International Journal of General Systems
Vol. 00, No. 00, Month 200x, 1–21

RESEARCH ARTICLE

Using concept lattices for visual navigation assistance in large databases

Jean Villerd[a*], Sylvie Ranwez[a], Michel Crampes[a] and David Carteret[b]

[a]*LGI2P – École des Mines d'Alès, Parc Georges Besse, Nîmes, France*; [b]*I-Nova, 11 avenue Albert Einstein, Villeurbanne, France*

(*Received 00 Month 200x; final version received 00 Month 200x*)

The increasing size of indexed document sets that are digitally available emphasizes the crucial need for more suitable representation tools than traditional textual lists of results. Many efforts have been made to develop graphical tools capable of providing both overall and local views of a collection when focusing on a particular subset of documents. However an overcrowded visual representation may not be useful to users if they are not guided in the navigation process. Our goal is to combine the classification features of FCA and existing visualisation techniques to suggest navigation paths in a visual representation through meaningful and progressive foci. While results are promising, in particular concerning interactions between local and overall views, some difficulties arise concerning non-binary attributes. Therefore we propose embedded Multidimensional scaling (MDS) projection that offers various abstraction layers. The test collection used for our study is an indexed patent database provided by our industrial partner.

Keywords: Formal concept analysis; Semantic navigation; Multidimensional scaling

1. Introduction

The size of digitally available indexed document sets increases every day. However, associated exploring tools are often based on the same traditional model: users send their query and are then answered with huge lists of results. There is a crucial need for suitable representation tools where the semantics of the documents are better exploited and may be used as guidelines for navigating through the database. Formal concept analysis (FCA) helps to crystallise conceptual structures from data. Such structures may be used to visualise inherent properties in data sets and to dynamically explore a collection of documents. Indeed, the associated mathematical formalisation of FCA is not only useful to organize the database but also draw inferences during the information retrieval process. This paper presents a method that combines FCA and information visualisation techniques to assist visual navigation in large collections. This research is done in response to industry demand: one of our partners is facing the problem of visualising and browsing a large collection of indexed patent data. We therefore apply our approach in their domain.

The following section presents the context of our research. The state of the art is presented in Section 4, particularly concerning information retrieval using FCA and

*Corresponding author. Email: jean.villerd@ema.fr

ISSN: 0308-1079 print/ISSN 1563-5104 online

DOI: 10.1080/0308107YYxxxxxxxx
http://www.informaworld.com

visualisation techniques to explore large databases. Section 5 develops our method that combines FCA and visualisation tools to assist users in browsing a large tagged collection. Section 6 deals with the application of this method on a real database provided by our industry partner and we comment on the results. The specific case of non-binary tags is analysed in Section 7. Embedded Multidimensional scaling (MDS) projections are described that offer conceptual abstraction layers during the navigation or the indexing process. The last section concludes with some of the limitations of our approach.

2. Context and problem setting

Searching for technical solutions to improve innovation within big companies, I-Nova, our industry partner, develops collaborative platforms to internally share some parts of the company's knowledge. This may be sets of ideas, patents, laws, policies, etc. The efficiency of this sharing relies on the participation of every type of employee. Therefore the sharing interface must be intuitive enough to favor the participation of people not familiar with information retrieval tools. The difficulty is to visualise and browse a large collection of indexed documents. The current platform makes use of a classical search engine which lists retrieved documents (patent records in this case) corresponding to a set of keywords. Two main problems arise: human operators cannot get an overall view of the entire collection and they cannot easily evaluate changes in the result sets when adding or removing a keyword, because a new list is displayed for each new query. These drawbacks are particularly noisy for the browsing process. We may note that this problem is likewise encountered using Web search engines like Google.

Much reasearch has been done to graphically represent indexed document sets in general (Tricot *et al.* 2006), of which several aim to represent patent databases. In MultiSOM (Lamirel and Al Shehabi 2006), keywords are divided into subsets corresponding to different aspects of the indexation (costs, techniques, etc.) and a self-organized map is computed for each subset, presenting different points of view on the database. When focusing on a specific item of the collection on one particular map, the user can switch to another aspect. This solution solves the problem of providing an overall view of the collection. However, users have lost the ability of selecting a subset of patent records through a set of keywords and they have no information about why one patent record is closer to another in the overall view.

Because on the one hand browsing an overall view may be unsuitable for focusing on a particular keyword, and on the other hand, displaying local results without any overall information causes users to loose their orientation in the information space, we present in this paper a method that assists user navigation from overall to local views. The idea is to "sum up" the collection by coherent local views corresponding to subsets of patent records/keywords. These views are ordered as a lattice, defining possible navigation paths that will be suggested to users.

The lattice used is actually a concept lattice (Ganter and Wille 1999). Using Formal concept analysis (FCA) and concept lattices for information retrieval is not new (Ducrou and Eklund 2008, Eklund *et al.* 2008) but in these approaches, content of nodes, i.e. local views, are still displayed as a list of results, retaining all the above mentioned problems. Studies on visualisation exploration processes have been done but mainly in the medical or scientific imagery domain, focusing more on optical transitions between visualisations (Jankun-Kelly *et al.* 2002) than on semantic aspects of the information that is displayed. Before going further towards the solution that we propose, let use provide some basic FCA definitions.

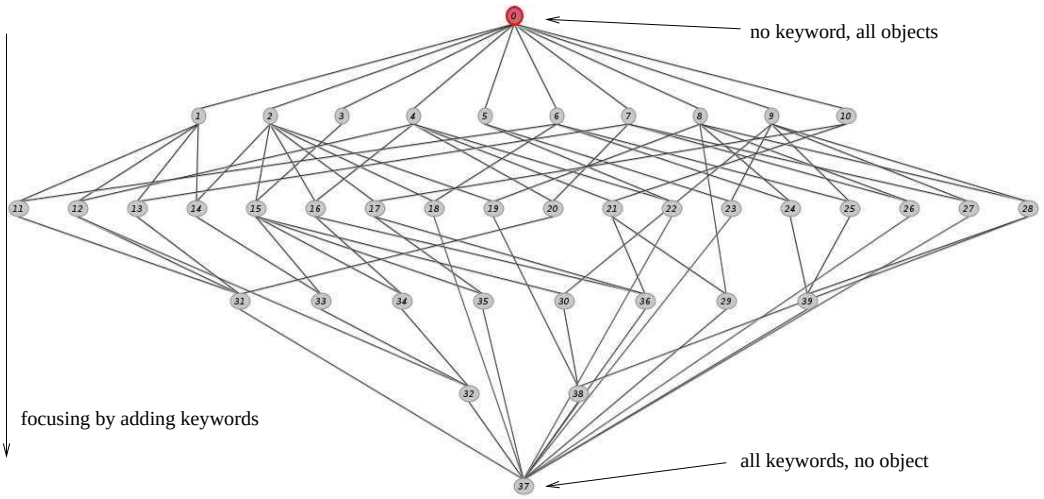


Figure 1. The concept lattice computed using Galicia from the patent test database (329 objects, 10 attributes).

3. Formal concept analysis background

In this section, we briefly recall FCA basic definitions from Ganter and Wille (1999). A *formal concept* is a triple $\mathbf{K} := (O, A, I)$ where O is a set of objects, A a set of attributes and I is a binary relation between the objects and the attributes, i.e. $I \subseteq O \times A$.
For a set $O_i \subseteq O$ of objects and a set $A_j \subseteq A$ of attributes, we define the set of attributes common to the objects in O_i by

$$f : 2^O \to 2^A \qquad f(O_i) = \{a \in A \mid \forall\, o \in O_i, (o, a) \in I\}$$

and the set of objects which have all attributes in A_j by

$$g : 2^A \to 2^O \qquad g(A_j) = \{o \in O \mid \forall\, a \in A_j, (o, a) \in I\}$$

The pair (f, g) is a Galois connection between $(2^O, \subseteq)$ and $(2^A, \subseteq)$.
A *formal concept* of the context (O, A, I) is a pair (O_i, A_i) with $O_i \subseteq O$, $A_i \subseteq A$, $A_i = f(O_i)$, and $O_i = g(A_i)$. A_i is called the *intent* and O_i the *extent* of the concept (O_i, A_i).
Let L bet the set of concepts of (O, A, I) and let $\leq_L$ be a partial order defined as follows, for $(O_i, A_i) \in L$, $(O_j, A_j) \in L$,

$$(O_i, A_i) \leq_L (O_j, A_j) \iff O_i \subseteq O_j \iff A_i \supseteq A_j$$

The pair $(L, \leq_L)$ is called the *Galois lattice* or *concept lattice* of (O, A, I). The *simplified extent* of a concept (O_i, A_i) is the set of objects which belong to O_i and do not belong to any lower level concept. In other words, the simplified extent denotes objects that do not have any other attributes than those in A_i.
In the following we denote objects as documents or patent records and attributes as terms or keywords. The concept lattice computed from the patent records/keywords matrix of our test database is shown in Figure 1.

4. State of the art

Searching for solutions to assist navigation through a large database, we focus particularly on two aspects in the following state-of-the-art survey: applications that use FCA techniques for information retrieval and visualisation techniques that may be used to graphically parse large sets of data.

4.1 *Formal concept analysis for information retrieval*

The powerful classification skills of FCA have found many applications in information retrieval. Some of them have been listed by Priss (2006). Since the early works of Godin *et al.* (1989) on an information retrieval system based on document/term lattices, a lot of research leading to significant results has been done. Carpineto and Romano (2004) argue that, in addition to their classification behaviors for information retrieval tasks, concept lattices can also support an integration of querying and browsing by allowing users to navigate into search results. Nowadays, several FCA-based applications like Credo (Carpineto and Romano 2004) or MailSleuth (Cole *et al.* 2000, 2003) are available. MailSleuth is an e-mail management system providing classification and query tools based on FCA. This tool allows users to navigate into data and intervene in the term classification by displaying concept lattices. Upstream research has studied the understandability of a lattice representation by novice users (Eklund *et al.* 2004). ImageSleuth (Ducrou and Eklund 2008, Eklund *et al.* 2008) proposes an interactive FCA-based image retrieval system in which subjacent lattices are hidden. Users do not interact with an explicit representation of a lattice. They navigate from one concept to another by adding or removing terms suggested by the system. This ensures progressive navigation into the lattice.

4.2 *Visualisation techniques to browse large databases*

Graphical solutions for visualising a mass of abstract information have been studied for several decades, leading to the emergence of the information visualisation domain (Card *et al.* 1999). Even with the use of a visual representation, the navigation into a large collection may not be obvious. Shneiderman (1996) has defined a visualisation information paradigm called "focus + context" that recommends to first provide an overall view, then to let the user identify an area of interest (focus), and finally to display locally contextual information (context). Starting from an overall view helps users to maintain a unique mental map but they still have to achieve the focus task on their own.

In the particular case of document visualisation, Tricot *et al.* (2006) defines two categories of solutions: on the one hand visualisations of the inner structure of a document, e.g. WebBook (Card *et al.* 2004), and on the other hand, visualisations of a document collection, e.g. DocCube (see Mothe *et al.* 2003). The work presented in this paper belongs to this second category which can also be divided into two parts depending on how the structure of the collection is managed. Some tools show this structure by representing clusters of documents, e.g. Grokker[1], using tiling based visualisation techniques such as TreeMaps (Shneiderman 1992). Other tools do not show the collection's structure but represent the collection by a set of points into two or three dimensions dispatched according to a semantic distance usually based on indexed vectors (e.g. DocCube).

[1] http://www.grokker.com

The choice between these two strategies depends on the user needs. Showing the structure allows progressive navigation through clusters but reduces the probability of visual insight which the obervation of similarity distances may offer. We have tried to benefit from both solutions by using the first for the overall view and the second for local views (see Section 5.3).

4.3 *Visualisation paths*

The visualisation exploration process has been studied for several decades. Upson *et al.* (1989) and Card *et al.* (1999) describe the visualisation process as an analysis cycle, starting from raw data transformation, mapping onto visual primitives, and view rendering provided to the user who performs feedback that restarts the cycle. The central idea of their model is that a description of the visualisation process has to focus on this interaction with the user who changes the visualisation parameters. The user influences all steps of the process by adjusting these parameters: from data filtering and transformation (by specifying a data threshold value), mapping onto visual primitives (by specifying another color or shape) and view rendering (by changing the orientation of the view). Upson and Card's models provide a characterisation of visualisation exploration as a general process but their granularity is not deep enough to detail and define a particular user's exploration path. Following their work, two approaches have arisen: visualisation space paths and derivation models.

Visualisation space paths (Tory *et al.* 2005) consider visualisation exploration as a navigation path in a multidimensional parameter space, a visualisation result being a single point in the parameter space. Based on this model, novel visualisation user interfaces such as Design Galleries (Marks *et al.* 1997) display an overview of the entire parameter space by random sampling. The main interest of this model is that it provides explicit relations between visualisation results and visualisation parameters. However, particular relations between results are not explicit. Derivation models have been introduced to overcome this problem.

Derivation models describe how new visualisation results are created from previous ones. Originally developed to provide visual assistance to users in scientific problem solving environments, the Gasparc system (Brodlie *et al.* 1993) builds a history tree of visualisation results in order to store solution parameters and related results. In Lee (1998) and Lee and Grinstein (1995), a graph structure is used to model the visualisation process for databases. Vertices stand for visualisation states and edges are based on similarities between structural attributes of the states. Jankun-Kelly *et al.* (2002) extend these models to be more general and extensible.

Due to the historic development of visualisation applications, the majority of these works concentrate on scientific visualisation (problem solving, medical imagery), rather than information visualisation. Unfortunately, the semantics of the visualised data, which is what we are interested in, are in this case not taken into account in the cycle of the visualisation exploration process.

4.4 *Knowledge maps and multidimensional scaling*

The tests and validation of our approach are performed through an agent oriented software environment that we developed. Molage (for Molecular Agents) allows visual manipulation of entities using several visual functionalities such as zoom, fisheye, semantic lenses, filtering, etc. Each entity may be typed and described by several attributes or descriptors. The collection of those entities may represent a multimedia database, e.g. music titles described by moods, textual documents

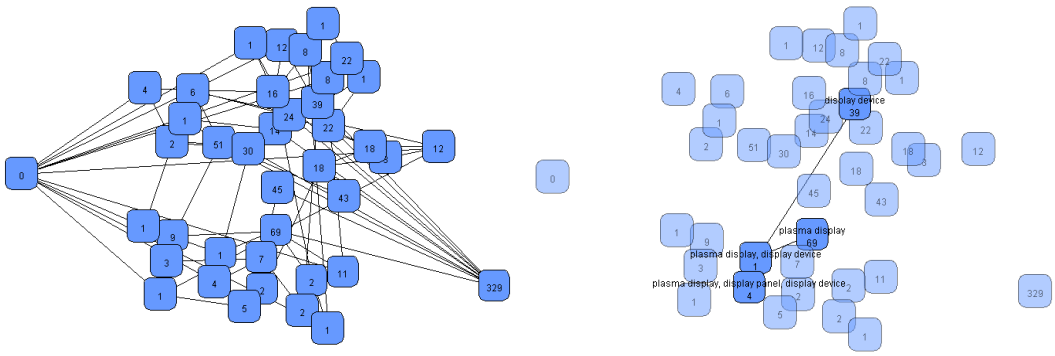
characterised by a keywords' vector, pictures that are classified according to their subject, etc.
Considering the set of descriptors, each entity is characterised in a m-dimensional space. The collection is projected onto a plane using a Multidimensional scaling (MDS) technique (see Kruskal and Wish 1978, Chalmers 1996). The method consists of minimizing a "stress" function between n points (originally described in the m-dimensional space) after those points have been projected onto a space with fewer than m dimensions. The minimization is performed through the compression or extension of the Euclidian distances between the points. This type of MDS approach is known as "Force directed placement" or "Spring embedding algorithm" (named after the work of Eades (1984)).
We detailed the Molage environment and its use for navigation through a music collection in Crampes *et al.* (2007). In that application, the musical landscape is used to semi-automatically index new music records by "drag and drop" of a new entity onto the map. It is also used to automatically build musical playlists. However some drawbacks were still present, in particular the lack of assistance for visual navigation through the map. For this reason, we proposed to give semantics to the map and to assist the user in construction of the map in Crampes *et al.* (2006). In the following, we broaden this visual assistance by proposing a method for browsing large databases.

5. A visual database browsing method

Within large databases, it is often possible to distinguish several layers that may be considered separately. Our idea is to keep the human operator's mental map as stable as possible and to give him or her some hints when he or she switches from one view to another one.

5.1 *Problem decomposition*

Our approach aims to provide adaptive visualisations for different abstraction layers. Information visualisation techniques aim to assist the user in the visual inference task. It is very hard to reach a compromise between a strong analytic strategy displaying results of classification methods on data, and a visualisation representing faithfully all raw data. The first approach may be clearer for users but may introduce a bias in their interpretation and preclude insight by hiding unexpected information. The second approach faces the dimensionality problem of visualising too much information simultaneously. Our work proposes to satisfy both approaches by finding a suitable compomise.
We assume that Shneiderman's overview, i.e. an overall view corresponding to the higher abstraction layer, has to emphasize the structure of the document collection because that is where the source of insight may reside at this level. This overview should be used as a kind of GPS for further navigation into different local views. For the lower abstraction level, i.e. Shneiderman's context views, the aim is to display concrete relations between local data using the most intuitive visualisation techniques. In the following we will describe the overall view and then local views but note that we propose a user interface that will show simultaneously the overall view and one particular local view (see Figure 3).



(a) (b)

Figure 2. The overall view is a MDS projection of the lattice concepts. All edges shown (a). A particular concept {*plasma display, display device*} and its neighbors emphasized in (b).

5.2 *Lattice-based overview*

The concept lattice computed from the document/term (patent record/keyword) matrix will be used both as support for navigation across the collection (as used in ImageSleuth (Ducrou and Eklund 2008, Eklund *et al.* 2008)), and as a visual overview emphasizing its structure.

Rather than using a traditional lattice representation such as line diagrams (see Figure 1) and graph drawing techniques (see Di Battista *et al.* 1994), lattices nodes are displayed using force directed placement techniques (see Section 4.4) according to a measure of semantic distance described in Ranwez *et al.* (2006).
Not all edges are represented as in Figure 2(a) to avoid visual overloading. We use a visual device called "topological lenses" (see Crampes *et al.* 2007) to show edges of the selected node and to emphasize nodes reached by these edges. In Figure 2(b), the selected node intent is {*plasma display*, *display device*} and its simplified extent cardinality is 1 (this cardinality indicates how many documents belong to this concept and not at any lower). Concepts that are linked and their intents are emphasized. Moving the cursor over a concept shows its intent. This visualisation allows users to browse nodes via their lattice order while evaluating the semantic distance covered at each step thanks to their semantic distance.
Several strategies could be applied in order to reduce the number of nodes to avoid visual overloading. Iceberg lattices (Stumme *et al.* 2002) and alpha lattices (Pernelle *et al.* 2003) consist of smaller concept lattices retaining significant concepts that respect a threshold criterion based on the extent cardinality. Considering the concept lattice as an overview of the collection, these lattice reduction techniques based on the frequency of the concepts are compatible with our approach. Another lattice reduction strategy would consist of decomposing the overall lattice into smaller sublattices, as done in ImageSleuth (Ducrou and Eklund 2008, Eklund *et al.* 2008). A partition of term space would be achieved by a domain expert or by identification of clusters of concepts with respect to the semantic distance.
Since an overview supporting navigation is already provided to users, we will present how a local view associated with a lattice concept is visualised and then how users interact with local and overall views.
As presented in Section 5.1 local views should represent relations and similarity between raw data.
A first approach consists of computing a distance between documents once and for all. Much research has been done concerning semantic distance computation

between textual documents, these are outside the scope of the present paper (see Ranwez *et al.* 2006). However, we assume that a distance is available between documents.

This distance may be independent from the information used as attributes in the formal context. Indeed it can be interesting to compare two indexation layers. For instance, in the case of multimedia documents, textual tags such as musical moods could be used to structure the collection and build the concept lattice whereas physical descriptors from signal processing analysis could be used to compute a similarity measure between documents, assuming that a search pattern consists of using high abstraction layer features to achieve Shneiderman's focus task, and then using low abstraction layer features, closer to raw data, to observe local organization of information. Proposals for handling numerical attributes are discussed in Section 7.

Using the same distance to represent all local views ensures the preservation of the users' mental map. Actually, moving from one lattice concept to another consists of making some documents visible and others invisible, so remaining documents' positions should not change even if the heuristic nature of our force directed placement approach may affect stability. In any event, Molage is able to fix positions for a subset of objects (here, the remaining documents), letting others (here, the newly retrieved documents) find a position with respect to semantic distance.

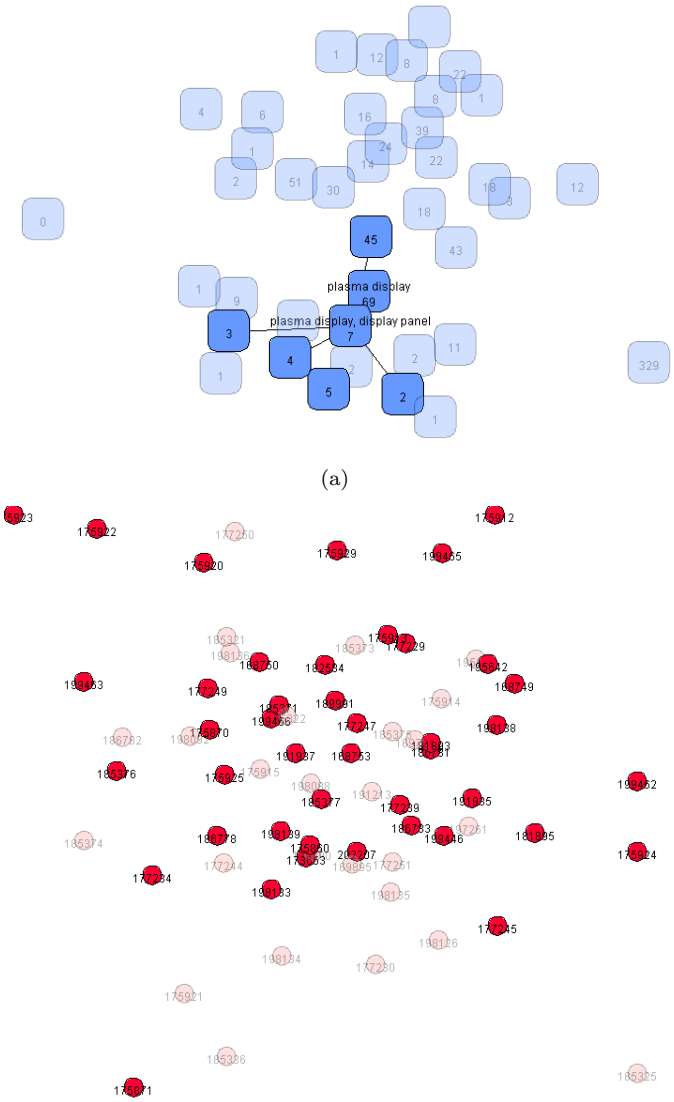
Another approach consists of computing a particular distance for each local view, i.e. for each lattice concept. Assuming that each term is associated with a numerical function, each document has a numerical vector on which a MDS projection can be performed, taking into account rows corresponding to terms present in the lattice concept intent. Moving from one lattice concept to another also consists of making some documents visible or not, however the positions of all documents change, because a new distance is computed by adding or removing vector rows in the MDS calculation. Even if this approach provides a more precise distance adapted to a local concept intent, its major drawback is that the users' mental map can become less accurate. However, Molage can dynamically perform such selective MDS projection (see Section 7.4), allowing users to observe objects' movements when updating the distance.

5.3 *Interactions between local and overall views*

Several use cases have been identified. The most basic consists of selecting the top node of the lattice in order to display the whole collection as a local view.

Thereafter two possibilities occur: users can continue their navigation across the lattice by choosing one of the top nodes that has been emphasized. When users move the mouse onto a particular node, documents that belong to this node are emphasized in order to preview the next visualisation and to observe the distribution of these documents in the current view (see Figure 3). When users select a new node, the local view is updated to remove items that do not appear in the new node extent. In this case, distances between documents in the different views are not recalculated and the map does not change; only the visible layer changes.

Another possibility is to select a document on the local view. The local view is then updated to represent documents belonging to the lowest level concept, i.e. the more precise concept, where the selected document appears. This concept is emphasized on the overall view, so that users see how far they have moved from the start node.



(a)

(b)

Figure 3. Overall (a) and local views (b). On the overall view, the node labelled {*plasma display*} has been selected. This node represents the formal concept containing only the attribute {*plasma display*} in its intent. The extent contains all objects tagged with this attributes. These objects are displayed in the local view (b). Their proximities reflect their semantic similarities. A Multidimensional scaling technique is performed in order to project the objects on the plane with respect to the given semantic distance. On the overall view (a), a child node of the selected node is labelled {*plasma display, display panel*}. This node represents a formal concept that contains one additional attribute in its intent, compared to the selected node. Consequently, its extent contains fewer objects than the extent of the selected node. This is shown on the local view (b), where objects that belong to the node {*plasma display*} but not to {*plasma display, display panel*} are faded. Through this mechanism, the user navigates in a coherent (by adding or removing keywords) and progressive way (by observing transitions between navigation steps).

6. Results

Our navigation method has been tested on real data, a collection of indexed patent records, provided by our industry partner with a similarity matrix between patent records. In order to reduce the size of the experimental collection, we extracted 329 patent records sharing the term *plasma*. These patent records own an average of 0.8 terms in addition to *plasma*. In other words, each patent record owns the term *plasma* and some of them own one or more additional terms. Ten different additional terms have been identified, that constitute the attribute set of the formal

context. The resulting lattice, computed by Galicia (Valtchev *et al.* 2003) using Bordat's algorithm, has 40 concepts. The top node extent contains patent records that do not own any supplementary attribute. We submitted several screenshots to our industry partner corresponding to different steps in a particular navigation use case. He was asked to compare the usability and user-friendliness of our method to their traditional list-based interface. Concerning the local view, he pointed out the usefulness of the edges: "*Edges between concepts and cardinalities of extents give us new information, make the navigation fun, and allow us to identify the coherence of the documents related to one concept*". Concerning the overall view, he appreciated the spatialization of the documents and "*the fact that one may see at first sight how many documents are displayed in a more direct and explicit manner compared to the display of the number of results above the list of results*". Concerning the navigation method itself, he agrees that "*it maintains a kind of mental map. That was the problem pointed out by users. The database analysis consists of two steps: an exploration step to discover and identify the content of documents, and then a rationalization of these discoveries. The problem is that these two steps have to be done at the same time and with a list-based interface, and consequently one may forget the different discovery paths used from the beginning. The visual aspect of [our] method will certainly facilitate tthe rationalization and the ability to save pertinent groups of documents*".
He also pointed out some drawbacks and suggested improvements: "*more information about documents should be displayed in the local view*", "*it would be interesting to fix a local view and, when selecting a new concept, to color additional documents on the local view, and to differently color documents in the intersection*". This last remark is particularly important because it matches our perspectives: to go further than navigation and to provide visual tools to assist data analysis process.
The results have been obtained using binary attributes to compute the overall view and a single given distance between objets. This article is an extended version of Villerd *et al.* (2007). As an extension of our previous work, and since both binary and non-binary attributes are usually available in databases, the next section will deal with handling non-binary attributes, which were out of the scope of the original article.

7. Handling non-binary attributes

In order to handle non-binary attributes, Ganter and Wille (1999) introduced the conceptual scaling mechanism, which consists of discretizing each non-binary attribute into a set of binary attributes using a value scale.

7.1 *Conceptual scaling*

In this subsection we first recall formal definitions of many-valued contexts and conceptual scales, and then present and discuss nested-line diagrams, a variant of line diagrams.
A *many-valued context* is a tuple $(O, A, (V_a)_{a\in A}, I)$ where V_a is a set of values for each attribute $a \in A$, and $I \subseteq O \times \bigcup_{a\in A}(\{a\} \times V_a)$ a relation such that, if $(o, a, v_1) \in I$ and $(o, a, v_2) \in I$, then $v_1 = v_2$.
A *conceptual scale* for an attribute $a \in A$ is a binary context $\mathbf{S}_a := (O_a, A_a, I_a)$ with $V_a \subseteq O_a$.

Table 1. Raw data.

	gender	age
Adam	M	21
Betty	F	50
Chris	?	66
Dora	F	88
Eva	F	17
Fred	M	?
George	M	90
Harry	M	50

Table 2. Conceptual scale for attribute *age*.

	< 18	< 40	≤ 65	> 65	≥ 80
17	×	×	×		
21		×	×		
50			×		
66				×	
88				×	×
90				×	×
?					

The binary context $\mathbf{R}_a := (O, A_a, J_a)$ with

$$(o, b) \in J_a \Longleftrightarrow \exists\, v \in V_a \mid (o, a, v) \in I \wedge (v, b) \in I_a$$

is a *realised scale* for the attribute a. The *derived context* with respect to conceptual scales $(\mathbf{S}_a)_{a \in A}$ is the binary context $\mathbf{D} := (O, B, J)$ with $B = \bigcup_{a \in A}(\{a\} \times A_a)$ and

$$oJ(a, b) \Longleftrightarrow \exists\, v \in V_a \mid (o, a, v) \in I \wedge (v, b) \in I_a$$

Conceptual scaling successfully extends FCA to non-binary attributes but can unfortunately increase dramatically the number of binary attributes in the derived context, and consequently the resulting lattice size, depending on the scales used. Since we are interested in interactions between two or more many-valued attributes, nested-line diagrams can be a solution both to address this complexity drawback and to visualise these interactions (see Vogt and Wille 1995). Nested-line diagrams consist of partitioning the attribute set into two sets, say first factor and second factor attributes, then computing the concept lattices of the two corresponding subcontexts, and finally replacing each node in the first factor lattice by the line diagram of the other lattice, showing how objects of the first factor's current node are distributed with respect to the second factor.

An example from Wolff (1993) illustrates the whole process on a basic many-valued context. Non-binary attributes from Table 1 are discretized using an ordinal scale for *gender* (see Section 7.3 for details on elementary scales) and a biordinal scale for *age* (see Table 2). The resulting derived context is shown in Table 3. In order to provide a nested-line diagram representation, *gender* is chosen as the first factor and *age* as the second factor.

Figure 4 depicts concept lattices computed from related subcontexts and Figure 5 shows the resulting nested-line diagram. Note that unoccupied nodes are not removed from embedded diagrams, since the aim is to preserve the scale structure.

7.2 *MDS projections as an alternative to embedded line diagrams*

Combining more than two scales may result in complex embedded diagrams, useful for deep analysis but not suitable for the first-glance navigation method we aim

Table 3. Derived context.

	M	F	< 18	< 40	≤ 65	> 65	≥ 80
Adam	×			×	×		
Betty		×			×		
Chris						×	
Dora		×				×	×
Eva		×	×	×	×		
Fred	×						
George	×					×	×
Harry	×				×		

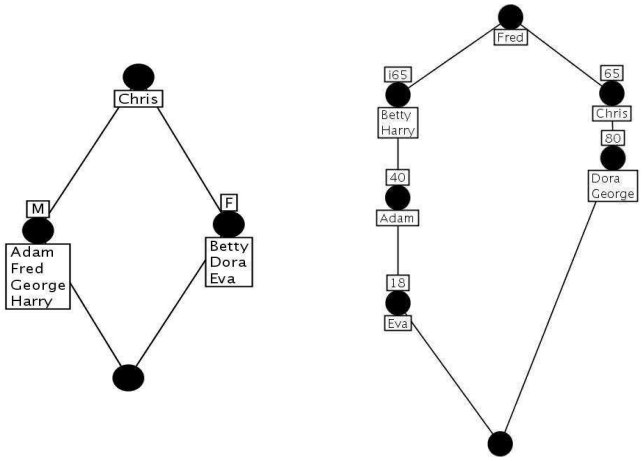


Figure 4. Concept lattices for first (*gender*) and second (*age*) factors.

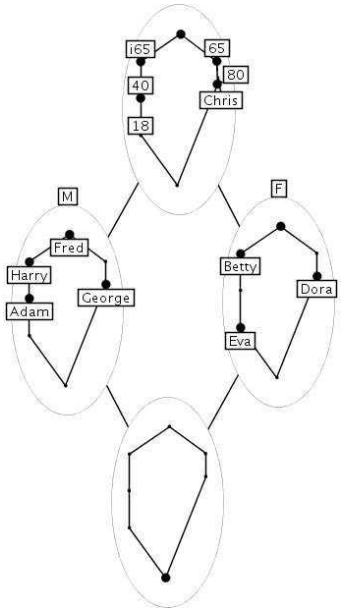


Figure 5. Resulting nested-line diagram. Smaller nodes represent unoccupied concepts.

to support. Even with simple embedded diagrams, some visual misunderstandings may occur. For instance, considering embedded diagrams in Figure 5 representing age, the partition between ≤ 65 and > 65 clearly appears but, going deeper in the diagram, age is decreasing in one branch while increasing in the other. This may be quite disrupting for users since they may expect a representation that reflects the natural order relation between integer values.

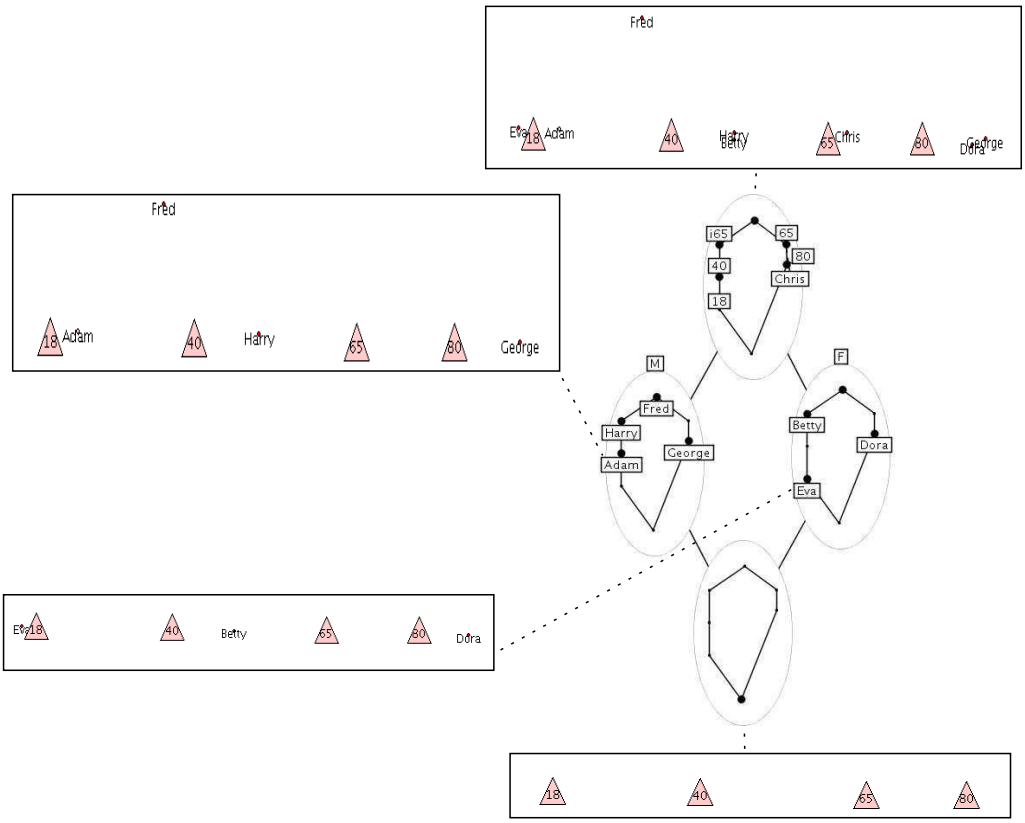


Figure 6. Embedded line diagrams representing the second factor (*age*) can be replaced by MDS projections.

In addition, discretizing numerical attributes leads to inevitable loss of precision in terms of proximities between objects. An alternative would be to consider binary attributes as a first factor set, and non-binary attributes as a second factor set. The overall lattice resulting from the first factor set will represent the database structure and since second factor attributes are scalable or comparable, embedded MDS projections will be used instead of embedded line diagrams to show distributions of one first factor's node extent with respect to non-binary attributes.
One may object that numerical attributes may have been discretized purposefully to observe the distribution of objects with respect to given thresholds that will disappear in MDS projections. A simple solution is to introduce virtual objects, these we call indicators, with threshold values used to represent a given scale. In our example, we create four indicators $i_{1...4}$ for the age scale: $age(i_1) = 80$, $age(i_2) = 65$, etc. Such a representation both reflects an order relation on integers and provides visual dichotomy with respect to thresholds. Figure 6 depicts interactions between the overall view and local MDS views. From this simple example, we have shown how to use MDS to represent a biordinal scale (partitioning values into two classes, here ≤ 65 and > 65, each one being ordered by an ordinal scale).

7.3 *MDS projection patterns for elementary scales*

In this subsection, we recall elementary scales presented by Ganter and Wille (1999) and precise how they can be represented using MDS projections.

7.3.1 Nominal and dichotomic scales

Nominal scales are used to scale attributes whose values exclude each other (e.g. masculine, feminine, neuter). They lead us to consider many-valued attributes that are not numeric. A preprocessing phase is required to compute MDS on these attributes. A numerical value is assigned to each nominal value, for instance considering three nominal values, $v_0 = 0$, $v_1 = 50$, and $v_2 = 100$. The resulting MDS projection effectively groups objects into three clusters since the Euclidian distance between objects sharing the same nominal value is null. For an optimal visual separation between clusters, numerical values defined in the preprocessing phase have to be chosen with respect to a regular scale in a given range, formally: $v_i = r(i/n)$ for $i \in \{0 \dots n-1\}$ for n values in range $\{0 \dots r\}$. Note that since nominal values are not ordered, matching between nominal and numerical values can be randomly performed. A dichotomic scale is a particular nominal scale where $n = 2$.

7.3.2 Ordinal and biordinal scales

Ordinal scales are used for attributes with ordered values. Biordinal scales are used when objects are assigned to one of two poles and this with a different degree. The example studied in Section 7.2 has already shown how to use a MDS projection to represent a biordinal scale. If an ordinal scale had been applied instead, the MDS projection would have been the same, since biordinal scales only differ from ordinal scales in that they partition the values into two poles with respect to a threshold value. This partition can be visually achieved thanks to indicators. Figure 7 compares the biordinal scale applied to the attribute age with its MDS representation which actually may also suit for an ordinal scale. The two poles are easily identified thanks to the indicator 65. More generally, from a MDS representation of an ordinal scale, all possible biordinal scales, generated by changing the attribute value used as threshold, are simultaneously visualised thanks to indicators. For instance, the MDS representation in Figure 7 can also be used to show ≤ 18 and > 18 poles.

7.4 *Visual interaction between non-binary attributes*

In the previous subsections we focused on MDS projections computed on a single non-binary attribute as a second factor. However raw data may contain many non-binary attributes. Users may be interested in studying interactions between these non-binary attributes. Using nested-line diagrams would lead us to compute complex embedded diagrams combining several conceptual scales that would have to be redrawn for each addition or removal of an attribute in the second factor set. On the other hand, computing MDS projection with respect to several attributes is very easy since it has been designed to represent multidimensional data. Each object has a numerical vector, each row containing a numerical value for the corresponding non-binary attribute. Our visualisation tool Molage allows users to dynamically select dimensions on which MDS is performed. Thus users can observe objects proximities, clusters and their changes while selecting or unselecting attributes (see Figure 8).

The following example (based on real data from Asuncion and Newman (2008)) illustrates our visual navigation method in a many-valued context. Raw data contains 205 objects and 24 attributes, dealing with imported cars in the United States. All attributes are non-binary: 14 are numeric and 10 are nominal. Two nominal attributes, *number of doors* and *number of cylinders*, can easily be ordered, so finally we have 16 numeric and 8 nominal attributes. The last are used as the first factor (their resulting lattice stands for the overall view) and the first as the second

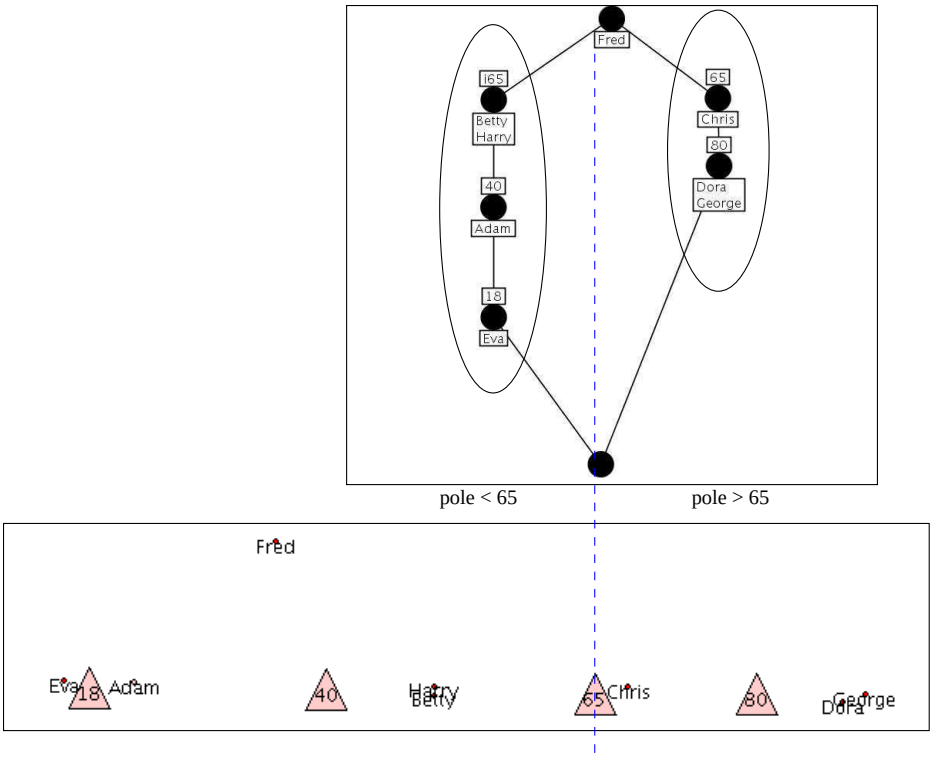


Figure 7. Representing a biordinal scale using a MDS projection.

factor (selectable dimensions in embedded MDS projections).

7.4.1 First factor attributes and overall lattice

The first factor attributes are eight nominal attributes such as *name of constructor*, *engine location*, etc. A nominal scale is computed for each first factor attribute, since their values exclude each other. The resulting derived context contains 47 binary attributes, generating a 551 node lattice. This shows how much conceptual scales increase the final number of binary attributes in the final context and consequently impact the lattice size. As previously explained, we aim at providing a visual navigation method that enables an overall analysis of a database and its structure; hence we reduce the lattice size by applying an iceberg lattice algorithm with 20% support (see Figure 9).

7.4.2 Second factor attributes and embedded MDS projections

The second factor attributes are 14 numerical attributes such as *height*, *width*, *weight*, *horsepower*, *price*, etc. MDS projection is suitable to compare objects with respect to such attributes. Each attribute is normalized in order to maintain equal weights between vector rows during Euclidian distance computation. Note that the axis have no semantics in a MDS projection, only distances between objects are meaningful since they reflects similarities between object vectors. Nevertheless it is possible to force the assignment of one attribute to x axis and another to y axis by using the following process. All objects are grouped at one position. Then their y position is fixed and a first attribute is selected so that objects find a position on the x axis with respect to their value on this first attribute. Objects' x positions are then fixed, first attribute unselected, y positions unfixed and finally second attribute selected. As a result, an object's position on x (resp. y) axis reflects its value according to the first (resp. second) attribute. Such a process can be automated in our visualisation tool (see Crampes *et al.* 2006).

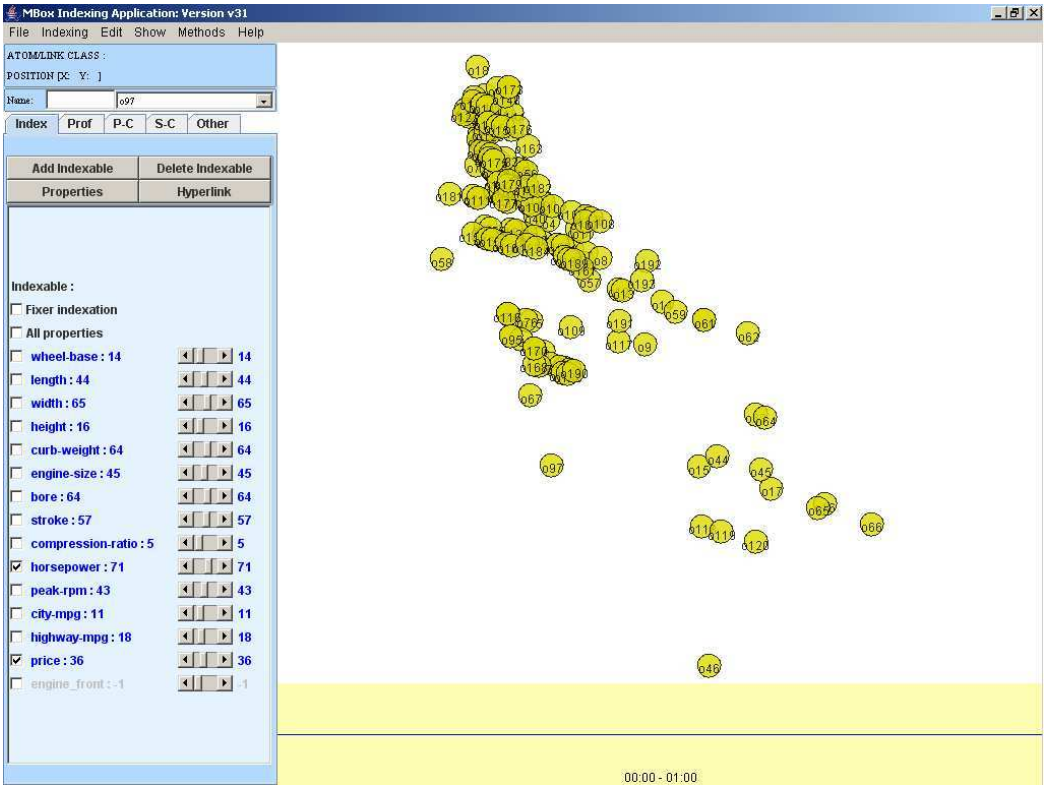
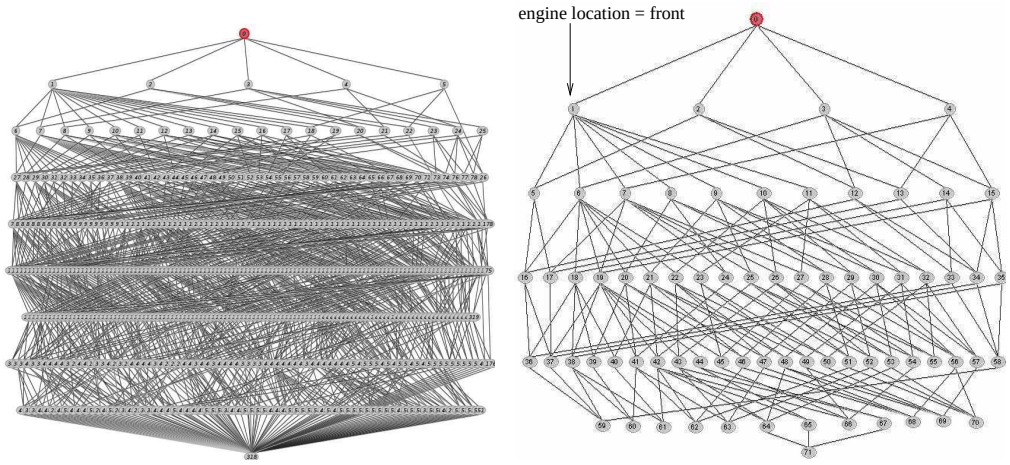


Figure 8. Selective MDS: Euclidian distances between objects are computed with respect to the attributes *horsepower* and *price* (see left panel).



(a) (b)

Figure 9. Complete lattice (551 nodes) (a). iceberg lattice with 20% support (72 nodes) (b).

Figure 10 shows the top node's embedded MDS projection on attributes *horsepower* and *price*, showing that these attributes are roughly correlated. We have seen how the first factor attributes have been used to build the overall lattice. Nevertheless they can also be visualised in embedded MDS projections. For instance, numerical values for the nominal attribute *engine location* have been defined following the preprocessing phase described in Section 7.3.1. The top node's embedded MDS projection on this attribute is depicted by Figure 11. Two clusters clearly appear, corresponding to each value defined for attribute *engine location*: *rear* and *front*. Users can observe at first sight that most of the cars have a front-located engine. In

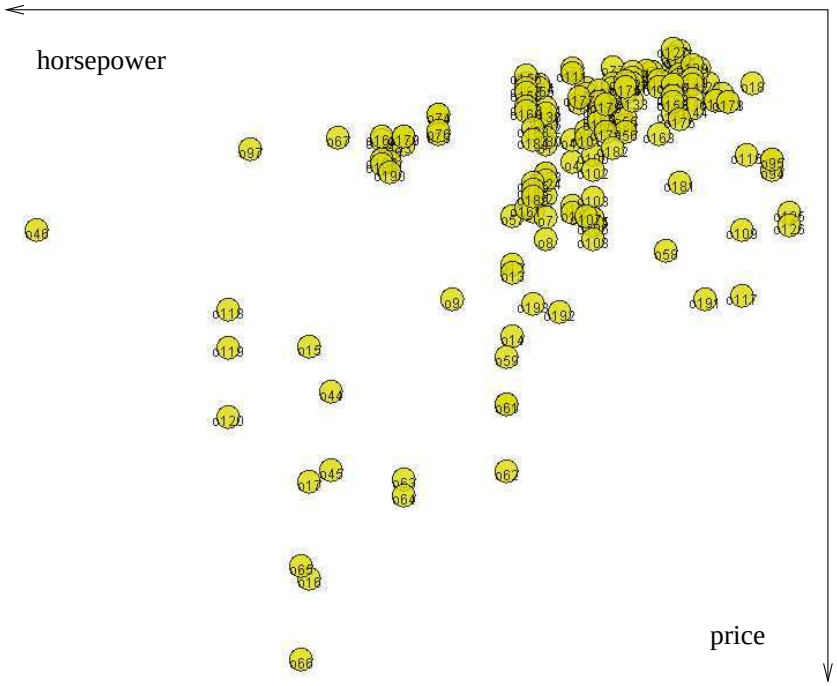


Figure 10. Projection of the top node extent with respect to the attributes *horsepower* and *price*. Although the axis are meaningless in a MDS projection, we can force the assignment of a particular attribute to a particular axis (see Section 7.4.2).

other words, the value *front* for the attribute *engine location* is shared by most of the objects and tends to group them. This observation is confirmed by the fact that the scaled attribute *engine location: front* is the intent of one of the four children nodes of the top node (see Figure 9(b)).

8. Conclusion and perspectives

This paper has presented a method for visual navigation in indexed document collections combining classification behaviors of FCA to produce overviews and intuitive visualisation techniques when focusing on local views. Initial feedback from our industry partner is promising. The approach seems pertinent and results match the users' needs. The challenge is now to broaden the scope of our research to emphasize the analysis process during database exploration. As mentioned in the introduction, we aim to use FCA techniques not only to organize the database but also to infer some reasoning during the visualisation navigation process. Research presented in this paper only deals with assisting users in their navigation and no information is inferred by the system itself. This is a proper problem solving behaviour in information visualisation. Indeed, quoting Robert Spence's definitions in Spence (2001) "*to visualise is to form a mental model or mental image of something. Visualisation is a human cognitive activity, not something that a computer does*", our goal is not to infer or deduce formal results from raw data because data analysis techniques such as FCA succeed in this without any visualisation need. We try to explore new techniques to bootstrap the cognitive activity of users.

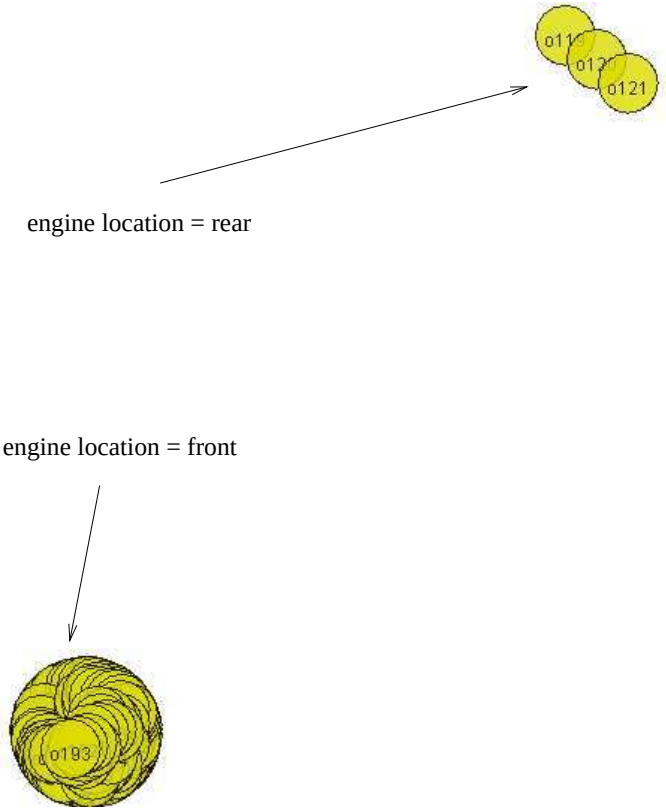


Figure 11. Projection of top node extent with respect to the attribute *engine location*.

Notes on contributors

Jean Villerd holds an MSc from the University of Montpellier (France) and a PhD in Computer Science from the École Nationale supérieure des Mines de Paris. He has published several research and technical papers on the application of Formal Concept Analysis techniques to Software Engineering and Information Visualisation issues.

Sylvie Ranwez is a research member at the LGI2P Research Center of the École des Mines d'Alès (France). She holds a PhD in computer science from the University of Montpellier (France), dealing with automatic composition of adaptive hypermedia documents based on ontologies and intentional request of a user. Her research focuses on technical assistance in indexation, navigation and information search using semantic support (ontologies) and visualisation. She has published research papers on semantic distance, conceptual maps, ontological and FCA based indexing and visual navigation.

Michel Crampes is a senior research member at the LGI2P Research Center of the École des Mines d'Alès (France) with a PhD in Computer Science from the University of Montpellier (France). He has directed several PhD thesis in his research topics, such as adaptive multimedia, knowledge maps, information visualization, adaptive HMI, Semantic Web, agent models, Formal Concept Analysis, conceptual indexing. He has initiated and managed several research projects in Europe.

David Carteret is the Managing Director and chief scientist of I-Nova, a software and professional services company specialized in idea and innovation management. He holds a MSc degree in Mathematics from the École Centrale de Paris and started his carrer at Peregryne System.

References

Asuncion, A. and Newman, D., UCI machine learning repository [online]. : University of California, Irvine, School of Information and Computer Sciences. Available from: http://www.ics.uci.edu/∼mlearn/MLRepository.html, [Accessed 28 September 2008].

Brodlie, K., Poon, A., Wright, H., Brankin, L., Banecki, G. and Gay, A., 1993. Grasparc: A problem solving environment integrating computation and visualization. *In*: G. Nielson and D. Bergeron, eds. *Proceedings of the 1993 IEEE conference on visualization* IEEE Computer Society, 102–109.

Card, S., Hong, L., Mackinlay, J. and Chi, E., 2004. 3Book: a scalable 3D virtual book. *Extended abstracts on Human factors in computing systems, CHI'04* ACM Press, 1095–1098.

Card, S., Mackinlay, J. and Shneiderman, B., 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.

Carpineto, C. and Romano, G., 2004. Exploiting the potential of concept Lattices for information retrieval with Credo. *Journal of Universal Computer Science*, 10 (8), 985–1013.

Chalmers, M., 1996. A linear iteration time layout algorithm for visualising high-dimensional data. *Proceedings of the 1996 IEEE conference on visualization* IEEE Computer Society, 127–133.

Cole, R., Eklund, P. and Stumme, G., 2000. CEM — A program for visualization and discovery in email. *In*: D. Zighed, J. Komorowski and J. Zytkow, eds. *Proceedings of the 4th European conference on principles and practice of knowledge discovery in databases, PKDD'00*, Vol. 1910 of *LNAI* Springer Verlag, 367–374.

Cole, R., Eklund, P. and Stumme, G., 2003. Document retrieval for email search and discovery using formal concept analysis. *Applied Artificial Intelligence*, 17 (3), 257–280.

Crampes, M., Ranwez, S., Villerd, J., Velickovski, F., Mooney, C., Emery, A. and Mille, N., 2006. Concept maps for designing adaptive knowledge maps. *Information Visualization*, 5 (3), 211–224.

Crampes, M., Villerd, J., Emery, A. and Ranwez, S., 2007. Automatic playlist composition in a dynamic music landscape. *Proceedings of the 2007 international workshop on Semantically aware document processing and indexing, SADPI'07* ACM Press, 15–20.

Di Battista, G., Eades, P., Tamassia, R. and Tollis, I., 1994. Algorithms for drawing graphs:

an annotated bibliography. *Computational Geometry: Theory and Applications*, 4 (5), 235–282.

Ducrou, J. and Eklund, P., 2008. An intelligent user interface for browsing and search MPEG-7 images using concept lattices. *International Journal of Foundations of Computer Science*, 19 (2), 359–381.

Eades, P., 1984. A heuristic for graph drawing. *Congressus Numerantium*, 42, 149–160.

Eklund, P., Ducrou, J. and Brawn, P., 2004. Concept lattices for information visualization: Can novices read line diagrams. *Proceedings of the 2nd international conference on formal concept analysis, ICFCA'02*, Vol. 2691 of *LNAI* Springer Verlag, 57–72.

Eklund, P., Ducrou, J. and Wilson, T., 2008. An intelligent user interface for browsing and search MPEG-7 images using concept lattices. *Proceedings of the 4th international conference on concept lattices and their applications*, Vol. 4923 of *LNCS* Springer Verlag, 1–22.

Ganter, B. and Wille, R., 1999. *Formal concept analysis*. Springer New York.

Godin, R., Pichet, C. and Gecsei, J., 1989. Design of a browsing interface for information retrieval. *Proceedings of the 12th annual international conference on research and development in information retrieval* ACM Press, 32–39.

Jankun-Kelly, T., Ma, K. and Gertz, M., 2002. A model for the visualization exploration process. *Proceedings of the IEEE conference on visualization*, 323–330.

Kruskal, J. and Wish, M., 1978. *Multidimensional Scaling*. Sage Publications.

Lamirel, J. and Al Shehabi, S., 2006. MultiSOM: A multiview neural model for accurately analyzing and mining complex data. *Proceedings of the 2006 international conference on coordinated and multiple views in exploratory visualization* IEEE Computer Society, 42–54.

Lee, J., 1998. A systems and process model for data exploration. Thesis (PhD). University of Massachusetts Lowell.

Lee, J. and Grinstein, G., 1995. An architecture for retaining and analyzing visual explorations of databases. *Proceedings of the 1995 IEEE conference on visualization* IEEE Computer Society, 101–108.

Marks, J., Andalman, B., Beardsley, P., Freeman, W., Gibson, S., Hodgins, J., Kang, T., Mirtich, B., Pfister, H., Ruml, W. *et al.*, 1997. Design Galleries: a general approach to setting parameters for computer graphics and animation. *Computer graphics*, 31, 389–400.

Mothe, J., Chrisment, C., Dousset, B. and Alau, J., 2003. DocCube: Multi-dimensional visualisation and exploration of large document sets. *JASTIS*, 54 (7), 650–659.

Pernelle, N., Ventos, V. and Soldano, H., 2003. ZooM: alpha galois lattices for conceptual clustering. *Proceedings of the managing specialization/generalization hierarchies (MASPEGHI) workshop*.

Priss, U., 2006. Formal concept analysis in information science. *Annual review of information science and technology*, 40, 521–543.

Ranwez, S., Ranwez, V., Villerd, J. and Crampes, M., 2006. Ontological Distance Measures for Information Visualisation on Conceptual Maps. *Proceedings of the OntoContent workshop*, Vol. 4278 of *LNCS* Springer Verlag, 1050–1061.

Shneiderman, B., 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics*, 11 (1), 92–99.

Shneiderman, B., 1996. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings of the 1996 IEEE symposium on visual languages* IEEE Computer Society, 336–343.

Spence, R., 2001. *Information Visualization*. ACM Press Books.

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N. and Lakhal, L., 2002. Computing iceberg concept lattices with Titanic. *Data & knowledge engineering*, 42 (2), 189–222.

Tory, M., Potts, S. and Möller, T., 2005. A parallel coordinates style interface for exploratory volume visualization. *Transactions on visualization and computer graphics*, 11 (1), 71–80.

Tricot, C., Roche, C., Foveau, C. and Reguigui, S., 2006. Cartographie sémantique de fonds numériques scientifiques et techniques. *Document numérique*, 9, 12–35.

Upson, C., Faulhaber Jr, T., Kamins, D., Laidlaw, D., Schlegel, D., Vroom, J., Gurwitz,

R. and van Dam, A., 1989. The application visualization system: a computational environment for scientific visualization. *Computer graphics and applications*, 9 (4), 30–42.

Valtchev, P., Grosser, D., Roume, C. and Hacene, M., 2003. Galicia: an open platform for lattices. *Proceedings of the 11th international conference on conceptual structures, ICCS'03* Shaker Verlag, 241–254.

Villerd, J., Ranwez, S., Crampes, M. and Carteret, D., 2007. Using concept lattices for visual navigation assistance in databases: application to a patent database. *Proceedings of the 5th conference on concept lattices and their application, CLA'07*, Vol. 331 CEUR, 88–99.

Vogt, F. and Wille, R., 1995. Toscana: a graphical tool for analyzing and exploring data. *Knowledge organization*, 22 (2), 78–81.

Wolff, K., 1993. A first course in formal concept analysis: how to understand line diagrams. *Proceedings of SoftStat'93* Gustav Fischer Verlag, 429–438.