Cubix: A Visual Analytics Tool for Conceptual and Semantic Data

Cassio Melo*, Alexander Mikheev*, Bénédicte Le Grand †, Marie-Aude Aufaure*
*MAS - École Centrale Paris
Châtenay-Malabry, France
{cassio.melo, alexander.mikheev, marie-aude.aufaure}@ecp.fr
†LIP6-CNRS
Université Pierre et Marie Curie
Paris, France
benedicte.le-grand@lip6.fr

***Abstract*—This paper presents Cubix, a Formal Concept Analysis (FCA)-based analytics tool for Business Intelligence. The main purpose of Cubix is to provide novel ways of applying visual analytics in which meaningful diagrammatic representations will be used for manipulating, filtering and visually querying complex data. We present its main features, typical applications and future steps towards an advanced FCA-based visual analytics.**

I. Introduction

The advances in technology for creation, storage and dissemination of data have dramatically increased the need for tools that effectively provide users with means of identifying and understanding relevant information. Formal Concept Analysis (FCA) may play an important role and help fill this gap by employing more intelligent means in the analysis process. FCA became popular in the early 80's as a mathematical theory of data analysis based on the philosophical notion of a concept [1]. Since then it has been applied to a variety of domains, to name a few; information retrieval [2], [3]; gene expression [4]; and machine learning [5].

FCA provides an intuitive understanding of generalization and specialization relationships among objects and their attributes in a structure known as a concept lattice. A concept lattice is traditionally represented by a Hasse diagram illustrating the groupings of objects described by common attributes. This hierarchical structure can provide reasoning for classification and clustering, implication discovery and rule learning.

One critical issue of the traditional lattice visualization is that it grows exponentially with the number of objects and attributes [6], [7], [8]. On the other hand, the analysis process can be greatly enhanced with aid of visual analytics techniques. Today, only a few number of tools are able to deal with large lattice visualization and the support for interactive analysis is quite limited.

In this demo article we present Cubix, a HTML5-compliant visual analytics tool for Conceptual and Semantic data. The techniques implemented allow selecting, comparing, filtering, detailing and overview of concept lattice features. The novelty of our work consists of (i) the combination of FCA with visual analytics data exploration techniques, (ii) new algorithms for condensing and filtering conceptual data, and (iii) integration with semantic data from semantic repositories (triple stores).

Typical uses of Cubix include semantic data analysis and pattern detection, anomaly detection, comparisons, information classification, and knowledge discovery. Cubix's workflow allows users to carry out an analysis starting from a real data set, converting it into a formal context, simplifying the context to make it manageable, and visualizing the result as a concept lattice (Figure 1 - 2). Cubix is a step forward in relation to the current FCA tools for the number of techniques employed and its ability to deal with much larger data. The tool is currently being developed with the active involvement of our three users groups and their use cases (space control operations, gene expression analysis and company recruitment analysis).

In the rest of this demo paper, we briefly introduce some of the popular similar tools on FCA in Section 2. In Section 3, Cubix user interface and its functionalities are introduced. Some real use cases of Cubix are presented in Section 4. Finally, the conclusion and future work in Section 5 followed by a demonstration plan in Section 6.

II. Similar Tools

Over the past decade a number of FCA analysis and visualization tools have appeared. Their purpose is to generate the concept lattice of a given formal context and the corresponding association rules. ToscanaJ[1], Galicia[2] and Concept Explorer (ConExp)[3] are among the most popular ones. They can compute and visualize concept lattices but are not designed to do so for large numbers of concepts and the support for interactive analysis is limited.

Recently, OpenFCA[4], an open source FCA-based web application has drawn attention in the area for its ability to create formal contexts, mine and visualize concepts and explore association rules. It is one of the first tools with a highly interactive layout for concept analysis. There are a few

[1] toscanaj.sourceforge.net
[2] iro.umontreal.ca/~galicia
[3] www.source-forge.net/projects/conexp
[4] www.code.google.com/p/openfca

Fig. 1. Cubix user interface displaying the adult data set. Its main components: 1) Toolbar; 2) Visualisation canvas; 3) Dashboard; 4) Selection & entities bar and; 5) Filter bar.

limitations however, for instance, the only method for reducing large lattices is based on defining a maximum tree depth.

Specialized tools like FCA Bedrock[5] can scale data into formal contexts and has become a essential step in order to employ FCA in a Business Intelligence context. Facettice, a tool for visualizing and navigating in concept lattices demonstrated that visualization and interaction techniques can greatly enhance the understanding of conceptual data. However, the tool was not designed to support interactive analysis.

Cubix extends the features of existing tools, from data scaling to the visual exploration of the concept lattice and association rules.

III. System Overview

In the following sections we describe some of Cubix features and techniques employed for organizing, visualizing, interacting and filtering conceptual data. Cubix is part of the CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) project[6].

A. Workspaces and Data Sources

Cubix allows grouping of different data such as files, databases and a data streaming API, on a single workspace to perform analysis. Each workspace provides options for filtering and merging data from different sources.

B. Data Scaling and Concept Mining

In FCA, scaling of data refers to the process of discretizing data in order to generate formal concepts. Cubix automatically identifies categorical, location and date/time fields, and it allows the filtering of attributes/objects and selection of sub-contexts. In some cases it is possible to specify the granularity level of attributes. For instance, in an attribute called "date" one can specify time periods to be analyzed as groups (e.g. weeks, months or years). This way, users can focus their analysis to pertinent attributes. A formal context is then generated from the previous selection and processed by the In-Close2 algorithm for generating formal concepts [9].

C. Concept Clustering

The number of concepts can grow exponentially with the number of objects and attributes, yielding poorly readable concept lattices [8]. Clustering of concepts can be useful to facilitate the analysis and to identify zones of interest. Cubix uses a K-means clustering algorithm to identify clusters. Some similarity measures are based on the concept lattice topology (e.g. counting the number of links between two concepts); Intent/extent similarity (e.g. Jaccard); or confidence between two pairs of concepts. Given a (formal) context $K = (G, M, I)$, where G is called a set of objects or $extent$, M is called a set of attributes or $intent$, and the binary relation $I \subseteq G \times M$ specifies which objects have which attributes, the $derivation\ operators$ $(\cdot)\prime$ are defined for $A \subseteq G$ and $B \subseteq M$:

$$\begin{aligned} A\prime &= \{m \in M | \forall g \in A\colon gIm\} \\ B\prime &= \{g \in G | \forall m \in B\colon gIm\} \end{aligned} \tag{1}$$

A formal concept is a closed set of object/attribute relations when $A\prime = B$ and $B\prime = A$. In other words, its extent contains

[5]www.source-forge.net/projects/fcabedrock
[6]www.cubist-project.eu

all objects that have the attributes in its intent, and its intent contains all attributes shared by all objects in its extent. (for details on the FCA terminology see [1]).

Concept similarity (Jaccard). It is a coefficient for calculating the ratio of shared attributes between concepts. We define concept similarity as:

$$CSim(A,B) = \frac{|m_a \cap m_b|}{|m_a| + |m_b|} + \frac{|g_a \cap g_b|}{|g_a| + |g_b|} \quad (2)$$

Proximity. Conceptual proximity is the topological distance between concepts A and B in the concept lattice.

$$prox(A,B) = 1 - \frac{shortestDistance(A,B)}{diameter(Lattice)} \quad (3)$$

Strength. It is the average concept similarity value (CSim) along the shortest path between a pair of concepts.

D. Transformations

This option allows users to transform the concept lattice in other structures like a tree or bi-partite graphs that are easier to navigate than lattices. Trees are common and have easily understandable visual representations. We consider them as a visualization alternative to large cluttered concept lattices, which preserves all lattice entities and some of its structure. In order for a tree visualization to be an effective alternative to a lattice, the extraction of the tree from the lattice needs to preserve the most essential features of the original structure. We consider various strategies for selecting parent concepts, including the stability and support indexes [8], [10], confidence, as well as topological features of the lattice (see [11] for a detailed explanation on the tree transformation process).

E. Concept Lattice Analytics

Traditional software in FCA makes little use of visualization techniques, producing poorly readable lattice graphs when the number of concepts exceeds a few dozen. To reduce the complexity of lattices, simplified diagrams can be produced by condensing or clustering concepts according to their intent frequency [8]. Visualisations can also be restricted to portions of the data [12], and concept size reduction is possible by incorporating conditions into the data mining process [13]. Finally, conceptual measures can be applied to identify the most relevant concepts and filter outliers [14].

Cubix combines the state-of-art of visualization techniques with data mining, allowing users to interactively analyse the conceptual data, by filtering and selecting, transforming and clustering concepts. Figure 1 - 2 displays the selected visualization for the concept lattice. Other visualization options are: hasse diagram, matrix, radial-filling (sunburst), icicle and tree.

A filter bar (Figure 1 - 5) has two functions, first it allows the filtering of concepts through the visual selection of attributes; second, it displays the current conceptual distribution for each attribute. The charts are automatically adapted for the current data, for instance, time related values are displayed in a timeline.

It is possible to perform text searches of attributes in concepts, with an auto-completion feature to help users easily search for both attributes and values in concepts. The results are dynamically highlighted as the user searches in the different concept nodes.

Based on recommendations from our users, we use the notions of support, stability and association as filters as well, since these strategies help our users answer questions of the form most frequent pattern. But we also represent these notions visually on the lattice to enhance data understanding. For instance, the power of implications of different concepts can be rendered by edge thickness. The concept node itself can be a visual metaphor for its intent and extent, depicting for instance, the ratio between the number of attributes and objects. In this way users can be guided in understanding and choosing criteria for reducing their lattices. Using these criteria we can also extract a tree structure to simplify the lattice representation [11].

F. Association Rule Analytics

Another feature of Cubix is a dashboard dedicated to association rules mining and analysis. Association Rules are of the form premise $\Longrightarrow$ conclusion: $m_1 AND m_2 AND ... m_n \Longrightarrow n_1 AND n_2 AND ... n_n$ for $m, n \in M$ and carry very little information about how they can be visualized. They are typically displayed as a list of logical sentences, unpractical to analyse when the number of rules is large.

Cubix provides three new visualizations combined with statistics and charts to enable progressive exploration of the set of rules. The visualizations are: A traditional matrix view where each rule is displayed in a row and the concerned pairs of attribute-value in columns. The confidence of each rule can be measured by the opacity of each cell. The second visualization is a radial graph showing how pairs attribute-value relate to each other. The confidence of a rule is represented by the thickness of the connecting line. Finally, a bubble graph visualization displays premises and conclusions as connected bubbles with the concerned attribute-value pairs inside each bubble. Users can navigate in the bubbles using a panning/zooming control.

Similar to the concept lattice analytics, all interface controls work analogously with association rules. The semantic dashboard adapts to the analysis of association rules, for instance, one chart displays support, stability and confidence on a scatterplot view. If an item from the chart is selected it filters the association rules to be displayed.

G. Semantic dashboard

In addition to the main concept lattice visualization, several charts display different aspects of the underlying conceptual structure such as co-occurrence of attributes, concepts distribution, stability vs. support, etc. (Figure 1 - 3). Some charts are updated when the user points the mouse over a concept, highlighting details of the concept. Similarly, a selection of a point/series in the chart will highlight the concerned concepts in the lattice. This technique is called *Linking and Brushing*.

H. Ontology integration

Cubix can be integrated with a RDF/OWL triple store which enables the conceptual analysis on top of SPARQL queries. The conceptual analysis of ontologies provides unique information on the semantic data, by grouping entities belonging to particular properties in a hierarchical fashion. The process starts when the user enters a SPARQL query, which is then sent to the triple store via Cubix. The results are automatically scaled to a formal context, formal concepts are then mined and visualized in the analytics view. Currently, Sesame and Owlim triple store are supported in Cubix.

IV. Cubix Use Cases

The features implemented in Cubix are results of a two-year project with active involvement of end users. As a first step we conducted interviews with three types of use cases attempting to find patterns within their data: a company conducting market intelligence, computational biology, and a space control centre operation monitoring. All users had in common large amounts of data of which they wanted to answer mainly the following types of questions: frequent pattern detection, anomaly detection and pattern comparison. An example of frequent pattern detection is "during the first stage of a mouse embryo development what are the genes expressed together most often?". An example of an anomaly detection is "what are the sensor and telemetry logs of a space load on the International Space Station that may be related to a specific instrument malfunction?". Finally a pattern comparison question would be "Are the jobs available in Liverpool similar to those in Manchester?" FCA allowed them to create semantic groupings of objects and attributes based on their co-occurrence and progressively explore the lattice to look for answers to their questions.

V. Conclusions and Future Work

Traditional methods for visualization of large FCA lattices are restricted to Hasse diagrams. Using a user-centered approach we developped Cubix, a tool that extends traditional FCA techniques using interactive visualization and data mining techniques for conceptual data.

Cubix is an attempt to bring conceptual analysis to Business Intelligence. It employs several alternatives for the visual exploration of lattices through searching, filtering and subselection of concepts and attributes; clustering and transforming concepts; visual display related attributes and the implications and; integration with semantic triple stores.

In the future we plan to explore other visual metaphors and more sophisticated navigation and interaction techniques for dealing with very large lattices (up to 10k concepts), and ways to navigate and zoom into different levels of concepts clusters. We also plan to release the tool under an open source license.

VI. Demonstration Plan

We will suggest to the audience three popular data sets and a corresponding list of analysis tasks for each of them: Traffic accidents in the UK, Wine ontology, and the Adult census dataset. For example, the first dataset contains traffic accident information which occurred in the UK in 2006. Participants will be asked to run an analysis to discover hidden patterns between different combinations of attributes (e.g. road surface, weather, accident severity, etc.). Examples of tasks are "What can be considered the main causes of 'accident severity: serious?" or "how many accidents have 'light conditions: darkness' and 'accident severity: serious?".

Next in the demonstration, we will show how to use Cubix to analyse the data by selecting and filtering, choosing the best visualization, etc. Multiple computers and tablets can access Cubix at the same time. To motivate participants to engage, we will track the time for each person to complete all tasks for a given dataset, and challenge them to outperform their peers. The participants can also run some analysis by using data of their own choice.

Acknowledgment

We would like to acknowledge the CUBIST project (Combining and Uniting Business Intelligence with Semantic Technologies), funded by the European Commission 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.

References

[1] B. Ganter and R. Wille, *Formal Concept Analysis*, mathematical foundations ed. Springer, 1999.

[2] C. Carpineto and G. Romano, "Using concept lattices for text retrieval and mining," in *Formal Concept Analysis*, 2005, pp. 161–179.

[3] K. S. Cheung and D. Vogel, "Complexity reduction in lattice-based information retrieval," *Inf. Retr.*, vol. 8, no. 2, pp. 285–299, apr 2005.

[4] E. Akand, M. Bain, and M. Temple, "A visual analytics approach to augmenting formal concepts with relational background knowledge in a biological domain," in *Sixth Australasian Ontology Workshop*, T. Meyer, M. Orgun, and K. Taylor, Eds., dec 2010.

[5] S. Kuznetsov, "Machine learning and formal concept analysis," in *Concept Lattices*, ser. Lecture Notes in Computer Science, P. Eklund, Ed. Springer Berlin / Heidelberg, 2004, vol. 2961, pp. 3901–3901.

[6] C. Carpineto and G. Romano, "Ulysses: a lattice-based multiple interaction strategy retrieval interface," in *Human-Computer Interaction*. LNCS 1015-Springer, 1995, pp. 91–104.

[7] U. Priss, "Lattice-based information retrieval," *KNOWLEDGE ORGANIZATION*, vol. 27, pp. 132–142, 2000.

[8] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal, "Computing iceberg concept lattices with titanic," *Data & Knowledge Engineering*, vol. 42, pp. 189–222, August 2002.

[9] S. Andrews, "In-close, a fast algorithm for computing formal concepts," in *the 7th International Conference on Conceptual Structures*, 2009.

[10] S. O. Kuznetsov, "On stability of a formal concept," *Annals of Mathematics and Artificial Intelligence*, vol. 49, no. 1-4, pp. 101–115, Apr. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10472-007-9053-6

[11] C. Melo, B. Le-Grand, M. Aufaure, and A. Bezerianos, "Extracting and visualising tree-like structures from concept lattices," in *15th Int. Conference on Information Visualisation (IV)*, july 2011, pp. 261 –266.

[12] Ducrou, P. J., Eklund, and T. Wilson, "An intelligent user interface for browsing and searching mpeg-7 images using concept lattices," in *CLA 2006*, vol. 4923. Springer-Verlag Berlin Heidelberg, 2008, pp. 1–21.

[13] M. Zaki and C.-J. Hsiao, "Efficient algorithms for mining closed itemsets and their lattice structure," in *IEEE Transactions on Knowledge and Data Mining*, vol. 17, no. 4. IEE Computer Soc., 2005.

[14] M.-A. A. Michel Soto, Benedicte Le Grand, "Spatial visualisation of conceptual data," *International Conference Information Visualisation*, pp. 57–61, 2009.