

From Visualization to Association Rules : an automatic approach

Gwenael Bothorel, Mathieu Serrurier, Christophe Hurter

▶ **To cite this version:**

Gwenael Bothorel, Mathieu Serrurier, Christophe Hurter. From Visualization to Association Rules : an automatic approach. SCCG 2013, Spring Conference on Computer Graphics, ACM, May 2013, Smolenice, Slovakia. pp 57-64, 10.1145/2508244.2508252 . hal-01078330

HAL Id: hal-01078330

https://enac.hal.science/hal-01078330

Submitted on 16 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From Visualization to Association Rules: an automatic approach

Gwenael Bothorel*
DSNA/French Air National Service Provider

Mathieu Serrurier†
IRIT/University of Toulouse

Christophe Hurter‡
ENAC/University of Toulouse

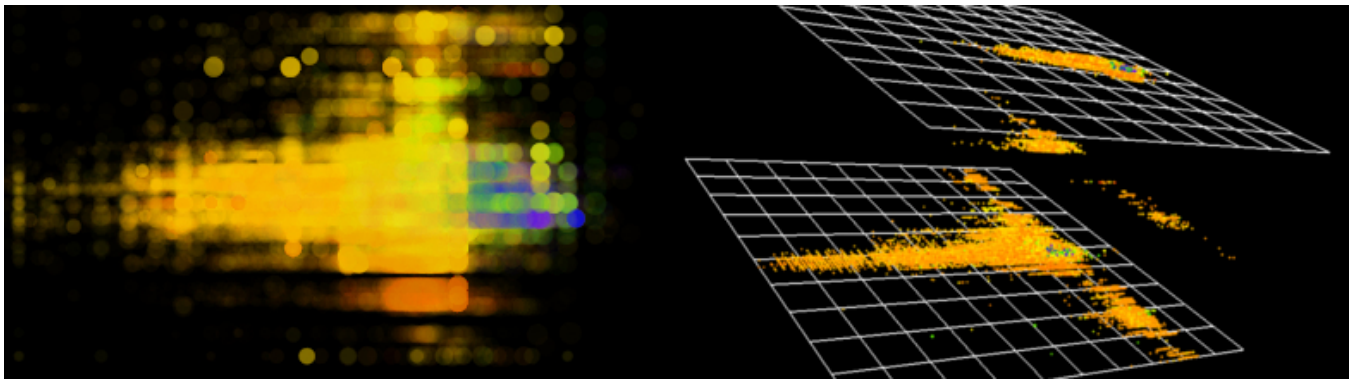


Figure 1: The quantity of involvement of the data in the rules is emphasized by the opacity (left) and on the Z axis (right).

Abstract

The main goal of Data Mining is the research of relevant information from a huge volume of data. It is generally achieved either by automatic algorithms or by the visual exploration of data. Thanks to algorithms, an exhaustive set of patterns matching specific measures can be found. But the volume of extracted information can be greater than the volume of initial data. Visual Data Mining allows the specialist to focus on a specific area of data that may describe interesting patterns. However, it is often limited by the difficulty to deal with a great number of multi dimensional data. In this paper, we propose to mix an automatic and a manual method, by driving the automatic extraction using a data scatter plot visualization. This visualization affects the number of rules found and their construction. We illustrate our method on two databases. The first describes one month French air traffic and the second stems from 2012 KDD Cup database.

CR Categories: H.5.m [Information Interfaces and Presentation (e.g., HCI]: Miscellaneous; H.2.8 [Database Management]: Database Applications—Data Mining

Keywords: association rules, Visual Data Mining, characterization of a visualization, aeronautical data, KDD Cup

1 Introduction

Data mining is a knowledge extraction process from a vast volume of data. One purpose of data mining is to find patterns that link this data. It can be done automatically, for instance by algorithms that find association rules such as $pizza, chips \Rightarrow beer$. The most famous one is Apriori [Agrawal and Srikant 1994]. The main interest of this approach is its completeness: every association rule that satisfies the constraints of a set of measures will be found. However, the volume of rules can sometimes be greater than the volume of initial data. Indeed the number of possible rules that have to be validated grows exponentially with the number of attributes. Since this number of attributes is usually high, we may have to face a new data mining problem of identifying subsets of relevant rules from a vast volume of rules. The research can be done by using filters over rules a posteriori [Lallich et al. 2006]. However, this approach is time consuming and the tuning of the filters is a tricky issue.

Another data mining process is visual data mining (VDM). It is a manual way of extracting patterns from the visual exploration of data thanks to interactive tools. One major advantage of VDM is its abilities to navigate and interact with a set of user defined data. The user is directly involved in the data mining process [Keim 2002]. Different techniques and tools have been the subjects of much research and many publications [Soukup and Davidson 2002] [Keim 2002] [Simoff et al. 2008]. The tool can be fully manual like FromDaDy [Hurter et al. 2009] where the user chooses the links between attributes and visual variables and adapts the visualization to his needs, or it can be assisted [Guettala et al. 2011]. The main advantage of the VDM approach is that it is driven by the specialist and it focuses on relevant patterns. However, this can also be a limitation since it is difficult to extract unsuspected new patterns with this kind of approach. It is more adapted for validating the specialist's hypotheses. Moreover, when the number of attributes is high, it is not always easy to build relevant visualizations.

It is also possible to mine or to explore frequent itemsets or association rules with visual tools in order to manage the large amount of extracted information. For instance, FIsViz [Leung et al. 2008] shows the frequent itemsets in a 2D space, by linking the items with connecting edges. Such a tool gives a global graph of the dataset. As for ARVis, it shows the rules and associated measures values in a 3D information landscape [Blanchard et al. 2003]. Each rule is presented as a combination of a sphere and a cone both at a specific po-

*e-mail: gwenael.bothorel@aviation-civile.gouv.fr
†e-mail: mathieu.serrurier@irit.fr
‡e-mail: christophe.hurter@enac.fr

Copyright © 2013 by the Association for Computing Machinery, Inc.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions Dept, ACM Inc., fax +1 (212) 869-0481 or e-mail permissions@acm.org.
SCCG 2013, Smolenice, Slovakia, May 1 – 3, 2013.
© 2013 ACM 978-1-4503-2480-9/13/0005 $15.00

sition in the landscape. Their characteristics give information about the measures. More recently, CBVAR is an integrated Clustering-based Visualizer of Association Rules, which creates rules clusters and visualizes them [Couturier et al. 2007].

In this paper we propose to combine an algorithmic and a manual approach. Our goal is to characterize a scatter plot visualization with a compact set of association rules extracted with Apriori. The specialist builds a visualization according to a specific issue, based on settings and selections, in order to obtain the best view of the data. Considering that the database is described as a set of values with associated attributes, the user matches the visual variables with data attributes. In agreement with the semiology of graphics [Bertin 1983], we deduce from this visualization a set of constraints for the Apriori algorithm. The visualization is also used for discretizing continuous values. What we obtain is a compact set of relevant association rules that focus on the displayed data and on the way it is represented. The main advantages of this visual-driven process are the possibility of finding local rules that would be difficult to detect by considering the whole database, and a shorter computation time. Moreover, the number of found rules is always acceptable. We also propose a method in order to emphasize a rule or a set of rules on the visualization.

The first section presents some background and context about the semiology of graphics and the association rules extraction. Then, we formalize the notion of visualization. In Section 4, we describe our approach, named Videam platform (VIsual DrivEn dAta Miner), and we explain how the visualization can drive the different steps of the data mining process and we discuss the impact of the visualization settings on the complexity and the number of rules generated. Finally we illustrate this concept with two databases. The first describes one month French air traffic through the description of the flights and the planes.The second is the 2012 KDD Cup database which describes a social network of microblogging.

2 Background and context

2.1 The Semiology of Graphics

In *Semiology of Graphics* [Bertin 1983] Bertin writes that the graphics depicts only the relationships established among components or elements. The choice of visual variables is a main factor that contributes not only to the readability of the graphics but also to its intelligibility. The information can belong to three levels of organization: the qualitative (nominal), ordered or quantitative model. The Bertin's semiology of graphics has been used by Card and Mackinlay in order to characterize visualizations [Card and Mackinlay 1997] and is the basis of visual data mining tools.

VDM tools must use a visual description language. Card and Mackinlay's model is operated in FromDaDy (FROM DAta to DisplaY) [Hurter et al. 2009], the primary purpose of which is the visual exploration of large volumes of data (see Figure 2). FromDaDy employs a simple paradigm to explore multidimensional data based on scatter plots, brushing (i.e. selection), pick and drop (i.e. direct manipulation), juxtaposed views and rapid visual configuration. Together with a mix between design customization and simple interaction, users can filter, remove and add data in an incremental manner until they extract a set of relevant data, thus formulating complex queries.

Figure 3 illustrates the selection steps to construct a visualization. With basic operations, the user can perform all kinds of Boolean operators, such as AND, OR, NOT and XOR.

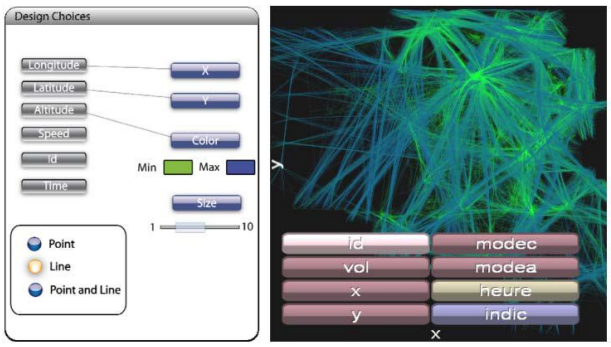


Figure 2: FromDaDy overview.

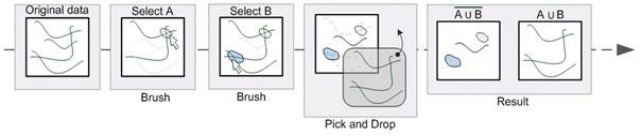


Figure 3: FromDaDy implements Boolean operators.

Figure 4 illustrates some advanced features of FromDaDy such as accumulation maps.

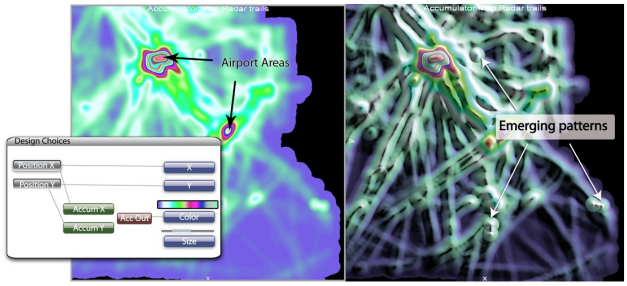


Figure 4: Design configuration and accumulation maps in FromDaDy, without shading (left), with shading (right).

Visual data mining takes into account the expertise of the user. As the specialist knows what he wants, he sets the layout of the HMI to find data, in order to answer a question. In fact, VDM tools are more used to validate hypotheses than to research knowledge. If the tool is configurable like FromDaDy, the settings will be different for another question or with another user.

2.2 Association rules extraction

One of the most important way of mining data is the association rules extraction. In this study, we use Apriori algorithm [Agrawal and Srikant 1994]. An association rule establishes a link between attributes which describe the data. The rule $A \Rightarrow B$ means: if the attributes contained in the head A are present in a tuple, then the attributes of the conclusion B are present too.

Two basic measures are used to find interesting patterns:

- Support $s(A \Rightarrow B) = P(AB)$: probability to have both A and B,
- Confidence $c(A \Rightarrow B) = P(B/A)$: probability for B to be true considering that A is true.

These measures are initial constraints for Apriori algorithm which guarantees the extraction of all association rules the confidence and support of which are above given thresholds. Many other quality measures, like the lift, are used to evaluate the rules (see [Lallich et al. 2006; Blanchard et al. 2005] for a discussion about them). They can report on other properties and are used to filter the rules afterwards. The algorithmic approach allows the user to have the exhaustive list of association rules, corresponding to preliminary settings. But the number of extracted rules and the duration of the process can be crippling. The complexity of Apriori algorithm is $\mathcal{O}(nm2^m)$, where n is the number of tuples and m the number of attributes. This shows that the time consumption of Apriori grows linearly with the number of tuples and exponentially with the number of attributes.

The exhaustiveness of this approach contrasts with VDM because of the specialist's choice on visualization. Thus, Apriori may find rules that are difficult to detect manually with VDM.

3 Formalization of data visualization

The Card & Mackinlay's model describes only the matching from the attributes to the visual variables and the organization of the visualization. The process from data to what is displayed on the user's device is not handled by this model. Thus, we formalize data and visual variables in order to understand the links between the data and the visualization. A database is a set of vectors over an attribute space. We name $\mathscr{A} = \{A_1, A_2, ..., A_m\}$ the set of the m possible attributes. Each attribute A_i can be numerical (integer or decimal) or nominal (a noun...). Depending on the type, an attribute can be ordered or not. The feature space $\mathscr{X}$ is the space of the possible vectors over the attributes. Given $x \in \mathscr{X}$ we have $x_= < a_1, a_2, ..., a_m >$ where a_j is the value for the attribute A_j. A database $\mathscr{D}$ is a subset of $\mathscr{X}$: $\mathscr{D} = \{x_1, x_2, ..., x_n\}$, where $\forall i, x_i \in \mathscr{X}$ and n is the number of the data. A piece of data $x_i =< a_{i1}, a_{i2}, ..., a_{im} >$ is a vector where a_{ij} is the value for the attribute A_j for the piece of data i.

The visualization of data is based on the instantiation of visual variables depending on the attributes values. We name $\mathscr{V} = \{V_1, V_2, ..., V_q\}$ the set of q visual variables. Visual variables can have different types such as the position in space, the size, the color or the transparency of a point. Note that a visual variable can also be ordered (i.e. position, size and alpha) or unordered (i.e. color). A point p is a vector of values over the visual variables. We denote $\mathscr{P}$ the point space. Let us now formally define a visualization.

Definition 1 *A visualization vis is a mapping from $\mathscr{X}$ to $\mathscr{P}$ (i.e. $vis : \mathscr{X} \to \mathscr{P}$) built as follows:*

- *vis defines a function from $\mathscr{V}$ to $\mathscr{A}$. We denote map the function which associates with the index of the visual variable and the index of the corresponding attribute. Note that it is not necessarily an injection, because one attribute can be associated with two different visual variables. Each visual variable V_i is then associated to an attribute $A_{map(i)}$. Thus we obtain a set of q pairs $(V_i, A_{map(i)})$.*
- *for each pair $(V_i, A_{map(i)})$ we have a function $f_i : A_{map(i)} \to V_i$.*

Then, given $x \in \mathscr{X}$ we have:

$$vis(x) =< f_1(a_{map(1)}), \ldots, f_q(a_{map(q)}) > .$$

So, given a visualization, a displayed point is characterized by its visual variable values, each one corresponding to the value of an attribute. It is not mandatory to use all the possible visual variables (for instance a user can choose to display data only in the plane), and the number of visual variables has to be less than or equals to the number of attributes. Note that the function f_i depends on the type of the visual variable and the attribute. It can be, for instance, a linear projection (when considering numerical attributes a space coordinate visual variable for example), a gradient function, etc. Given a visualization, what can be displayed will depend on the restriction that we apply over the visual variable and to the selection of the user. In this scope, we define the scene.

Definition 2 *A scene sc is a subset of $\mathscr{P}$ that determines which point can be effectively displayed on the device. We have:*

$$sc \subseteq \mathscr{P}.$$

Then, the combination of a visualization *vis*, applied to a set of data and a scene *sc*, defines an image that can be displayed on a device (by a projection of the considered point on a plane). In practice the scene depends on the different actions of the user. It can be, for instance, pan & zoom interaction or camera position which restricts the area of the space being considered. Data selection (by brush, bounding box...) is also handled by the scene.

Thus, the Card & Mackinlay's model is handled by the *map* function of the visualization. The choice of relevant *map* and f functions can be driven by the semiology of graphics. A visual data mining process consists in a series of visualizations $vis_1, \ldots, vis_p$ related to a series of scenes $sc_1, \ldots, sc_p$ which corresponds to a series of views and selections in these views. In the following, we are only interested in the characterization of the last visualization and scene.

4 Visual driven Data Mining

The purpose of our approach is to characterize a visualization of data, built by a specialist with a visual data mining tool, by a set of association rules. In order to answer a question, the specialist elaborates a visualization. Since the number of visual variables is limited, he has to restrict the set of attributes to the most relevant ones (even if the whole database and the whole attributes set are still available). Visual data mining is based on this principle, because the specialist builds the visualization by trying to establish relations between the attributes. In our approach, we assume that the visualization process suggests the type and the shape of association rules that may interest the specialist. The goal is then to extract and to present only these rules by involving the visualization at each step of the Apriori process (see Section 2.2). The advantages of this approach are:

- The visualization drives the whole process through the visual data mining tool. It make it easier to focus the Apriori algorithm on the considered issue without tuning the parameters. Indeed, this last step may be tricky to handle manually.
- It is possible to find local rules. If the algorithm is applied to the whole database, some rules won't be discovered. For example, the support may be too small with respect to the whole database, or the rule may be only true (high confidence) on a specific subset of the data. By selecting a part of the database, Apriori will search rules that specifically apply on this subset.
- The processing time is drastically shortened because of the lighter volume of data and attributes. Thus, the computation always remains acceptable for interactive visual tools. This point is critical since the volume of data usually overcomes millions of tuples.

We have developed the Videam platform which is a visual and automatic data mining platform. The rules are generated by Apriori algorithm. The HMI stems from FromDaDy project [Hurter et al. 2009] which was presented in Section 2. The platform takes advantage of GPU optimization in order to have an efficient representation of millions of data. With Videam, the matching is done between attributes and visual variables which are the location, the color, the transparency, and the point size, on an orthographic projection or on a 3D view.

4.1 Data selection and discretization

There are two ways of selecting data. They define the perimeter of the scene *sc*.

- With pan & zoom interaction, only displayed data is selected, regardless to possible occlusions.
- Manual selection of the data in different views by the use of brush or bounding box for instance.

According to his requirements, the user links some attributes with the visual variables. The most obvious selection is to restrict on the data for which visual variable values are in the displayed range. This can be the result of pan and zoom or more simply the restriction that applies to the display device. The second type of selection corresponds to manual data picking. As mentioned in Section 2.1, VDM tools allow manual selection of a subset of the displayed data. It is important to notice that these selections can be made on different views and then can be applied to attributes that are not represented in the final visualization. Then, the Apriori algorithm will be applied on the subset $\mathscr{D}' \subseteq \mathscr{D}$ of the data that is displayed on the visualization. Given a visualization *vis* and a scene *sc*, $\mathscr{D}'$ is defined as follows:

$$\mathscr{D}' = \{x \in \mathscr{D} | vis(x) \in sc\}.$$

This selection has some consequences on the Apriori process:

- The pan & zoom and selection tools act as data filters. Since confidence and support are computed on this subset, Apriori can then extract rules that are locally relevant, even if it is not the case for the whole dataset.
- As the number of tuples is decreased, the complexity of Apriori linearly decreases by a range related to the selected data ratio. This is due to the fact that, since $\mathscr{D}' \subseteq \mathscr{D}$, we have $n' = |\mathscr{D}'| \leq n$.
- If the selection can be simply described, adding this description in the head of a local association rule makes it globally relevant. However, it is not always easy, or even possible, to clearly describe this selection, particularly when it is drawn with a brush tool.

In order to produce association rules, the numerical data has to be discretized. It means that it has to be grouped in different categories corresponding to ranges of values. Since the goal of the approach is to characterize the visualization, we apply automatic clustering to the visual variable range rather than to the initial domain. More formally, it means that, for an attribute A_i, the clustering is made on $f_i(A_i)$. Since f_i can be non-linear and dynamically modified by the user in VDM algorithm, it ensures that the produced rule will have a meaningful representation in the visualization.

4.2 Frequent itemsets

If the algorithm is applied to the whole set of attributes, there will be a combinatorial explosion of the number of found association rules and so a high computation time consumption. As we have seen previously, the number of visual variables constrains the number of attributes and then the frequent itemsets size and scope. From the user's point of view, this is not a real limitation since it is cognitively difficult to tackle more than six dimensions at the same time [Miller 1956]. Then, the Apriori algorithm will only take into consideration a subset $\mathscr{A}' \subseteq \mathscr{A}$ of attributes. Given a visualization *vis*, $\mathscr{A}'$ is defined as follows:

$$\mathscr{A}' = \{A_i \in \mathscr{A} | \exists V_j \in \mathscr{V}, map(j) = i\}.$$

The complexity is reduced by a $2^{m-|\mathscr{A}'|}$ factor and the number of rules is also naturally reduced. Since *map* is not necessarily an injection and the visual variables are limited to six in our platform, we have $|\mathscr{A}'| \leq 6$.

This allows us to only present the rules that are interesting for the user in an acceptable time. This first point is critical since, when considering all the attributes, the number of rules can be very high and even overcoming the number of initial data. The problem of filtering rules is a well known issue in data mining. Sometimes, finding interesting rules can be as difficult as a data mining process. So the choice of the visual variables by the user will determine the data, more precisely the attributes, that are submitted to the algorithm, and then will focus it on the rules that may interest the user. This choice corresponds to the function *map* which connects the index of a visual variable to the index of the corresponding attribute.

4.3 Association rules restriction

We have seen that the rules are strongly dependant on the selected data and attributes. However the way the matching is done between the visual variables and the attributes, and the way the data is shown (orthographic or 3D projection) may be used to constrain the rules structure. It is based on the organization of visual variables [Bertin 1983]. We have to deal with the question of which attributes will appear in the head and which ones in the conclusion of the association rules. In the following, we assimilate the visual variable to the corresponding attribute.

The visual variables that we presently use are the position, the color, the transparency and the point size. As the dimensions of the plane have the three levels of organization (see 2.1), they are generally the starting point of a visualization. Indeed two components data is usually displayed in a plane. Then the color, for instance, can be added to this data in a third dimension. The same principle is applied to space visualization. In 2D, the natural basic approach is to represent the ordinate as a function of the abscissa. In 3D, the usual representation is the depth function of the plane. We assume that it is the starting point for our rule representation. In 2D, we consider that the attribute associated with the abscissa is necessarily in the head of the rule. The size, like the dimensions of the plane or of space, has the three levels of organization. Thus the ordinate or the size can be in the head or in the conclusion of the rule. However, they are the two only components that can appear in the conclusion. This can be extended in 3D by considering the plane instead of the abscissa.

The color doesn't have the three levels of organization, and discriminating with this visual variable is less easy without ordered and quantitative levels. That is why it will only appear in the head of the rules as a complement of the abscissa (or the plane in 3D).

In the same way, the transparency, which is not studied in the semiology of graphics, is also difficult to discriminate even if it is an ordered visual variable. So it will only appear in the head too.

Considering the visual variables belonging to the visual variables ensemble $V_V = \{X, Y, Z, S, C, A\}$ (S is the size, C the color and A the transparency), the association rule restriction, based on the attribute associated with the visual variable, thanks to *map* function, can now be formalized in Equation 1 (planar presentation) and 2 (spatial presentation) depending on the user's construction of the visualization:

$$X \wedge \{Y, S, C, A\}^* \Rightarrow \{Y, S\}^+. \quad (1)$$

$$X \wedge Y \wedge \{Z, S, C, A\}^* \Rightarrow \{Z, S\}^+. \quad (2)$$

The $*$ means that zero, one or more values can be used. The $+$ means that at least one value must be used. Note that a visual variable cannot appear twice in the rule.

4.4 Rule representation

In order to emphasize a rule or a set of rules in the visualization, we propose to compute a new variable *em* which corresponds to the contribution of a data with respect to the rule. Given an association rule $A \rightarrow B$ and a piece of data x_i, if x_i doesn't satisfy A we have $em(x_i) = 0$, if x_i satisfy A and not B (counter example of the rule) we have $em(x_i) = 1$ and if x_i satisfy A and B (example of the rule) we have $em(x_i) = 2$. For the set of rules, an operator such as the sum or the max can be used. By assigning the variable *em* to a visual variable (Z for instance) we can emphasize the area where the rule applies and where the rule makes error. If we consider a large set of rules, we obtain a map which emphasizes the areas where we can easily extract information.

5 Illustrations

In this section, we illustrate our approach with two databases:

- The first, which is not public, belongs to the French civil aviation authority. It describes the French air traffic in August 2010. It contains data about flights and planes.
- The second is public and stems from the 2012 KDD Cup database. It describes a social network of microblogging.

5.1 Aeronautical database

This database represents about 280 000 flights. It has also been successfully tested on the whole 2010 traffic with more than 2 500 000 flights. A preliminary Apriori calculation, applied to the August traffic, with 16 attributes, support equal to 0.10 and confidence equal to 0.85, gives 2 800 000 rules in 35 minutes. The number of rules is then too high for manual exploration by a specialist.

We consider the visualization shown in Figure 5, where we associate the aircraft speed to X, the arrival airport longitude to Y, the altitude to the point size and the departure airport latitude to the color. The Z axis is not assigned.

The calculation time of the algorithm applied to this displayed data and attributes is 3 seconds with 41 rules found. We obtain rule such as:

$$Altitude \in [360, 380] \Rightarrow Arrival\ Longitude \in [-35, 3].$$

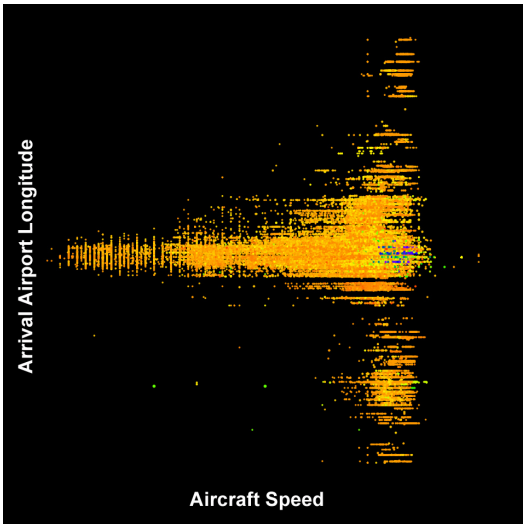


Figure 5: Visualization of four attributes of French air traffic data in August 2010.

It illustrates that an aircraft that flies at a medium altitude goes to Europe. Notice that the discretizations are obtained by clustering over the visual variables and transformed into the initial variables domain.

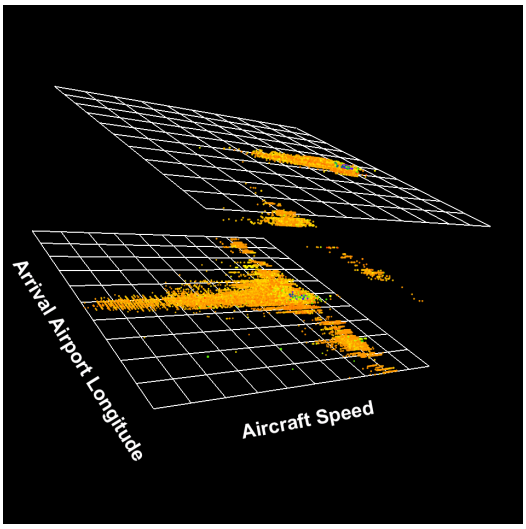


Figure 6: One rule has been selected. The concerned data is raised.

Figure 6 shows the initial data with a perspective view, in order to give an overview of the whole data with an extra dimension. On this picture, we emphasize the rule previously described by assigning the variable *em* to Z. It allows us to determine immediately where the rule applies (high level of Z) and where it fails (medium level of Z). The low level of Z corresponds to data that is not concerned by the rule.

In Figure 7, all the rules have been selected and combined by using sum operator. As it is difficult to have a correct perspective view with many Z values, we use the *alpha* to point out the amount of rules that applies to the data. Event if it is difficult to render transparency on the snapshot, we can observe that the area where X is high and Y is medium has the highest opacity. It means that

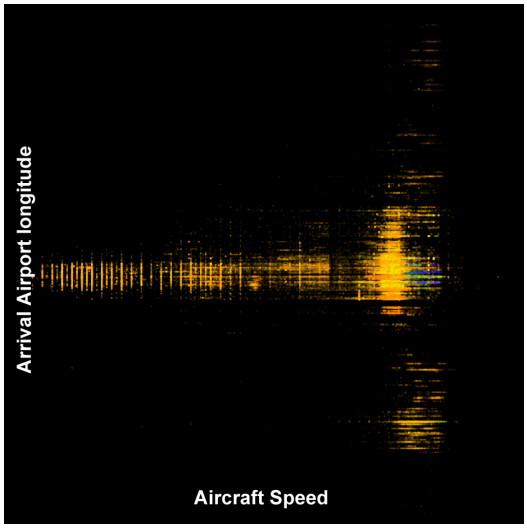


Figure 7: All the rules have been selected. The concerned data is enhanced by opacity. The other data is more transparent.

a lot of information can be extracted with data mining algorithm in this area. One reason is that it corresponds mainly to local French flights which are the most represented in the database. The biggest points are in the lowest levels of *alpha* which means that we cannot extract too much information from the aircraft that fly at high altitude.

5.2 2012 KDD Cup database

The second illustration deals with the 2012 KDD Cup database. This database describes a social network of microblogging. "Users" are people in the social network and "items" are famous people or objects that the users may follow. Users may be friend with other users. The task of this challenge is to predict if a user will accept or not to follow an item proposed by the system.

There are 10 millions of users described by their age, their gender, some keywords and their friends, and 50 000 items described by keywords and tags. The database contains 70 millions of propositions to follow an item with a label that indicates if the user has accepted the proposition or not. The problem is then a binary classification problem. We define 9 attributes based respectively on the percentage of users of the same gender of the target one that follow the item, the percentage of users of the same age category of the target one that follow the item, the session time, the number of friends of the target user that follow the item, the distance between the items followed by the user and target item, the number of users that follow the target item, the number of items that are followed by the user, the number of times the item has been proposed to the user and the class (accepted or not). We build a dataset of around 1.9 million of propositions.

In order to avoid the sparsity of the data, we transform the visualization of the data (formalized by the use of the function f for the visualization of each attribute) in order to have a more uniform and homogeneous distribution of the data in the visualization. Since the discretization is performed on the visualization, after the application of f, we obtain very homogeneous clusters. Then, we have choosen to have 5 clusters per visual variable, each cluster associated to a nominal intuitive label:

- *vs*: very small (for the 20% lowest values)
- *s*: small (for the 20% next values)
- *med*: medium (for the 20% next values)
- *h*: high (for the 20% next values)
- *vh*: very high (for the 20% highest values)

A first Apriori calculation is applied to the whole dataset, with the 9 attributes. The support threshold is equal to 0.005 and the confidence one is equal to 0.80. We obtain 3647 rules in 37 seconds. Even the number of data is greater than for the aeronautical dataset, the computation time is shorter and the number of rules is lower, due to the lower number of attributes. In this scope the interest of our approach is not to reduce the time of computation, but to focus on the attributes, or on specific areas of data we want to consider.

5.2.1 Experiment 1

In a first time, we consider the visualization shown in Figure 8. Here we associate the number of items that are followed by the user to X, the number of users that follow the target item to Y, the session time to the *color* (blue for the lowest values, red for the highest values), the number of times the item has been proposed to the user to the *pointsize*. In this example we use the Z axis to represent the class (accepted at the top and not accepted at the bottom). We can observe that it is tricky to extract information directly from this noisy visualization, due to the large amount of high density data. For instance, it is difficult to distinguish different point sizes.

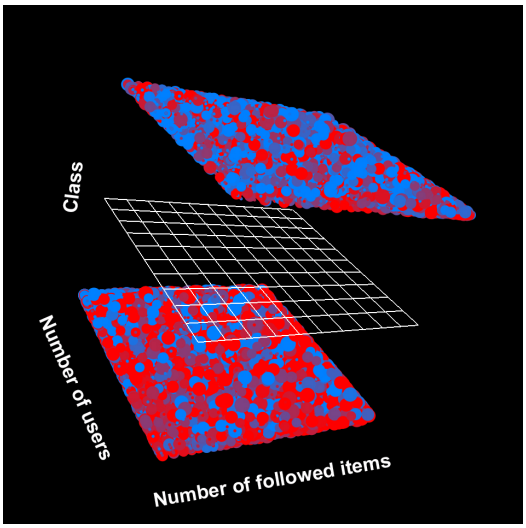


Figure 8: Perspective view of 5 attributes of the 2012 KDD Cup dataset: experiment 1.

The calculation time of the algorithm applied to this displayed data and attributes is 5 seconds with 112 rules found. We obtain rule such as:

$$SessionTime = vs, item = vs, class = 1 \Rightarrow NbPropositions = vs.$$

It illustrates that if a new user, which follows few items, accepts a proposition, then he accepts it the first times it is proposed by the system.

We have also the two following rules:

- $NbPropositions = vh \Rightarrow class = 0$,
- $SessionTime = vh \Rightarrow class = 0$.

In a complementary way to the first rule, these last rules show that propositions are no longer accepted after too many propositions or too long sessions. This corresponds intuitively to the fact that a user is more open to proposition at the beginning of the session. It is interesting to remark that these rules cannot be directly identified in the visualization.

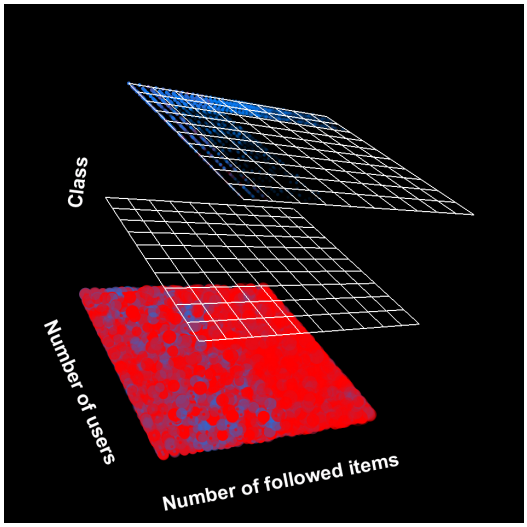


Figure 9: The contribution of data in the rules is assigned to the *alpha*. It shows for instance that it is easier to predict when a proposition is refused (very large majority of low level values).

Figure 9 shows the same visualization in which we emphasize all the rules on the *alpha* visual variable (Z is already used for the representation of the class). In this view, we can immediately observe that very few information can be extracted for the accepted proposition. Indeed, only the points corresponding to the lowest values of session time (blue point) and the number of items that are followed by the user, and the highest values of the number of users that follow the target item are discernable for the class 1. It exactly corresponds to the examples of the first described rule.

On the contrary, we see that there is much information for class 0. This can be explained by the fact that it is easier to predict when a proposition will be refused than when it will be accepted. Moreover, the domination of red points shows that the most relevant information deals with longer sessions.

5.2.2 Experiment 2

In a second experiment, we consider the visualization shown in Figure 10. We have first selected the data corresponding to class 1, that is to say the accepted propositions. Here we associate the percentage of users of the same age category of the target one that follow the item to X, the percentage of users of the same gender of the target one that follow the item to Y, the distance between the items followed by the user and target item to the *color* (blue for the lowest values, red for the highest values), and the number of friends of the target user that follow the item to the *pointsize*. The Z axis is not used.

The calculation time of the algorithm applied to this displayed data and attributes is 2 seconds with 14 rules found. We obtain rule such as:

$$DateScore = vs, GenScore = vs, ItemDist = vs \Rightarrow FriendScore = vs.$$

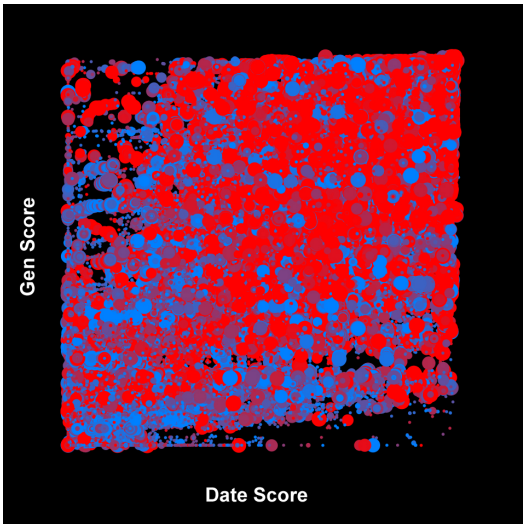


Figure 10: Orthographic view of 4 attributes of the 2012 KDD Cup dataset: experiment 2.

It illustrates that low values of the percentage of users of the same age category of the target one that follow the item, the percentage of users of the same gender of the target one that follow the item and the distance between the items followed by the user and target item are highly correlated to the low values of the number of friends of the target user that follow the item. This rule is emphasized in the third dimension in Figure 11. We can remark that it corresponds to a very precise area of all the visual variables which cannot be identified directly in Figure 10. Figure 12 presents the same visu-

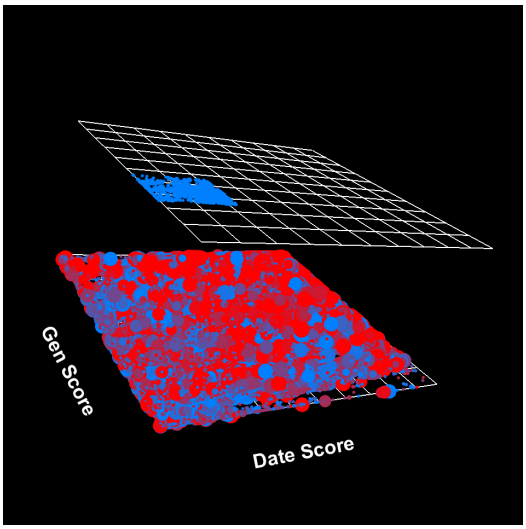


Figure 11: One rule has been selected. The concerned area of data is raised on Z axis.

alization as Figure 10 in which we emphasize all the rules on the *alpha* visual variable. It shows that only small values of variables (corresponding to the low left area) leads to the learning of rules. Although the size of the point seems to increase when x and y increase, no rules of this types are found. It means that, even if such relations may exist, they do not have enough support.

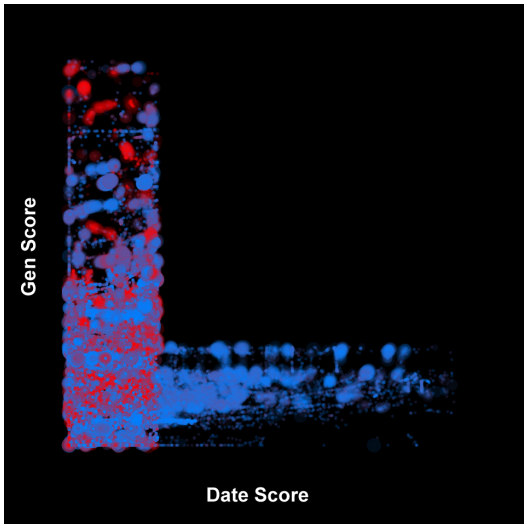


Figure 12: The contribution of data in all the rules is assigned to the *alpha*. It enhances the most involved data and reduces the other.

6 Conclusion

In this paper we have presented an automatic approach to characterize a visualization by a set of association rules. We have proposed a formalization of the visualization, from the data structure to the final displayed image, including Card & Mackinlay's characterization of the visualization. With Videam platform, the user builds the representation of data depending on his needs, thanks to visual data mining techniques. Then he generates the association rules that correspond to this visualization, according to the semiology of graphics. These rules depend on the displayed and selected data, on the matching between the attributes and the visual variables, and on their organization. Our approach allows the user to extract rules that focus on his interest and that may apply only on a subset of data rather than the whole data. Moreover, constraints induced by the visualization decrease significantly the computation time of the Apriori algorithm and make the extraction of the rule compatible with interactive exploration. We also proposed a method for emphasizing the rules found on the visualization. It allows us to automatically determine the area of application of a rule. When combining more rules, we obtain a map that represents the amount of information extracted in the area. We apply our approach on two databases. On the aeronautical one, we show that our approach can be used for finding information on specific variables and area. With the 2012 KDD Cup database we show that our approach can automatically emphasize information that is not easily detectable due to the large amount of data represented in the visualization.

In the future, it would be interesting to establish a backward link from the rules mining space to the data visualization space, by using an algorithm that would optimize the representation of data. We also have to determine which kind of visual patterns can be effectively characterised by one (or a set of) association rules.

References

AGRAWAL, R., AND SRIKANT, R. 1994. Fast algorithms for mining association rules in large databases. In *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, Morgan Kaufmann, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds., 487–499.

BERTIN, J. 1983. *Semiology of graphics*. University of Wisconsin Press.

BLANCHARD, J., GUILLET, F., AND BRIAND, H. 2003. A user-driven and quality-oriented visualization for mining association rules. In *Proceedings of the Third IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, ICDM '03, 493–.

BLANCHARD, J., GUILLET, F., BRIAND, H., AND GRAS, R. 2005. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proceedings of the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, ENST, 191–200.

CARD, S. K., AND MACKINLAY, J. 1997. The structure of the information visualization design space. In *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, IEEE Computer Society, Washington, DC, USA, 92–.

COUTURIER, O., HAMROUNI, T., YAHIA, S. B., AND NGUIFO, E. M. 2007. A scalable association rule visualization towards displaying large amounts of knowledge. In *Proceedings of the 11th International Conference Information Visualization*, IEEE Computer Society, Washington, DC, USA, 657–663.

GUETTALA, A. E.-T., BOUALI, F., GUINOT, C., AND VENTURINI, G. 2011. Premiers rsultats pour un assistant utilisateur en fouille visuelle de donnes. *18 mes Rencontres de la Socit Francophone de Classification*, 1986, 71–74.

HURTER, C., TISSOIRES, B., AND CONVERSY, S. 2009. Fromdady: Spreading aircraft trajectories across views to support iterative queries. *IEEE Transactions on Visualization and Computer Graphics 15*, 1017–1024.

KEIM, D. A. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8* (January), 1–8.

LALLICH, S., TEYTAUD, O., AND PRUDHOMME, E. 2006. *Association rules interestingness: measure and validation*. Quality Measures in Data Mining. Springer, Heidelberg, Germany, 251–275.

LEUNG, C. K.-S., IRANI, P. P., AND CARMICHAEL, C. L. 2008. Fisviz: a frequent itemset visualizer. In *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining*, Springer-Verlag, Berlin, Heidelberg, PAKDD'08, 644–652.

MILLER, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review 63*, 2 (March), 81–97.

SIMOFF, S. J., BÖHLEN, M. H., AND MAZEIKA, A., Eds. 2008. *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, vol. 4404 of *Lecture Notes in Computer Science*. Springer.

SOUKUP, T., AND DAVIDSON, I. 2002. *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, 1st ed. John Wiley & Sons, Inc., New York, NY, USA.