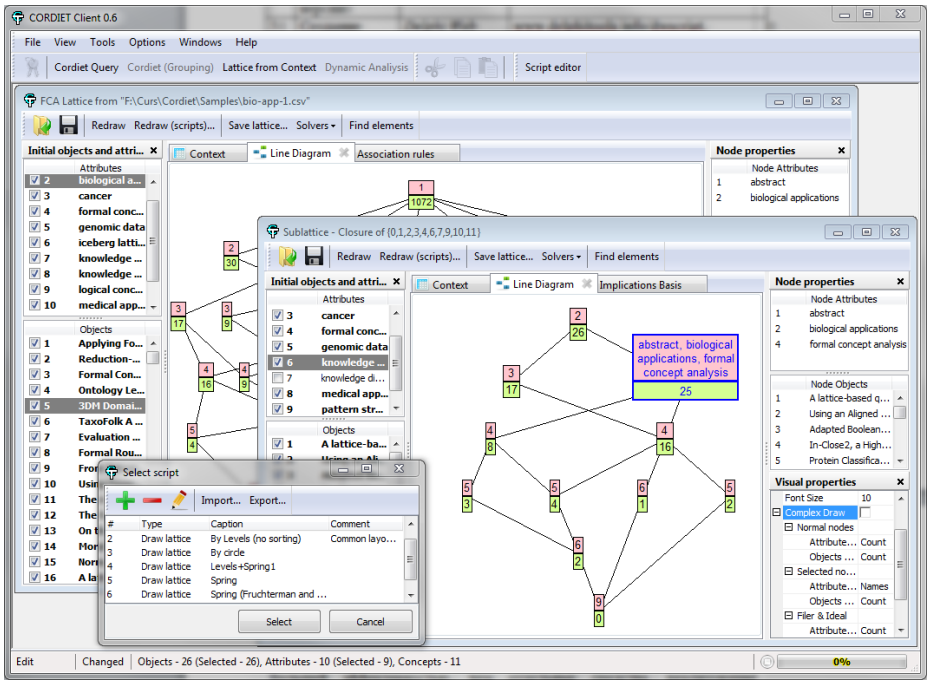

Fig. 2. Concept lattice browser

The user can customize the lattice browsing settings. The user can specify whether the nodes corresponding to concepts show the numbers of all (or only the new) objects and all (or only the new) attributes in extent and intent respectively, or the names of all (or only the new) objects and all (or only the new) attributes. Separate settings can be specified for the selected concept, the concepts in the order filter and the remainder of the lattice. If the user presses shift and at the same time selects a concept,

the order filter generated by this concept is shown. The colors of concepts and edges can be customized also.

A right click on the name of an attribute shows the user several options: the user can choose to build a sublattice containing only objects having the selected attribute, to build a sublattice containing only objects which do not have the selected attribute, or to find the concept in which the attribute first occurs starting from the supremum of the lattice.

3.3 Validation and applications

Main solvers of the system were validated on the classical test sets (from Frequent Itemset Mining Dataset Repository and UCI Repository). Because of constant improving of basic algorithms and data structures we don't have a good comparable set of benchmarks now.

We used Cordiet-FCA in the research work of the Laboratory of Intelligent Systems and Structural Analysis in NRU HSE and in some applied tasks connected with medical informatics, crime investigations, etc. Fig. 3 demonstrates an example of analyses of a pedophile behavior (it is based on information from chat conversations in Internet).

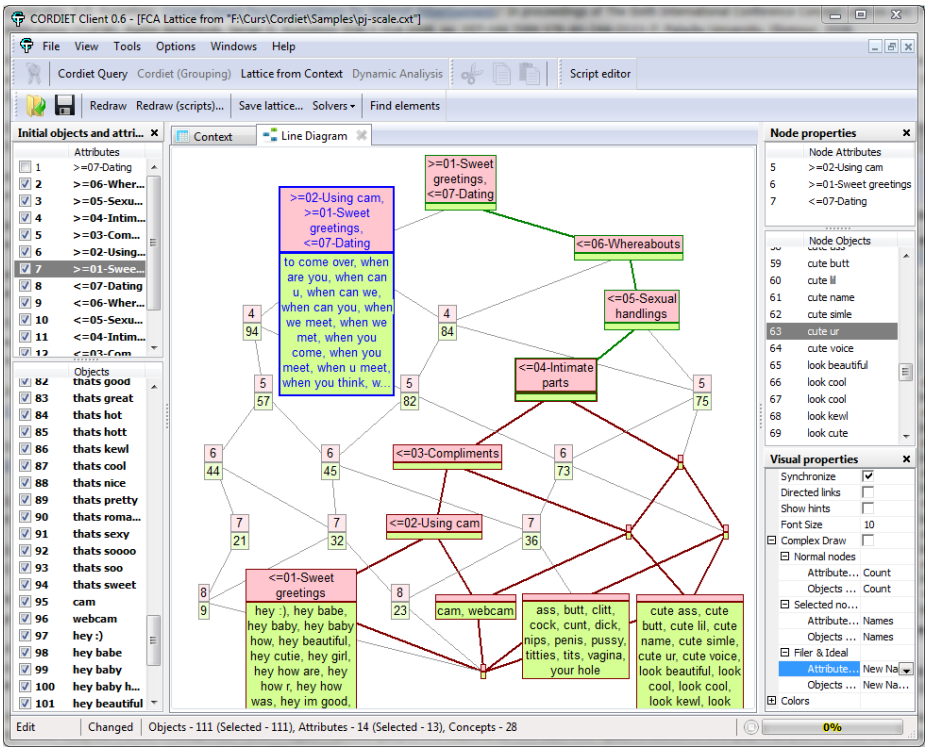

Fig. 3. Sample of concepts exploration (filter, ideal and selection of attributes)

4 Comparison with existing well-known FCA software

Comparing Cordiet-FCA with big analytic software like IBM i2 Analyst's Notebook or QSR NVivo shows that the latter do not have a normal set of FCA tools and have a completely different methodology of data analysis. We also compare basic functionality of the system with well-known tools for building and visualizing FCA artifacts (table 1).

Table 1. Some well-known FCA software tools

Program title	Author	Web-site
Concept Explorer	S.A.Evtuchenko	conexp.sourceforge.net
FcaStone	U. Priss et al	fcastone.sourceforge.net
Conflexplore	P.Borza, O.Sabou	code.google.com/p/openfca
Galicia	P.Valtchev et al	www.iro.umontreal.ca/~galicia
ToscanaJ	University of Queensland, Technical University of Darmstadt	toscanaj.sourceforge.net

All of the tools from Table 1 have unique features. For example, Concept Explorer has interesting modes of visualization of a lattice and good default layout, Galicia introduces the generic MultiFCA approach to deal with a set of contexts, ToscanaJ can visualize nested lattices and involves an editor of conceptual schemas on relational databases, FcaStone was primarily intended for file format conversion and other low level operations. Unfortunately, most useful tools for end-user (ConExp and ToscanaJ) did not have official updates from 2006.

The main problem of compared tools is low limits of size of interactively analyzed artifacts (for example, lattices with more than 8000 concepts can hardly be operated and visualized on modern hardware). This is mainly due to the use of Java and cross-platform GUI or different goals of developing. The current version of Cordiet-FCA can manipulate bigger lattices. After all planned optimizations we will present comparison of implementations of all basic algorithms in the form of compiled components and scripts (Cordiet-DMS platform has built-in tools for benchmarking).

5 Conclusion and future work

Cordiet-DMS is a powerful platform for developing applied software tools, for example, Cordiet-FCA for analyzing data with FCA. This analysis can give us insights into underlying conceptual structure of the data. For the dynamic text collections we can prepare several profiles and iteratively check the sequence of concept lattices.

We assume to improve methodology, extend the set of solvers, optimize some algorithms and use proposed system in different data mining tasks. Some of new solvers will be based on concept stability [5] and similarity [6] calculation algorithms. Also we will extend our platform with triadic concept analysis and noise-robust triclustering methods [7]. Also brand new lattice visualization technique is almost

done with antialiasing, scaling, iceberg concept lattices drawing and more. The next major release of the software (1.0) is planned for November 2012.

It's important to us to provide a freeware version of Cordiet-FCA, that can be extended by community and used in various application fields.

Acknowledgements

The results of the project "Mathematical Models, Algorithms, and Software Tools for Intelligent Analysis of Structural and Textual Data", carried out within the framework of the Basic Research Program at the National Research University Higher School of Economics in 2012, are presented in this work.

Jonas Poelmans is an aspirant at the Research Foundation Flanders.

References

1. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations, Springer, 1999.
2. Poelmans J., Elzinga P, Neznanov A., Viaene S., Kuznetsov S., Ignatov D., Dedene G. Concept Relation Discovery and Innovation Enabling Technology (CORDIET) // CEUR Workshop proceedings Vol-757, CDUD'11 – Concept Discovery in Unstructured Data, 2011.
3. Grange E. DelphiWebScript Project (http://delphitools.info/dwscript)
4. Apache Lucene (http://lucene.apache.org)
5. Kuznetsov S.O. On Stability of a Formal Concept // Annals of Mathematics and Artificial Intelligence, Vol. 49, pp.101-115, 2007.
6. Klimushkin M.A., Obiedkov S., Roth C. Approaches to the Selection of Relevant Concepts in the Case of Noisy Data // 8th International Conference on Formal Concept Analysis (ICFCA 2010),. pp. 255-266, 2010.
7. Ignatov D.I., Kuznetsov S.O., Magizov R.A., Zhukov L.E. From Triconcepts to Triclusters // Proceedings of 13th International Conference on rough sets, fuzzy sets, data mining and granular computing (RSFDGrC 2011), LNCS/LNAI Volume 6743/2011, Springer, pp. 257-264, 2011.