Supporting Data Analysis Through Visualizations

Paolo Buono, Maria Francesca Costabile, Francesca A. Lisi

Dipartimento di Informatica, Università di Bari, via Orabona 4, 70125 Bari, Italy
{buono, costabile, lisi}@di.uniba.it

Abstract. Information visualization is a process that transforms information into a visual form, thus enabling the user to observe it. By graphically presenting data, the user may discover new and useful properties, their correlations, and also detect possible deviations from the expected values. In this paper, after discussing some ideas about possible fruitful use of visualization for data mining, we present a visualization module we are developing in the context of a project funded by the European Union. The project aims at offering on-line innovative services to support the business processes of trade fairs, both real and/or Web-based virtual fair. This module generates data visualizations on the WWW, which are exploited to facilitate human-computer interaction, to allow easy access to the stored data, and to present the retrieved information in appropriate ways, thus helping users in their data analysis activities.

1 Introduction and Motivation

Visual representation have the capability of shifting load from the user's cognitive system to the perceptual system. Indeed, information needs to be visualized in an information space in order to be retrieved by users. This visualization can either be carried out by the users in their own mind, in which case it is essentially the users' conceptualisation of that information, or it could be accomplished by the system, in which case the visualization is generated on the display screen. The latter is actually called *information visualization*, and is defined as "a process of transforming information into a visual form enabling the user to observe information" [1]. Recent research has proved that a suitable visualization can reduce the time to get information, and to make sense out of it. In the field of information systems, visualizations have a wide range of applications: they can be used for visualizing various types of meta-information, as well as queries and retrieved results. Moreover, by allowing dynamic user control of the visual information through direct manipulation principles, it is possible to traverse large information spaces and facilitate comprehension with reduced anxiety. In a few tenths of a second, humans can recognize features in mega-pixel displays, identify patterns and exceptions, recall related images. The use of proximity coding, colour coding, size coding, animated presentation, and user-controlled selections enable users to explore large information spaces rapidly and with fun.

Because of these characteristics, we believe that information visualization techniques may be very fruitful for data mining [2]. They may especially support post-processing activities necessary to fully understand the results of a data mining application [3]. Information visualisation can be also considered in some way a data mining technique itself, and the one that involves more interaction between users and system. Indeed, graphically presenting data may allow user to discover new and useful properties, their correlations and also detect possible deviations from the expected values. Use of colour may highlight data aggregations, use of animation may allow to quickly go through multiple levels of details (see for example [4]). A further advantage of this type of techniques is that users do not need to know what kind of phenomena they should observe in order to discover anything interesting or unusual. We are confident that useful techniques can be derived from the information visualization area to support the work performed by machine learning and data mining communities for analyzing huge amount of data.
In this paper, we describe how some visualization techniques can be advantageously applied in FAIRWIS (trade FAIR Web-based Information Services), a project funded by the European Union, which aims at offering on-line innovative services to support the business processes of trade fairs, both real and/or Web-based virtual fair. More specifically, the paper is organized as follows. In Section 2, we briefly survey some information visualization prototypes that primarily inspired our work. We then discuss some information visualization guidelines in Section 3. Section 4 describes the FAIRWIS project, while Section 5 illustrates the visualization techniques that are exploited to allow easy access to the stored data, and to present the retrieved information in appropriate ways, thus helping users in their data analysis activities.

2 Information Visualization Prototypes

We are all familiar with direct manipulation interfaces; their success testify the power of using the computer in a more visual manner. Direct manipulation is based on some fundamental concepts, such as the visualization of actions and objects of interest, the use of fast, incremental and reversible actions, and the immediate visualization of the result. Visual displays give the possibility of showing relationships by proximity, containment, connected lines, color coding, etc. Highlighting techniques, like blinking, brightening, reverse video can be used to focus the attention to specific items among thousands of items. Rapid selection can be performed by pointing to a visual display.
By visually presenting information, we exploit the potentiality of visual perception of human beings. Visual presentations are particularly useful since they allow users to activate perceptual procedures to quickly obtain the desired results. Such procedures substitute the logical inferences the user should perform without a visual presentation.
Exploring large multi-attribute databases is greatly facilitated by presenting information visually. Among different visualization techniques of databases proposed in the literature, Ahlberg and Shneiderman have proposed starfield displays [5], that plot items from a database as small selectable spots (either points

or small 2D figures) using two of the ordinal attributes of the data as the variables along the display axes. The shown information can be filtered by changing the range of displayed values on either axes. If this is done incrementally and smoothly, the result is zooming in and out on the starfield display, and the user can track the motion of the spots without getting disoriented by sudden, large changes in context.
The values of other attributes of the database can also be varied by the user through appropriate widgets that allow performing dynamic queries [6]. This is a very interesting visual query formulation technique (see [7] for a classification of such techniques), based on range selection, i.e. it allows a search conditioned by a given range on multi-key data sets. The query is formulated through direct manipulation of graphical widgets, such as buttons, sliders, and scrollable lists, with one widget being used for every key. The user can either indicate a range of numerical values (with a range slider), or a sequence of names alphabetically ordered (with an alpha slider). Given a query, a new query is easily formulated by moving the position of a slider with a mouse; this is supposed to give a sense of power but also of fun to the user, who is challenged to try other queries and see how the result is modified.
Higher usability is ensured if the query results fit on a single screen and are displayed quickly, i.e. within a second [8]. Moreover, input and output data are of the same type and may even coincide. As a consequence, dynamic query applications typically encode multi-attribute database items as dots or colored polygons on a starfield display.
A first application of dynamic queries is shown in [6] and refers to a real-estate database. There are sliders for location, number of bedrooms, and price of houses in the Washington, D. C. area. The user moves these sliders to find appropriate houses. Retrieved ones are indicated by bright points on a Washington, D. C. map shown on the screen. Another interesting application that combines dynamic queries and starfield displays is FilmFinder [5]; it allows information about movies to be retrieved by providing names of actors, actresses, or movie directors through Alphasliders, or values of other attributes through appropriate range sliders and buttons. The user can select some values by using a slider, and this first choice determines the set of values that can be selected with the remaining widgets. For example, if the user has selected a specific movie director, only names of actors and actresses who worked with that director can be selected next. This strategy is called tight coupling and it is aimed at preventing users from specifying null sets. In other words, query widgets and their related query formulation mechanisms are designed to interact with each other to avoid empty query results; this is achieved by restricting users to specify query criteria that lead to non-empty results. A tightly coupled query is then a series of filters selecting a subset of a database. For each new filter that is set, users can only select values of the remaining filters that let through at least one database object still existing after the last filter.
Dynamic queries are also called direct-manipulation queries, since they are based on the same fundamental concepts of direct manipulation illustrated above. One of the big advantages of such interaction technique is that it allows focusing the attention on the task users have to perform. Objects of interest are all displayed so that actions occur in the high level semantic domain. Each command is a comprehensible action in the problem domain whose effect is immediately visible; this relieves the user from the burden of decomposing tasks into syntactically complex sequences, thus reducing user load in problem solving. The sliders are a

good metaphor for the operation of entering a value for a field in the query: changing the value is done by a physical action instead of entering the value by a keyboard. Such action is easily reversible by moving the drag box, if the obtained results are not what users expected. No action is illegal, hence error messages are not needed. More references to work on dynamic queries can be found in [9].
At Xerox PARC in the last ten years a group of researchers has developed several information visualizations, with the aim of helping the users understand and process the information stored into the system [10, 11, 12, 13]. They have created the "information workspaces", i.e. computer environments in which the information is moved from the original source, such as networked databases, and where several tools are at disposal of users for browsing and manipulating the information. One of the main characteristics of such workspaces is that they offer graphical representations of information that facilitate rapid perception of the overall patterns. Moreover, they use 3D and/or distortion techniques to show some portion of the information at a greater level of detail, but keeping it within a larger context. These are usually called fisheye techniques [14], or alternatively focus + context, that better gives the idea of showing an area of interest (the focus) quite large and with detail, while the other areas are shown successively smaller and in less detail. Such an approach is very effective when applied to documents, and also to graphs [15]. It achieves a smooth integration of local detail and global context. It has more advantages of other approaches to filter information, such as 1) zooming or 2) the use of two or more views, one of the entire structure and the other of a zoomed portion; the former approach shows local details but looses the overall structure, the latter requires extra screen space and forces the viewer to mentally integrate the views. In the focus + context approach, it is effective to provide animated transitions when changing the focus, so that the user remain oriented across dynamic changes of the display avoiding unnecessary cognitive load. The Perspective Wall [10] provides a good example. For other techniques developed at Xerox PARC see [11].
Numerous prototypes have been proposed for information visualization. The ones mentioned above are among those providing novel ideas that have inspired our work. A very good reference for a survey of information visualization techniques is [16].

3 Supported Tasks in Information Visualization

There are many visual design guidelines. A central principle for information visualization might be summarized in the Shneiderman's Visual Information Seeking *Mantra "Overview* first, *zoom* and *filter*, then *details on demand*" [17]. The overview allows the user to grasp the content of the application and its distribution across the different attributes. Providing an overview is particularly useful in WWW interfaces for information systems that give users direct access to the content and interconnections within an information domain. WWW navigation should be stimulating and attractive for the users; unfortunately, due to the large amount of accessible information, the search of some detailed information can often become a long and complex activity for the user. One of the main problem is the

difficulty users have in generating their mental model of the system they are interacting with; it can be difficult for them to grasp the kind of information stored and the modality for managing it. Such a problem is particularly serious since WWW interfaces are mostly used by occasional users, who are not willing to perform an in-depth study, but need to easily grasp the kind of information they can have and want to get it quickly.
Zooming is another interesting task, since users typically have an interest in some portion of a collection, and they need tools to enable them to control the zoom focus and the zoom factor. A satisfying way to zoom in is to point to a location and to issue a zooming command. Smooth zooming helps users to preserve their sense of position and context. Another popular approach for keeping the context while zooming some areas of interest is the fisheye strategy [14]; the fisheye distortion magnifies one or more areas of the display.
Users may filter out uninteresting items, so that they can quickly focus on item of interest. Dynamic queries applied to the items in the collection constitute one of the key ideas in information visualization [5]. Sliders, buttons, or other control widgets coupled to rapid display update are used for the filter task.
We can select an item or a group of items to get details. Once we have obtained a few dozen of items, it should be easy to browse the details about the group or individual items. The usual approach is to simply click on an item to get a pop-up window with values of each attribute. In Spotfire [18], the details-on-demand window can contain text with links to further information.
Besides the four tasks explicitly mentioned in the Shneiderman's Mantra, three other tasks are very useful in information visualization, namely *relate*, *history*, *extract*. Referring to the first, users can view relationship among items. In the FilmFinder details-on-demand window [5] users could select an attribute, such as the film's director, and cause the director Alphaslider to be reset to the director name, thereby displaying only films by that director. The Table Lens emphasizes finding correlations among pairs of numerical attributes [11].
We can keep a history of actions to support undo, reply, and progressive refinement. Information exploration is inherently a process with many steps, thus keeping the history of actions and allowing users to retrace their steps is important.
Once the users have obtained the item or the set of items they desire, it would be useful for them to be able to extract that set and to store into a file in a format that would facilitate further uses, such as sending by e-mail, printing, inserting into a presentation package. As an alternative to saving the result set, they might want to save the settings for the control widgets.

4 Data Analysis in Trade Fairs: the FAIRWIS Project

FAIRWIS (Trade FAIR Web-based Information Services) is an on going project at the University of Bari, funded by EU this project aims to offering on-line innovative services to support the business processes of real trade fairs as well as providing information services to a great number of exhibitors organised in a Web-based virtual fair. FAIRWIS has a real time connection with an underlying database to guarantee coherence of data and up-to-date status.

Traditionally, information media for supporting trade fair events is paper-based: booklets, flyers, maps, etc. are the means used to exchange information. In recent years, some Web-based information sites have been made available, providing information both on trade fair events and on companies participating in these fairs. However, these data are not organised in an integrated, homogeneous and comprehensive way, since are usually presented in a rigid pre-designed company oriented style. Moreover, currently available Web sites exploit static data that it is difficult to update and to put on-line in an appropriate format.

Presenting data on the WWW in a convincing and understandable way requires a lot of work when data change dynamically; in particular it is difficult to modify the graphical layout without disorienting the users. We describe a module of the FAIRWIS system, devoted to the generation of data visualizations that are exploited to facilitate human-computer interaction and to allow easy access to the stored data. By presenting the retrieved information in appropriate ways, specific categories of user to whom FAIRWIS is primarily addressed, namely fair organisers, exhibitors, and professional visitors (people who visit the fair for business reasons and not only for fun), will get a valuable help in the different phases of the decision making processes they may undergo to improve their own business.

FAIRWIS aims to both support real trade fairs and offer on-line innovative services regarding a virtually unlimited number of companies, products and events. The whole concept of trade fairs is transferred into an electronic form, and visualisation techniques, including virtual reality, are used in order to provide “reality” feelings to the users of trade fair information systems. The project does not aim at substituting, but at enhancing the existing traditional approach of getting people together.

The software module here described presents a WWW interface that allows users to easily retrieve information useful for their marketing activities. The aim of the FAIRWIS marketing component is to manage and improve interactive relationships among the FAIRWIS users. More specifically, fair organisers and company managers may forecast their company activity on the basis of history data available in the database. To this purpose, they need techniques that enable them to discover specific trends of the stored data. To provide support to the users in this process, we have developed a prototype that exploits an information visualisation technique known as query previews [19]. It is presented in the next section.

5 Analyzing FAIRWIS Data Through Query Previews

Within the FAIRWIS project we examined hundreds of Web sites related to trade fair events and we discovered that most of them do not provide any mechanism for analysing the data they may possibly store in underlying databases. A web site that makes some kind of data analysis is CIBUS, a fair event of Fiera di Parma trade fair, which stores in a database data of the exhibiting companies, and makes them available for the final user on the WWW [20]. The users can make query by choosing the filter criteria of several attributes, thus retrieving the companies that fit their needs.

As shown in Figure 1, the user may choose among the attributes such as the

company sector (*settore di produzione*), the product type (*tipologia*), the company income (*classe di fatturato*), the number of employees (*numero dipendenti*) and so on. When the user has selected some of these attributes, he or she clicks on the button "*inizia ricerca*" to retrieve the companies of interest. This kind of interface has some drawbacks. It does not provide the user any hints about the actual content of the database, for example the user doesn't know if there are exhibitors whose company sector is *carni congelate* or if there are exhibitors coming from a certain town. As a consequence, after the user has spent time for inputting several attribute values, very likely he or she gets as result an empty dataset, as shown in Figure 2, where the message in Italian says that the search did not find any company.

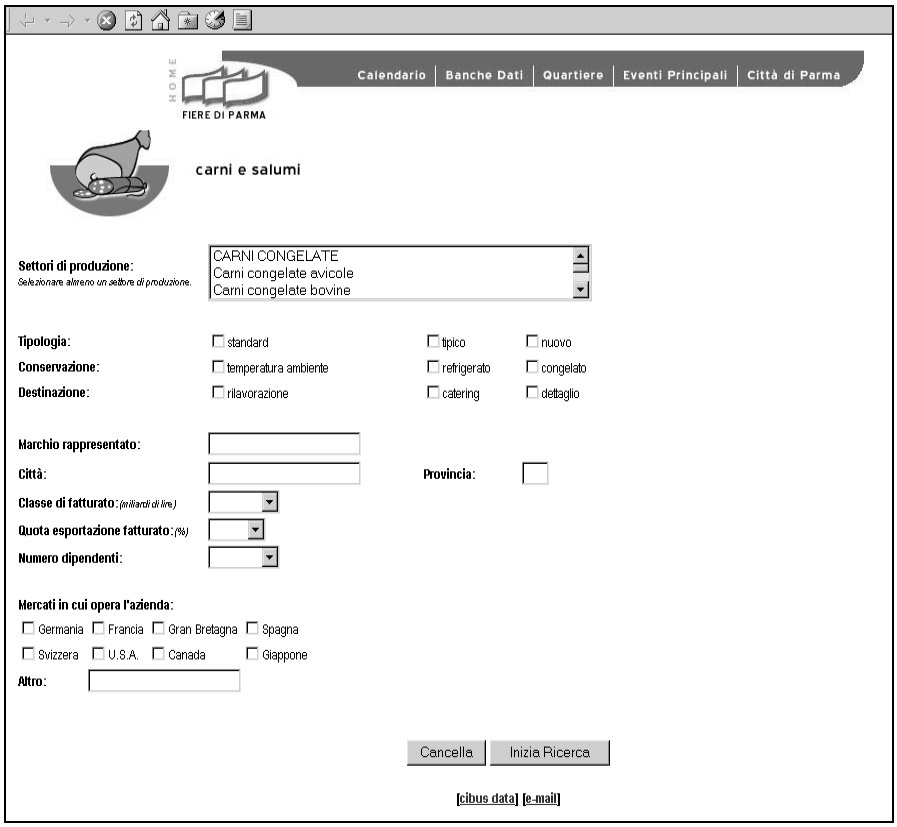

Fig. 1. A screenshot of CIBUS: the user can input values of several attributes to find the companies of interest.

Fig. 2. The result of a query in CIBUS often gives an empty data set.

To overcome these problems, in FAIRWIS we designed a visualization tool that provides users a rapid overview of the information stored in the fair database in order to support the users' data analysis. This overview shows the data distribution along some major attributes. Then, we use dynamic queries and query previews to support efficient query formulation [19]. As we said in Section 2, dynamic query user interfaces apply the principles of direct manipulation and imply: 1) visual representation of the query and of the results; 2) rapid, incremental and reversible control of the query; 3) selection by pointing (no typing); 4) immediate and continuous feedback.

Query preview interfaces provide the possibility of easily getting preliminary information about data interesting for the user, making visible the problems or gaps in the metadata that are undetectable with traditional form fill-in interfaces. In this way, the user may rapidly eliminate undesired datasets and also preview the size of the result set to avoid the so-called zero-hit queries, i.e., queries that provide an empty set as result.

In order to see how a query preview interface works, let us refer to the FAIRWIS prototype. Let us suppose the organiser of a fair on agriculture wants to perform a segmentation of the exhibitors of the last edition of the fair. In order to help the users (in this case fair organisers) in their analysis, we allow them to some major attributes for generating a data overview and then perform some query previews.

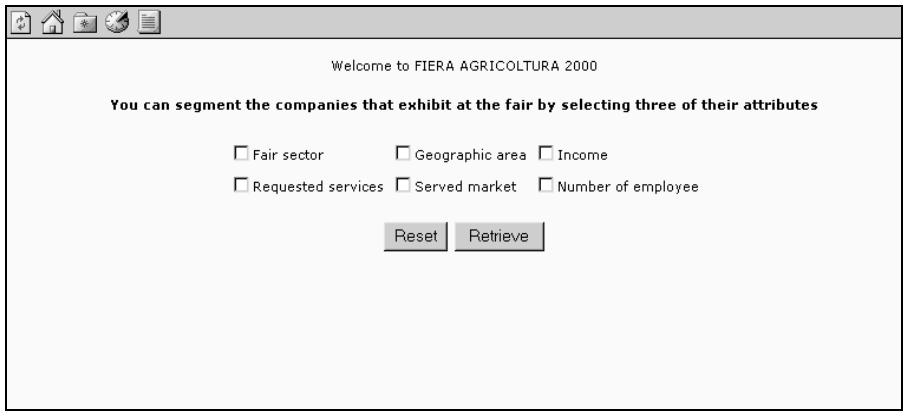

Fig. 3. A first Web page in which a user is invited to select three major attributes.

In Figure 3, the user can select some attributes (three in this prototype) of interest. Let us to suppose that the user selects the attributes *Fair sector*, *Geographic Area*, *Requested services*. The resulting overview is shown in Figure 4. In the overview, data distribution is displayed along these attributes.

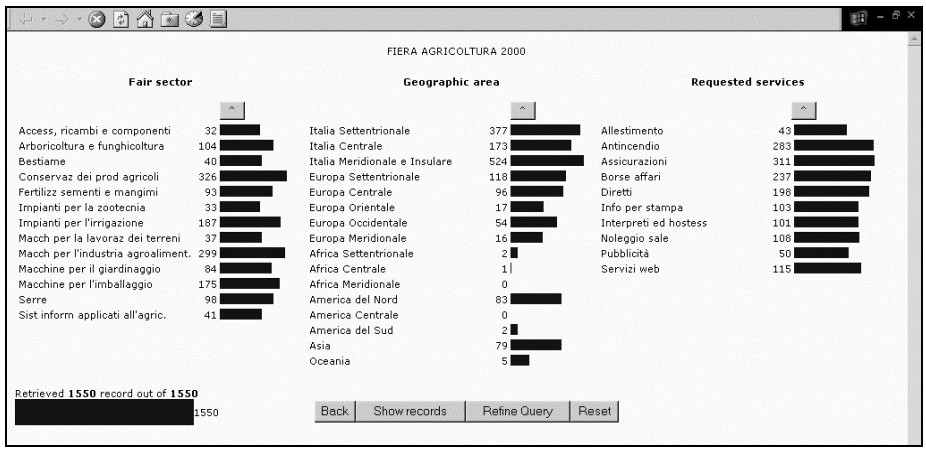

Fig. 4. A Query Preview interface for Fiera Agricoltura 2000.

The user immediately gets a lot of information from the overview: for example, no company comes from South Africa (*Africa meridionale*) so that it is useless to perform any query with *Africa meridionale* as *Geographic Area* since it will return a empty data set. The user also sees that 96 exhibitors come from Central Europe (*Europa centrale*), and so on.
The interface allows the user to perform previews of data, for example by clicking on the value *Conservaz dei prod agricoli*, only the records with this attribute value

will be selected and the number of retrieved data is updated consequently, as shown in Figure 5. We also provided the possibility of sorting the shown elements. The user can sort the values in numerical order by clicking on the icon on top of the numerical value of the retrieved record.
If the query preview shows too many records, it would not be useful to visualize all of them. In this case, the system allows query refinement by clicking on the button "Refine query". The query refinement phase supports dynamic queries over other relevant attributes of the database. In this way, the user can get a reduced set more meaningful for his or her interests. Once we get the list of the retrieved records, details on a specific record can be obtained by clicking on an element of the list, and a window with all available information on that specific company appears on the screen.

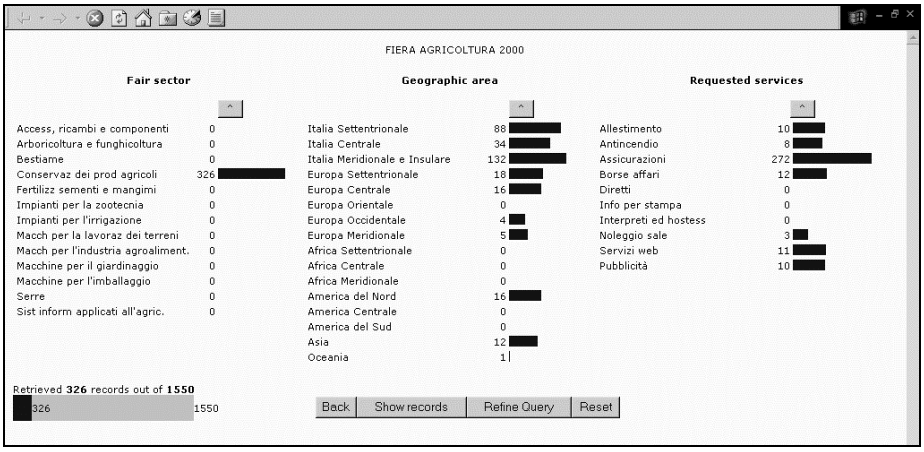

Fig. 5. Results of a query preview. The user has selected the value *Conservaz dei prod agricoli* for the attribute *Fair sector* and only the 326 companies exhibiting in that sector are retrieved.

6 Conclusions

The prototype presented here is part of the on-going project FAIRWIS carried out with other partners in the European Union. FAIRWIS primarily addresses the needs of professional users in the fair context, namely fair organisers, exhibitors, and visitors attending the fair for business reasons.
We have shown some functionalities of a data analysis engine module that allows users to easily retrieve information useful for their marketing activities. The module exploits information visualization techniques that are capable to give to the users useful hints on meaningful patterns of data. As Klösgen says, "knowledge discovery in databases can be divided into paradigms such as search, visualization, navigation, and low level strategies for searching and evaluating patterns. Visualization gives a

feeling for the contents of the data and presents findings" [21]. Our work is in accordance with this perspective.
Other tools are under development in order to allow the FAIRWIS users to manage their interactive relationships, and to improve their company activity by letting them to directly take full advantage of the data available in the system database.

Acknowledgement

We are grateful to Marco di Fonzo for his help in implementing the prototype. The support of European Commission through grant FAIRWIS IST-1999-12641 is acknowledged.

References

1. Card, S., Eick, S. G., and Gershon, N. (1997) "Information visualization", CHI97 Tutorial Notes, Atlanta, GA, March 1997, pp. 22-27.
2. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds). Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, 1996.
3. B. Moxon, "*Defining Data Mining*", DBMS online, 1996, http://www.dbmsmag.com/9608d53.html
4. Silicon Graphics, "*MineSet*", http://www.sgi.com/software/mineset, 2001.
5. Ahlberg, C., Shneiderman, B. "Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays", Proceedings ACM CHI'94: Human factors in computing systems, pp. 313-317.
6. Shneiderman, B., Williamson, C., Ahlberg, C. "Dynamic Queries: Database Searching by Direct Manipulation", Proceedings ACM CHI'92: Human Factors in computing Systems, pp. 669-670.
7. Catarci, T., Costabile, M. F., Levialdi, S., Batini, C. "Visual Query Systems for Databases: a Survey", Journal of Visual Languages and Computing, Vol. 8, 1997, pp. 215-260.
8. Ahlberg, C., Williamson, C., Shneiderman, B. "Dynamic Queries for Information Exploration: An Implementation and Evaluation", Proceedings ACM CHI'92: Human Factors in computing Systems, pp. 619-626.
9. Shneiderman, B. "Dynamic Queries for Visual Information Seeking", IEEE Software, 11, 1994, pp. 70-77.
10. Robertson, G.G., Card S.K., Mackinlay, J.D., "Information Visualization Using 3D Interactive Animation", Communications of the ACM, 36 (4): 7-71.
11. Rao, R., Pedersen, J.O., Hearst, M.A., Mackinlay, J.D., Card, S.K., Masinster, L., Halvorsen, P.-K., and Robertson, G.G. (1995) "Rich interaction in the digital library", Communications of the ACM, 38 (4): 29-39.
12. Card S. K., G.G. Robertson, W. York "The WebBook and the Web Forager: An Information Workspace for the World-Wide Web", Proceedings CHI'96, April 13-18, 1996, pp. 111-117.
13. Card S.K., "Visualizing Retrieved Information: A Survey", IEEE Computer Graphics and Applications, March 1996, pp. 63-67.
14. Furnas, G.V., "Generalized Fisheye Views", Proceedings CHI'86 Conference: Human Factors in Computing Systems, ACM Press, New York, pp. 16-23.

15. Sarkar M., Brown M.H., " Graphical Fisheye Views", Communications of the ACM, 37(12): 73-84.
16. S. Card, J. MacKinlay, B. Shneiderman. Readings in Information Visualization. Morgan Kaufmann Publisher Inc., 1999.
17. Shneiderman, B. (1996) The eyes have it: A task by data type taxonomy for information visualization, *Proc. 1996 IEEE Symposium on Visual Languages*, Boulder, Colorado, September 1996, IEEE Computer Society Press, Los Alamitos, California, pp. 336-343.
18. C. Ahlberg, IVEE Development AB, "*Spotfire*", http://www.ivee.com, 2001.
19. Shneiderman B., C. Plaisant, K. Doan, T. Bruns, "Interface and Data Architecture for Query Preview in Networked Information Systems", *ACM Transaction on Information System*, vol. 17, No.3, 320-341, July 1999.
20. http://www.fiere.parma.it/CibusData/irpc0.htm
21. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds). Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, 1996, pp. 249-271.