An Interactive Approach to Display Large Sets of Association Rules

Olivier Couturier[1], José Rouillard[2], and Vincent Chevrin[2]

[1] Centre de Recherche en Informatique de Lens (CRIL) – IUT de Lens,
Rue de l'université, SP 18, F-62307 Lens Cedex, France
couturier@cril.univ-artois.fr
[2] Laboratoire Trigone/LIFL, CUEEP, Bâtiment B6, Cité Scientifique,
F-59655, Villeneuve d'Ascq Cedex, France
{jose.rouillard,vincent.chevrin}@univ-lille1.fr

Abstract. Knowledge Discovery in Databases (KDD) is an active research domain. Due to the number of large databases, various data mining methods were developed. Those tools can generate a large amount of knowledge that needs more advanced tools to be explored. We focus on association rules mining such as "If Antecedent then Conclusion" and more particularly on rules visualization during the post processing stage in order to help expert's analysis. An association rule is mainly calculated depending on two user-specified metrics: *support* and *confidence*. All current representations present a common limitation which is effective on small data quantities. We introduced a new interactive approach which combines both a global representation (2D matrix) and a detailed representation (Fisheyes view) in order to display large sets of association rules.

Keywords: Knowledge Discovery in Databases (KDD), Human Computer Interaction (HCI), Visualization.

1 Introduction

In front of the increasing number of large databases, extracting useful information is a difficult and open problem. This is the goal of an active research domain: Knowledge Discovery in Databases (KDD). KDD techniques have been proposed and studied to help users to understand better and scrutinize huge amounts of collected and stored data [10]. KDD is a new hope for companies which use methods (e.g. statistical methods) that do not allow to tackle large amounts of data. Currently, commercial KDD tools panel develops quickly and some of these tools are marketed such as Purple Insight[1]. However, they are generally complex to use and they are not flexible depending on the user's problem. We focused on association rules mining (ARM) [1] but we oriented our work about the interaction between the user and a KDD process.

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

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 258–267, 2007.

Precursory works are rather old because one of the first methods of mining correlations between Boolean values is the GUHA (General Unary Hypotheses Automaton) method [9]. Association rules interest was started three decades later thanks to the first large databases including commercial transactions [1]. The aim is to obtain rules such as "If *Antecedent* then *Conclusion*". This problem is also called *market basket analysis* and it is the starting point of ARM. In this case, each basket is relevant for one customer[2] depending on his needs and desires but if the supermarket tackles all baskets simultaneously, useful information can be extracted and exploited. All customers are different and buy different products in different quantities. However, market basket analysis consists in studing customer's behaviors as well as the factors which push them to carry out a kind of purchase. It allows to study what kind of products are bought together time with other products and consequently, to adapt a corresponding promotional campaign. The following simple example "If *Smoker* Then *Cholesterol* (75%)" means that a person who smokes has 75% of risk to have too much cholesterol. Although this method is initially planned for the great distribution sector, it can apply to other fields. The approach remains identical whatever the studied field: to propose models, tools and transdisciplinary methods in order to help the expert's analysis[3] in order to take the good decision.

Several works on Human Computer Interaction (HCI) are focused on Visual Information-Seeking Mantra illustrated by Shneiderman: *"Overview first, zoom and filter, then details on demand"* [22]. First of all, graphical tools must provide a global view of the system in order to deduce the main important points. The user must be able to determine the starting point of his analysis thanks to this global view. During his analysis, he can explore in-depth particular areas if he wishes so. Indeed, all details don't need to be displayed at the same time. Unfortunately, all current representations do not respect this crucial point, which is necessary in order to visualize large sets of knowledge. In addition, we need to display all metrics thanks to various colors pallets. Current tools still appear limited to display more than two metrics. Our purpose is how to obtain a representation adapted to visualization of large sets of association rules by jointly presenting a general view (the global) and a sight targeted on one or more particular elements (the detail)? This question constitutes the starting point of this work. The key to the success of a visualization prototype is a full compliance with such recommendations. We propose a hybrid visualization, which is composed of a 2D matrix to display an overview of our rules, and a fisheyes view to detail particular information. In the following, we will present our work corresponding to this research area.

This article is structured as follows: the second section introduces the framework of this work and presents in details ARM (Association rules mining) which constitutes the heart of our work. The third section describes our work which merges both HCI and KDD concepts based on human factors. This merging is crucial for

[2] Until the end of this paper, we consider one customer both man or woman

[3] In our work, the final user is a domain's expert. We will use independently the terms "expert" and "user".

relevant decisions support systems success. Finally, we conclude this article and propose some research ideas as perspectives.

2 Problem

This section formally introduces the association rules problem within a KDD process and it describes our motivations.

2.1 Knowledge Discovery in Databases (KDD)

In front of the increasing number of large databases, extracting useful information is a difficult and open problem. This is the goal of an active research domain: Knowledge Discovery in Databases (KDD). Nowadays, information which circulates in the whole world is mainly stored in digital form. Indeed, few years ago, a Berkeley University project estimated that in the world, the volume of annually generated data is equal to about one exa-byte[4] (i.e. 1 billion of gigabytes). Among these data, 99,997 % are available in digital form, according to [13] quoted in [18]. In commercial companies such as bank, insurances or distribution field, large amounts of customers' data are collected and they are not always exploited thereafter. How to make these data profitable in shorter running time? Indeed, current traditional request, as SQL (Structured Query Language) or OLAP[5] (On-Line Analytical Processing) are now

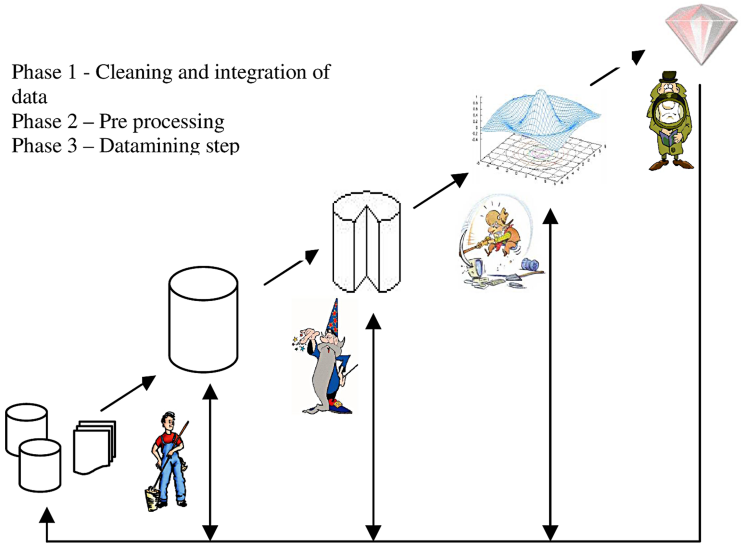

Fig. 1. KDD process

[4] 1 exa-byte (Eo) = 2^{60} bytes ; 1 zetta-byte (Zo) = 2^{70} bytes ; 1 yotta-byte (Yo) = 2^{80} bytes.

[5] Decision support software that allows the user to quickly analyze information that has been summarized into multidimensional views and hierarchies.

limited due to the increasing collection of large databases. To answer this problem, KDD is a new hope for companies which use methods (e.g. statistical methods) that do not allow to tackle large quantities of data (see Figure 1). Thanks to KDD techniques, large databases became rich and reliable sources for the generation and the validation of knowledge. Data mining is the main step of KDD process which consists in applying intelligent algorithms in order to obtain predictive models (or *patterns*). We focus on association rules mining which is a specific data mining task.

2.2 Association Rules Mining (ARM)

Association Rules Mining (ARM) [1] can be divided into two subproblems: the generation of the frequent itemsets lattice and the generation of association rules. The complexity of the first subproblem is exponential. Let $|I| = m$ the number of items, the search space to enumerate all possible frequent itemsets is equal to 2^m, and so exponential in m [1]. Let $I = \{a_1, a_2, \ldots, a_m\}$ be a set of items, and let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions establishing the database, where every transaction t_i is composed of a subset $X \subseteq I$ of items. A set of items $X \subseteq I$ is called *itemset*. A transaction t_i contains an itemset X in I, if $X \subseteq t_i$. Several ARM published papers are based on two main indices which are support and confidence [1]. The support of an itemset is the percentage of transactions in a database where this itemset is one subgroup. The confidence is the conditional probability that a transaction contains an itemset knowing that it contains another itemset. An itemset is frequent if support $(X) \geq minsup$, where *minsup* is the user-specified minimum support. An association rule is strong if confidence(r) $\geq$ *minconf,* where *minconf* is the user-specified minimum confidence. Left part of an association rule is called *antecedent* and right part is called *conclusion*. Our motivations are described hereafter.

2.3 Motivations

The number of generated rules is a major problem on association rules mining. This number is too significant and leads to another problem called *Knowledge mining*. The human cycles spent in analyzing knowledge is the real bottleneck in data mining. This issue can limit the final user's expertise because of a strong cognitive activity. To solve it, visual data mining became an important research area. Indeed, extracting relevant information is very difficult when it is hidden in a large amount of data. Visual data mining attempts to improve the KDD process by offering adapted visualisation tools which allow to tackle various known problems. Those tools can use several kinds of visualization techniques which allow to simplify the acquisition of knowledge by the human mind. It can handle more data visually and extract relevant information quickly.

During the last few years, several graphical approaches were proposed to display association rules in order to help experts' analysis. The first works were done in a text-mode. Their efficiency is restricted to the database size. For instance, if an expert searches for particular information, he can occult some essential information for his analysis. To answer it, several works were proposed in order to present this rules set such as graphs and trees, 2D and 3D matrix or virtual reality. Currently, these graphical representations present advantages and drawbacks. One limitation of the

aforementioned representations is to display all rules in the same screen space. Indeed, the user's visibility and understanding are proportionally reduced according to the number of generated rules. The common problem of the representations is that they are not simultaneously global and detailed. Indeed, global representations are quickly unreadable, whereas detailed representations do not present all information.

3 Visual Data Mining

The rise of KDD revealed new problems as *knowledge mining*. These large amounts of knowledge must be explored with specific advanced tools. Indeed, expertise requires an important cognitive work, *a fortiori*, a harmful waste of time for industrial. Extracting nuggets is a difficult task when relevant information is hidden in a large amount of data. In order to tackle this issue, visual data mining was conceived to propose visual tools adapted to several well-known KDD tasks. These tools contribute to the effectiveness of the processes implemented by giving understandable representations while facilitating interaction with experts. Visual data mining is present during all KDD process: upstream to apprehend the data and to carry out the first selections, during the mining, downstream to evaluate the obtained results and to display them. Visual tools became major components because of the increasing role of the expert within KDD process. Visual data mining integrates concepts resulting from various domains such as visual perception, cognitive psychology, visualization metaphors, information visualization, etc.

We focus on visualization during the post processing stage and we are interested by ARM. Independently of both context and task, ARM has a main drawback which is the high number of generated rules. Several works on filtering rules were proposed and a state of the art was presented in [4]. Although reducing the whole of generated rules significantly, this number remains however important. Expert must be able to easily interact with an environment of data mining in order to more easily understand the displayed results. This point is essential for the global performance of the system. Visual tools for association rules were proposed to reduce this cognitive analysis but they remain limited [4].

3.1 Visual Association Rules Mining

Various works already exist to help expert analysis in text-mode [16]. Several works on visual rules exploration were published [1], [2], [3], [25]. The main beliefs of our interactive ARM are described hereafter. All these tools use several methods which are textual, 2D or 3D way. The choice of one of them proves to be a difficult work. Moreover, their interpretations can vary according to the expert. Each one of these techniques presents advantages and drawbacks. It is necessary to take them into account for the initial choice of the representation. The effectiveness of these approaches is dependent on the input data files. These representations are understandable for small quantities of data but become complex when these quantities increase. Indeed, particular information can not be sufficiently perceptible in the mass. The common limitation of all the representations is that if they are global, they

quickly become unreadable (size of the objects in 2D, occlusions in 3D) and if they are detailed, they do not provide an overall picture on these data to the expert.

3.2 Hybrid Representation of Association Rules

Data mining is effective if the tools are able to represent the great masses of results obtained. Moreover, various functions of interaction were proposed in HCI (overall picture, zooms, data filtering, visualizations of relations between posted graphic objects) in order to facilitate the task of the user who must know this information and decide the level of relevance of an element among others. The work presented in [4] showed that, even for experts of a field, tools facilitating research and navigation in large sets of information are unavoidable. We successively proposed several representations (summarized textual of decision rules, 2D visualizations, then colored 3D, see Figure 2) in order to reduce the cognitive effort of the expert. Among various known representations, allowing to apprehend a great mass of information, such as the hyperbolic lenses or trees [14], perspective walls [17], fish eye view (FEV) [8] or superpositions of transparent sights [12], the FEV constitutes one possible solution adapted for the fields that we studied (bank, health, etc).

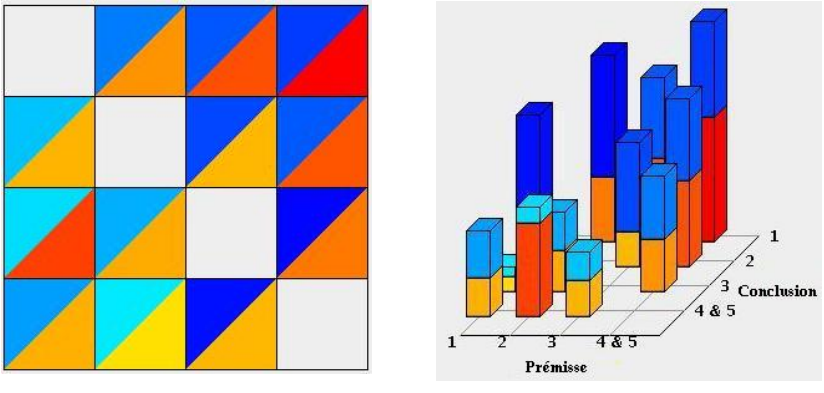

Fig. 2. 2D and 3D association rules visualization with *LARM*

In this paper, we present our study based on the continuation of the work started in [24] and [21] in order to interpret results in a visual way while preserving the context (see Figure 3b). For our study, we also tested *InfoVis* (see Figure 3c) [6]. However, several elements are not really adapted to our needs. For instance, with *InfoVis*, a rule is represented with the intersection of its metric. The number of metrics can not be higher than two. In our case, there can be much more. Consequently, legibility is reduce because of several rules can overlap (see Figure 3c). To answer this issue, we propose a representation in which the rules would be drawn in an allocated area (see Figure 3d). This kind of visualization makes it possible to represent several metrics in this area, thanks to various pallets of colors. In our example, we use two metrics and the allocated space is divided into two. With *N* metrics, the same space will be divided into *N* equal parts. Our assumption consists in supposing that we will have

better results by hybridizing a semantically colored view [7] and a FEV. The user will be able directly to point the polygon (colored in a gradual way according to the value of one or more metric associated to the rule) of the FEV which appears to him most relevant according to its task.

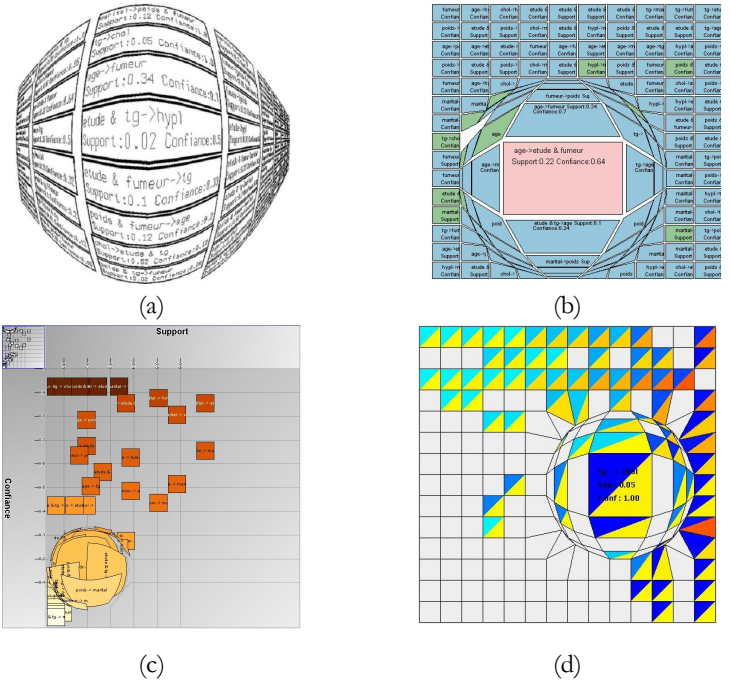

Fig. 3. Visualizations with (a) aiSee, (b) JAVA applet, (c) InfoVis, et (d) LARM

3.3 Current Tools Using FEV

We studied tools, allowing using our data files input and making it possible to visualize the rules using FEV. A system completely adapted to our problem does not exist, according to the literature related to this issue. Nevertheless, we studied the case of IDL (Interactive Language Dated) (www.rsinc.com) which is dedicated to the processing and the visualization of data (time series, images, cubes,…). IDL is based on compiled modules from the C language (there is a documentation that explain how to write such modules). It also makes it possible to write procedures and functions but also to make directed programming object. Since version 5, the programming of widgets became possible. The main aim of this language is to handle and display data with little investment in programming.

Then, we tested *aiSee* software (www.aisee.com), based on GDL (Graph Description Language) (near to IDL, but free). The *aiSee* software allows to read a data input file, (see Figure 4) resulting from a file (.gdl) and to display these data in

```
graph: {
       node: { title:  "A"  color: blue }
       node: { title:  "B"  color: red  }
       edge: { source: "A"  target: "B" }
      }
```

Fig. 4. Input format which is interpreted by GDL language

various forms, in particular in FEV. GDL describes a graph in term of nodes, edges, sub graphs and attributes. These attributes can be color, size, etc.

These two solutions seem adapted to our problems (see Figure 3a), but in real case, the following limitations appear: (a) Management of any field in a too generic way, however within the framework of the visualization of association rules, it is necessary to display several colors on the same node; (b) Difficult implementation for the developer. Indeed, the problems arising previously imply to code new modules in C language and that represents an important effort of implementation; (c) Usability reduced for the end-user. It should not be forgotten that an expert is a specialist within his field, but not necessarily within the tools (software) provided in order to achieve the task. It is essential to propose intuitive, simple and ergonomic interface. These languages appear to us more adapted to the scientists wishing to handle and display data, without need for producing code.

In order to start a footbridge between ECD and IHM, we conceived and developed our own display system exploiting a FEV (see Figure 3d). This implementation is carried out in JAVA and it is completely integrated into the LARM (Large Association Rules Mining) system which was presented in [5]. We thus proposed a display system of association rules allowing reducing the cognitive analysis while increasing the expert's efficiencies. According to our investigations, it seems that our approach is the only one able to propose a detailed and general sight simultaneously of association rules. Indeed, expert is in front of a general sight of the gradually colored rules and thanks to the FEV, he can obtain information on a highlight rule.

4 Conclusions and Further Works

After a study about current tools including FEV, we propose a first implementation which is integrated within an existing visualization rules platform (LARM). This approach can tackle large sets of rules in a same screen space. We show that our solution allowing to manage simultaneously both a global and a detailed representation. This is not yet the case in recent researches around KDD. This work is included in the *LARM* platform. We are testing it on both several banking and medical benchmarks. The first results are interesting and relevant. They show that human factors must play the first role at any time of the process during the decisional systems designing. Their efficiency depends on a good mix between KDD and HCI directly.

This work is a first outline to visualize large sets of association rules. However, it is necessary to evaluate our visualization system. In order to realize it, we will work in two times. Firstly, we wish to evaluate it thanks to the *discount usability testing* method [23]. The aim is to highlight the main advantages and the major drawbacks of

our visualization method. Secondly, as soon as the first evaluation will be validated, we will evaluate in-depth our system thanks to experts intervention. Currently, we focus on the first point. In another way, we focus on clusters visualization. In our approach, the fisheyes view focus point is a detailed rule. We wish to use it to visualize a set of clusters thanks to a 3D representation. Finally, we wish to invest about a new active research domain: haptic system [19] which can be significant in human computer interaction. Indeed, haptic system can be profitable in a knowledge management data mining system in order to help KDD actors during the analysis. On a more global point of view, we are planning to integrate multimodal features to our system. In output, Text to Speech (TTS) and haptic feedback will be used to give information to the user. In input, Automatic Speech Recognition (ASR) will be an interesting modality in order to command the system (example: "zoom in", "zoom out", "save this subset on the disc", etc.) but also to interact more naturally with it, in a natural language manner (example: "what are the best combinations for the year 2006?"). Freehand manipulations on interactive surfaces are already used to retrieve geographical information, for instance [20]. It provides more friendly interaction with the system, and, according to us, nobody has adapted those techniques in a KDD context. We will propose different kinds of multimodality: first, exclusive multimodality will bring the opportunity to switch from a modality to another (speech instead of keyboard/mouse, for example), then we will propose synergic multimodality, in which the user could, for example, pronounce "Zoom 63%" while he/she will manipulate graphical object with the mouse.

Acknowledgments. This work has been partly supported by the "Centre National de la Recherche Scientifique" (CNRS), the "IUT de LENS" and the "Université d'Artois". Moreover, The authors are thankful to the MIAOU and EUCUE programs (French Nord Pas-de-Calais Region) and the UE funds (FEDER) for providing support for this research.

References

1. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Toivonen, H., Verkamo, A.I.: Fast discovery of association rules, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, pp. 307–328 (1996)
2. ben Yahia, S., Mephu, N.E.: Emulating a cooperative behavior in a generic association rule visualization tool. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'04), Boca Raton, Florida, USA (2004)
3. Blanchard, J., Guillet, F., Briand, H.: Exploratory Visualization for Association Rule Rummaging. In: Proceedings of the 4th International Workshop on Multimedia Data Mining MDM/KDD2003, Washington, DC, USA, pp. 107–114 (2003)
4. Couturier, O.: Contribution à la fouille de données : règles d'association et interactivité au sein d'un processus d'extraction de connaissances dans les données, PhD Thesis, Université d'Artois, CRIL, Lens, France (2005)
5. Couturier, O., Mephu, N.E., Noiret, B.: A formal approach to occlusion and optimization in association rules visualization. In: Proceedings of International Symposium of Visual Data Mining (VDM) of IEEE 9th International Conference on Information Visualization (IV@VDM'05), Poster, London, UK (2005)

6. Fekete, J.D.: The InfoVis Toolkit. In: Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis'04), pp. 167–174. IEEE Press, New York (2004)
7. Fekete, J.D., Plaisant, C.: Interactive Information Visualization of a Million Items. In: INFOVIS 2002. IEEE Symposium on Information Visualization, Boston, pp. 117–124 (2002)
8. Furnas, G.W.: Generalized Fisheye Views. Proceedings of ACM Conference CHI'86, ACM SIGCHI Bulletin 17(4), 16–23 (1986)
9. Hajek, P., Havel, I., Chytil, M.: The GUHA method of automatic hypotheses determination. In Computing (1), 293–308 (1966)
10. Han, J., Kamber, M.: Data Mining; concepts and techniques. Morgan Kauffman, San Francisco (2001)
11. Harisson, B.L., Vicente, K.J: An experimental evaluation of transparent menu usage. In: Proc of ACM Conference CHI'96, pp. 391–398. ACM Press, New York (1996)
12. Keim, D.: Visual Exploration of large data Sets. Communications of the ACM 44(8), 39–44 (2001)
13. Lamping, J., Rao, R., Pirolli, P.: A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. In: Proceedings ACM Conference on Human Factors in Computing Systems (CHI'95), Vancouver, Canada, pp. 401–408 (1995)
14. Liu, B., Hsu, W., Wang, K., Chen, S.: Visually aided exploration of interesting association rules. In: Zhong, N., Zhou, L. (eds.) Methodologies for Knowledge Discovery and Data Mining. LNCS (LNAI), vol. 1574, pp. 380–389. Springer, Heidelberg (1999)
15. Mackinlay, J.D., Robertson, G.G., Card, S.K.: Perspective Wall: Detail and Context Smoothly Integrated. In: Proc. ACM Conference CHI'91, pp. 173–179. ACM Press, New York (1991)
16. Nigay, L.: Modalité d'Interaction et Multimodalité, Habilitation à Diriger des Recherches, spécialité Informatique de l'Université Joseph Fourier - Grenoble I (2001)
17. Pietrzak, T., Martin, B., Pecci, I.: Information display by dragged haptic bumps. In: Proceedings of the 2nd International Conference on Enactive Interfaces, Genoa, Italy (2005)
18. Rekimoto, J.: SmartSkin: An Infrastructure for Freehand Manipulations on Interactive Surfaces CHI2002, Conference on Human Factors in Computing Systems took place in Minneapolis, Minnesota (April 20-25, 2002)
19. Rouillard, J.: Navigation versus dialogue sur le web, Une étude des préférences. IHM'99, Montpellier (1999)
20. Schneiderman, B.: The eyes have it: A task by data type taxonomy for information visualization. In: Proceedings of IEEE Symposiumon Visual Languages, Boulder, Colorado, USA, pp. 336–343 (1996)
21. Schneiderman, B., Plaisent, C.: Designing the user interface, 4th edn., International Edition, Boston, Addison-Wesley, Reading, MA (2005)
22. Vernier, F., et Nigay, L.: Représentations multiples d'une grande quantité d'information, IHM'97, Futuroscope de Poitiers, France (1997)
23. Wong, P.C., Whitney, P., Thomas, J.: Visualizing Association Rules for Text Mining. In: Proceedings of the 1999 IEEE Symposium on Information Visualization (INFOVIS'00), Salt Lake City, Utah, USA, pp. 120–128 (2000)