A framework for Visual Data Analysis

by

Paolo Buono
buono@di.uniba.it

Dipartimento di Informatica
Università degli Studi di Bari

Promotor: Prof. Maria Francesca Costabile

A dissertation submitted in partial satisfaction of the requirements
for the degree of
Doctor of Philosophy in Computer Science
in the Graduate Division of the
University of Bari, Italy

Credits

This dissertation was typeset using these shareware programs:

- TeXnicCenter 1 Beta 6.21
  available at: `http://www.texniccenter.org/`
- MikTEX 2.1
  available at: `http://www.miktex.de`

Submitted *January 2005*

Contents

List of Figures

Acknowledgments

I would like to express my gratitude to my advisor Prof. Maria Francesca Costabile for her wonderful guidance, support and for the energy that she brings everyday.

I would like to thank Rosa, Tatiana, Carmelo, Tony for the good times and many discussions we had together, and also for all the days they brought me food or we shared lunch, brunch, dinner, or breakfast.

I am also grateful to all the members of the Human-Computer Interaction Laboratory at the University of Maryland where I spent four months in 2004 working on TimeSearcher, which became part of this thesis. In particular I thank Catherine Plaisant who provided feedback and encouragements during the writing of the thesis.

I thank also Annalisa for her encouragments, and the pleasure to work with her. Even thought she has many complaints, we know that she cares about us.

LAPI, because whenever I arrive at the office.. he is already there! and Stefano, because when I leave the office.. he is still there! Thanks to the ever-hungry Michelangelo, and DM - that one could think is Data Mining but is really Donato Malerba, and also Oriana, Nico, Mara, Oronzo, Michele, Marco G. and everybody I don't list here, because of the space, not because I forgot to mention them.

I also thank Marco and Pasquale, who for the second year we are doing a nice co-residency.

I would like to acknowledge the support of my colleagues from the FAIRWIS and FairsNet projects in which I was involved from 2000 to 2004, they were crucial to the making of this thesis.

Of course, I am grateful to my parents for their patience and love.

Bari, Italy *Paolo Buono*
January 14, 2005

Abstract

The evolving demands of the market and the increasing competition force companies to support their business plans with tools that help managers to make decisions in a rapid and more effective way. Presenting data in a convincing and understandable way requires a lot of work when data changes dynamically.

Business Intelligence refers to concepts and methodologies of different disciplines whose aim is to provide decision support by transforming the information usually stored in huge distributed databases into knowledge useful to optimise the company processes as well as the customer relationships. Business Intelligence includes data mining techniques that aims at extracting patterns or models or relationships among the data that can be easily interpreted and understood.

In the data mining research, very little attention has been devoted so far to human-computer interaction (HCI). HCI might allow the user to get inside the data or to steer the data mining algorithm; it might also help to represent prior knowledge, so that the data mining algorithm does not rediscover what is already known.Visual Data Mining (VDM) is emerging area in explorative data analysis and mining that may provide a good contribution along this direction. VDM refers to methods for supporting exploration of large data sets by allowing users to directly interact with visual representations of data and dynamically modify parameters to see how they affect the visualized data. The graphical presentation of the data allows users to discover specific patterns, as well as new and useful properties in the data, their correlations, and also detect possible deviations from the expected values.

Using good visualizations to present the information hidden in various company repositories can improve the decision process. Advances in information visualization offer promising techniques for presenting knowledge structures and for permitting explorative analysis of the data.

This thesis illustrates a framework for VDM that we have developed as a result of the work within the European funded project FairsNet (IST-2001-34290). The aim is to provide various visualization techniques to assist the users in their decision making processes. Even if FairsNet focused on a specific application domain, namely trade fair management, the framework is quite general and applicable to different domains.

Chapter 1

Background and motivation

1.1 Introduction

The enormous amount of data available on the web must be adequately exploited by company managers in order to improve their business. An important role is played by various techniques that are capable of extracting useful information. Traditionally, these are data mining techniques based on either statistical methods or machine learning methods. The approach adopted in this thesis enhances the use of classical data mining algorithms with the use of visualization techniques with various purposes, according to an emerging area in explorative data analysis and mining that is Visual Data Mining (VDM) [186, 129, 10, 86, 185, 64].

Shneiderman calls computational tools for discovery both data mining tools and information visualization tools [179]. They have advanced dramatically in recent years, but they have been developed by largely separate communities with different philosophies. Data mining and machine learning researchers tend to believe in the power of statistical methods to identify interesting patterns without human intervention. Information visualization researchers tend to believe in the importance of user control by domain experts to produce useful visual presentations that provide unexpected insights.

VDM is essentially a combined approach that exploits both classical data mining algorithms and visual representation techniques. This combined approach can lead to new tools that enable effective data navigation and interpretation, preserve user control, and provide the possibility to discover anything interesting or unusual without the need to know in advance what kind of phenomena should be observed.

In this thesis we present a framework for VDM that we have developed as a result of the work within two successive European funded projects: FAIRWIS (IST-1999-12641) and FairsNet (IST-2001-34290). Trade fairs is the application domain of both projects. FairsNet is actually a trial that has been funded with the objective of deriving a system ready for the market from the prototype developed as result of FAIRWIS. In the following, we will refer only to FairsNet. Its aim is to offer on-line innovative services to support the business processes of trade fairs and to provide information services organised in a web-based system. The project is described in the following sections.

1.2 The FairsNet project

1.2.1 Trade fair domain

Advances in technology give the possibility to virtually be ubiquitous, using several medium for communication, from the 3G mobile to video conference systems. However, people still need to physically meet and shake hands with other persons who are customers, competitors, suppliers, etc. To do this, one possibility is to participate to a trade fair. Trade fairs represent an indispensable medium for exchanging ideas, relationships and information within a given business area.

Developing a Web application for supporting a trade fair business process poses the main challenge of how the information, communication, and knowledge management functionalities that are required should be designed and efficiently implemented to best support the trade fair business process. Due to the particularly dynamic global market in which the trade fair business process is positioned, the requirements toward the Web application change as frequently as the market's tastes and trends do [143]. In Barbini et al. [16] there are some indications on how important trade fairs are in the objectives of companies. In the European market, trade fair investment ranks as 10% of total advertising investment, and in the case of companies operating in the business to business area this figure reaches 40%. Trade fairs are an integral communication facility, which allows simultaneously commercial promotion, advertising, face to face contacts, selling actions, public relations and market research.

Nowadays, information media for supporting trade fair events are still mostly paper-based. Booklets, flyers, maps, etc. are usually the means used to exchange information. Web sites have been created to provide information both on trade fair events and on companies participating in these fairs. However, these data are often not organized in an integrated, homogeneous and comprehensive way, since they are usually presented in a rigid pre-designed company oriented style. Moreover, currently available web sites exploit static data that is difficult to update and to put on-line in an appropriate format.

FairsNet has a real time connection with an underlying information system that guarantee coherence of data and up-to-date status. FairsNet aims at enhancing the existing traditional approach of getting people together by means of trade fair, not at replacing the traditional business process. The system gives a technological support that facilitates and speeds up contacts among companies. Indeed it provides a means for allowing visibility of small and medium enterprises within the market.

In FairsNet, the whole concept of trade fairs is transferred into an electronic form, and visualisation techniques, including virtual reality. Environments generated by a graphical engine are used in order to provide “reality” feelings to the users of trade fair information systems, to allow the users to grasp the knowledge stored in the database, and improve human-computer interaction.

FairsNet primarily addresses three types of users: fair organizers, exhibitors, and professional visitors, that are described in Section 1.2.3. The data analysis functionalities we have developed aim at providing a valuable help to these users in different phases of the decision making process they may undergo to improve their own business. Some relevant information in the trade fair database is related to data about companies involved in the fair. Such data are exploited to provide company managers a valuable support for improving their activities during the various phases of the fair business process. FairsNet supports users’ activities strictly related to the trade fair, such as trade fair preparation, stand planning, and also to marketing activities that can be carried out during and even beyond the fair. For example, a company manager may be interested in some business data stored in the trade fair database in order to retrieve a set of companies with similar with mutual interests. Once an appropriate set of companies has been identified, the decision maker can start some contact activity or a marketing campaign. We developed several software components that support the exhibitor in this kind of analysis.

1.2.2 FairsNet overview

FairsNet is a flexible, Web-based IT system, designed to support a trade fair organizer’s objectives to seamlessly manage on-line activities related to real trade fairs as well as to provide information services to a large number of fair customers.

FairsNet is composed of four main components:

1. Core system, which permits to generate and manage an on-line trade fair web-site;

2. User Profile Engine (UPE), which provides system personalization;

3. 3D Engine, which provides virtual reality representations;

4. Data Analysis Engine (DAE), which provides various tools for the analysis of the stored data.

The Core component is composed by a set of tools that supports the activities of managing and running the trade fair. A good deal of current research in Web-based applications is aimed at enabling an application to adapt its own behavior to the users characteristics, such as goals, tasks, interests, that are stored in user profiles. We performed a survey on the trade fair domain and we discovered that personalization is not available in any of the hundreds of trade fair web sites we analyzed. Personalization applied to a Web site is a process of gathering and storing information about visitors of that web site, analyzing the stored information, and, based on this analysis, delivering the right information to each visitor at the right time. It is increasingly used as a mean to make the site useful and attractive in order to stimulate the visitor and consequently to tempt it to visit the Web site again. Our approach integrates the data the system collects about users, both explicitly and implicitly in order to provide recommendations to the user during the visit of the on-line fair catalog. More details about UPE are available in [38, 41, 40, 37, 35, 36].

Three factors characterize the FairsNet system architecture:

1. database independence of the system and tools for a comfortable integration of existing databases into the Web application;

2. the introduction of a powerful and flexible form-based user interface model;

3. the use of the Web services as paradigm for components interaction.

These factors facilitate integration with legacy systems and existing application database, as well as enable low-cost deployment, which is a must for the targeted trade fair market. Database independence is achieved by relying on an architecture that uses standards like ODBC for accessing the database and by restricting database operations to those supported by all commercially available relational and object-relational databases. Furthermore, a database independent RDF-based domain model is introduced together with tools for its navigation, inspection, and adaptation. This provides an additional layer of abstraction that eases the effort of integrating other types of databases like native XML databases already present in the managers IT infrastructure. DAE is integrated in the overall system through this model. The introduction of a powerful and flexible form-based user interface model makes the FairsNet Web application flexible, effective and responsive [207, 27]. The form-based user interface model is based on XForms [1]. Its innovative features are:

- extended client side interactions enabling the construction of dynamic forms that adapt to the current business step requirements, e.g. by dynamically adding new form elements;
- client-site validation of structural as well as value-based constraints on the user interface instance data;
- a clear conceptual separation of data, control, and layout following the model view controller approach, which enables the support of different UI agents and even of service interfaces, and the systematic support for flexibly grouping form elements, including declarative approaches for dynamic collections of instance data, as well as for the grouping of entire forms into larger, meaningful units with respect to the implemented business process (multi-form dialogs).

For component interaction the Web service paradigm is chosen [107]. System components are wrapped as Web services that can be activated using the Internet Infrastructure. Relying on standards like XML and SOAP[2], the Web service paradigm enables dynamic, platform independent component interaction and integration and, thus, contributes to the openness and extensibility of the system. Furthermore, it facilitates the integration of components built on other platform on external third party components.

FairsNet is a system to support trade fair events both in the real fair environment and in the virtual one. In fact, one of the main objec-

[1] XML application that represents the next generation of forms for the Web. By splitting traditional XHTML forms into three parts—XForms model, instance data, and user interface—it separates presentation from content, allows reuse, gives strong typing—reducing the number of round-trips to the server, as well as offering device independence and a reduced need for scripting [210].

[2] Short for Simple Object Access Protocol, a lightweight XML-based messaging protocol used to encode the information in Web service request and response messages before sending them over a network. SOAP messages are independent of any operating system or protocol and may be transported using a variety of Internet protocols, including SMTP, MIME, and HTTP.

tives of FairsNet is to create a web based virtual system that operates simultaneously with the real (physical) event. Moreover, a system like FairsNet can be a valuable support for each of the three typical phases of a trade fair event: the organization before the event, during the event, and after the trade fair is ended [34, 40, 39]. The FairsNet partners performed various activities in order to define the general requirements of the FairsNet system. They are described in Section 5.4. Next section describes the types of users that emerged from the user analysis as the most relevant for FairsNet.

1.2.3 Users in FairsNet

Before the development of FairsNet we performed a deep study of the requirements. We performed several activities to define the user requirements described in [16]. Detailed analysis of users involved in the business of trade fairs, the environment in which they operate and the tasks they perform have been carried out. The study has been performed by the various project partners working with users in Italy, Spain, and the United Kingdom. This section describes the types of users that emerged from the user analysis.

Venue owner/manager. The venue owner/manager is the organization or individual who owns the hall, the exhibition center, etc. They sell space in the venue to the event organizer together with a number of specific services. They are not usually involved in the organization nor in the running of the event. This type of user was primarily investigated in UK, where venue owners are typically local authorities or commercial companies for medium to large venues. The venue owner/manager could use DAE to search event organizers and/or events which could be interested in using the venue. In particular it could use DaeQP described in Section 4.9, which is a tool that allow to segment potential exhibitors that have some specific characteristics.

Organizer. The organizer is the organization or individual responsible for organizing the events. It may be a venue owner or a different user. The organizer may be a professional company specialized in this activity or one for which organizing events in a particular market area is complementary to its main business. FairsNet could be of great benefit to this user in many ways, some examples are:

- Helping in planning and organizing trade fairs;
- Searching venues for events and exhibitors;
- Managing mailing lists and mail shots;
- Managing the build-up and end-up processes.

FairsNet also has the potential to provide information and services to fair visitors and to the press before, during and after the event. Due to the effects of the Internet on global economy, organizers nowadays work, compete, and cooperate on a worldwide scale. They must be able to cope with the fast pace of worldwide competition by being highly innovative in the products and the services they offer. Because of the dynamic nature of involved processes, there is a need for new frameworks and integrated computational environments to model, develop and support business processes in all their steps. Officiating a trade fair event is only one part of the tasks of a trade fair organizer. This most visible activity is preceded by other activities such as exhibitor acquisition and resource reservations, and followed by a variety of other activities, such as business evaluation and marketing analysis phases. Fundamentally, the organizer's tasks are geared toward planning and managing trade fairs on behalf of the exhibitors by satisfying their desire to increase visibility on the market improving the image of the company and successfully attracting new economic partners. In the trade fair domain there are a lot of activities that an organizer does. In the fair business process three basic phases are found and analyzed: pre-event, running, and post-event. For each phase, the involved users may perform data analysis for different goals.

Exhibitor. The exhibitor is the organization or individual responsible for putting together and running an individual stand or exhibit at an event. Exhibitors will vary from major corporations to small/medium companies or sole traders, depending on the nature of the trade fair/event. DAE helps the exhibitors in searching events to attend and in planning their participation, managing mailing lists and mails, etc.

Visitor/Delegate. Visitors and delegates are also beneficiaries of DAE/FairsNet services (conferences and seminars are often held during the trade fairs). The system could help in providing information in order to select an event to attend, to register for the event, to book additional functions, and to provide information after the event. Visitors range from business/professional visitors (also called *professional visitors*) to the general public and students (called *generic visitors*). They will necessarily span a very large range of characteristics in age, IT-literacy, social class, interests, intelligence and buying power, depending on the focus of the event. In FairsNet we focused on professional visitors because they are the most suitable users for the developed tools.

Contractor/Service Provider. The contractors and service providers are the individuals or organizations responsible for building and setting up the stands. They include: audio-visual equipment providers, catering services, electricians, carpenters, florists, furniture providers, carpet fitters, graphical designers, photographers, stand erectors, telecommunications, etc. They work with the organizer or the individual exhibitors.

Fair workers. This type of user includes all individuals who work in a fair, usually when running a fair. Examples of workers are: hostesses, language translators, florists, etc. Often they get a job through some agencies or service providers, but in some cases they apply directly to the fair organizer, or contact the exhibiting companies. They could be interested in using the system in order to speed up the process of applying for a job at the fair, with the possibility of quickly sending their possible "multimedia" curriculum (including photos and other information) and obtain rapid feedback.

Press and Media The press and media are special types of visitors who could be interested in the information that can be supplied by the system.

Sponsor Sponsorship of events is an important factor, although sponsors are not considered as principal beneficiaries of FairsNet services. Prestigious or at least credible sponsorships can definitely determine the success of an event. Sponsors may typically be banks or financial institutions, major corporations, government bodies, local authorities, etc.

Power user. Members of the trade fair Organizer enterprises are empowered to be special users of the Web application, not only to use the application but also to adapt it to changing requirements. In fact, the task of adapting the Web application can be shifted from a programming task, for which a software expert is necessary, to a design task, which can be performed by members of the trade fair Organizer enterprises, the so-called power user. In this way, time-consuming software re-implementation cycles and the associated communication overhead are avoided, and the stakeholders of the domain expertise (the members of the trade fair Organizer's enterprise) are factually involved in the Web application development life cycle.

As a power user of the FairsNet system, a trade fair organizer is capable of dynamically design and deploy a Web-based application as it is necessary to support the specific business process characteristics of the users of the Web-based application, like trade fair exhibitors, professional visitors, visitors (called the End-User to be distinguished from the Power Users). For example, the power user can design and customize Web application modules to support the trade fair service booking process so that the trade fair exhibitors can navigate and book specific services from the service offering of the trade fair. This customization includes administrative tasks in the area of taxonomy and content management (e.g. defining a taxonomy of services, managing and classifying specific service offers), and by performing Web design tasks to customize the Web pages for the on-line booking of services. It is needed a user that builds the analysis and sets which tools are suitable for a particular type of analysis. The power user should be able to decide what kind of analyses may be performed by what user.

We performed a deep analysis of all the users involved in the trade fair. At the end of the analysis the FairsNet consortium agreed on addressing the system to the three most relevant users: organizer, exhibitor, professional visitor. A comparison of the user involved in trade fair in several European countries (Germany, England, Italy, Spain) and their use of Internet in 2000 was also performed. Results are that in Spain there is a low Internet penetration. The number of active Internet users in Spain is about 4,5 million people (around 12% of Spanish population). According to IDC's forecast, only 18% of Internet users perform online shop. In the UK factors as age, computer/web attitude and experience will vary over a wide range. But it should be also noted that the use of IT and now increasingly web-based services is the norm rather than the exception. Survey in the UK (by Age Concern and Microsoft) indicated that 4 million people over 50 own a computer and one in four of these used it in their spare time. 30% of the UK, population of more than 60,000,000 have access to, and use, the Internet.

The trade fair organizer plays a key role in the trade fair world. Fundamentally, the organizer 's goal is to plan and manage trade fairs that will target specific business sectors. This activity is carried out on behalf of exhibitors, whose goal is to obtain increased visibility and successfully attract new visitors (or economic partners). Organizers provide exhibitors catalogs and information about scheduled events so that visitors are aware of the exhibitors' and organizers' offerings. In addition, the organizer provides services and material resources so that exhibitors can create the best environment to support visitors in their buying and decision-making activities.

Within FairsNet, the organizer assumes an additional role - that of co-designer of the Web application that support the trade fair business through the offer of on-line innovative services. A fair organizer can dynamically incorporate trade fair business specific factors, (such as target market specific business logic, or purchasing options) into the creation and deployment of tailored Web application that fair customers (exhibitors and visitors) can employ for planning and preparing their participation to the trade fair event (for example, exhibitors procuring resources necessary for disseminating information to potential Visitors in a trade fair).

1.3 Thesis main contribution and organization

1.3.1 Main contribution

This thesis illustrates a framework for Visual Data Mining that we have developed as a result of the work carried out within the FairNet project. It is a software component called DAE (Data Analysis Engine) whose aim is to provide various visualization techniques to assist the users in their decision making processes. Even if FairsNet focused on a specific application domain, namely trade fair management, the framework is quite general and applicable to different domains. Several data visualizations are generated to browse among the data and to present the retrieved information in appropriate ways for each user category.

DAE supports several activities involved in the analysis, starting from the data selection through data transformation, until the presentation of the results. DAE is not intended for replacing the current activity of a decision maker, but for integrating his activities by allowing him to directly analyze with alternative tools the data in order to confirm analyses and/or discover new and unexpected insights.

The goal of DAE is to give users the possibilities of putting their “hands on data”. DAE main users are not computer science experts nor they want to become such. They are experts in their application domains, they know about the data of their companies. As designers of interactive software systems, our challenge is to develop systems that domain experts are capable to use for their activities with no intermediary and possibly no training. DAE has been developed with this challenge in mind. In this respect, DAE does not aim to compete with any of the big commercial products of business intelligence that are briefly reviewed in this paper, such as Cognos or Microstrategy. These are complex systems that need specialized computer and human resources to be used. DAE main users are the managers of companies who can directly use some visualization sools that DAE provides so that, thanks to their expertise on the application domain, they can best exploit the visualized data to extract useful knowledge.

A further contribution of this work is that some of the tools developed within DAE have been designed to be general enough and can be used to visualize the results of data mining algorithms when they generate tabular outputs with multidimensional data.

1.3.2 Organization

The rest of the thesis is organized as follows: Chapter 2 presents some background knowledge in all the topics considered in the thesis: Business Intelligence, Data Mining, Information Visualization, Visual Data Mining. Chapter 3 focuses on the related work. Chapter 4 describes the work of this PhD research, there is the presentation of framework for data analysis and the description of all the developed tools, including some examples of interaction with them. Chapter 5 describes the steps followed during the development of the framework according to a user-centred and participatory design approach; it also reports the usability evaluations performed on some tools that are part of the framework.

Chapter 2

Visual Data Mining and related disciplines

2.1 Introduction

There is still confusion about what VDM is and what is not. One of the reasons is because this is a young research area that combines two consolidated research areas: Data Mining and Information Visualization. Depending on the research area to which researchers belong or prefer, the definition may change.

We will refer to VDM as the use of visualization to help data miners in their task of discovering knowledge from data and help decision makers to understand data they interact with.

This means that information visualization techniques should be used to support the knowledge discovery process, from data selection to data presentation. These concepts and methods are applied in several contexts, one of the most relevant domain is Business Intelligence (BI). This section starts describing BI, giving definitions and explaining how the BI process works, then a brief history follows, in order to explain the motivation that make BI important, especially in the business domain. Section 2.3 shows how BI is related to Knowledge Discovery in Databases (KDD). Next two section describe the two main area mentioned above, namely Data Mining and Information Visualization.

2.2 Business Intelligence

2.2.1 Introduction

The need for trade fair, and in general for companies to analyze the data they have in order to make previsions and to perform strategic decisions lead decision makers to use appropriate tools that allow them to do that. Decision makers need to explore data, apply filters, transform data in an appropriate form, perform analyses and look at the results in order to take decisions, this is similar to what business intelligence process is.

Business intelligence refers to concepts and methodologies of different disciplines whose aim is to provide decision support by transforming the information usually stored in huge distributed databases into knowledge useful to optimize the company processes as well as the customer relationships. In this section an overview of Business Intelligence is presented. In particular are presented the relevant terms and task that typically are performed in Business Intelligence, at the end of the section there is a brief history of Business Intelligence, in order to better understand how this research area is born and how it's growth has been done. After the presentation of relevant tasks and definitions in the Section 2.2.3 it is illustrated a brief history of this discipline.

2.2.2 Relevant activities in Business Intelligence

The word Business Intelligence cannot explain the set of activities and tools involved in this area. In this section are presented relevant activities and terms that will be used in the rest of the thesis.

Business Intelligence involves several activities like Enterprise Resource Planning, whose aim is to integrate all the activities of the company in a single application. The existence of different data sources is a real situation in many companies, there are tools that allow the data exchange among different company data sources.

Data need to be analyzed, there are tools that allow users, typically company managers, to query data sources and produce reports that summarize the data status in a given moment. The need for improvements in ERP systems, that basically are built to organize transactional systems has led to the development of new tools that allow to perform quick analyses and to match characteristics that belong to different dimensions. These are DSS tools, that often work with an underlying architecture called data warehouse. Often data warehouse are organized to allow quick analysis along different dimensions. The activity of organizing data with a multidimensional structure and it's analysis is called On-Line Analytical Process.

Enterprise Resource Planning

The ERP acronym means Enterprise Resource Planning (ERP), which should consist in all the activities whose goal is to organize enterprise resources and allow users to make plans for the growth of the company.

Today ERP software, differ from the meaning of its acronym. Planning and resource does not fit to the tools that managers have in their hands. This The relevant part is the enterprise part. The ambition of ERP consists in the attempt of integrating all departments and functions across a company to create a single software program that runs on a single database.

In order to better understand this concept an example is described in the following. A company with several department and a customer buys a products (perform an order). Each of those departments, like finance or human resources, typically has its own computer system, each optimized for the particular department. After the customer places an order, the order begins a mostly paper-based journey from in-basket to in-basket around the company, often being typed and retyped into different computer systems along the way. All that lounging around in in-baskets causes delays and lost orders, and all the typing into different computer systems potentially introduce errors. Meanwhile, no one truly knows the order status.

ERP automates the tasks necessary to perform a business process, such as order fulfillment, which involves taking an order from a customer, shipping it and billing for it. With ERP, when a customer service representative takes an order, he or she has all the necessary information, such as the customer's credit rating and order history, the company's inventory levels and the shipping dock's trucking schedule. Everyone in the company can view the same information and have access to a single database that holds the order. When one department finishes with the order the ERP system automatically route it to the next department. To find out where the order is at any point, one need only login into the system.

In order to have the presented advantages, the company needs to change it's current business process. The changes do not come without pain (migrating to ERP may take years). It's critical to figure out if the current business process will fit within a standard ERP package before doing any change. The migration to ERP has a strong impact in the business process and all the people in the enterprise needs to be trained on the new business process. In addition there is a long list of expenses before the benefits of ERP appear.

Extraction Transform Load

The job for Extraction, Load and Transformation (ETL) tools is to take the data from the source ERP or an OLTP (On-Line Transaction Processing) system then clean and transform them to match the data warehouse schema, and finally load the data in the data warehouse or in a data mart[1]. Many data warehouses also incorporate data from other sources, such as text files, legacy systems, and spreadsheets; such data also requires extraction, transformation, and loading.

[1] see pag. 28 for data mart definition

In its simplest form, ETL is the process of copying data from one database to another. This simplicity is rarely, if ever, found in data warehouse implementations; ETL is often a complex combination of process and technology that consumes a significant portion of the data warehouse development efforts and requires the skills of business analysts, database designers, and application developers.

When defining ETL for a data warehouse, it is important to think of ETL as a process, not a physical implementation. ETL systems vary depending on the particular data warehouse or the particular department data mart within a data warehouse. A monolithic application, regardless of whether it is implemented in Transact-SQL or a traditional programming language, does not provide the flexibility to change to ETL systems. A mixture of tools and technologies should be used to develop applications that each perform a specific ETL task.

Since new data is added to a data warehouse periodically, ETL is used periodically too. Typical periodicity may be monthly, weekly, daily, or even hourly, depending on the purpose of the data warehouse and the type of business it serves. Because ETL is an integral, ongoing, and recurring part of a data warehouse, ETL processes are automated and operational documented procedures. ETL also changes and evolves as the data warehouse evolves, so ETL processes must be designed for ease of modification.

Data warehouses evolve to improve their service to the business and to adapt to changes in business processes and requirements. Business rules change as the business reacts to market influences—the data warehouse must respond in order to maintain its value as a tool for decision makers.

A tight integration between the ETL and the end-user analysis has the potential to provide better insights. Companies want possibly buy one product set to do all these activities, this has led vendors to expand their tool sets in order to be more competitive. There are today several tools on the market: Oracle Discoverer (Oracle); DecisionStream (Cognos); PowerAnalyzer (Informatica); Data Integrator (Business Object) are some examples. In the Section 3.2 are analyzed the most relevant systems for Business Intelligence.

Query and reporting

This activity consists in querying the data source and format the result for the analysis task. During this activity users may query a data warehouse, a data mart or they may query a transaction system. The definition is not clear, because it is possible to see the query and reporting as an operational reporting, but if we consider that the purpose of Business Intelligence is to explore and analyze data to improve profitability or to manage costs. With ever decreasing business cycles and increased urgency for action, decision making has been pushed down throughout organizations. From the executive suite, down through line workers, more and more people need quick access to accurate information so they can make good decisions and take optimal actions.

Even with the proliferation of information and reports it is possible to hear the cry of users throughout organizations because they cannot get the information they need quickly. Data is scattered in multiple locations. It is not easily available nor in usable formats. Each time answers are needed, decision makers must turn to IT for help.

Some Business Intelligence tools allow users to store standard reports in a centralized repository, the stored reports are then accessed to different users of the company and they can take advantage from the work made by others and avoid to do again the same work. Query and reporting capabilities should empower users across an organization not only to answer to specific questions, but to gain the insights needed to create optimal business strategies. Rather than just finding out what happened, the users need to be able to make informed predictions about what will happen next.

Data warehouse

Every little and medium sized company has some level of mechanization of the basic activities of the company management (purchases, accountancy, suppliers, furnishing planning and control, billing, credit management, customer management). The mechanization of those activities has reduced the every day work load of the employees, but few advantage from the market side. During their decision making activity, managers have an ever growing need for an easy and quick access to the information available in management systems.

Decision makers need both the analytic and synthetic information[2].

[2] Kant explains the a logical distinction between analytic and synthetic propositions: in analytic

In order to have the needed information, managers ask to the system administrators to produce reports with the data they need. This activity is not the natural activity of the system administrator, but they are the only people that can satisfy the needs of the managers. In the latest years this need has been kept in the IT market and there have been produced solutions and products for the company decision makers. They take the name of Decision Support Systems (DSS). The infrastructure that supports those systems is called data warehouse. Sean Kelly reports that in 1997 almost all the most important companies in the United States were equipped with a data warehousing strategy [124]. He gives several interesting examples in different sectors. For instance in manufacturing sector the Toyota Motor Sales case that adopted several data warehousing application to improve the management of the car sales. It's worst to mention that the sales are over a million cars each year, the data warehouse approach has reduced the costs and has speeded-up the delivery of the cars. The introduction of data warehouse has led many companies to discover waste on many customers and high profits concentrated in a smaller set of customers, this changed the organization of their work and has permitted to focus on the actual relevant customers. In the sectors of delivery and retail data warehouse are used often to select suppliers, to analyze the demand, to point on the borders for products and customers. An example is McKesson Corp., the biggest distributor for pharmaceutical products in the USA declared that one year after the introduction of data warehouse had a profit of 2 millions dollars. Today data warehouse is widely used in many sectors from medicine to tourism, there are applications in the accountancy too.

Data warehouse definition. Despite data warehouse is used in several context and this term is common today, there are no commonly agreed definition for it. Literally it means warehouse for data. Inmon was the first to talk about data warehouse, defined as a collection of data that are: subject oriented, integrated, time variant, not volatile [111].

Subject oriented. Data in traditional database are organized to optimize the most common operations in a company (adding a new order,

propositions, the predicate-concept is implicitly or explicitly contained in the subject-concept (for example, "A bachelor is unmarried" or "An unmarried male is male"), so the proposition conveys no new information and is true by identity alone; in synthetic propositions, the content of the predicate is clearly not contained in the subject-concept (for example, "Bachelors are unhappy"), so the proposition conveys new information and cannot be true by identity alone [155].

billing information, stock management). These operations are called transaction operations. Data warehouses are organized to support the subjects that influence managers in decision making. Subjects may be customers, suppliers, sales, purchases. While the usual systems focus on the applications available on the company, data warehouse focus on objects, facts, information, not on their manipulation.

Integrated. This is probably the most important characteristic of the data warehouse. This property follows the need of giving coherence to the different data models used by the applications. It is reasonably that the data model are different for each application, since each application is intended for a different goal and use different data, at least they are organized differently. The problem of giving a unique meaning to data that is expressed in different form but represent the same object may be solved with the adoption of a data warehouse. In order to have this homogeneity there are some problems to face that will be presented with some examples. The presented activities are typically covered by ETL processes.

- let us suppose that there are four data source, in each of them the nationality of a customer is stored in a different way; what format is the valid one? which of them is better to use as a valid format?

- applications may use different unit measures to perform computations, for instance in a data source there is an order whose currency is in euros while into another source in the orders dollars are used. In order to guarantee integrity only one currency must be used all the orders must be expressed in that currency;

- more than a source may contain the description of an item, in this case it is needed a decision on what should be the most suitable and comprehension description to be stored in the data warehouse or if it is better to store all the description into a single one (the one of the data warehouse);

- attributes that refers to the same argument may be defined in different ways, for instance a customer ID may be expressed as a string or a number or as an alphanumeric code. There is the need to choose the best way and then to commute all the code to the adopted code.

Time variant. The data stored in a data warehouse have a bigger time range (5-10 years) than data stored into a transactional system

(60-90 days). In a data warehouse are stored all relevant information that show the situation in a given moment or that show a phenomena that last for a very long period. In a traditional database data represent the ultimate situation of the company (the real-time status of data) and often they don't provide the historical evolution of the analyzed phenomenon. This means that the operational data represents the current situation of the company while in the data warehouse data are "snapshots" of the company in a given moment in the past, since they have to support the decision makers and the analysts.

Not volatile. This property indicates that there is no possibility to change data in the data warehouse, since it is read-only, because data warehouses are used to perform analysis about the data, not to manage them. This means that it is easier to design the data warehouse database than a transactional database. Data warehouse designer don't care about the anomalies due to data update and don't care of complex techniques to keep data integrity or techniques to manage access during updates. Typically data are loaded during batch session and then are available to the final users.

There are several reasons for a company to use a data warehouse. Most relevant are:

- Growth opportunity, data warehouse allows the managers to drive and control in the company life, finding new opportunity to increase the market, improving the services for customers, improving the demand of solutions. Indeed, data warehouse allow to deeply know customers and their purchase behavior, thanks to the possibility to perform very detailed and fine tuned analysis on huge amount of data.

- Internal efficiency creation, one of the most common problems in the decision support domain is the representation of the reality. Very often, during the company meetings there are heated discussions and each of the participants has his vision of the truth, and usually this truth is far from the truth of the other participants. This is not weird because we make information from data but information is not unique and it is possible to have different information about the same data, depending on the goal we want to reach and the knowledge we have. A solution to this problem comes with data warehouse, by using a single source to get information and exploiting the collaborative work that is very difficult with transactional systems.

- Costs reduction, a good use of data warehouse allow to reduce operative costs, i.e. typically analysis and decision processes are very expensive and complex; with data warehouse those processes are make easier and costs are reduced. Data warehouse allow also to reduce marketing costs; with a careful use of the data warehouse it is possible to better analyze campaigns, like a new product launch, before the commercialization of the product a good marketing campaign permits to avoid a total failure. The legacy of the company is a mine with plenty of information from whom change new knowledge.
- Strategic decision support, data warehouse is a tool to highlight synthetic information for decision makers, because of the ability to grasp relevant information from huge quantity of data in information systems. Data warehouse is also a powerful way to support knowledge workers, people that do business analysis, that use data warehouses to support their analytic and decisional work.

Company data warehouse and data mart. There are two main data warehouse categories: company data warehouse and data mart.

Company data warehouse, include information about the overall company. Typically they are composed by several topics and subjects, like customers, sales, purchases, etc. Their dimensions ranges from dozens of Gigabytes to several Terabytes, and they are very expensive in terms of design and management. The average time needed to the develop a data warehouse is about one year.

Data marts are data warehouses specific for a company sector, they contain only a subset of the data available in the company. The average time for their development is about weeks or months instead of years, and costs are about dozens of thousand dollars instead of millions dollars. The current trend on data warehouse development is to design and develop big projects, paying attention to the development of data marts first.

Data warehouse architecture. Data warehouse physically stays in the central information system because it is a secondary system. Transactional platforms continuously perform update activities, while, a decision support platform is optimized to accomplish a limited number of complex queries. A data warehouse architecture is presented in Figure 2.1, it is possible to see both operational systems and analytical applications, the data warehouse is built on the operational systems. Data marts may be of two type, those built for reporting purposes also called Operational Data Store and those build for analysis purposes [125].

When companies adopt a virtual data warehouse approach this architecture is not applicable because, in this case, data are not manipulated or copied but directly accessed from transactional systems. Data are still accessed like they are in a data warehouse but the platform is not a real data warehouse. This solution avoid duplicates but does not exploit advantages of using a real data warehouse, in particular for performing analysis on historical data.

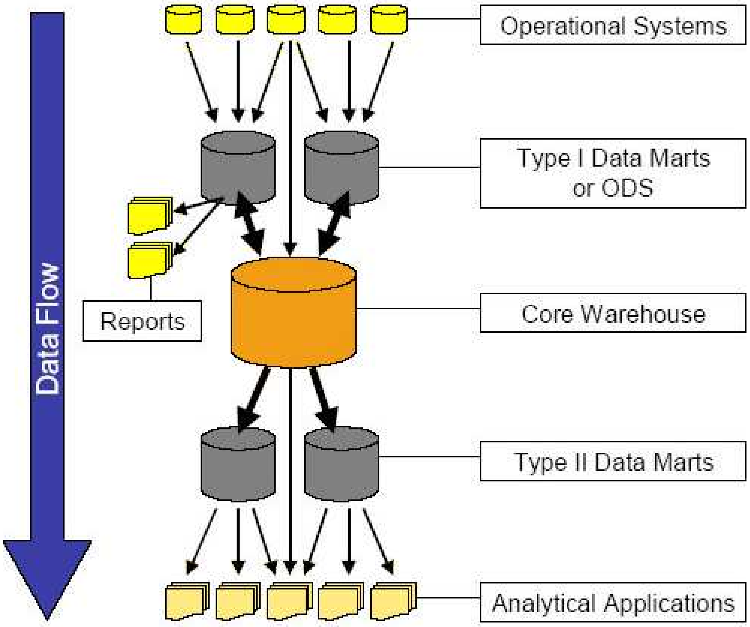


Figure 2.1: A data warehouse architecture

A basic data warehouse organization is composed of four layers:

- Data quality: the acquisition and validation of data in the data warehouse;
- Data preparation and store: data delivery to users and to analytic applications;
- Data analysis and interpretation: data transformation into information;

- Presentation: presentation of the results to the users.

In order to properly have these layers there should exist an organization that supports this process with well defined users and roles and there should exist a technology that supports the process, based on choices and functional needs of the process itself. The choice of the right technology is very important because of the data integration.

In the architecture of data warehouse it is possible to distinguish the following main components:

- Data filter: checks the correctness of data coming from different transactional systems. In this component there is also the data cleaning process, whose goal is to remove wrong data and to detect/correct inconsistency in the data;
- Export: extract data from the transactional systems;
- Data loading: loads data in the data warehouse.

As presented before, these components are the set of applications that are called ETL. These applications are used cyclically; the cycles may be daily, weekly or monthly. There are other components inside a data warehouse that perform specific tasks:

- data alignment: synchronize data in the different transactional systems;
- data access: query data for analyzing them;
- data mining: perform complex searches inside data in order to find information hidden in data;
- export: export data from a data warehouse to another allowing a hierarchical architecture.

There are also other modules depending on the needs, like modules for the design and management of the data warehouse, or the data dictionary, that may be helpful to understand better the content of the data warehouse.

On Line Analytical Processing (OLAP)

The acronym OLAP stands for On-Line Analytical Processing, and it is a technology that provides multidimensional analysis on different dimensions at different levels of detail. Typical OLAP functions are: drill-down, rotate, and swap. The definitions for OLAP where given by E.F.

Codd in 1993. OLAP may be further divided in relational (ROLAP), multidimensional (MOLAP) and desktop (DOLAP). Those approaches are different depending on how data processing is performed. if computation is done in a relational database then it is called ROLAP; if calculations are performed in a server-based multidimensional database then we call it MOLAP. Typically, cubes are available to input data or to do analysis; Finally, we say it DOLAP if cubes are read-only and calculations for building them are performed on the desktop/client or Web midtier. The main component of OLAP is the OLAP server, which sits between a client and a database management systems (DBMS). The OLAP server understands how data is organized in the database and has special functions for analyzing the data. There are OLAP servers available for almost all the major database systems.

The term OLAP is more than a definition, and it is not a clear description of what OLAP means. It certainly gives no indication of why a company would want to use an OLAP tool, or even what an OLAP tool actually does. And it gives no help in deciding if a product is an OLAP tool or not.

Nigel Pendse [156] considered this problem in late 1994 because he needed to decide which products fell into the category. This is a difficult task, because as more and more vendors claim to have 'OLAP compliant' products, whatever this mean (often they don't even know). Pendse created a vendors own definition table in which are listed the definition for OLAP that vendors give their own. There is also an OLAP Council, but the membership to the OLAP Council is not a good criteria because several significant OLAP vendors were never members or resigned, and several members were not OLAP vendors.

The Pendse work group built the FASMI test in order to classify OLAP applications without dictating how they should be implemented. There are many ways of implementing OLAP compliant applications, and no single piece of technology should be officially required, or even recommended. They suggest in which circumstances one approach or another might be preferred, and have identified areas where we feel that all the products currently fall short of what they regard as a technology ideal.

They summarized the OLAP definition in just five key words: Fast Analysis of Shared Multidimensional Information (FASMI).

Fast means that the system is targeted to deliver most responses to users within about five seconds, with the simplest analyses taking no more than one second and very few taking more than 20 seconds. Independent research in The Netherlands has shown that end-users assume that a process has failed if results are not received with 30 seconds, and they are apt to hit 'Alt+Ctrl+Delete' unless the system warns them that the report will take longer. Even if they have been warned that it will take significantly longer, users are likely to get distracted and lose their chain of thought, so the quality of analysis suffers. This speed is not easy to achieve with large amounts of data, particularly if on-the-fly and ad hoc calculations are required. The full pre-calculation approach fails with very large, sparse applications as the databases simply get too large (the database explosion problem), whereas doing everything on-the-fly is much too slow with large databases, even if exotic hardware is used. Even though it may seem miraculous at first if reports that previously took days now take only minutes, users soon get bored of waiting, and the project will be much less successful than if it had delivered a near instantaneous response, even at the cost of less detailed analysis. The OLAP Survey found that slow query response is consistently the most often-cited technical problem with OLAP products, so too many deployments are clearly still failing to pass this test.

Analysis means that the system can cope with any business logic and statistical analysis that is relevant for the application and the user, and keep it easy enough for the target user. Although some pre-programming may be needed, it is not acceptable if all application definitions have to be done using a professional 4GL. It is certainly necessary to allow the user to define new ad hoc calculations as part of the analysis and to report on the data in any desired way, without having to program, so probably should be excluded applications (like Oracle Discoverer) that do not allow adequate end-user oriented calculation flexibility. All the required analysis functionality be provided in an intuitive manner for the target users. This could include specific features like time series analysis, cost allocations, currency translation, goal seeking, ad hoc multidimensional structural changes, non-procedural modeling, exception alerting, data mining and other application dependent features. These capabilities differ widely between products, depending on their target markets.

Shared means that the system implements all the security requirements for confidentiality (possibly down to cell level) and, if multiple write access is needed, concurrent update locking at an appropriate level. Not all applications need users to write data back, but for the growing number that do, the system should be able to handle multiple updates in a timely, secure manner. This is a major area of weakness in many OLAP products, which tend to assume that all OLAP applications will be read-only, with simplistic security controls. Even products with multi-user read-write often have crude security models; an example is Microsoft OLAP Services.

Multidimensional it is a key requirement. This is the one-word definition of OLAP. The system must provide a multidimensional conceptual view of the data, including full support for hierarchies and multiple hierarchies, as this is certainly the most logical way to analyze businesses and organizations. It is not set a specific minimum number of dimensions that must be handled as it is too application dependent and most products seem to have enough for their target markets. This is a conceptual view, this means that this applies on every database model or tecnology, it is important to have a truly multidimensional conceptual view.

Information is all of the data and derived information needed, wherever it is and however much is relevant for the application. It is the capacity of various products in terms of how much input data they can handle, not how many Gigabytes they take to store it. The capacities of the products differ greatly — the largest OLAP products can hold at least a thousand times as much data as the smallest. There are many considerations here, including data duplication, RAM required, disk space utilization, performance, integration with data warehouses.

FASMI is a reasonable and understandable definition of the goals OLAP is meant to achieve.

Data Mining

In a Business Intelligence tool is it possible to find a component whose task is to apply a data mining tool or technique in order to discover hidden and unexpected insight from data. Data mining is the process of discovering relationships, patterns and information previously unknown and potentially useful in big databases. This topic is presented in Section 2.4.

Time Series analysis

Time series is present in many domains, it is very interesting in the business domain. There are time series into sales, purchases, stock trends, customer profiling and customer behavior analysis. Everything changes with the time are feasible to be represented as time series. More details are in Section 4.10.

Analytic applications

This term was coined by Henry Morris of IDC (International Data Corporation). It refers to a software that works independently of the transaction or source system, allow time-based analysis and extract, transform, and integrate data from multiple sources; automate a group of tasks related to the optimization of particular business processes.

Morris [141] distinguishes analytic applications from business intelligence technologies or tools. There is a long-standing market for business intelligence tools such as OLAP, query/reporting and data mining. Analytic applications incorporate such technologies, but are fundamentally different in terms of specialization, segmentation and structure.

Specialization: Analytic applications are specialized into a particular business process or function, while business intelligence tools are generic.

Segmentation: Analytic applications can be segmented by business function (such as finance or marketing), while business intelligence tools can be segmented by technology (such as data mining or OLAP).

Structure: Analytic applications build and manage business activities to achieve a particular result (such as producing a budget or assessing the performance of key suppliers); business intelligence tools can support ad hoc query and analysis that are not predefined.

Analytic applications expand the objectives of business intelligence to an extended user base, packaging these technologies in a business context; but analytic applications will not displace business intelligence tools. They will continue to be a need for ad hoc analysis for exploring new types of issues and questions that they arise.

2.2.3 History of business intelligence

Business intelligence is born in the early 1990s, a relatively young history despite the fact that the need of information is old in people that makes strategic decisions for a company. Before business intelligence the access to true data was an expensive task in terms of time and money. In the 1980s the tools that decision makers could count on where basically *printed reports*, *spreadsheets* and *instinct of the analyst/decision maker.*

Typically the reports were generated periodically on mainframe-based systems. When an important information was missing in the report it could take months to create a new custom report.

Spreadsheets provided more flexibility than printed reports, but at that time, putting data into a spreadsheet was not an easy task (even today, often it is not immediate), analysts and accountants put data manually, that led to huge amounts of human errors.

Instinct is still a good requirement for decision makers. Managers are, and were, close to the markets and the customers, and markets did not change at the same speed that they do now. In the past when manager accessed to quantitative numbers, the high noise produced low reliability on data, so low that they often did not trusted on data.

Because of those problems, Decision Support Systems (DSSs) and Executive Information Systems (EISs) were created in order to improve the reliability on data and speed-up their analysis. The first task that was optimized was the data retrieval. Instead of having fixed reports, by entering some parameters (typically time intervals, geographic areas, customer or category of customer, product or category of products) the users were able to retrieve the results, typically in a tabular format. The DSS enabled the user to get the data from the mainframe-based transaction system and look at them according to the specified parameters.

Even though the DSSs were fine most of the time, the the problem of visualizing the data was not solved. In general it was difficult (often impossible) to view data according to a subject not designed before. Typically DSS were designed to have an own custom transaction system, making it almost impossible to share information across functions. If we consider a company with different departments, i.e. accountancy, sales and supply chain. Typically the sales sold the product to a specific customer, the supply chain delivered the product to the customer address and the accountancy produced the billing information. This process looks fine but the problem is that information typically was not shared among departments, so the information needed about the customer was typed each time the product moved from a department to another. Also unique codes for customers were not shared among different department. Each DSS was intended for the particular department. This made impossible to cross information about the sales data with the orders, this scenario was typical.

Moreover, the proprietary nature of DSS led to the birth of EISs. The aim of those systems was to provide graphical dashboards based on a larger set of information. Often they got data from external sources, those systems had high implementation costs and they were focused only on executive data, but the need to analyze data was felt very soon from all decision makers, not only on executive people. The main problem to solve was to introduce a new concept of data organization and design and changing the business process is always hard for any company. It was harder to accept also because until the 1990s data warehousing was not considered a good technology, and just browsing inside the old organization of information system was almost an impossible task.

The most used products were: Pilot Software Inc.'s Lightship [158]; Platinum Technology's Forest and Trees [83]; Comshare Inc.'s Commander Decision [62], most of them are sold now by successful companies or have been acquired by other companies.

Several big historical factors led to the need to get information faster than in the past: the increased free trade, the unification of Germany, the signing of North American Free Trade Agreement (NAFTA), the growth of the globalization and the possibility to operate to a wide range of new markets. The possibility to operate to a global market pushed decision makers to perform cross analysis among different sections of the company. In distributed market the main analysis was made along the region-based DSSs, but this type of analysis was still not satisfying.

The need of analysts and decision makers were increasing while costs for buying of the devices was decreasing and the PCs were becoming more powerful and common office tools. Users started using spreadsheets and PC-based graphics programs to better understand data. This pushed the demand of more powerful and robust reports and for distributed computing (first client/server, then multilayered architectures). The need to reduce the number and the complexity of custom transaction systems, met the business demand for growth and globalization, derive the productivity and cost benefits of business process reengineering, led to the development of the Enterprise Resource Planning (ERP) systems. ERP systems are composed by modular components that share common data, and each module follows its own business rule. This mechanism permits, for instance, to track all the activities for order shipping, i.e. the customer buy a product, the data of that customer are stored in a common repository, the shipping module takes the subset of those data to ship the ordered product to the customer address, and the accountancy department may produce the invoice with the details about the order and the customer.

Having an ERP system running in the company, the common data are typed once, the consistency of the information about the company is preserved. One of the biggest reasons that settled the decision of many companies to move to ERP logic is the year 2000 problem that pushed companies to change their business process and definitely leave the legacy systems. An ERP system does not guarantee to perform Business Intelligence tasks because ERP is focused on data consistency, performance and low customization costs. The Business Intelligence tools had room to coexist with ERP systems, because just sharing data is not enough to analyze them.

Moreover, the adoption of the Business Intelligence solution was pushed by the possibility of implementing a client/server architecture. Most of the rendering of previous mainframe-based reports could be obtained on almost every desktop. The success of Business Intelligence tools came from the adoption of data warehouse. A data warehouse extracts information from the ERP and aggregates it to allow fast analysis on huge amount of data. The best exploitation of data warehouse comes from building what Inmon defined subject-oriented data marts. Data marts allow to produce quickly analysis on data. In general a central data warehouse acts as a platform, populates data marts, the transactions are safely isolated from the reporting system. Compared to an ERP transaction system in a data warehouse the data is combined into a subject area or a business view, allowing users to perform analysis across multiple business processes. Moreover, a data warehouse can contain years of history, since data are aggregated, and users can analyze trends and patterns. This is the main difference with ERP and transaction systems that typically contain current data at the most granular level of detail.

Despite business intelligence tools included advanced report, OLAP technology, and enterprise-wide deployment capabilities, the executives (the user category that most exploited BI) still needed something similar to the dashboard they where used to work with and former DSS users still wanted to be able to log into the system from any terminal and see the data at the push of a button. Internet has been the solution to these needs.

Initially most BI vendors allowed users to access to standard report in HTML format. As BI vendors redesigned their products, Internet promised to fulfill two main needs: push-button access to key performance indicators, and flexibility of interactive, ad hoc access when required [13].

2.3 Business Intelligence and Knowledge Discovery in Database

There are many analogies between BI process and KDD process, a typical BI process pyramid is the one shown in Figure 2.2. As shown at the bottom of the figure, row data are usually extracted from various sources. They may be stored in single files or organized in different databases, they may also come from paper documents or from information providers. These data are then organized in data warehouses and data marts. Multidimensional databases are used for data exploration and analysis along different attributes. Data mining techniques are exploited to extract knowledge that is presented to decision makers in order to support their decision process for their strategic activities. Indeed, decision makers should be allowed to easily browse inside the presented data and to be helped in discovering interesting patterns on the basis of which they will take their decision.

Visualization techniques can be profitably used in the data exploration level, but also for data mining as well as at the data presentation level. They are used for data exploration activities to better select sets of interesting data. In the data mining and data presentation levels, visualizations may especially support post-processing activities necessary to fully understand the results of a data mining application [142]. Moreover, visualization techniques can sometimes be considered as data mining technique themselves, which actually 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 [121].

Knowledge Discovery in Databases (KDD) is defined as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data by Fayyad at al. [81]. There are usually several steps in a KDD process: data selection, preprocessing, transformation, data mining, and interpretation/evaluation of the results, as shown in Figure 2.3. Despite there is a real distinction between data mining and KDD (data mining is a step of the KDD proceess) many researcher use those two terms interchangeably.

Other researchers propose more detailed processes for KDD, Sarker et al. [170] propose 13 steps for KDD to be efficiently automated:

- problem definition and determining the mining task;
- data description and selection;
- data conversion;
- data cleaning;
- data transformation;
- data reduction and projection;
- domain-specific data preprocessing;
- feature selection;
- choosing the mining algorithm;
- algorithm-specific data preprocessing;
- applying the mining algorithm;
- analyzing and refining the results;
- knowledge consolidation;

The base is that each step is defined either in terms of time span or technical knowledge and human effort. In data description and selection, suitable files are selected as well as suitable data fields. There may exist many files/tables in the database. However, not all are suitable for knowledge discovery. After an initial study, suitable and unsuitable data should be identified. The selection here may involve selection of fields that are potentially useful for multiple mining tasks. In data

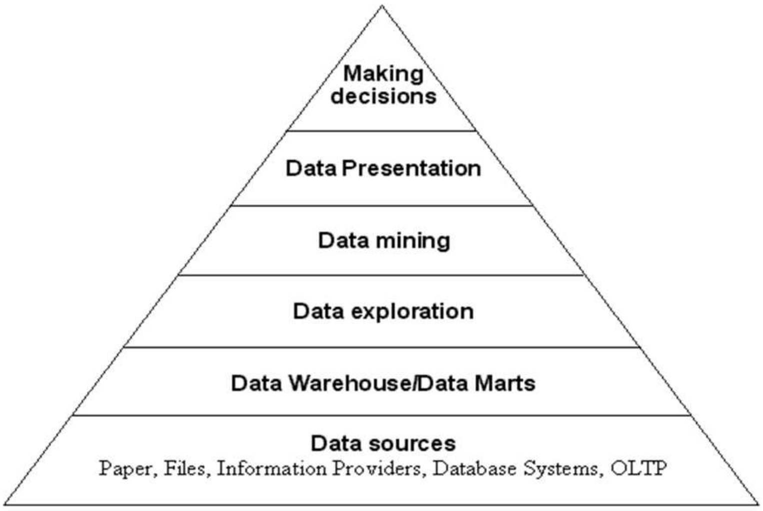

Making decisions
Data Presentation
Data mining
Data exploration
Data Warehouse/Data Marts
Data sources
Paper, Files, Information Providers, Database Systems, OLTP

Figure 2.2: A typical Business Intelligence pyramid

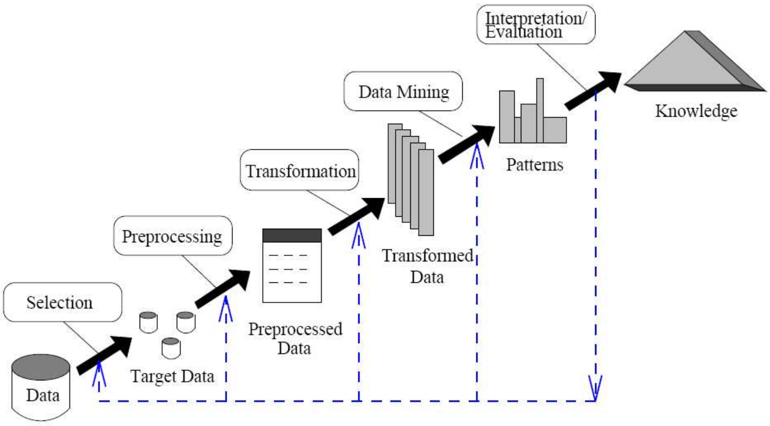


Figure 2.3: Steps of the KDD process

conversion, the database file system suitable for the mining process is identified and the data is converted from its original stored format to the selected one. Data cleaning removes (or reduces) noise and errors in the data. If the proportion of inconsistent or noise data is quite small, their deletion may have little effect on the results. Inconsistency may also cause sufficient trouble for the mining algorithm and small SQL-like statements may track this type of inconsistency and facilitate the mining task. With data transformation we mean reflecting the logical relations between the tables into a single table that contains all the information needed for the mining process. Many of the mining algorithms do not work on multiple-tables and therefore we need somehow to combine the tables into one. Redundancy may arise as a result and one needs to be careful here as this combination of tables may change the class frequency, which will in turn affect the mining process.

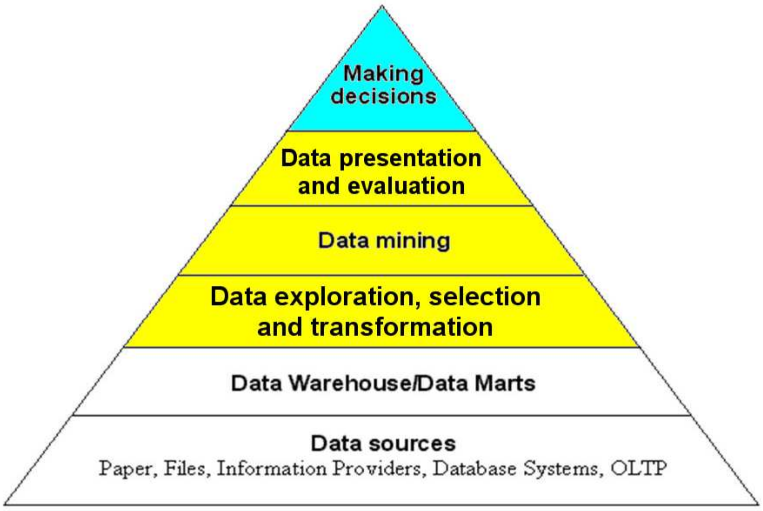

Making decisions
Data presentation and evaluation
Data mining
Data exploration, selection and transformation
Data Warehouse/Data Marts
Data sources
Paper, Files, Information Providers, Database Systems, OLTP

Figure 2.4: The BI pyramid integrated with the KDD process

Where and how the two process may overlap? If we look at the BI pyramid, the two bottom layers are basically the organization of the data source of a company. In particular, the activities at the bottom are those described in Section 2.2.2 related to data warehouse.

In Figure 2.4 is shown a single process covering both KDD and BI, in which is possible to distinguish the KDD after the first two layers, typical of data warehouse.

The use of data mining techniques in Business Intelligence process is visible today in many platform.

2.4 Data Mining

Nowadays, there are some domains (remote sensor on satellites, telescope scanning skies, complex scientific simulations, microarrays generating gene expression data, etc.) that produce/collect and store data at enormous speeds (Gbyte/hour). Traditional techniques are infeasible for raw data, this domains needs more than others the help of data mining tools and techniques. Data mining is used for several goals, for data reduction, cataloging, classifying, segmentation, that helps scientists in hypothesis formulation.

In data mining, there are three primary components: model representation, model evaluation and search. There are two basic types of search methods used: parameter search and model search [80].

In parameter search, the algorithm searches for the best set of parameters for a fixed model representation that optimizes the specific model on the data set. The optimal solution is reachable for relatively simple problems. For more general models this is more difficult, so it is needed to apply other methods, like gradient descend method of back-propagation [102].

In model search parameter are fixed and loop occurs over the search methods. The model representation is changed in a family of models. With this criteria the use of heuristic is very important and parameter are used to evaluate the quality of a particular model. Here, heuristic search plays a key role in finding good solutions. More information about modern heuristic techniques is provided in Abbass [2].

In order to give a simple taxonomy of data mining a first possible distinction is between tasks and techniques. The tasks are the goals that a data miner have in mind, they answer to this question: do I want to describe what is in data or I want to predict some event basing on the data stored in the database? Another question may be related to the types of technique available to reach the goal.

2.4.1 Data mining tasks

There are two may types of data mining tasks: descriptive and predictive, a classification is presented in [1]. In the following are listed the data mining tasks, in parenthesis is specified if they are descriptive or predictive, for each task the methods used are listed:

- Features selection (descriptive) - Dependency models (Association

Rules).

- Point prediction/estimation (predictive) - Regression methods; Neural networks; Regression decision trees; Support vector machines.
- Classification (predictive) - Statistical Regression models; Neural networks; Decision trees; Support vector machines.
- Rule Discovery (descriptive) - Decision trees; Learning classifier systems.
- Clustering (descriptive) - Density estimation methods; Neural networks; Clustering techniques.
- Association methods (descriptive) - Association rules; Density estimation models.

The most relevant methods are summarized below:

Dependency models/Association rules determine how are related various characteristics of data (attributes). Dependency modeling exists in two levels: the structural level of the model that specifies which variables are locally dependent on which, the quantitative level of the model specifies the strengths of the dependencies using numerical representation [82]. More about association rules is described in Section 4.5.

Feature Selection is concerned with the identification of a subset of features that allow to discriminate or predict the problem. The feature is found using fast iterative linear-programming-based algorithm that terminates in a finite number of steps [135]. The feature selection problem is defined as a mathematical program with a parametric objective function and linear constraints [25].

Summarization methods find a compact description of a subset of data. Summarization can be performed using a bar chart or statistical analysis. This is useful for understanding the importance of certain attributes when compared against each other [108]. More sophisticated methods involve the derivation of summary rules [6], multivariate visualization techniques, and the discovery of functional relationships between variables [213].

Clustering identifies a finite set of categories or clusters to describe the data [115]. The categories may be mutually exclusive and exhaustive, or consist of a richer representation such as hierarchical or overlapping categories [81]. Unlike classification, in general the number of desired groups is unknown, the clustering problem is usually treated as a two-stage optimization problem. First, is determined the number of clusters then the data is fitted to the most probable cluster. The first step (determining the number of clusters) may be automatic or manual. Outliers are always possible. By using the sequential optimization techniques or setting the number of clusters the optimality of the overall problem is not guaranteed.

Regression modeling minimize the error function with an unconstrained optimization.

Artificial Neural Networks are widely used for prediction, estimation and classification [15, 176, 214]. In terms of model evaluation, the standard squared error and cross entropy loss functions for training artificial neural networks can be viewed as log-likelihood functions for regression and classification respectively [88, 167]. Regression Trees and Rules are also used for predictive modeling, although they can be applied for descriptive modeling as well.

In *classification*, the basic goal is to predict the most likely state of a categorical variable (the class) given the values of the other variables. This is fundamentally a density estimation problem [174]. A number of studies have been undertaken in the literature for modeling classification as an optimization problem [25] including discriminant analysis for classification which uses an unconstrained optimization technique for error minimization [162].

Rule Discovery (RD) is one of the most important data mining tasks. The basic idea is to generate a set of symbolic rules that describe each class or category. Rules should usually be simple to understand and interpret. RD can be a natural outcome of the classification process as a path in a decision tree from the root node to a leaf node that represents a rule. However, redundancy is often present in decision trees and the extracted rules are, in general, simpler than the tree. It is also possible to generate the rules directly without building a decision tree as an intermediate step. In this case, Learning Classifier Systems play a key method to rule discovery.

2.4.2 Data mining techniques

In the rest of this section, we will present some of the most commonly used techniques for data mining. This list should not be considered complete, but rather a sample of the techniques for data mining.

Bayesian methods is a powerful class of techniques for data mining. It can be proven that bayesian methods work under uncertainty. The ability of these techniques to capture the underlying relationship between the attributes and their use of probabilities as their method of prediction increase their reliability and robustness in data mining applications. The main problem with bayesian methods is that they are not scalable and the research is trying to overcome this problem.

Feedforward Artificial Neural Networks are one of the most commonly used artificial neural networks architectures for data mining. Feedforward artificial neural networks are non-parametric regression methods, which approximate the underlying functionality in data by minimizing a loss function. Artificial neural networks are known to be slow, recent advances in this field present fast training algorithms as well as adaptive networks [212]. Artificial neural networks are largely used in real life application since, by rule of thumb, they have been demonstrated that are more accurate than many data mining techniques.

Decision trees may be univariate or multivariate [2]. Univariate decision trees approximate the underlying distribution by partitioning the feature space recursively with axis-parallel hyperplanes. The underlying function, or relationship between inputs and outputs, is approximated by a synthesis of the hyper-rectangles generated from the partitions. Multivariate decision trees have more complicated partitioning methodologies and are computationally more expensive than univariate decision trees. The split at a node in an multivariate decision trees depends on finding a combination of attributes that optimally (or at least satisfactorily) partitions the input space. The simplest combination of attributes is taken to be linear. Even in the simple case, the process is very expensive since finding a single linear hyperplane that optimally splits the data at a node is an NP-hard problem. A path from the root node to a leaf node in both univariate decision trees and multivariate decision trees represent the rule for defining the class boundary of the class present at the leaf node.

Support Vector Machines [42, 56, 196] are powerful tools for both classification and point estimation. They classify points by assigning them to one of two disjoint half spaces that are either in the original input space of the problem for linear classifiers, or in a higher dimensional feature space for nonlinear classifiers. Support vector machines represent a good example of data mining techniques that are based on optimization theory.

Optimization provide another alternative set of techniques that produce robust results [25]. The main problem with these techniques is scalability and slow convergence. Global optimization can be combined with heuristics to overcome the slow performance of optimization techniques [14]. However, conventional heuristics are much faster by many orders of magnitude than conventional optimization. Those methods applies to all the methods previously presented.

2.5 Information Visualization

2.5.1 Introduction

McCormick says that "Visualization is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights" [137].

Visualization may be thought as an adjustable mapping from data to a visual form that a human may perceive. In the Figure 2.5 is described a process that is needed to have this mapping. Raw data are available in some format, then the data transformation process maps data to data tables or relational descriptors; then visual mappings transforms tables into visual structures, combining spatial substrates, marks and graphical properties. Finally, the view transformation process maps the visual structures to create views of the visual structures by specifying graphical parameters such as position, scaling, clipping.

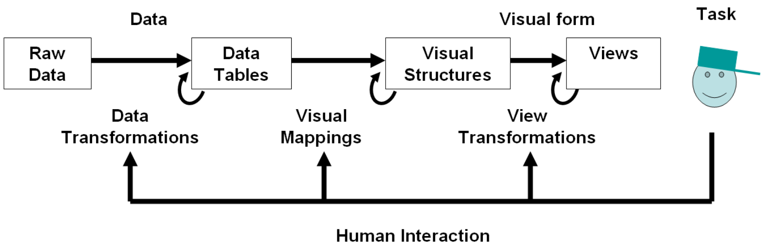


Figure 2.5: Reference model for visualization. Visualization can be described as the mapping of data to visual form that supports human interaction in a workspace for visual sense making

A definition for information visualization is "the use of interactive visual representations of abstract data to amplify cognition" [45, 204, 187, 18].

Information visualization should not be confused with scientific visualization, the difference is in the term *abstract*. In scientific visualization people wants to represent on the screen real events and real object (i.e. the health of a brain, a weather status or forecast, an EKG, stock trends..), the information visualization's goal is to graphically represent information that is in some way related to the knowledge of the user, not to a physical event. There are many visualization techniques and they may be classified in different ways.

Moreover, in scientific visualization, 3D may be necessary, in particular when typical questions involve continuous variables and volumes, surfaces, inside/outside, left/right, and above/below. In information visualization, typical questions involve more categorical variables and the discovery of patterns, trends, clusters, outliers, and gaps in data such as stock prices, patient records, or social relationships [53, 54, 44, 24].

We are now all familiar with direct manipulation interfaces; their success testify the power of using the computer. 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 allow to show relationships by proximity, containment, connected lines, color coding, etc. In order to focus the attention to specific items among thousands of items highlighting techniques (blinking, brightening, reverse video) can be used. 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 result. Such procedures substitute the logical inferences the user should perform without a visual presentation. 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 megapixel displays, identify patterns and exceptions, recall related images. The use of proximity coding, color coding, size coding, animated presentation, and user-controlled selections enable users to explore large information spaces rapidly and with fun.

Today the research is focused on interaction techniques that, combined with information visualization techniques, permit to reach the goal of information visualization in a more effective way. The goal may be synthetically expressed with this slogan: Find what you need and understand what you find. The sense is obvious, the user needs to find information, but even if it have the results of the search he needs to well understand them. So "find" is the keyword for information visualization. Finding information may be obtained through exploration, that is supposed to be a nice experience for the user, but this is not always true because one of the main problems with information visualization is the information overload and anxiety [172].

2.5.2 Tasks in Information Visualization

There are many visual design guidelines. Perceptual psychologists, statisticians, and graphic designers offer valuable guidance about presenting static information [208, 192, 60, 21]. The challenge for information visualization researcher today are dynamic displays. 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" [178]. In this section is presented a classification of tasks in information visualization according to Shneiderman's classification in [181].

Overview

The overview allows the user to grasp the entire content of the application and its distribution across the different attributes. Overview strategies include zoomed-out views of each data type that allow users to see the entire collection plus an additional detail view. This detail view (also called field-of-view box) is displayed into the overview that gives the user the feedback on what area is observed in the detail window and often allow the user to move inside the overview area in order to quickly focus on another data set. 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. 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.

Zoom

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 sens of position and context. Another popular approach for keeping the context while zooming some areas of interest is the fisheye strategy [85]; the fisheye distortion magnifies one or more areas on the display. Typically, zoom factors range from 3 to 30 while fisheye views factor is limited to about 5. Semantic zooming is an interesting feature, that allows to focus always on relevant information depending on the level of detail. Some example are Jazz and its successor Piccolo toolkits [17]. Zooming is crucial in applications for small displays, like application for mobile phones or PDAs.

Filter

Users are not interested to all the displayed information so they may want to filter out uninteresting items in order to quickly focus on the items of interest. Dynamic queries applied to the items in the collection are one of the key ideas in information visualization [7]. Sliders, buttons, or other control widgets coupled to rapid display update are used for the filter task.

Details-on-demand

Users want to see the details of data they are interacting, because they are able to fully evaluate information looking to numbers or words. After having performed some interaction the information visualization tool should provide the possibility to show the details to the users. 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 attributes. The details-on-demand window can contain information that are related to the items of interest but not necessary visible on the working window, an example is Spotfire [43], that may contain HTML text with links.

Relate

Users may want to view relationship among items. relationships can be displayed by using proximity, by containment, by connected lines, or by color-coding. In the FilmFinder details-on-demand window [7] 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 or categorical attributes [165].

History

We can keep the history of actions to support undo, replay, 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. Currently, many prototypes fail to deal with this requirement.

Extract

It is also useful to allow extraction of sub-collections of the query parameters. Once 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 it into a file in a format that would facilitate other 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. At the moment, few prototypes support this task.

Shneiderman replaces the well known sentence "a picture is worth a thousand words" to "an interface is worth a thousand pictures" [181], this easily give the sense of information visualization goals. Many people use computers everyday, digital environment are ubiquitous. New techniques are used and each is better than the one before. While before the development of a complex interface was a long process nowadays designers and developers have powerful tools to develop them in a very short time. The most relevant think is not anymore how to do but what to do. The main goal is to give the user a better experience with the machine.

2.5.3 Techniques in Information Visualization

There are many classifications and taxonomies for information visualization techniques, Keim uses three main dimensions [121]:

- Data type (1, 2, 3, multidimensional, text/web, hierarchies/graphs, algorithms/software);
- Visualization techniques, (standard 2D and 3D display, geometrically transormed display, iconic display, dense pixel display, stacked display);
- Interaction and distorsion techniques, (standard, projection, filtering, zoom, distortion, link/brush).

He combines those dimensions into a cube and put the tool in the space defined by these three dimensions.

Shneiderman classify information visualization along two dimensions, the tasks (presented in Section 2.5.2) and the data type.

In the Shneiderman's taxonomy [181], there are four basic data types (1, 2, 3, or multidimensional) plus three more structured data types (temporal, tree, and network). This classification is useful to describe the visualizations that have been developed and to characterize the classes of problems that users encounter.

1-D Linear Data

This big class of data includes program source code, textual documents, dictionaries, and alphabetical lists of names, all those data that can be organized sequentially. Are excluded those sequential data that depend strictly on time, for those there is a dedicated section (2.5.3). An example in this category is [74]. In order to represent this class of data type issues like colors, size, layout to use, what overview, scrolling, or selection methods are of interest for the user.

2-D Map data

Also called planar data include geographic maps, floorplans, and newspaper layouts. Items are represented on the plane, sometimes there is a layer logic in order to simulate a 3-D approach, but is not 3D, it is only a multi layered 2-Dimensional representation. There is a huge collection of techniques and tools about maps in [65].

3-D World

Those are data type that come from real-world such as molecules, the human body, and buildings that have volume and complex relationships with other items. Computer-assisted medical imagery, architectural drawing, mechanical design, chemical structure modeling, and scientific simulations are built to handle these complex three-dimensional relationships. Representing the 3^{rd} dimension in a 2D screen is not an easy task because there are well known problems such as occlusion, navigation, comparison, color. Solutions to the typical problems have been found, such as overviews, landmarks, teleportation, enhanced 3D technique. There are many examples and many information visualization researchers discuss a lot about the opportunity to display the third dimension, and it is difficult to justify in many cases.

Temporal data

Time series are very common. The characteristic in this type of data is the time, this means that data are ordered (and this order cannot be changed) and often are cyclic. Examples are meteorological data, that are influenced by season, and during the day by the sun (if we consider temperature, during the night is lower than during the day). Silva and Catarci [184] made a survey on temporal data while more details about time series are in Section 4.10. An interesting work about the combination space-time data is the work of Andrienko et al. [8], in the domain of geovisualization.

Multidimensional data

There is an active research about multidimensional data. The possibility to analyze multidimensional data is a goal of several research area, not only in information visualization. In any case the challenge is to represent those dimensions on the screen and researcher uses several techniques. A typical domain is the one described in Section 2.2.2, related to the business domain. The representation of multidimensional data can be dynamic and bidimensional. The user have some widgets that allow him to move into other dimensions. In Section 4.8, 4.6 and 4.9 are presented tools that allow multidimensional data visualization and analysis.

Tree data

Hierarchies or tree structures are collections of items, in which each item (except the root) has a link to one parent item. Items and the links between parent and child can have multiple attributes. One of the most known tree representation is the Windows Explorer, some examples of application that visualize tree data are [160, 46, 130, 19].

Network data

Trees are a sub-case of graphs. The difference between a tree and a graph is that a node in a tree have only one parent, while in a graph a node may have more than one parent. There are many types of networks (acyclic, lattices, (un)rooted, (un)directed), but it is convenient to consider them all as one data type. Network visualization is an old but still imperfect art because of the complexity of relationships. In Section 4.5 is presented a tool that uses graphs to visualize association rules.

The presented data types may be combined and other types may be used (four-dimensional data, mutitrees). The taxonomy here are to help understanting the range of problems in information visualization, there are other problems to face in order to create successful tools. *Importing data* is often an underestimated task, one of the problem is the selection and the cleaning of data, but also dealing with missing data [73] is not an easy task; Is very important to provide *Textual information* using labels or screen tips; additional information is often needed, so it is important to give access to *related information* about a specific item; another challenge is to visualize *large volume of data*, often visualization are not scalable; also the integration with *data mining* tools is an activity that is more present in the most recent information visualization tools. Data miners needs to visualize their results and to interact with them, DAE tools aim to this result, more in detail see Section 4.5 and 4.6. The *cooperation* is another topic to take into account, since collaborative work and distance learning are activities that have been gained popularity; finally the *universal usability* is a challenge for information visualization, since there are communities that cannot use standard devices, in order to give an example, the mobile users community may be compared to “blind” users, at least those that use non graphic mobile phones.

2.6 Visual Data Mining

Currently there are several several definition for Visual Data Mining (VDM), it may be defined as an approach to explorative data analysis and knowledge discovery that is built on the extensive use of visual computing. Many researcher does not have a clear idea of what VDM is and what is not.

Shneiderman presents the advantages of the use of information visualization combined with data mining, explaining that computational tools such as data mining and information visualization have advanced dramatically in recent years, but they have been developed by largely separate communities having different philosophies. Data mining and machine learning researchers tend to believe in the power of their statistical methods to identify interesting patterns without human intervention, while information visualization researchers tend to believe in the importance of user control by domain experts to produce useful visual presentations that provide unanticipated insights [179].

Niggerman talk about VDM but gives no definitions [152].

Keim says that the basic idea of visual data mining is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data. It is also the process of searching and analyzing databases to find implicit but potentially useful information [121]. More formally: VDM is the process of finding a subset D' of the database D and hypoteses $H_u(D', C)$ that a user u consider useful in an application context C [121].

Many researcher agree that Visual Data Mining is also considered as a collection of interactive reflective methods that support exploration of data sets by dynamically adjusting parameters to see how they affect the information being presented [185, 138, 151, 91].

VDM is the use of visualization techniques to allow data miners and analysts to evaluate, monitor, and guide the inputs, products and process of data mining. It can help introduce user insights, preferences, and biases in earlier stages of the data mining life-cycle to reduce its overall computation complexity and reduce the set of uninteresting patterns in the product [86].

VDM could be related with all the previously tasks described in Section 2.4. The goal should be to provide a synthesis of visualization and data mining, to enhance the effectiveness of the overall data mining process.

Ankerst defines VDM as "a step in the KDD process that uses visualizations as a communication channel between the computer and the user to produce novel and interpretable patterns" [10]. He also presents a possible combination of information visualization techniques with data mining tools [9].

VDM involves the use of visual representations that can be applied in the three data-mining life cycle stages: data preparation, model application and validation of the results. This may involve a partitioning of VDM in three fields, each one targeted on producing visual representations that will enhance each single stage, plus another set of visualization tools at an higher abstraction level that allow the interaction with the overall process.

VDM could be exploited in the preparation to visually define and guide the pre-processing tasks. This involves the visual manipulation of row data according to the requirements imposed by the data mining tasks. Visual manipulation means the ability to use visualizations to handle problems such as missing data, data transformations, sampling and pruning, data inconsistencies, all tasks related to this stage. Such ability enable the data miner to formulate accurate hypotheses and objectives in KDD, and select carefully only the relevant and useful data to be sampled and extracted during the data pre-processing. This is the case shown in Figure 2.6(c). Data are immediately visualized without running a sophisticated algorithm before. By interacting and operating on the visualization, the user has full control over the search in the search space. The patterns are obtained by exploring the data. This stage may be used also to perform visual data mining without the actual use of data mining algorithm. Looking to the definition of data mining, if the user discovers novel and unexpected insights from data this may be seen as a way to perform data mining.

VDM may imply the specification of data mining models using visualizations. Selection of the training data set and model, definition of its parameters, training process specification and outcomes storage are the general tasks of this stage. Moreover visualization techniques may provide a visual overview of the whole model. This implies evaluation, monitoring and guidance of the data-mining module. Evaluation includes the validation of training samples, test-samples, and learned models against the data in the database plus the appropriateness of data and learning algorithms for specific data-mining activities. Monitoring includes activities such as tracking the progress of the data-mining algorithms, evaluating the continued relevance of learned patterns in the context of database updates, etc. Guidance includes activities such as user-initiated biasing or altering inputs, learned patterns and other system decisions. Moreover often data mining algorithms are iterative and the results of an iteration are the input for the next iteration. The intermediate result can be appropriately visualized, then the user retrieves the interesting patterns in the visualization of the intermediate result and may also change them in order to influence the next step. One basic motivation for this approach is to make the algorithmic part independent from an application. This scenario is depicted in Figure 2.6(b).

The results of a data mining model application should allow users to discuss and explain the logic behind the model with colleagues, to explain results to customers, or, in general to decision makers. If the user can understand what has been discovered, he or she will trust it and use it. A comprehensible data mining model is also a trusted model, even often, accuracy is traded off for understandability. Advanced visualization techniques expand the number of models that domain experts can understand. From this point of view having visualizations for the presentation of results is the most common choice. In most cases the results of data mining algorithms are difficult to be understood by humans who are accustomed to perceive information by their visual senses or, even if they are experts the quantity of data may produce cognitive overload and they may not well understand the results or, at least they may find very hard to extract useful knowledge from the results. This case is the one depicted in Figure 2.6(a).

In this context, VDM may be seen as the graphical presentation of data, whether the data is row data, summary data, or mined outcomes extracted from data. This is a type of visual data analysis, where the analytic component is shifted from the cognitive system to human perception. Based on the visualization, the user may want to return to the data mining algorithm and run it again with modified input parameters;

This classification provides a possibility to distinguish between different approaches for visual data mining and refines the definition of the KDD process by focusing on the data mining and evaluation.

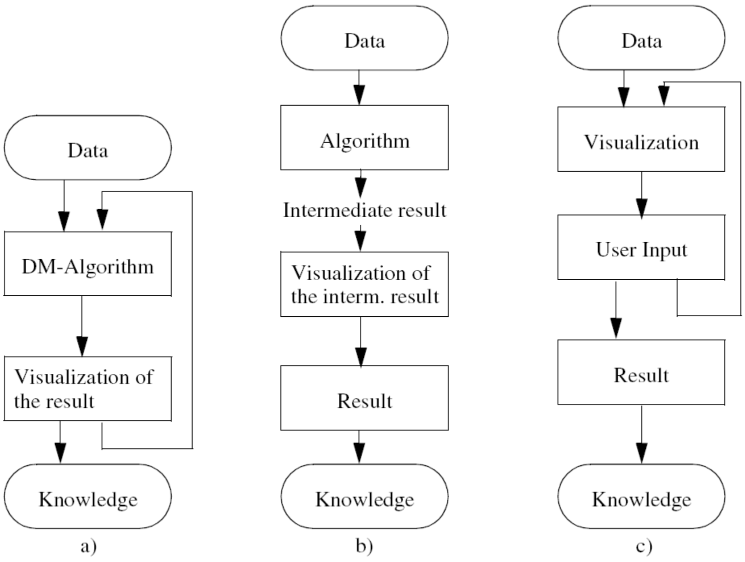


Figure 2.6: Different approaches to visual data mining

The goal of VDM is to move information hidden in the data space to the visualization space, in this way the is easier for the knowledge worker (or decision maker) to grasp insight from data. To do that there is the need to provide the user with more information possible, but on the other hand, the risk for information overload is high, so it is needed that only relevant information should be presented to the user. The balance those two opposite needs makes hard the production of new visualization models. Another issue related to data presentation is that from raw data to the result it is needed to visualize also relationships and how original data have been transformed into results.

2.6.1 Explanatory example

The classic example to explain the usefulness of visualization is attributed to Dr. J. Snow. In the nineteenth century, there were several outbreaks of cholera in London. In the 1849 outbreak, a large proportion of the victims received their water from two water companies. Both of these water companies had the source of their water on the Thames River, just downstream from a sewer outlet. In an 1854 outbreak, most of the deaths occurred within the area of the Southwark and Vauxhall Water Company. Fortunately, just before the outbreak, the Lambeth Water Company relocated their water source to a less polluted point so fewer deaths occurred among their customers. The distribution of deaths was one of the primary factors which proved that the deaths were caused by ingestion. Dr. Snow plotted the distribution of deaths in London on a map 2.7. He determined that an unusually high number of deaths were taking place near a water pump on Broad Street. Snow's findings led him to petition the local authorities to remove the pump's handle. This was done and the number of cholera deaths was dramatically reduced. The work of Doctor Snow stands out as one of the most famous and earliest cases of geography and maps being utilized to understand the spread of a disease [61].

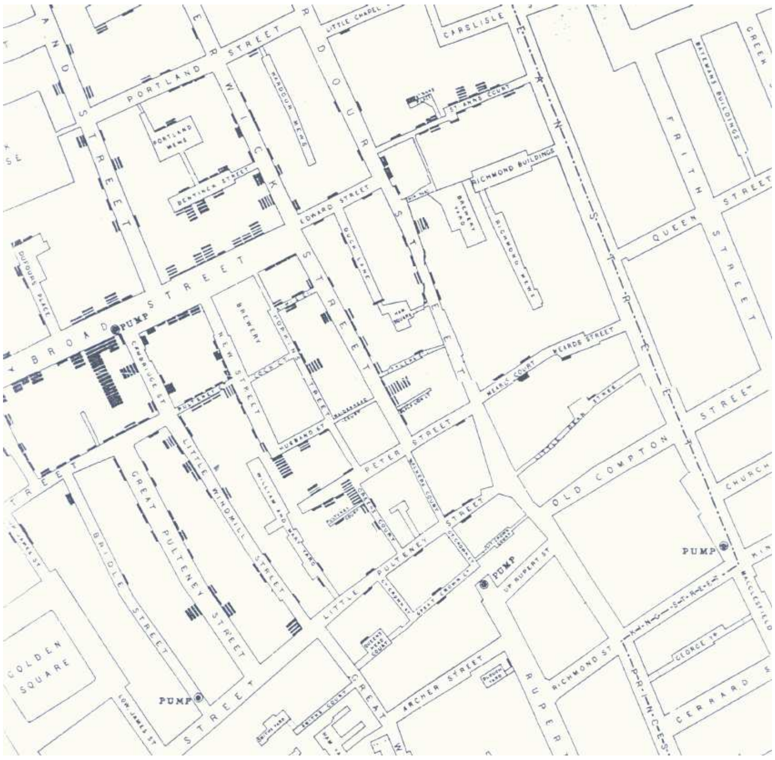


Figure 2.7: Snow's map. In the map are drawn the death and the wather wells of London in 1849

Today computers are powerful means that help human in managing information processing systems. Humans are limited in scaling and are easily overwhelmed by big data volumes. Data mining help to reduce the data volumes and complement human capabilities. Visualization makes data mining near the humans presenting results of this process and giving the human more possibility to grasp insights from data. Data mining is primarily centred on computation and on algorithms, visualizations emphasize user interaction and data manipulations. Algorithms that works stand alone may miss the possibility to take advantage of to the human knowledge, and then the users can lose opportunity in high-dimensional spaces. The combination of the two approach provide opportunity to take advantage each other and hence to solve more difficult analysis problems.

The missing thing is that the human perception enables the users to analyze complex event in a very short time, hence to recognize relevant information and then make decisions. The human perception system is able to immediately spot unusual properties and ignore well-known properties. Humans are able to handles imprecise knowledge easier and better than a complex current computer system, and easily provide complex conclusion. It is worth to note that most data mining techniques work fully automatically but need to have a-priori defined tasks. The tasks are a specific type of hypothesis and the goal of the algorithms are to find quantitative rules that make the hypotheses more specific and allow the user to confirm or reject them. Task-oriented data mining is important but it is also important to develop techniques for data-driven hypotheses generation. For this purpose, it is necessary to include the human in the data mining process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers [121].

For those reasons visual data mining could be very important to make the decision maker part of the data mining process and take advantage of human's perceptual system.

Chapter 3

Related work

3.1 Introduction

In this section a survey on research and commercial systems that work in the area of Business Intelligence, Information Visualization, and Visual Data Mining is presented. We have seen that these three areas are related, this means that some of the systems presented could be positioned in more than one area. We put those system in the area that is more representative for the system taken into account. The survey may also be extended, there is a huge presence of systems both in the research and in the commercial sectors.

3.2 Business Intelligence systems

The current scenario is characterized by an increasing competition, the company data are more and more used as a strategic resource, and the business is more and more based on the strategic support. Many people says that now is the right moment to exploit the Business Performance Management strategies [51, 206].

Business Performance Management solutions help companies to measure their own performance in order to improve the management of processes, gain efficiency and competitiveness, through a faster and easier access to the company information, everywhere and when needed.

Many information useful for divine market changes, customer behavior, products and services demand, in order to make business plan are stacked in transactional systems, in the so many worksheets and into log files produced by web applications. Without the right tools to move information to the right people at the right time companies may loose market positions.

The scenario is rapidly changing and we have already seen in Section 2.2.2 that the stakeholders understand that the actual value for a company is measured by the capacity to access to information.

Business Intelligence provides the users with technologies and basic services needed to develop powerful Business Performance Management application for a specific company. With the Business Intelligence tools the users should be able to interact with data and intuitively explore them in real-time.

In this section is presented a little survey on the most relevant Business Intelligence tools used today. A particular attention is dedicated to the user interface. The order in which they are presented is not relevant. At the end of the section it is illustrated a short comparison table, which compares relevant features on those tools.

3.2.1 Microstrategy

According to Microstrategy[1] Business Intelligence systems should have the following characteristics:

Business Intelligence software allows users to find answers in their data. In order to trigger business decisions, they must be presented in an intuitive way, possibly using graphical forms, grids, scorecards, maps, or a combination. Reports must be descriptive, visually appealing and accessible to the entire organization.

To maximize value, true enterprise reporting must address the following capabilities:

- Rich Data Access: BI solutions must be able to read large volumes of data from disparate data sources.
- Complex Formatting: Users should be able to develop and produce reports with a wide variety of precise formats, from dashboards to invoices to detailed sales or inventory reports to the most sophisticated graphical presentations.
- Scalable, Secure Platform Architecture: Reports should be deployable to a large number of users with central administration and without compromising performance or data security.

Microstrategy has a rich set of presentation styles. The data can be presented in common formats such as grids or graphs. The XML architecture enables data to be transformed into other presentation media such as maps, Gantt charts and scorecards. Customers can visualize the data in a format that makes the most sense for their business needs. More in detail, Microstrategy has:

- Support for 80 graph types (including 3-D rendering options) with title formatting, axes formatting, grid wall formatting and graph background formatting;
- A library of standard grid formatting templates that can be customized using the the XSL (eXtended Style Language) language;

[1] http://www.microstrategy.com/Solutions/5Styles/enterprise_reporting.asp

- Formatting features including subtotals, banding and custom grouping;
- Creation of documents that combine grid, graph and text and that may be viewed over the Web or distributed via email;
- Batch production and distribution of reports;
- Ability to combine multiple reports into a single "report dashboard"

Microstrategy, born in 1989, is one of the most representative company in the business intelligence area. Basing on the historical development of applications and technology for business intelligence today Microstrategy is promoting the five common styles of that have evolved during the past decade; each style represents a characteristic usage and function by end users [139]:

Enterprise Reporting Broadly deployed pixel-perfect report formats for operational reporting and scorecards/dashboards targeted at information consumers and executives.

Cube Analysis OLAP slice-and-dice analysis of limited data sets, targeted at managers and others who need a safe and simple environment for basic data exploration within a limited range of data.

Ad Hoc Query and Analysis Fully investigative query into all data, as well as automated slice-and-dice OLAP analysis of the entire database – down to the transaction level of detail if necessary. Targeted at information explorers and power users.

Statistical Analysis and Data Mining Full mathematical, financial, and statistical treatment of data for purposes of correlation analysis, trend analysis, financial analysis and projections. Targeted at the professional information analysts.

Alerting and Report Delivery Proactive report delivery and alerting to very large populations based on schedules or event triggers in the database. Targeted at very large user populations of information consumers, both internal and external to the enterprise.

In Figure 3.1 is shown the relationship between the user interactivity and the number and type of users compared to all each of this five styles. In the middle there are business managers, which are those users who

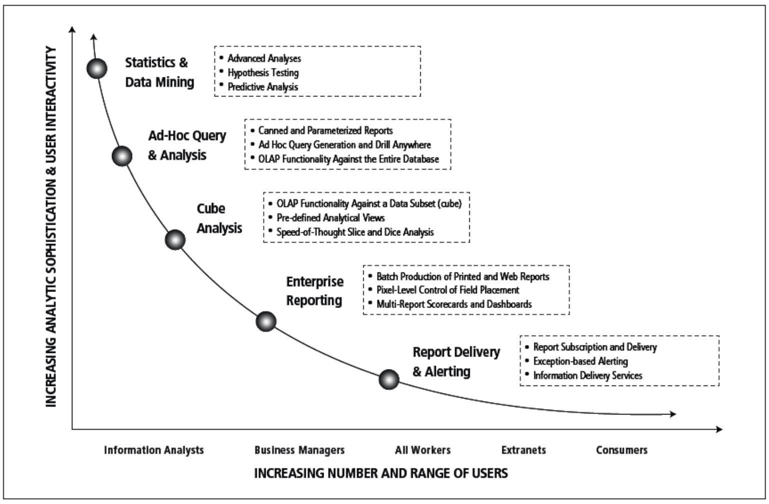

INCREASING ANALYTIC SOPHISTICATION & USER INTERACTIVITY
Statistics & Data Mining
• Advanced Analyses
• Hypothesis Testing
• Predictive Analysis
Ad-Hoc Query & Analysis
• Canned and Parameterized Reports
• Ad Hoc Query Generation and Drill Anywhere
• OLAP Functionality Against the Entire Database
Cube Analysis
• OLAP Functionality Against a Data Subset (cube)
• Pre-defined Analytical Views
• Speed-of-Thought Slice and Dice Analysis
Enterprise Reporting
• Batch Production of Printed and Web Reports
• Pixel-Level Control of Field Placement
• Multi-Report Scorecards and Dashboards
Report Delivery & Alerting
• Report Subscription and Delivery
• Exception-based Alerting
• Information Delivery Services
Information Analysts
Business Managers
All Workers
Extranets
Consumers
INCREASING NUMBER AND RANGE OF USERS

Figure 3.1: The 5 Styles of Business Intelligence have evolved to support different needs, from advanced professional analysis to basic information consumption.

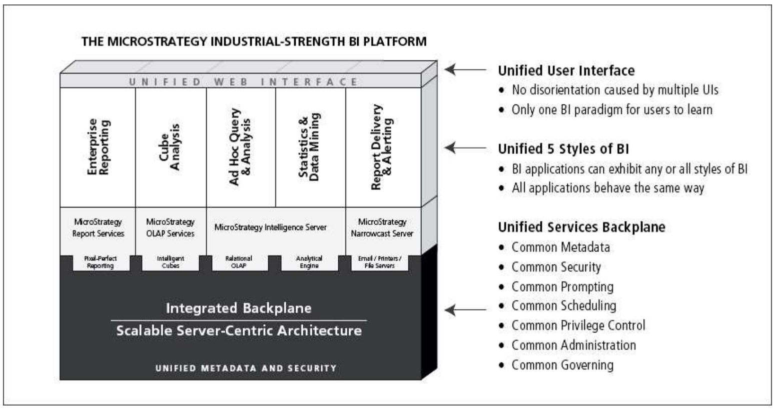

THE MICROSTRATEGY INDUSTRIAL-STRENGTH BI PLATFORM
UNIFIED WEB INTERFACE
Enterprise Reporting
Cube Analysis
Ad Hoc Query & Analysis
Statistics & Data Mining
Report Delivery & Alerting
MicroStrategy Report Services
MicroStrategy OLAP Services
MicroStrategy Intelligence Server
MicroStrategy Narrowcast Server
Pixel-Perfect Reporting
Intelligent Cubes
Relational OLAP
Analytical Engine
Email / Printers / File Servers
Integrated Backplane
Scalable Server-Centric Architecture
UNIFIED METADATA AND SECURITY
Unified User Interface
• No disorientation caused by multiple UIs
• Only one BI paradigm for users to learn
Unified 5 Styles of BI
• BI applications can exhibit any or all styles of BI
• All applications behave the same way
Unified Services Backplane
• Common Metadata
• Common Security
• Common Prompting
• Common Scheduling
• Common Privilege Control
• Common Administration
• Common Governing

Figure 3.2: The MicroStrategy architecture delivers any or all Styles of BI through a unified user interface and leverages an integrated backplane of unified services

makes decisions. They typically use scorecards and dashboards, the read also reports, that are mostly static.

Microstrategy architecture

During time the MicroStrategy architecture was completely rebuilt from the ground up from 1996 through 2000 to achieve precisely this range of flexibility, along with unparalleled scalability – all the things that companies need. In the 7i version Microstrategy developed a unified user interface in order to give the user always the same environment even using different services or modules. the application layer is composed of all applications belonging to each of the 5 BI styles (five group of services). Those five group of services are performed from four

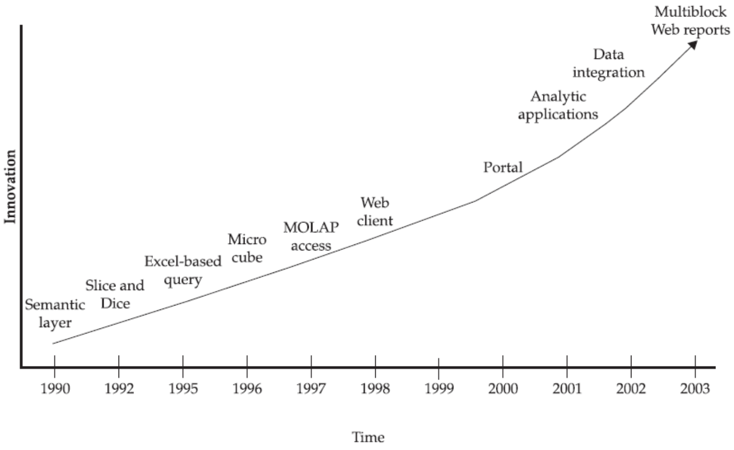


Figure 3.3: Evolution of offered services during time in Business Objects

components:

- Microstrategy Report Service
- Microstrategy OLAP Services
- Microstrategy Ingelligence Server
- Microstrategy Narrowcast Server

At the bottom is visible an integrated backplane that is the data layer.

3.2.2 Business Objects

Overview

In the Figure 3.3 is shown the relatively long history (compared with others BI tools) of Business Objects (BO). Like other BI vendors, BO offers a wide set of products in order to satisfy the need of the companies. Today BO is at the version 6.5 and improved a lot of previous services. Considering the different BI areas, BO has been distinguished in the past for the advanced reporting system. Even today, that offers a global solution, reporting is one of the best services that is possible to have with this tool.

BO now is Web-based and is composed of:

- Query and analysis - tools that allow to query data source (Relational or OLAP) and then analyze them;

- Performance management - to monitor business metrics, analyze performances and choose objectives;
- Analytics - allow business analysis like the customer profiling and ROI[2] analysis;
- Data integration - access, integration, transformation, and distribution among data sources;
- Business Intelligence infrastructure - a framework for BI that includes a portal, a data access interface, an SDK[3] and systems for data access control, monitor and alerting.

3.2.3 COGNOS

Cognos has a set of business intelligence capabilities that allow people across a company to create, modify, and distribute reports. Cognos contains OLAP software that allow to easily perform multidimensional analysis and data visualization software to communicate complex information intuitively. ETL capabilities let the user unite different data sources.

These capabilities help companies execute an overall business strategy of driving breakthrough performance across the enterprise.

3.2.4 Hyperion

Hyperion Business Intelligence Platform combines OLAP[4], query and reporting, ad-hoc analysis, data integration and application development tools. Hyperion platform integrates different types of data from multiple systems through a set of integration technologies. This integration provides a common view of business structures and data across the enterprise, including data from transactional applications such as billing, collections, manufacturing, sales force automation and call center; and data warehouses, spreadsheets, flat files, and Web logs. Hyperion platform leverages common tools such as work flow, modeling, reporting and OLAP within and across Business Performance Management applications to support all aspects of the management cycle. Hyperion platform enables in-house and third-party developers to con-

[2] Return Of Investment

[3] Software Development Kit

[4] On-Line Analytical Processing

centrate on building new applications or extending existing ones rather than on how to integrate new applications once they are developed.

Hyperion uses Visual Explorer a visual tool used by: *brand managers* to understand top and bottom performing products, and increase profitability by correlating product performance to different customer demographic, geographic and channel variables; *Business analysts* to understand trade promotion effectiveness by correlating lift with events, products, customers, and channels. They can allocate the spending of trade dollars based on execution and profitability; *Category managers* to analyze store sales, view the many dimensions of retail sales categories in a simple interface, and sort and filter category views to improve merchandiseplanning; *Quality assurance managers* to quickly "slice and dice" production data and locate problematic outliers, and easily produce interactive pictures of production line performance by dragging and dropping dimensions; *Financial analysts* to analyze, cluster, and categorize their customers along a variety of key metrics; *call center managers* to understand how to up-sell and cross-sell to customers based on various customer attributes and past product purchase history.

3.2.5 Microsoft

A white paper [69] perform a comparison among three of the main OLAP platform that support BI, namely Microsoft SQL Server 2000, IBM DB2 OLAP Server and Hyperion Essbase. Essbase has a strong OLAP engine and IBM has made significant improvements in the capabilities of DB2 OLAP Server, Analysis Services seems to be much more scalable, comprehensive, and affordable solution for Business Intelligence. The database architecture of Analysis Services has set a new standard in the OLAP marketplace, and seems to be superior to that of DB2 OLAP Server/Essbase for three main reasons:

- Its ability to manage data explosion. Analysis Services does not store sparse data, and contains support for virtual dimensions, dynamic calculated measures, and partial aggregation. Analysis Services also applies powerful data compression algorithms to reduce the final size of the cube that gets stored.

- Its ability to store data using MOLAP, ROLAP, or HOLAP across multiple partitions in a single cube, thereby allowing for very large applications.

- Its ability to provide an audit trail and instant availability of aggregations from write-back operations.

Analysis Services can be implemented and deployed more quickly because:

- It provides a series of wizards that walk users through the key steps of designing, building, and tuning the database.
- It can build meta data structures directly from the star and snowflake schemes in the data warehouse, without the need to purchase additional software.
- It uses an open API architecture that enables more rapid development of custom applications.

Analysis Services total cost of ownership is lower because:

- The software licensing fees are simply not as high.
- Implementation costs are reduced due to shorter implementation times. Fewer consulting hours will be required to perform database tuning.
- Cost of support is lower since Analysis Services provides powerful scheduling and data transformation tools.
- Analysis Services has demonstrated that it is the best-in-class OLAP engine.

3.2.6 Oracle

Oracle Business Intelligence 10g[5] features ranges from ad-hoc query, reporting and analysis to Extract, Transform and Load (ETL) to business intelligence application development, in a single package. It tries to accomplish the need of companies that prefer a single environment to perform all the activities related to management and query of information that present in the company.

[5] http://www.oracle.com/appserver/bi_home.html
http://www.oracle.qassociates.co.uk/oracle-9i-business-intelligence.htm

Discoverer [6] is an intuitive ad-hoc query, reporting, analysis, and Web-publishing tool that empowers business users at all levels of the organization to gain immediate access to information from data marts, data warehouses, on-line transaction processing systems and Oracle E-Business Suite. Casual users (those that perform a random access to the database) can view and navigate through pre-defined reports and graphs. Discoverer allow to hide the complexity of the underlying data structure.

In order to consolidate disparate data sources, perform any required data transformations, manage the warehouse life cycle, and integrate with the analysis tools, there is DS Warehouse Builder that first maps transactional sources to a target data warehouse using an extraction, transformation, and loading (ETL) process. Then DS Warehouse Builder generates the code to extract, transform and load data. Once the consolidation take place and data are loaded into a target warehouse, the multidimensional design is shared with Discoverer, another tool of the suite.

Analytic Functions allow users to answer sophisticated business questions on demand, analytics consist of ranking, period-to-period comparisons and moving averages, that are available to administrators and end users. Users can sort, pivot, and drill on the data to meet their analysis needs. Oracle 9i DS Reports will permit multiple queries in the same report, where each query can be based on a different data source.

Reports can publish data from the database using different formats: PDF, XML, HTML, HTML/CSS, Postcript, PCL, Delimited text, and RTF. Users can publish data using industry-standard JSP's [7].

Also developers are supported, they are able to create their own java-based extensions to the Reports Server to open it up to previously unsupported destinations such as Fax, FTP, etc. The broadcast of this information can be done on-demand or scheduled, or even as a reaction to an event that has occurred within an Oracle9 i Database.

Oracle9i Data Mining allows companies to build advanced business intelligence applications that mine company's databases to discover new insights and integrate those insights into business applications. The Oracle9i Database has embedded data-mining functionality like classifications, predictions, and associations.

This allows application developers to integrate data-mining capabil-

[6] http://www.oracle.com/technology/products/discoverer/index.html

[7] Java Server Pages: http://java.sun.com/products/jsp/

ities into their business intelligence applications [8].

Oracle9 i AS Portal is a complete framework for development and deployment of web-based portals. It includes user administration, security, content customization, and development features to create and maintain basic reports, charts, and form-based applications. Creating a Business Intelligence dashboard personalized by job role is easy with Oracle9 i AS Portal. Charts and/or reports representing key performance indicators (KPI's) can be rapidly developed. These charts and reports are deployed as portlets. Individual users may customize their portal presentation by selecting the KPI portlets that are most relevant to their management focus.

3.2.7 SAS

The SAS slogan is "A wealth of knowledge at your fingertips", now is at version 9, and offer an integrated Intelligence Platform that is composed by:

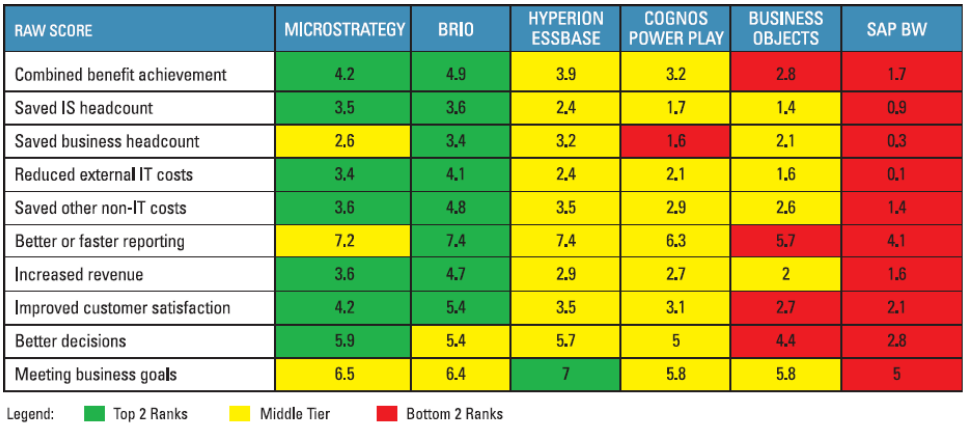
- an ETL Server that includes different access engines, an integrated metadata management, data cleansing and a graphical interface;
- a way to store information o that all the application can access to the same information (like in a data warehouse);
- the SAS Enterprise BI Server empowers users by giving them access to information in the format they need, when they need it. It provides appropriate interfaces for various user skill levels and needs, enabling users to generate their own answers;
- a set of analytics, algorithms, mathematical data manipulation and modeling capabilities. SAS takes the mystery out of these high-end statistical techniques by coupling them with a wide range of user interfaces and graphics. More in detail, it provides data visualization with maps, charts and plots. Users are enabled across the enterprise to visually present their ideas and findings using a huge variety of business maps, charts, plots and 3D relationship graphs. Choropleth, prism, block and surface maps can be created in many colors and patterns. Available charts include vertical and horizontal bar, pie, donut, sub grouped pie and donut, star and

[8] Typical activities supported by data mining are: Preventing customer attrition; Cross-selling to existing customers; Acquiring new customers; Detecting fraud; Identifying the most profitable customers; Profiling customers with more accuracy

block, and many more. Plots include scatter, line, area, bubble, multiple axis and overlay. Data also can be displayed within a three-dimensional coordinate system using response surfaces or scatter plots.

3.2.8 Concluding remarks

Considering the tools that implement OLAP in Figure 3.4 is clear that Microstrategy and Brio have the best solution for the company. Those tools implement also a set of user interface, including wizard and customized to the type of user that helps the user using that tool. The analysis considers several factors, among those factors there is the reporting (the leader is Brio), the improved customer satisfaction, how the system helps making better decisions (the leader is Microstrategy).



RAW SCORE	MICROSTRATEGY	BRIO	HYPERION ESSBASE	COGNOS POWER PLAY	BUSINESS OBJECTS	SAP BW
Combined benefit achievement	🟩 4.2	🟩 4.9	🟨 3.9	🟨 3.2	🟥 2.8	🟥 1.7
Saved IS headcount	🟩 3.5	🟩 3.6	🟨 2.4	🟨 1.7	🟨 1.4	🟥 0.9
Saved business headcount	🟨 2.6	🟩 3.4	🟨 3.2	🟥 1.6	🟨 2.1	🟥 0.3
Reduced external IT costs	🟩 3.4	🟩 4.1	🟨 2.4	🟨 2.1	🟨 1.6	🟥 0.1
Saved other non-IT costs	🟩 3.6	🟩 4.8	🟨 3.5	🟨 2.9	🟨 2.6	🟥 1.4
Better or faster reporting	🟨 7.2	🟩 7.4	🟨 7.4	🟨 6.3	🟥 5.7	🟥 4.1
Increased revenue	🟩 3.6	🟩 4.7	🟨 2.9	🟨 2.7	🟨 2	🟥 1.6
Improved customer satisfaction	🟩 4.2	🟩 5.4	🟨 3.5	🟨 3.1	🟥 2.7	🟥 2.1
Better decisions	🟩 5.9	🟨 5.4	🟨 5.7	🟨 5	🟥 4.4	🟥 2.8
Meeting business goals	🟨 6.5	🟨 6.4	🟩 7	🟨 5.8	🟨 5.8	🟥 5

Legend: 🟩 Top 2 Ranks 🟨 Middle Tier 🟥 Bottom 2 Ranks

Figure 3.4: Table showing the comparison of 6 BI products

Figure 3.5 represent an possible software configuration of BI in a medium-sized company. Typically are the head of departments of the company that choose their customized solutions department automation basing on which product has the greatest strength in one targeted BI usage area. No products today does have a global solution, even though some of them, like Microstrategy, have a vision that covers all the aspect of the company. As a consequence is that in the company there are have several applications, each specialized in a context. One of the current challenges of BI products is to offer a global solution to companies. The need comes from the need to have available the truth of the company's data.

Applying very flexible tools in one of the sector of the company will guarantee that even the architecture changes those tools may still work well on data extracted from the company data source. One of the ideas for the framework presented in the thesis is to handle the result of the query or of the report produced from those systems and offer to the decision maker an alternative way to put the hands on data. Looking to the technical specifications many of those systems are open and offer the possibility to integrate third party tools.

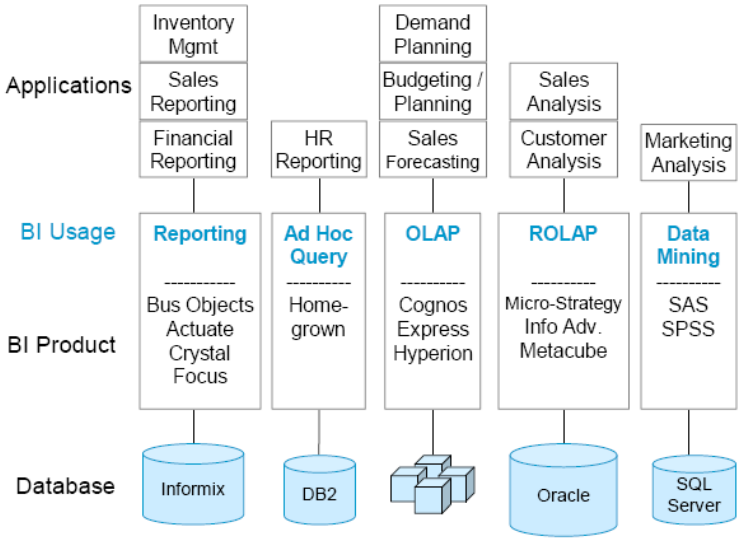


Figure 3.5: Multi-platform analytic application framework

3.3 Information Visualization systems

3.3.1 Information Visualization Prototypes

In this work, we are particularly interested to designing visualization tools that provide users a rapid overview of the content of an information system. Recently, many visual query systems have been developed [49]; such systems use visual representations to depict the domain of interest and express related requests. Indeed, 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 [7], 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 displayed 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 to perform dynamic queries [183]. This is a very interesting visual query formulation technique (see Catarci et al. [49] 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 [183]. 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.

An application of dynamic queries is shown in [183] 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 [7]; 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 tasks the 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 domain of the problem 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 [180].

At Xerox PARC in the last 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 [168, 165, 47, 130]. 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 characteristic 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 [85], 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 [169]. 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 remains oriented across dynamic changes of the display avoiding unnecessary cognitive load. A good example is provided by the Perspective Wall [134]. For other techniques developed at Xerox PARC see [165].

Numerous prototypes have been proposed for information visualization. The ones mentioned above are among those providing the most novel ideas. Shneiderman provides in [45] a very good survey.

3.3.2 CommonGis

CommonGis is a software tool for spatio-temporal data visualization [89, 8]. This class of software on the one hand exploits the opportunities provided by modern computer technologies and, on the other hand, incorporate the legacy from the conventional cartography. Andrienko et al. [8] have undertaken a study with the aim to enumerate the basic set of techniques (regardless of the implementation peculiarities) devised to support exploratory analysis of spatio-temporal data through visualization or in connection with visualization and evaluate these techniques from two perspectives:

1. what types of spatio-temporal data they are applicable to;

2. what exploratory tasks they can potentially support.

The result of the study is a structured inventory of existing techniques related to the types of data and tasks they are appropriate for. This result is potentially helpful for data analysts, in particular for users of geo-visualization tools: it provides guidelines for selection of proper exploratory techniques depending on the characteristics of data to analyze and the goals of analysis.

3.3.3 InfoVis toolkit

InfoVis Toolkit, under many aspects is similar to CommonGis, it is designed to support the creation, extension and integration of advanced 2D Information Visualization components in interactive applications. The InfoVis Toolkit provides specific data structures to achieve a fast action/feedback loop required by dynamic queries. It comes with a large set of components such as range sliders and tailored control panels required to control and configure the visualizations. These components are integrated into a coherent framework that simplifies the management of rich data structures and the design and extension of visualizations. Supported data structures currently include tables, trees and graphs. Supported visualizations include scatter plots, time series, Treemaps, node-link diagrams for trees and graphs and adjacency matrix for graphs. All visualizations can use fisheye lenses and dynamic labeling. The InfoVis Toolkit supports hardware acceleration when available through Agile2D, an implementation of the Java Graphics API based on OpenGL, achieving speedups of 10 to 60 times.

3.4 Visual Data Mining systems

There is a large presence of the VDM systems/tools, there are tools closer to data mining and those closer to information visualization. Some VDM systems are:

- VidaMine [127, 128]

- Interactive Parallel Bar Charts [57]

- Spotfire [188]

- VizMiner [129]

- Visual Analytics [197]

The most comprehensive list of the tools is out of the scope of this thesis, there are many others, among them is worst to mention: AVS/Express [4], CrossGraphs [20], Data Desk [67], DataScope [66], DEVise [140], ADVIZOR [198], JWAVE [199], Open Visualization Data Explorer [109], VisualMine [13].

3.4.1 VidaMine

VidaMine[9] [127, 128] is a visual data mining system designed to support the KDD process. This is a work very close to DAE that first presents a discussion on some systems/tools that offer a reasonably large and diverse number of data mining and visualization functionalities. The authors also acknowledge that effective visual strategies can be used in extracting useful information from data and propose VidaMine, as a VDM environment that can support the entire discovery process. In particular VidaMine is focused on the design and development of several data mining techniques and the respective interfaces.

The key features of VidaMine include:

- an open architecture for the system, which is split into a user layer and a data mining layer;

- a set of infrastructural services allowing interaction between the various software components and providing a clean interface for forthcoming system extensions;

- a consistent, uniform, flexible visual interface based on the goal of supporting the user across the entire data mining process;

- a real user-centered user interface design, equipped with usability studies. Usability methods have been progressively employed in the development life cycle. The usability studies involved expert users, such as data miners, statisticians and data analysts, and casual users, such as managers;

- the uniform presentation of different mining techniques, in order to reach the desired level of integration between system components. In this system there is the first attempt to present a uniform framework for the clustering algorithms and to offer a user interface specially for the visual construction of metaqueries. The system currently supports, but is not limited to, clustering, metaqueries, and association rules;

- a careful definition of visual syntax and formal semantics; each kind of data mining algorithm considered has been carefully analyzed in order to point out the precise meaning of each user choice.

[9] VidaMine is an acronym for VIsual DAta MINing Environment

The type of user of VidaMine is a data miner, that is an expert user, since only techniques such as clustering, association rules and metarules are included, even if the architecture is, in principle, open to the integration of other techniques. The use of such techniques is still very difficult for users like company managers. This is why it was realized a system that, beside classical data mining techniques, includes tools that can be used by managers with no need of any intermediary. More details will be provided in Section 4.

3.4.2 Interactive Parallel Bar Charts

Interactive Parallel Bar Charts [57] proposes an approach for visual data mining on temporal data in the medical domain, i.e. the management of hemodialysis, where clinicians have to deal with huge amounts of data automatically acquired during the hemodialytic treatment of patients suffering form renal failure (a medium-sized hemodialysis center collects about 228 millions of patients' parameter values per year). The approach is based on the integration of 3D and 2D information visualization techniques, such as 3D bar charts and parallel coordinates, and is very much user-centred. Indeed, the first prototype was a VRML-based visualizer of the collections of time-series acquired during hemodialytic treatments, which was discarded because the exploited metaphor was not effective for the end-users. Once the successive prototype, based on 3D bar charts, was evaluated with end-users, it emerged that the visualization and its interactive features were very quickly learned and remembered by clinicians. User evaluation also pointed out a number of needed usability improvements and new functionalities required by clinicians. For example, it emerged that the clinician, who is studying a parallel bar chart illustrating a parameter, needs to know the value of several other parameters, but only for a few selected time instants: Therefore, in order to permit more flexibility in the analysis of several parameters on the same screen, a new version of the prototype included another type of visualization, based on the well-known technique of Parallel Coordinates, that allows the clinician to relate the parallel bar chart he/she is considering with many other.

3.4.3 Spotfire

Spotfire DecisionSite [188] is a guided analytic application and platform for rapidly generating analytic applications for any business process and data source. A complete guided analytic application and platform providing a user-configurable environment that speeds interactive, multi-variant data analysis; integrates customer business processes; provides the infrastructure for communications and collaborative decision-making; easily integrates existing data and application infrastructure.

The Spotfire DecisionSite is composed by:

- Visual, Interactive Analysis

- Guided Processes

- Collaborative Decision-making

Visual, Interactive Analysis provides users a single user environment to rapidly relate and continuously ask questions about complex data. It includes an interactive application framework and a set of analysis tools for interacting with data using visualizations and direct manipulation in order to allow the decision maker to make decisions based on analysis data from multiple sources. For enhanced statistical analysis, the DecisionSite Statistics application adds a library of interactive analytic statistics, mathematical functions, and data mining capabilities for any DecisionSite application.

Guided Processes. Spotfire pioneered the ability to capture analysis processes and generate a configured application called a "Guided Analytic Application". A Guided Analytic Application combines customer unique analysis expertise and processes with the DecisionSite visual, interactive application environment so all users have a consistent business process framework for performing analysis and making decisions. This capability enables customers to have one application that includes visual interactive analysis, data access, and support for expert internal analysis workflows. With this capability, it provides the ability to integrate different people with different expertise.

Collaborative Decision-making. Spotfire have also an infrastructure to broadly publish, capture, and share live analysis insights and interactively make decisions. DecisionSite Posters enables customers to maximize the ROI[10] in using analytic applications by making it fast and easy to involve teams, groups, organizations, or companies in the analysis and decision making process to make each decision the highest quality. The entire set of related results, including the analysis data

[10] Return Of Investment

and annotated expertise, are saved on a Poster and stored in a internet accessible library.

3.4.4 VizMiner

Kopanakis and Theodoulidis propose some visual data-mining models on which they have constructed graphical representations of the outcomes produced by common data mining processes [129]. The goal of VizMiner is to equip the knowledge engineer with a tool that would be utilized on his/her attempt to gain insight over the mined knowledge, the tool presents as much information extracted in a human perceivable way. Additionally, VizMiner is built on guidelines that led the construction of each data mining model, as long as the definition of the underlying representational ideas. The different models have distinctive advantageous characteristics, addressing the commonly tedious issues that the knowledge engineer handles during the exploitation of the mining outcomes. Furthermore, the possibility to combine forces and enhance the information flow among the different models and the user, brings the users one step closer to make human part of the data mining process, in order to exploit human's unmatched abilities of perception.

3.4.5 Visual Analytics

Visual Analytics Inc. is a commercial product that uses several VDM techniques in order to perform link analysis, information sharing, and collaboration technology and services [197]. This tool is specialized in the fraud analysis, Visual Analytics exploits graph visualizations and matrix visualization in order to highlight relationships and periodicity. In this way it facilitates to spot potential fraud. The behavior of a typical user is taken into account and the potential fraud may be recognized from a non consistent behavior or considering relationships with suspicious persons. The features of the tool integrate a wide variety of data sources, permitting the building of a virtual data warehouse, facilitating a collaborative environment, enabling the analysis, reporting and exporting data. It is primarily uses as an analytical tool for pattern discovery, link analysis, data visualization and network-centric analyses.

Chapter 4

The framework for data analysis

4.1 Introduction

Companies support their business plans with tools that help managers to make decisions in a rapid and more effective way. In previous sections we have seen that there are many tools that allow this in different forms. There is still a lot of work to do in order to present data in a convincing and understandable way in particular when data change dynamically; it is also difficult to modify the graphical layout without disorienting the users.

As we have pointed out in previous chapters, Business Intelligence is a good way to help companies to perform the right choices and to exploit all information present in the company itself. We have also seen how Business Intelligence integrates with the Knowledge Discovery in Databases process, those (BI and KDD) are two distinct areas that have different users and different peoples working on it. ERP systems and data warehouses glue all the information in order to have a unique source that have the true of the information in a company. Over this unique data source people can build processes and applications that access and get information from data in order to satisfy the different needs of the different users of the company. Using good visualizations to present the information hidden in various company repositories can improve the decision process.

Advances in information visualization offer promising techniques for presenting knowledge structures [53] and for permitting explorative analyses of the data [68, 121]. Knowledge visualization has various interpretations depending on the authors [59, 55]. It can be defined as the visual explication of conceptual knowledge [59] based on understanding the domain knowledge, applying cognitive principles, and encoding important features graphically by exploiting the visual parameters. It is considered the intersection of three main areas that are cognitive science, graphic design, and information graphics. Cognitive science helps to understand the cognitive processes underlying perception, categorization, visual reasoning, communication, creativity, and motivation. Graphic design exploits the rich legacy of art and illustration. Finally, information graphics refers to the various graphs and diagrams visualizing quantitative information. Often, visual representations are used to identify single elements in a large knowledge base and also to show explicit relationships between elements. The benefits of visual representations comes from their ability to shift some of the 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 understood by users. This visualization can either be carried out by the users in their own mind, in which case it is essentially the users' conceptualization of that information, or it could be aided by the system by generating visualization on the display screen, thus reducing the users' cognitive load.

In the next sections will be presented how visualizations can be provided to assist users in their decision processes, presenting a modular framework that support organizations. This framework is called DAE (acronym for Data Analysis Engine) and was a component of the FairsNet system described in 1.2.1. After the project ended we continued in the development of the ideas in that framework, improving the modules, and adding newer, in order to have more than one visualization technique that allow the data analysis.

The aim of the framework is to assist the users in their decision making processes using visualizations. Even if FairsNet focused on a specific application domain, namely trade fair management, the framework is quite general and applicable to different domains. Several data visualizations are generated to explore the data and to present the retrieved information in appropriate ways for each user category.

4.2 DAE: Data Analysis Engine

FairsNet and the users involved in it have been described in Section 1.2.1. The framework for VDM we have developed within FairsNet is called DAE (Data Analysis Engine). DAE is primarily related to decision support tasks, its main aim consists in:

- managing and improving interactive relationships among the trade fair users;
- segmenting exhibitors and visitors on the basis of various characteristics;
- finding relationships, if exist, between data of the trade fair database.

To address the needs of specific types of users, primarily fair organizers and exhibitors, by allowing them to easily retrieve information useful for their marketing activities, DAE exploits appropriate visualization techniques, in accordance with the Visual Data Mining goals.

4.2.1 DAE architecture

The framework is general enough to be used in contexts different from trade fairs. DAE architecture is shown in Fig. 4.1. DAE is a module directly accessible by the users. In the FairsNet context DAE is connected to the FairsNet database, we can currently connect DAE to any database.

The application server is composed by a local database and the DaeMine module. This is the data mining component which implements descriptive data mining algorithms to find interesting patterns inside data. We included algorithms for association rules generation and algorithms for clustering, which will be described in Section 4.4.

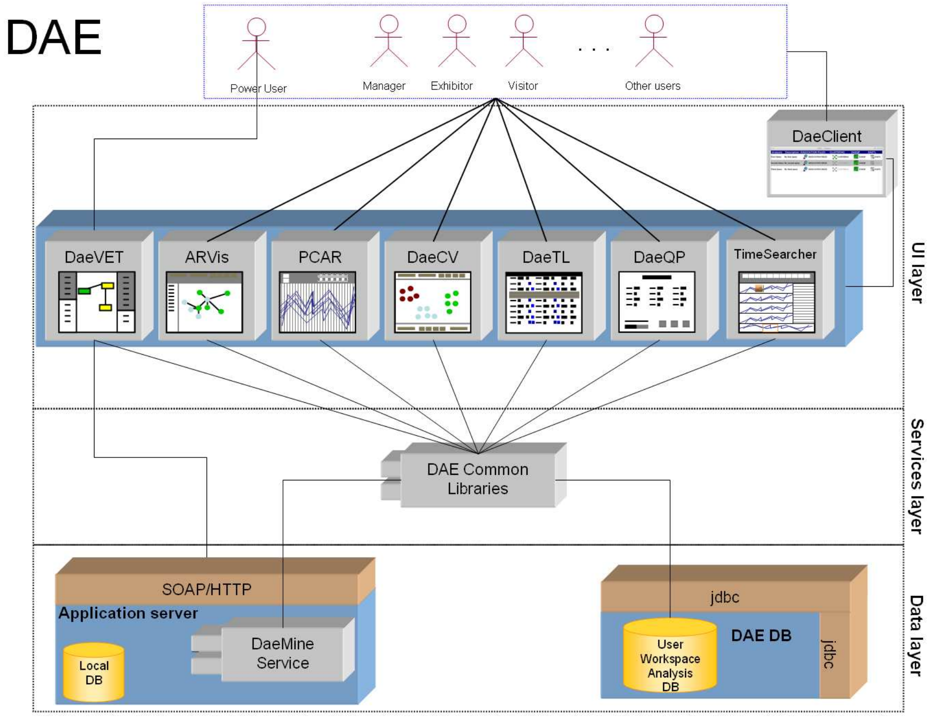


Figure 4.1: DAE architecture

DaeDB stores the results of data mining algorithms as well as the results of queries on the application domain databases. DAE Common Libraries includes all libraries that permits the communication among the DAE components. DAE may include several visual modules that are accessible by users to allow them to analyze the stored data. They compose the UI (User Interface) layer in the architecture in Fig. 4.1. Currently, in the framework there are the following visual modules:

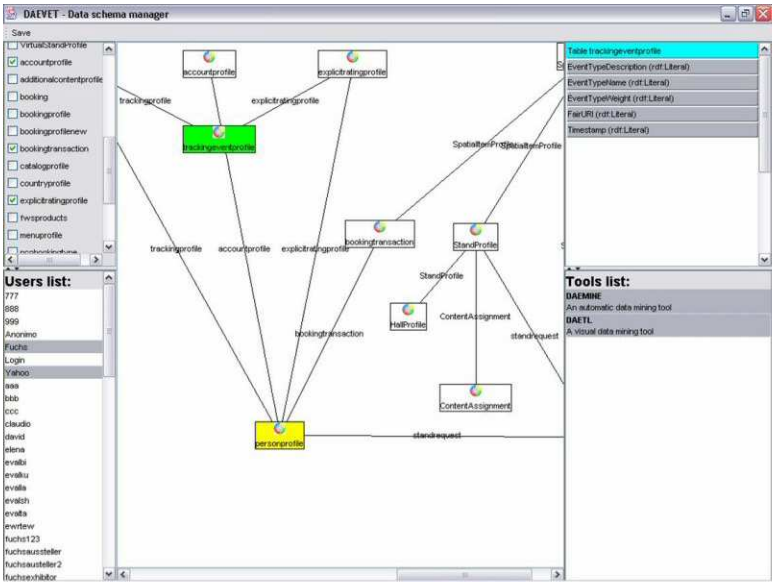
- DaeVET enables the user to select relevant data and to create analysis [33];

- ARVis visualizes association rules using graphs and allows the user to interact with them [32, 28];
- PCAR uses parallel coordinates to visualize association rules in a way that is complementary to ARVis [28];
- DaeCV allows the user to analyze the result of the clustering algorithms used in DaeMine;
- DaeTL allows the user to analyze the data by using a technique based on Table Lenses [164];
- DaeQP allows the user to analyze the data by using a technique based on Query Preview [182];
- TimeSearcher allows the user to interact with time series data and to perform multivariate analysis by exploring and searching in time series.

4.3 DaeVET: a visual tool to extract information from databases

Following the framework for Web-based Information, Content and Knowledge Management (ICKM) system approach [147], the final user may customize the system in which he is interacting with to tailor the system to the own needs. This implies that several analyses the user might perform are not known in advance. The data present in the system may also change, and may even change the structure of data. The analysis tools should be general enough to be adapted to a particular configuration of the system and to satisfy the need of the user.

The decision maker needs to analyze data in the company's database in order to make the right decisions. The process of making decisions is made of different steps. We have seen the knowledge discovery process in Section 2.3, the first action to do before any analysis can be done is to select the data from the database. This is often a long and difficult task, and it is sometimes needed to iterate it because in the selection some data may be missed. Moreover, people that need data are not those that can provide them (i.e. database administrators provide data for the marketing division to produce reports). Improving this task will result in a reduced time to acquire data and to perform the knowledge discovery process. DaeVET is a module used by the so-called power user, an expert user that knows the system and the data of the particular database. Figure 4.2 shows the DaeVET interface.



- □ VirtualStandProfile
- ☑ accountprofile
- □ additionalcontentprofile
- □ booking
- □ bookingprofile
- □ bookingprofilenew
- ☑ bookingtransaction
- □ catalogprofile
- □ countryprofile
- ☑ explicitratingprofile
- □ fwsproducts
- □ menuprofile

Users list:
777
888
999
Anonimo
Fuchs
Login
Yahoo
aaa
bbb
ccc
claudio
david
elena
evalbi
evalku
evalla
evalsh
evalta
ewrtew
fuchs123
fuchsaussteller
fuchsausteller2
fuchsexhibitor

Table trackingeventprofile
EventTypeDescription (rdf:Literal)
EventTypeName (rdf:Literal)
EventTypeWeight (rdf:Literal)
FairURI (rdf:Literal)
Timestamp (rdf:Literal)

Tools list:
DAEMINE
An automatic data mining tool
DAETL
A visual data mining tool

Figure 4.2: DaeVET interface the user may choose among different tools and analysis

DaeVET visualizes the data that will be used for the analyses. Once the structure of the database is displayed to the screen the user may interact with it performing dragging, selection, zooming and panning to chose the data in which he is most interested in. The power user may select tables and attributes and once he has finished this task he may select the users and the group of users that will be able to view that data. In order to finalize the task the user selects which tools is appropriate for the selected analysis.

For some analysis, in particular for the data mining tasks, it is not possible to offer to the user the data and the results of the task in real-time, because they need some computational time. In this case, when the power user stores the analysis, the data mining algorithm runs and, at the end of the computation, the results are stored on the server. Then the results will be available for the end-user.

4.4 DaeMine: running data mining algorithms

In this section are presented two DAE modules that perform data mining tasks, the first produces association rules and the second cluster data. In DaeMine there is a data mining engine that is called each time a data mining task is selected for the analysis.

4.4.1 Association Rules

The demand for visual and interactive analysis tools is particularly pressing in the Association Rules context where often the user needs to analyze hundreds of rules in order to grasp valuable knowledge. The proposed VDM framework includes a visual strategy to face this problem; it exploits a graph-based technique and parallel coordinates to visualize the results of the association rules mining algorithm used by DAE. The combination of the two approaches allows both to get an overview on the association structure hidden in the data and to deeper investigate inside a specific set of rules selected by the user. In the following we provide a brief description of association rules, while in Section 4.5 and 4.7 will be presented the visual modules.

Association rules can be defined as follows: let $I = i_1, i_2, \ldots, i_n$ be a set of items called literals (in the market basket analysis the items could be the products sold in the supermarket). The database consists of a set of transactions $\mathrm{T} = T_i, A_p, \ldots, A_q$, where $A_i \in I$ for $i = p, \ldots, q$ and T_i is the identifier of the transaction. Each transaction $T_i \in \mathrm{T}$ is a set of items, such that $T \subset \mathrm{T}$. An association rule is a condition of the form $X \rightarrow Y(s, c)$, where $X \subseteq I$ and $Y \subseteq I$, $X \cap Y = \emptyset$, s and c are called respectively support and confidence. The support s of the rule R is $s = n_R/n$, where n_R is the number of transaction in T holding $X \cup Y$ and n is the total number of transaction. The support represent the proportion of transactions containing both the antecedent and the consequent, and don't care about possible relationships between the antecedent and the consequent. The confidence c of a rule R is $c = n_R/n_X$ where n_X is the number of transactions with X in the left side of the implication. The confidence is a measure of the conditional probability of the consequent, given the antecedent. The confidence expresses the strength of the logical implication described by the rule. The main goal of researchers working in this area is to mine association rules, i. e., to produce as many significant rules as possible. This means that they have to produce as many rule as possible, discarding those with low meaning. The task of discarding rules is called pruning. Typically rules are pruned if they don't reach a minimum support or confidence threshold. Some authors introduced other parameters to improve the pruning phase and keep interesting patterns, see for example the Difference of Confidence in [106] or the Item Utility in [29]. In order to improve data mining task, researcher can reduce the number of interesting rules generated, or it is possible to produce good visualization tools that allow to perform explorative analysis of ARs.

4.4.2 Clustering

Clustering is a process through which the target dataset is divided into groups of similar objects. Each group, called cluster, contains objects that are similar each other. The objects in a particular group/cluster are dissimilar to objects in another or other groups/clusters. Clustering is sometimes referred to as unsupervised learning or unsupervised classification. The statistics, machine learning and data mining literature contains a huge body of work on clustering [99, 115, 78, 116]. Clustering is applicable in many areas such as: astronomical data, demographics, insurance, urban planning, and Web applications.

Classification of Clustering Methods

Traditionally, clustering methods have been classified into a taxonomy having two broad groups: hierarchical and partitional [116].

Hierarchical clustering

Hierarchical methods produce a sequence of nested partitions; which is a tree of clusters. The tree, which is referred to as a dendrogram, shows the hierarchical relationships among the clusters. It is therefore possible to explore the underlying dataset at various levels of granularity. Hierarchical methods are further subdivided into agglomerative and divisive [115, 120].

Agglomerative

This is a bottom-up approach in which the clustering starts with singletons (each cluster containing exactly one point). The clustering then recursively merges two or more most appropriate clusters. The process goes on until a stopping criterion is fulfilled (such as the number of clusters chosen by the user).

Examples of agglomerative algorithms include: CURE (Clustering Using REpresentatives) [92], and CHAMELEON [119].

Divisive

This is a top-down approach in which the clustering starts with one single cluster containing all the objects, and recursively subdivides the most appropriate cluster. The process goes on until some criterion is met. The PDDP (Principal Direction Divisive Partitioning) algorithm is an example of divisive algorithms [23]. Also in this category of divisive clustering are approaches based on the k-means algorithm [99, 100] such as the bisecting k-means algorithm [189, 201].

Partitional Clustering

Partitional methods attempt to identify clusters directly either: by iteratively relocating points between subsets, or by associating clusters with the areas that are densely populated with data. Consequently, partitional methods fall into two categories: relocation methods and density-based methods. Relocation methods focus on how well points fit into their clusters. Such methods intend to ensure that the built clusters have the proper shapes. Relocation methods are further subdivided into: probabilistic, k-medoids, and k-means. The probabilistic clustering model is based on the assumption that, data has been independently drawn from a mixture model of several probability distributions. The results of probabilistic clustering are often easy to interpret. Probabilistic clustering algorithms include: SNOB [200], AUTOCLASS [52], MCLUST [84].

In clustering methods that adopt the k-medoids approach, a cluster is represented by one of its points. When the medoids are selected, clusters are considered to be subsets of points close to respective medoids. Algorithms based on the k-medoid approach include: PAM (Partitioning Around Medoids) and the algorithm CLARA (Clustering LARge Applications) [120], CLARANS (Clustering Large Applications based upon RANdomized Search) [146]. In k-means [99, 100], a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster. Although the k-means algorithm does not work well with a dataset that has categorical attributes, the algorithm is an appropriate choice for datasets with numerical attributes.

Density-based methods aim at identifying connected components/areas in the dataset that are dense with data. In this respect, a cluster therefore corresponds to a connected dense component. Density-based methods can be further divided into two main categories: density-based connectivity and density functions. Density-based connectivity approach reduces density to a training data point. Algorithms that use this approach include: DBSCAN [77], OPTICS (Ordering Points To Identify the Clustering Structure) [11], DBCLASD (Distribution Based Clustering of Large Spatial Databases) [211]. Density functions approach reduces density to a point in the attribute space. DENCLUE [103] is an example of an algorithm based on density functions. In fact, DENCLUE is a blend of density-based clustering and grid-based preprocessing. There exist many other clustering techniques that do not fit well in one of foregoing categories. For instance: grid-based techniques, co-occurrence techniques, etc.

Grid-based techniques work indirectly with data by constructing summaries of data over the attribute space subsets. They segment the space and then aggregate appropriate segments. On the one hand, grid-based methods often use hierarchical agglomeration as a phase in their processing. Algorithms BANG [173], STING (STatistical INformation Grid-based method) [202], WaveCluster [175]. On the other hand, the idea behind grid-based methods is exploited by other types of clustering algorithms (such as CLIQUE (Clustering In QUEst) [5], MAFIA (Merging of Adaptive Finite IntervAls) [90, 144]) as an intermediate phase in their processing.

Co-occurrence techniques are meant to handle such special requirements when it comes to clustering categorical data. Algorithms ROCK [93], SNN (Shared Nearest Neighbors) [76], and CACTUS (Clustering Categorical Data Using Summaries) [87].

The Proposed Clustering Taxonomy and Framework

The traditional categorization of clustering methods into two broad groups: hierarchical and partitional [116] is technically sound and relevant to various application domains.

However, such categorization does not highlight similarities and differences between the various definitions of a cluster that are implicit in the methods. For instance, Ward's minimum-variance method [203] and the PAM method PAM (Partitioning Around Medoids) [120], are similar. However, the former is hierarchical, whereas the latter is partitional. As an alternative to the traditional approach of categorizing clustering methods, clustering can be seen as an optimization problem, in which the function to be optimized is a mathematical measure of homogeneity or separation [95]. Such a perspective enables one to categorize clustering methods according to a taxonomy of homogeneity or separation functions. Therefore such a perspective provides recourse for categorizing clustering methods.

Moreover, such a taxonomy expresses cluster definitions implicitly. Such a categorization is most likely more effective in capturing different behaviors in practice. It therefore provides a more natural avenue for the process of selecting a clustering algorithm which is most suited to a particular application or domain.

4.5 ARVis: visualizing mined association rules using graphs

4.5.1 Introduction

The main purpose of ARVis is to allow a data miner to interact with association rules. The best use of ARVis is when coupled with a data mining tool that produces association rules. To design ARVis it has been made a survey on tools visualizing association rules, some related work is presented in the Section 4.5.5. We have integrated ARVis with a tool working with spatial database. Spatial databases suffer of the problem of the large quantity of data to process, this because typically spatial database contain raster images and/or vector maps. Satellite or remote sensing systems nowadays are of large use, those have paved the way for advances in spatial databases. A spatial database contains (spatial) objects that are characterized by a geometrical representation (e.g. point, line, and region in a 2D context), a relative positioning with respect to some reference system as well as several non-spatial attributes. The widespread use of spatial databases in real-world applications, ranging from geo-marketing to environmental analysis or planning, is leading to an increasing interest in spatial data mining, i.e. extracting interesting and useful knowledge not explicitly stored in spatial databases.

Spatial association rules discovery is an important task of spatial data mining that aims at discovering interactions between reference objects (i.e. unit of observation in the analysis) and one or more spatially referenced target-relevant objects or space dependent attributes, according to a particular spacing or set of arrangements. This task presents two main sources of complexity that is the implicit definition of spatial relations and the granularity of the spatial objects. The former is due to geometrical representation and relative positioning of spatial objects which implicitly define spatial relations of different nature, such as directional and topological. The second source of complexity refers to the possibility of describing the same spatial object at multiple levels of granularity. For instance, United Kingdom census data can be georeferenced with respect to the hierarchy of areal objects ED → Ward → District → County, based on the internal relationship between locations. This suggests that taxonomic knowledge on task-relevant objects may be taken into account to obtain multi-level spatial association rules (descriptions at different granularity levels).

A full-fledged system that copes with both these issues is ARES (Association Rules Extractor from Spatial data) [12] that integrates SPADA (Spatial Pattern Discovery Algorithm) [132] to extract multi-level spatial association rules by exploiting an Inductive Logic Programming (ILP) approach to (multi-) relational data mining [72]. ARES assists data miners in extracting the units of analysis (i.e. reference objects and task-relevant objects) from a spatial database by means of a complex data transformation process that makes spatial relations explicit, and generates high-level logic descriptions of spatial data by specifying the background knowledge on the application domain (e.g. hierarchies on target-relevant spatial objects or knowledge domain) and defining some form of search bias to filter only association rules that fulfill user expectations.

Nevertheless, ARES may produce thousands of multi-level spatial association rules that discourage data miners to manually inspect them and pick those rules that represent true nuggets of knowledge at different granularity levels.

While a lot of research has been conducted on designing association rules exploratory visualization [30], no work, in our knowledge, properly deal with multi-level spatial association rules. ARVis visualize, allow the navigation and interpretation of multi-level spatial association rules by exploiting both the knowledge embedded on hierarchies describing the same spatial object at multiple levels of granularity and the number of refinement steps performed to generate each rule.

4.5.2 Multi-level spatial association rules

The problem of mining multi-level spatial association rules can be formally defined as follows: *Given* a spatial database (SDB), a set S of reference objects, some sets R_k, $1 \leq k \leq m$, of task-relevant objects, a background knowledge BK including some spatial hierarchies H_k on objects in R_k, M granularity levels in the descriptions (1 is the highest while M is the lowest), a set of granularity assignments ψ_k which associate each object in H_k with a granularity level, a couple of thresholds minsup[l] and minconf[l] for each granularity level, a language bias LB that constrains the search space; *Find* strong multi-level spatial association rules, that is, association rules involving spatial objects at different granularity levels.

The reference objects are the main subject of the description, namely unit of observation, while the task-relevant objects are spatial objects that are relevant for the task in hand and are spatially related to the former. Both the set of target object S and the sets of target relevant objects R_k typically correspond with layers of the spatial database, while hierarchies H_k define *is-a* (i.e., taxonomic) relations of spatial objects in the same layer (e.g. regional road is-a road, main trunk road is-a road, road is-a transport net). Objects of each hierarchy are mapped to one or more of the M user-defined description granularity levels in order to deal uniformly with several hierarchies at once. Both frequency of patterns and strength of rules depend on the granularity level l at which patterns/rules describe data. Therefore, a pattern P (s%) at level l is frequent if $s \geq minsup[l]$ and all ancestors of P with respect to H_k are frequent at their corresponding levels. The support s estimates the probability $p(P)$. An association rule $A \rightarrow C$ (s%, c%) at level l is strong if the pattern $A \cup C$ (s%) is frequent and $c \geq minconf[l]$, where the confidence c, estimates the probability $p(C|A)$ and A (C) represents the antecedent (consequent) of the rule.

Since a spatial association rules is an association rule whose corresponding pattern is spatial (i.e. it captures a spatial relationship among a spatial reference object and one or more target-relevant spatial object or space dependent attributes), it can be expressed by means of predicate calculus.

An example of spatial association rule is "*is_a(X, town), intersects(X, Y), is_a(Y, road)* → *intersects(X,Z), is_a(Z, road), Z≠Y* (91%, 100%)" to be read as "if a town X intersects a road Y then X intersects a road Z distinct from Y with 91% support and 100% confidence", where X denotes a target object in town layer, while Y and Z some target-relevant object in road layer. By taking into account taxonomic knowledge on task-relevant objects in the road layer, it is possible to obtain descriptions at different granularity levels (multiple-level spatial association rules). For instance, a finer-grained association rules can be "*is_ a(X, town), intersects(X, Y), is_ a(Y, regional_ road) →intersects(X, Z), is_ a(Z, main_ trunk_ road), Z≠ Y* (65%,71%)", which states that "if a town X intersects a regional road Y then X intersects a main trunk road Z distinct from Y with 65% support and 71% confidence."

The problem above is solved by the algorithm SPADA that operates in three steps for each granularity level: i) pattern generation; ii) pattern evaluation; iii) rule generation and evaluation. SPADA takes advantage of statistics computed at granularity level l when computing the supports of patterns at granularity level $l + 1$.

In the ARES system [1], SPADA has been loosely coupled with a spatial database, since data stored in the SDB Oracle Spatial are preprocessed and then represented in a deductive database (DDB). Therefore, a middle layer is required to make possible a loose coupling between SPADA and the SDB by generating features of spatial objects. This middle layer includes both the module RUDE (Relative Unsupervised DiscrEtization) to discretize a numerical attribute of a relational database in the context defined by other attributes [133] and the module FEATEX (Feature Extractor) that is implemented as an Oracle package of procedures and functions, each of which computes a different feature. According to their nature, features extracted by FEATEX can be distinguished as geometrical (e.g. area and length), directional (e.g. direction) and topological features (e.g. crosses) [12]. Extracted features are then represented by extensional predicates. For instance, spatial intersection between two objects X and Y is expressed with *crosses(X,Y)*. In this way, the expressive power of first-order logic in databases is exploited to specify both the background knowledge BK, such as spatial hierarchies and domain specific knowledge, and the language bias LB. Spatial hierarchies allow to face with one of the main issues of spatial data mining, that is, the representation and management of spatial objects at different levels of granularity, while the domain specific knowledge stored as a set of rules in the intensional part of the DDB supports qualitative spatial reasoning. On the other hand, the LB is relevant to allow data miners to specify his/her bias for interesting solutions, and then to exploit this bias to improve both the efficiency of the mining process and the quality of the discovered rules. In SPADA, the language bias is expressed as a set of constraint spec-

[1] http://www.di.uniba.it/ malerba/software/ARES/index.htm

ifications for either patterns or association rules. Pattern constraints allow to specify a literal or a set of literals that should occur one or more times in discovered patterns. During the rule generation phase, patterns that do not satisfy a pattern constraint are filtered out. Similarly, rule constraints are used do specify literals that should occur in the head or body of discovered rules.

In SPADA, users can specify exactly both the minimum and maximum number of occurrences for a literal in a pattern (head or body of a rule) and the maximum number of literals to be included in the head of a rule. In this way users may define the head structure of a rule requiring the presence of exactly a specific literal and nothing more. In this case, the multi-level spatial association rules discovered by ARES may be used for sub-group discovery tasks.

4.5.3 Multi-level spatial association rules graph-based visualization

A set R of multi-level spatial association rules can be naturally partitioned into $M \times N$ groups denoted by R_{ij}, where i ($1 \leq i \leq M$) denotes the level of granularity in the spatial hierarchies H_k, while j ($2 \leq j \leq N$) the number of refinement steps performed to obtain the pattern (i.e. number of atoms in the pattern). Each set R_{ij} can be visualized in form of a graph by representing antecedent and consequent of rules as nodes and relationships among them as edges.

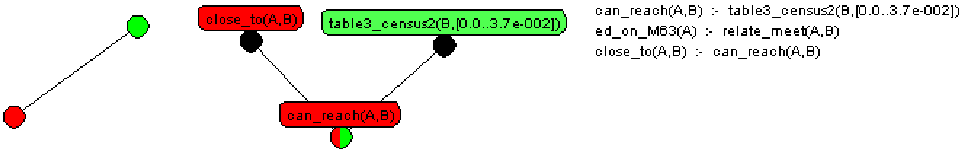
This graph-based visualization can be formally defined as follows: Given an association rules set R, a directed (not completely connected) graph $G = (N, E)$ can be built from R, such that:

- N is a set of couples (l, t), named *nodes*, where l denotes the conjunction of atoms representing the antecedent (A) or consequent (C) of a rule $A \rightarrow C \in R$, while t is a flag denoting the node role (i.e. antecedent, consequent or both of them).

- E is a set of 4-tuples (n_A, n_C, s, c), named edges, where n_A is a node with the role of antecedent; n_C is a node with the role of consequent, while s and c are the support and confidence of the rule $n_A.l \rightarrow n_C.l \in R$ respectively.

Each node of G can be visualized as a colored circle: a red circle represents a node n with the role of antecedent ($n.t = antecedent$) while a green circle represents a node n with the role of consequent

($n.t = consequent$). If the node has the role of antecedent for a rule and consequent for a different rule, it appears half red and half green. The label $n.l$ can be visualized in a rectangular frame close to the circle representing n. Conversely, each edge in G can be visualized by a straight segment connecting the node n_A with the node n_C. It corresponds with the rule $n_A.l \rightarrow n_C.l$ that exists in R. The confidence of this rule is coded by the length of the edge, the greater is the confidence, the longer is the edge. Conversely, the support is coded by color saturation of the edge: from light blue (low support) to black (high support). Support and/or confidence can be also visualized in a text label close to the edge (see Figure 4.3).



```
can_reach(A,B) :- table3_census2(B,[0.0..3.7e-002])
ed_on_M63(A) :- relate_meet(A,B)
close_to(A,B) :- can_reach(A,B)
```

Figure 4.3: Visualizing the graph of spatial association rules

As suggested by [28], this graph representation appears beneficial in exploring huge amount of association rules in order to pick interesting and useful patterns, since it takes advantages from human perceptual and cognitive capabilities to immediately highlight which association rules share the same antecedent or consequent with respect to the overall distribution of rules. Filtering mechanisms which permit to hide a sub-graph of G (i.e. subset of rules in R) according to either minimal values of support and confidence or the absence of one or more predicates in the rule provide a better interaction.

To explore multi-level spatial association rules discovered by ARES, this graph-based visualization should be further extended in order to enable data miners to navigate among several graphs G_{ij} according to either the levels of granularity i or the number of refinement steps j. In the former case, for each pair of granularity levels (i, h) with $1 \leq i < h \leq M$ $(1 \leq h < i \leq M)$ and number of refinement steps j $(2 \leq j \leq N)$, a specialization (generalization) operator $\rho_{i\downarrow h,j}$ $(\delta_{i\uparrow h,j})$ can be defined as follows:

$$\rho_{i\downarrow h,j} : R_{ij} \rightarrow \wp(R_{hj}) \ (\delta_{i\uparrow h,j} : R_{ij} \rightarrow \wp(R_{hj})),$$

where $\wp(R_{hj})$ denotes the power set of R_{hj}. For each spatial association rule $A \rightarrow C \in R_{ij}$, $\rho_{i\downarrow h,j}(A \rightarrow C) = \{A_1 \rightarrow C_1, \ldots, A_w \rightarrow C_w\}$, such that each $A_k \rightarrow C_k \in R_{hj}$ $(k = 1, \ldots, w)$ and $A_k \rightarrow C_k$ is a down-specialization (up-generalization) of $A \rightarrow C$.

To formally define the relation of down-specialization (up-generalization) between two spatial association rules, we represent each spatial rule $A \rightarrow C$ as $A_S, A_I \rightarrow C_S, C_I$, where A_S (C_S) includes all atoms in A (C) describing either a property (e.g. *area*$(X,$ [10..15]) or *cars*$(X,$ [150..1000])), a relationship (e.g. *intersect*$(X,\ Y)$) or an inequality (e.g. $X/ = Y$). Conversely, A_I (C_I) includes all is_a atoms (e.g. $is\_a(X, road)$). Therefore, $A' \rightarrow C' \in R_{hj}$ is a down-specialization of $A \rightarrow C \in R_{ij}$ iff there exists a substitution θ (i.e. a function that associates a variable with a term) that renames variables in $A' \rightarrow C'$ such that $A_S = A'_S\theta$, $C_S = C'_S\theta$, and for each is_a atom of $A_I(C_I)$ in the form $is\_a(X, v_i)$, where X denotes a target relevant object in R_k and v_i is a node at level i of the spatial hierarchy H_k, there exists an atom $is\_a(X, v_h)$ in $A'_I\theta$ $(C'_I\theta)$ with v_h a node in the sub-hierarchy of H_k that is rooted in v_i. The up-generalization differs from down-specialization only in requiring that v_i is a node in the sub-hierarchy of H_k that is rooted in v_h and not vice-versa.

Example: Let us consider the spatial association rules:
$R1$: intersects($X1$, $Y1$), cars($X1$, [25, 120]),is_a($X1$, town), is_a($Y1$, road) $\rightarrow$ mortality($X1$, high).
$R2$: intersects($X2$, $Y2$), cars($X2$, [25, 120]), is_a($X2$, town),
is_a($Y2$, main_trunk_road)$\rightarrow$ mortality($X2$, high).
where $R1.A_S$ is "intersects($X1$, $Y1$), cars($X1$, [25, 120])" and $R1.C_S$ is "mortality ($X1$, high)", while $R1.A_I$ is "is_a($X1$, town), is_a($Y1$, road)" and $R1.C_I$ is empty. Similarly $R2.A_S$ is "intersects($X2$, $Y2$), cars($X2$, [25, 120])" and $R2.C_S$ is "mortality ($X2$, high)", while $R2.A_I$ is "is_a($X2$, town), is_a($Y2$, main_trunk_ road)" and $R2.C_I$ is empty. R2 is a *down-specialization* of R1 since there exists the substitution $\theta = \{X2/X1, Y2/Y1\}$ such that $R1.A_S = R2.A_S\theta$, $R1.C_S = R2.C_S\theta$, and main_trunk_road is a specialization of road in the corresponding hierarchy. Conversely, R1 is an *up-generalization* of R2.

A different specialization (generalization) operator $\rho_{i,j\rightarrow h}$ $(\delta_{i,j\leftarrow h})$ can be further defined, for each granularity level i and pair of refinement step numbers (j, h) with $2 \leq j < h \leq N$ $(2 \leq h < j \leq N)$, such that:

$$\rho_{i,j\rightarrow h} : R_{ij} \rightarrow \wp(R_{ih})\ (\delta_{i,j\leftarrow h} : R_{ij} \rightarrow \wp(R_{ih})),$$

In this case, for each spatial association rule $A \rightarrow C \in R_{ij}$, $\rho_{i,j\rightarrow h}(A \rightarrow C) = \{A_1 \rightarrow C_1, \ldots, A_w \rightarrow C_w\}$, where $A_k \rightarrow C_k \in R_{ih}$ $(k = 1, \ldots, w)$

and $A_k \rightarrow C_k$ is a right-specialization (left-generalization) of $A \rightarrow C$. More formally, a spatial association rule $A' \rightarrow C' \in R_{ih}$ is a right-specialization (left-generalization) of $A \rightarrow C \in R_{ij}$ iff there exists a substitution θ such that $A\theta \subset A'$ and $C\theta \subset C'$ ($A'\theta \subset A$ and $C'\theta \subset C$).
Example: Let us consider the spatial association rules:
$R1$: is_a($X1$, town), intersects($X1$, $Y1$), is_a($Y1$, road) → mortality($X1$, high)
$R2$: is_a($X1$, town), intersects($X1$, $Y1$), is_a($Y1$, road), extension ($Y1$, [12..25])
→ mortality($X1$, high).
R2 is a *right-specialization* of R1, since there exists the substitution $\theta = \{X1/X2, Y1/Y2\}$ such that $R1.A\theta \subset R2.A$ and $R1.C\theta \subset R2.C$. Conversely, R1 is a *left-generalization* of R2.

Consequently, by combining a multiple graph visualization with operators of both specialization and generalization defined above, data miners are able to navigate among the graphs G_{ij}. This means that it is possible to down(right)-specialize or up(left)-generalize the portion of the graph G_{ij} representing a specific rule $R \in R_{ij}$ and visualize the corresponding sub-graph of spatial association rules extracted at a different level of granularity or number of refinement steps.

This graph-based visualization has been implemented into a visualization tool, named ARVis (multi-level Association Rules Visualizer), which actively supports data miners in exploring and navigating among several graphs of multi-level association rules G_{ij} by highlighting the portion of graph that represents the down (right)-specialization or up(left)-generalization of a rule, zooming rules, dynamically filtering rules according to minimal values of support and/or confidence as well as presence or absence of some relevant predicate and visualizing details about a rule (e.g. support, confidence, patterns, rules).

4.5.4 Mining geo-referenced data

ARVis has been used with the multi-level association rules produced by the previously mentioned system ARES. The system allow mining and exploring multi-level spatial association rules for geo-referenced census data interpretation. We considered census and digital map data stored into an Oracle Spatial $9i$ database provided in the context of the European project SPIN! (Spatial Mining for Data of Public Interest) [136]. This data is related to Greater Manchester, one of the five counties of North West England, which is divided into censual sections or wards, for a total of two hundreds and fourteen wards. Spatial analysis is enabled by the availability of vectorized boundaries of the 1998 greater Manchester census wards as well as Ordnance Survey digital maps where several interesting layers are found (e.g. urban area or road net). Census data, geo-referenced at ward level, provide socio-economic statistics (e.g. mortality rate that is the percentage of deaths with respect to the number of inhabitants) as well as some measures describing the deprivation level (e.g. Townsend index, Carstairs index, Jarman index and DoE index). Both mortality rate and deprivation indices are all numeric. They can be automatically discretized with ARES. More precisely, Jarman index, Townsend index, DoE index and Mortality rate are automatically discretized in (low, high), while Carstairs index is discretized in (low, medium, high).

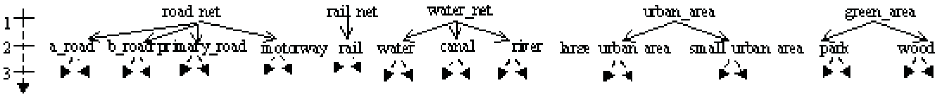


Figure 4.4: Spatial hierarchies defined for five Greater Manchester layers: road net, rail net, water net, urban area and green area.

For this application, we decide to employ ARES in mining multi-level spatial association rules relating Greater Manchester wards, which play the role of reference object, with topological related roads, rails, waters, green areas and urban areas as task relevant objects. Therefore, we extract 784,107 facts concerning topological relationships between each relevant object and task relevant object stored in the spatial database for Greater Manchester area. An example of fact extracted is *crosses*(*ward_135*, *urbareaL_151*). However, to support a spatial qualitative reasoning, we also express a domain specific knowledge (BK) in form of a set of rules. Some of these rules are:

crossed_by_urbanarea(X, Y) : − crosses(X, Y), is_a(Y, urban_area).

crossed_by_urbanarea(X, Y) :- inside(X, Y), is_a(Y,urban_area).

Here the use of the predicate is_a hides the fact that a hierarchy has been defined for spatial objects which belong to the urban area layer. In detail, five different hierarchies are defined to describe the following layers: road net, rail net, water net, urban area and green area (see Figure 4.4). The hierarchies have depth three and are straightforwardly mapped into three granularity levels. They are also part of the BK. To complete the problem statement, we specify a language bias (LB)

both to constrain the search space and to filter out uninteresting spatial association rules. We rule out all spatial relations (e.g. crosses, inside, and so on) directly extracted from spatial database and ask for rules containing topological predicates defined by means of BK. Moreover, by combining the rule filters *head_ constraint([mortality_ rate(_),1,1)* and *rule_ head_ length(1,1)* we ask for rules containing only mortality rate in the head. In addition, we specify the maximum number of refinement steps as $J = 8$ and the minimal values of support and confidence for each granularity level as: *minsup[1]=0.1, minsup[2]=0.1, minsup[3]=0.05, minconf[1]=0.3, minconf [2]=0.2* and *minconf [3]=0.1*.

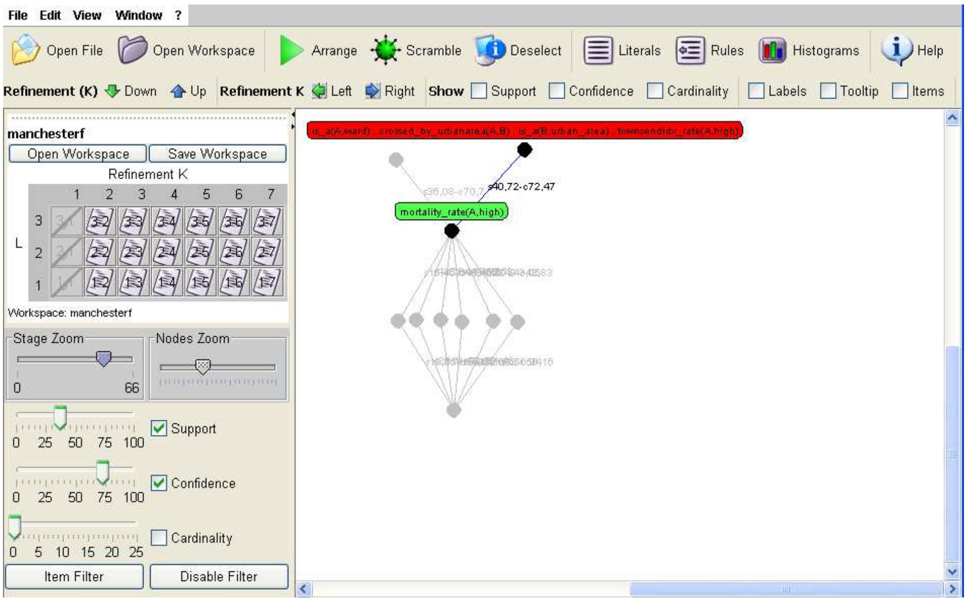


Figure 4.5: Visualizing the graph of spatial association rules using ARVis

ARES generates 239 strong rules at first granularity level, 1140 at second granularity level and 15 at third granularity level. These rules are extracted from a set of 28496 frequent patterns describing the geographically distributed phenomenon of mortality in Greater Manchester at different granularity levels with respect to the spatial hierarchies we have defined on road, rail, water, urban area and green area layers. To explore this huge amount of multi-level spatial association rules and find which rules can be a valuable support to good public policy, we exploit the multiple graph-based visualization implemented in ARVis. In this way, we are able to navigate among different graphs G_{ij} $(i = 1, \ldots, 3$ and $j = 2, \ldots, 8)$ representing the group of rules R_{ij} discovered by ARES at i granularity level after j refinement steps. For instance, Figure 4.5 shows the graph of spatial association rules G_{15}. By graphically filtering rules in G_{15} according to confidence value, we identify the most confident rule $R1$ that is: *is_ a(A, ward), crossed_ by_ urbanarea(A, B), is_ a(B, urban_ area), townsendidx_ rate(A, high) → mortality_ rate(A, high)* (c=39.71%, s=70.24%). This rule states that a high mortality rate is observed in a ward A that includes an urban area B and has a high value of Townsend index. The support (39.71%) and the high confidence (70.24%) confirm a meaningful association between a geographical factor such as living in deprived urban areas and a social factor such as the mortality rate. The same rule is highlighted in the graph G_{15} by filtering with respect to increasing value of support. Moreover, by left-generalizing $R1$, we navigate from the graph G_{15} to a portion of the graph G_{14} and identify the rule $R2$ that is *is_ a(A,ward), crossed_ by_ urbanarea(A,B), is_ a(B, urban_ area) → mortality_ rate(A, high)* (54.67%, 60.3%). This rule has a greater support and a lower confidence. The same rule is highlighted in the entire graph G_{14} by graphically filtering with respect to increasing values of support and confidence. These two association rules show together an unexpected association between Townsend index and urban areas. Apparently, this means that this deprivation index is unsuitable for rural areas.

Conversely, we may decide to up-generalize $R1$ and move from the graph G_{15} to the portion of the graph G_{25} representing association rules which are up-generalization of $R2$ mined by ARES at second granularity level after four refinement steps. In this way, we discover that, at second granularity level, SPADA specializes the task relevant object B by generating the following rule which preserve both support and confidence: $R3$: *is_ a(A, ward), crossed_ by_ urbanarea(A, B), is_ a(B, urban_ areaL), townsendidx_ rate(A,high) → mortality_ rate(A, high)* (39.71%, 70.24%). This rule clarifies that the urban area B is large. Similar considerations are suggested when we explore graphs of multi-level spatial association rules generated after more refinement steps.

We may explore spatial association rules characterizing low mortality wards. By visualizing G_{15} and moving the confidence filter slider, we discover that the highest confident rule with low mortality in the consequent is: *is_ a(A, ward), crossed_ by_ urbanarea(A, B), is_ a(B, urban_ area), townsendidx_ rate(A, low) → mortality_ rate(A, low)* (19.15%, 56. 16%), stating that a low valued Townsend index ward A that (partly) includes an urban area B presents a low mortality.

4.5.5 Techniques for Association Rules Visualization

Despite the big amount of algorithms and methods for the production and management of association rules (ARs), there are not so many tools that provide a good visualization of the discovered rules. We briefly present some tools divided in two categories: those visualizing AR using 2D and those using 3D. With respect to previous surveys [31, 121], more 3D tools are here reported.

Tabular visualization. The most immediate way to visualize ARs is using a table in which each row is a rule and the columns represent the item set. The last two columns are typically the value of support and confidence of the rules. The advantage of this approach is that it is easy to order rules according to an attribute. If, for instance, the user orders the rules by confidence, the rules with the high will be easily identified. The main disadvantage is that the user may analyze few rules at once and it is very hard to look at the overview of the rule set or relations among them. Even if there are few rules to analyze but for each of them there are many items it is hard to look at them for the user.

Twokey plot. Unwin et al. [193] use a scatter plot in which they represent the rules as colored discs. The position of the rules in the scatter plot is represented by the combination of support and confidence of the rule. The Y axis represent the confidence values and the X axis the support values. The color represent the cardinality of the rules (how many items has the pattern), this technique is effective to represent the relationship among support and confidence and cardinality. It is also possible to see relationships between ancestor and descendants of the rules (a rule is children of another one if the former has one more item than the latter). Even thought the user can get some information about the rules it is difficult to analyze the rules in detail, the items of the rules are not easily visible so this technique need to be associated to other techniques to better analyze ARs.

Pixel grid. Kian-Huat et al. developed a tool that use a technique similar to the twokey plot. Here the user can use a panel to interact with the rules and can set some parameters to filter inside the rule set. Moving the mouse over a rule it is possible to get more information about rules (i.e. it is possible to see how the rules are composed) [153].

Double-Decker plots. The main idea of this approach is to represent all possible permutation of the items in the rule on the LHS (Left-Hand Side) and RHS (Right-Hand side) as a bar chart. The Y value is the value based on the specific permutation. Each row is an item. Support is highlighted and confidence is the proportion of the highlighted area. The technique was born to analyze a rule in detail, for this reason it is difficult to analyze many rules at once. To overcome this limitations the author provided a visualization with a matrix of double-decker plots in order to analyze relations among rules, but it is still hard for the user to analyze several rules in the same screen [106].

Circular graph. The circular graph approach [163] adopts a circular graph layout where items involved in rules are mapped around the circumference of a circle. Associations are then plotted as lines connecting these points, where a gradient in the color of the line, from blue(dark) to yellow(light) indicates the direction of the association from the antecedent to the consequent. Each point of the circle is an item. Associations are represented by lines connecting those points. The color coding is used to show the direction of the link and there is a specific color to show the bi-directional links. This technique has been successfully used to highlight relationships among items and it is able to display high information volumes. Unfortunately with this technique it is difficult to represent support and confidence, this imply that it is needed another technique combined with this to be able to perform a good analysis on the ARs.

Directed graph. This is a reticular representation of the ARs. The graph technique may represent ARs in different way, representing relationships among items (as nodes of the graph) or relationships among rules (as edges of the graph). If graph is used to represent associations among items, with rules with many nodes the graph may become cluttered. A tool that uses this technique is DBMiner, that represents two kind of relationships among the nodes: *inter-attribute* association and *intra-attribute* associations. The former is an association among different attributes; whereas the latter is an association within one or a set of attributes. Since typically there are many rules in many cases this technique may be not useful.

Bi-dimensional matrix. The leading idea of a 2D matrix is the AR representation having the RHS (Right-hand side) and the LHS (Left-hand side) on the x-axis and y-axis respectively. The 2D representation of this technique uses the color to represent support and confidence, in this case there are a class of colors for the confidence and the color shade is used to represent the support. In order to face the confusion generated by the color shade and color class, hence to better highlight the different values of support and confidence DBMiner [96] uses a 3D version in which it exploit the third dimension (histogram height) to represent confidence and the color is used to represent the support. The best use of this technique is for one-to-one ARs, because the user may be confused about which item is in the RHS and which in the LHS. Another problem with the 3D representation is related to the classical problems of 3D: occlusion, comparison between objects on foreground and background, etc.

Matrix visualization. The application, developed by Wong et al [209], visualizes AR using a Text Mining engine. The system visualizes relationships using a matrix in which the rows represent the rules and the columns represent the items. In the cross there is a colored cell that is blue in case the item is on the LHS and is red if the item is on the RHS. In the background there are two histograms that represents the confidence and the support for each rule. This tool gives an overview of all the rule set. In order to avoid other problems related to the 3D the system presented by Wong et allow to rotate the matrix and to zoom and pan. This system work best with few rules, with a big number of rules (thousands) it may become difficult to understand the matrix. This technique is very good in giving the overview of the rules and it is possible to represent also a big number of rules, but it is difficult to see relationships among rules and between antecedents and consequents.

3D graph visualization. In the Ming C. et al. [97] approach a 3D graph is used. Items are represented as spheres linked with arrows. Items that are strongly correlated are grouped in elliptical clusters. The support is represented with the length of the edge and confidence using colors. The choice of the length of the edge is not the best, since it is difficult to represent a real proportional length. This problem comes from the need to have different conflicting conditions. One is the non overlap of the items, another one is that the length of the edge should be fixed, another one is that many items should be grouped. It may happen that not all those condition fit, then it is necessary a compromise. Another issue is that in this approach the relations among the rules are not clear, only relationships among the items of the rules are visible.

Arena. An alternative visualization is the work presented by Julien Branchard et al. [22]. In this work the authors use a metaphor of virtual arenas. There are some walls with steps on which the authors put the AR that are represented as spheres. The radius of the spheres is proportional to the support. The spheres are set on a cone whose height represent the confidence. By selecting the rules it is possible to focus on the set of rules related to the items that form the selected rule. This is a new approach, it is not recommended for a big number of rules.

ARVis 3D We developed also a 3D version for ARVis in order to exploit the third dimension to better organize elements in the screen. In figure 4.6 is shown a set of six association rules that share the same antecedent (red sphere) visualized with ARVis 3D. In the figure, the users has clicked on a rule (edge connecting the highlighted spheres) and the details of that rule is shown in the semi-transparent label.

We found that 3D interface is very attractive at glance, but for a deep analysis is not appropriate because the advantage of the third dimension is in contrast with the disadvantages of the 3D mentioned in Section 2.5.3. Comparing pros and cons we concluded that 3D is not appropriate for the goals of data miners so currently we are not going in that direction.

4.6 PCAR: visualizing mined association rules using Parallel Coordinates

The approach of parallel coordinates [112] to visualize ARs has been presented in [29]. They found a measure to indicate the utility of an item in a rule, that they call Item Utility (IU). The IU indicates how good is an item into a rule, for instance, if there is a rule of the form: $x\&y \Rightarrow z$ with confidence CR it is possible to get the value of the confidence of the rule removing the item y $CR(y)$. The formula to compute IUy lead to three different cases:

- if $IUy \in]-1;0[$ then y is dangerous for the rule;

- if $IUy = 0$ then y is redundant (neutral);

- if $IUy \in]0;1[$ then y is useful for the rule;

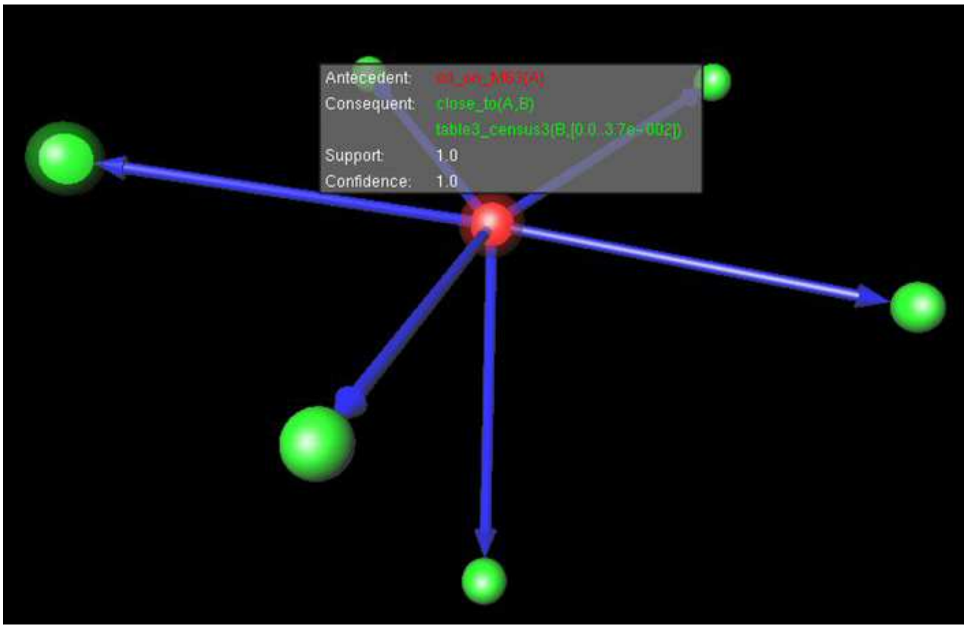
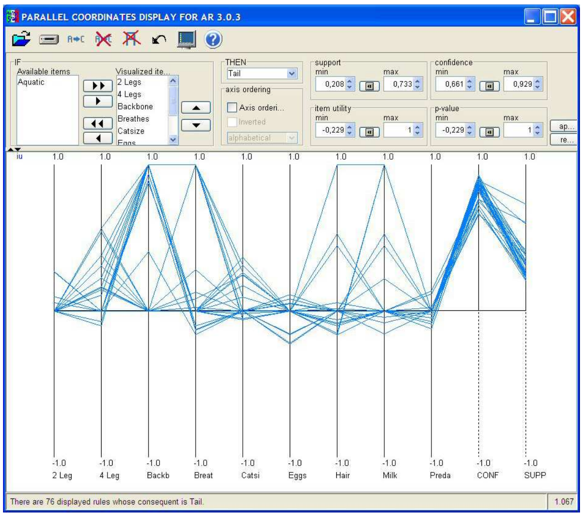


Figure 4.6: Visualizing association rules using the 3D version of ARVis


PARALLEL COORDINATES DISPLAY FOR AR 3.0.3

IF — Available items: Aquatic

Visualized ite...: 2 Legs, 4 Legs, Backbone, Breathes, Catsize, Eggs

THEN: Tail

axis ordering: □ Axis orderi... □ Inverted alphabetical

support min 0,208 max 0,733

confidence min 0,661 max 0,929

item utility min -0,229 max 1

p-value min -0,229 max 1

There are 76 displayed rules whose consequent is Tail.

1.067

Figure 4.7: AR visualization using parallel coordinates

In order to have this representation the axes are normalized to $[-1; 1]$ (IU range). Each item is an axis and each rule is a line, the line cross the axes according to the value of the IU for each item. One of the advantages using this technique is the possibility to perform some operation like the filter on dense rules, highlight interesting rules according to some parameters specified by the user, prune uninteresting rules according to some criteria (i.e. those whose items have a IU below a specific threshold).

Both these tools allow the user to browse and interact with association rules to perform different tasks. Graph visualization is very useful to describe the overview of the rules and the relationships among the items, while Parallel Coordinates work better if used as a visual pruning tool by exploring the strength of the discovered association. Due to these different goals, the two visualizations can be profitably used together by exploiting their synergic power. First the user may visualize the overview of all the rules produced by a classical data mining algorithm using the graph-based technique. Once the user selects a subset of rules of interest he pass them to the parallel coordinates tool. Once the user filters inside the rules, he may return to the graph visualization to explore the filtered rules.

4.7 DaeCV: visualizing mined clusters

4.7.1 Tools visualizing clustering results

Currently DaeCV visualizes textual results of the clustering algorithm used in DaeMine. The rules are produced with the k-means algorithm. DaeCV is a module that will be able to visualize the results of one or more clustering methods described in Section 4.4.2. A partial survey on tools that visualize results of clustering methods was conducted in order to understand the needs of users that currently perform analysis using visual tools.

The analyzed tools in the survey are:

- IAFC [145]

- gCLUTO [166]

- Starclass [191]

- StructureMiner [152]

- Majorclust [152]

- Cluster and Calendar [195]

- Intelligent Miner [113]

- 3-D Graph-based [55]

Some paper prototype have been shown to the data miners, that are the main users this tool is intended for. After some presentation of tools and having seen the proposal for the prototype the users provided us some useful hints to better understand their needs. Some relevant features were found. The interface should provide:

- configuration of input parameters;

- visualization and management of outliers;

- most important items highlighting and less relevant hiding;

- availability of widgets for browsing inside data and results.

Those are some basic and generic features, at the end of the analysis we will have a complete set of characteristics and needs of the tool. The first running prototype is expected by next summer.

4.8 DaeTL: dynamic tabular data analysis

DaeTL is basically intended to support the organizer in data analysis providing a simple way to identify interesting patterns. Based on some of the most powerful visualization techniques, it is able to enrich the traditional data mining results with information discovered by performing direct exploration among the data. Usually the data in databases are shown to the user in tabular format, the table lenses technique [164] has been proved that is a valuable technique for this type of problem, and this is the baseline for this tool; DaeTL merges the power of the table data view with some graphic representation that the human can easily understand.

This module may be used in two different ways for two different activities:

- Explorative analysis, usually performed when the user doesn't have information or hypothesis about data he is interacting with. The module is able to visualize data and the relationships among them so that it is easy for the user to grasp information simply observing data and their distribution.

- Confirmative analysis, when the user has already some hypothesis about data he may want to check them.

DaeTL provides users with an overview of the data they are analyzing. A fisheye effect [85] allow them to zoom into elements of interest without loosing the overview. In order to visualize quickly the information zoomed out, it has been implemented an automatic zoom that show the details while moving mouse. The user can choose to use or not the automatic zoom and tooltips, that are labels that appears while moving on items and show additional information about them. If the user needs to zoom more than one item in the same time, DaeTL allow him to zoom intervals instead of only one item, just dragging the expanded row until the item he/she wants to zoom. Since the type of data that the user will use it is not predictable at design time DaeTL automatically can manage numbers, strings and categories, using the most suitable way to show them on the screen without asking the user how to do it. The user can change parameters in order to customize the tool using a control panel.

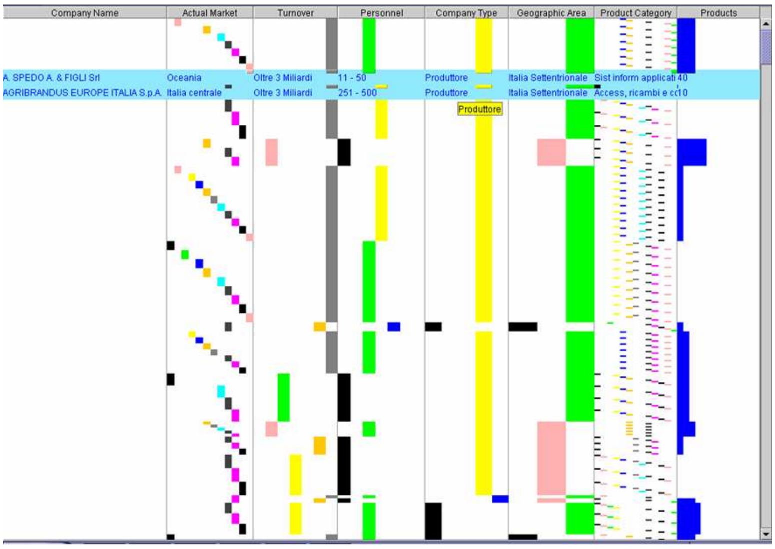
DaeTL is able to present in a very short way a number of attributes and a number of rows resulting from a query on the database. The input of the module is the result of an analysis available in DAE DB.

The tool has several features that allow the user to perform analyses to find interesting regularities (or irregularities) in the distribution of the data. With this tool the user can immediately see the overall distribution of the data along different attributes and simply moving the mouse over the interested data he/she can see the row expanded with the details about data.

In order to better explain how the tool works we show an example. Let us consider an organizer who wants to see some correlations in the data to perform further marketing actions. The organizer wants to use one of the DAE tools since he knows that the information he need are in the data stored in the system. The organizer connects with trade fair web site. After the login DaeClient shows the list of available analysis associated to the organizer.

In Figure 4.8, it is clear that the highlighted company “Aspeedo & Figli srl” doesn’t serve the following markets: “Africa”, represented by the black colour, “Europa Meridionale”, represented by the green color, "Europa Centrale", represented by the yellow color, "Europa Settentrionale", represented by the gray color. This company does not serve Europe but serves the rest of the world. Moreover, even if the company has a high turnover (over three billions ITL), it has a very little employee staff, in particular it falls in the range [11-50] people.

The other company highlighted in Figure 4.8, named "Agribrandus Europe Italia s.p.a." serves only the Italian market, more specifically, it serves the middle of Italy, and this is one of the companies with the highest turnover. This time, the employees number falls in the range [251-500], as shown by the yellow bar of the column related to personnel. Moreover, this is a big company whose products are all types but not "Bestiame", represented with the green color, "Impianti per la zootecnia", represented with the gray color, "Macchine per l'imballaggio", represented with the black color, and "Sist. informatici applicati all'agric.", represented with the green color.

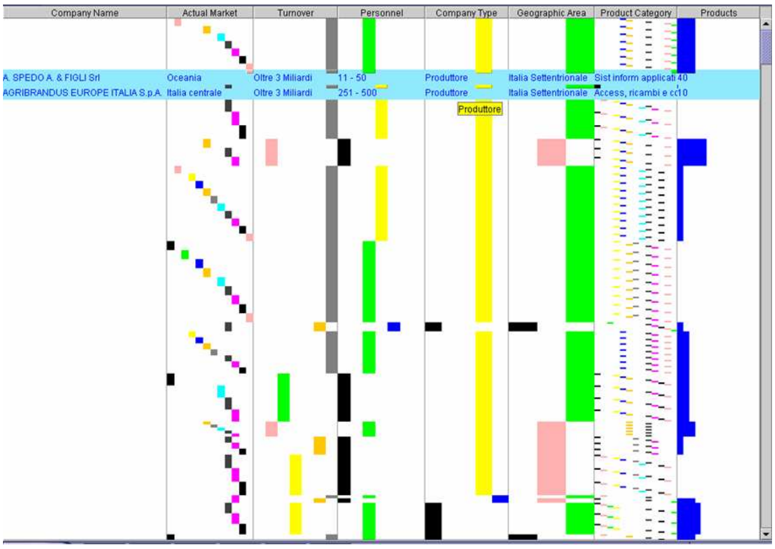


Company Name	Actual Market	Turnover	Personnel	Company Type	Geographic Area	Product Category	Products
A. SPEDO A. & FIGLI Srl	Oceania	Oltre 3 Miliardi	11 - 50	Produttore	Italia Settentrionale	Sist inform applicati	40
AGRIBRANDUS EUROPE ITALIA S.p.A.	Italia centrale	Oltre 3 Miliardi	251 - 500	Produttore	Italia Settentrionale	Access, ricambi e cc	10

Figure 4.8: A visualization of a trade fair data using DaeTL that shows two companies present in the database

Then the user sorts the data by Turnover column, as shown in Figure 4.9. The dark yellow color (second last) in the turnover column indicates that the company turnover is more than 3 billions (ITL). The organizer then may easily see that more than a half of the companies in database have that characteristic.

When the user sorts first on the column turnover then on the column geographic area, the result of this interaction are shown in the Figure 4.10. Most of the companies come from "Italia Settentrionale", showed



Company Name	Actual Market	Turnover	Personnel	Company Type	Geographic Area	Product Category	Products
A. SPEDO A. & FIGLI Srl	Oceania	Oltre 3 Miliardi	11 - 50	Produttore	Italia Settentrionale	Sist inform applicati	40
AGRIBRANDUS EUROPE ITALIA S.p.A.	Italia centrale	Oltre 3 Miliardi	251 - 500	Produttore	Italia Settentrionale	Access, ricambi e cc	10

Figure 4.9: The data set with the companies that have more that 3 billions (ITL) turnover

with green color. In such area, there are most companies with turnover more than 3 billions, while the companies coming from the middle Italy (black color in the Geographic Area column) are small or big, since the companies in the range [50 millions (ITL) - 1 billion (ITL)] are not present.

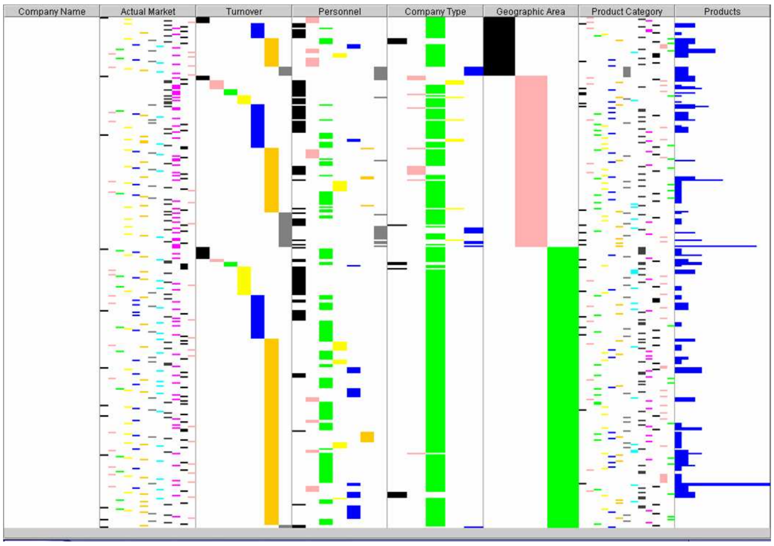


Figure 4.10: Data set sorted first on the column turnover then on the column geographic area

A note about colors, with gray scale figures is not possible to easily distinguish all the colors; the color choice has been made in order to help people with color-blind problems and there is always the possibility to see the tooltips or zooming in order to read the name of each item just moving the mouse.

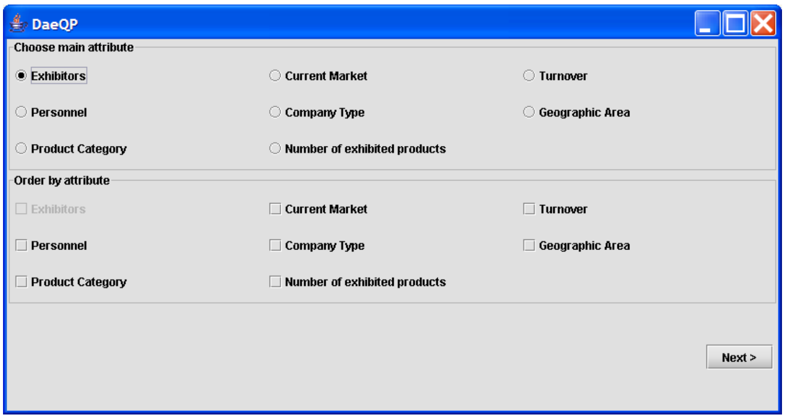
4.9 DaeQP: query preview for target selection

DaeQP is a tool inspired to the Query Preview technique [71, 190] and provides users with rapid multidimensional overviews of information about the data source, in order to perform appropriate data analysis along selected major attributes (or dimensions). One of the main advantage for this visualization is that allow a direct comparison between different variables at once. The overview shows the data distribution along the selected attributes. Then, we use dynamic queries and query previews to support efficient query formulation. Query previews provide the possibility of easily getting preliminary information about data of interest, making visible problems and gaps in the meta data that are difficult to detect with the traditional form fill-in interfaces. With this tool, the user may rapidly eliminate undesired data and also preview the size of the result set to avoid the so-called zero-hit queries, i.e., queries that provide an empty result set.

The question that this tool answers is: “How many elements and who (users, products, peoples, customers, items in general) in the data source have the following characteristics?”

In order to better understand how this module works, let us refer to this scenario: the organizer of an Italian trade fair on agriculture wants to perform a segmentation of the exhibitors of the last edition of the fair. Let us suppose that the organizer wants to find out a group of companies with some characteristics to which sending customized advertising when sending the invitation for the next event. The objective of the organizer is to increase the trade fair income by selling more services or producing services with a better quality. Therefore he is interested in selecting company segments to start appropriate marketing campaign to promote the fair services.

DaeQP first displays an overview of data visualized along some major attributes, the attribute shown are those previously selected by the organizer. As shown in Figure 4.11 the user is first asked to select the main attribute and some secondary attribute in which he is interested in. The reason for selecting the main attribute is because the analysis is focused on a single attribute and in a single analysis the main attribute depends on the preference of the user.



DaeQP

Choose main attribute

● Exhibitors ○ Current Market ○ Turnover

○ Personnel ○ Company Type ○ Geographic Area

○ Product Category ○ Number of exhibited products

Order by attribute

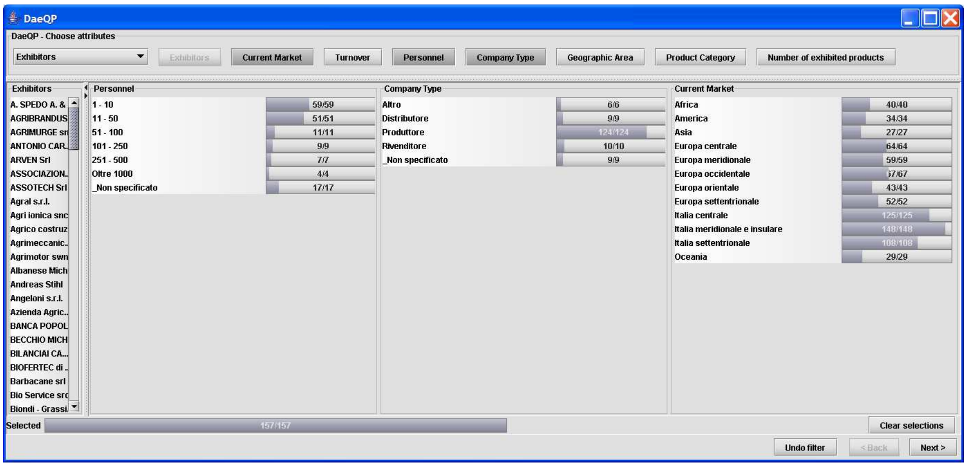
□ Exhibitors □ Current Market □ Turnover

□ Personnel □ Company Type □ Geographic Area

□ Product Category □ Number of exhibited products

Next >

FIGURE 4.11: Different attributes along with the user can perform the analysis



DaeQP

DaeQP - Choose attributes

Exhibitors

Exhibitors

A. SPEDO A. &
AGRIBRANDUS
AGRIMURGE sr
ANTONIO CAR..
ARVEN Srl
ASSOCIAZION..
ASSOTECH Srl
Agral s.r.l.
Agri ionica snc
Agrico costruz
Agrimeccanic..
Agrimotor swn
Albanese Mich
Andreas Stihl
Angeloni s.r.l.
Azienda Agric..
BANCA POPOL
BECCHIO MICH
BILANCIAI CA...
BIOFERTEC di ..
Barbacane srl
Bio Service src
Biondi - Grassi.

Personnel	
1 - 10	59/59
11 - 50	51/51
51 - 100	11/11
101 - 250	9/9
251 - 500	7/7
Oltre 1000	4/4
_Non specificato	17/17

Company Type	
Altro	6/6
Distributore	9/9
Produttore	124/124
Rivenditore	10/10
_Non specificato	9/9

Current Market	
Africa	40/40
America	34/34
Asia	27/27
Europa centrale	64/64
Europa meridionale	59/59
Europa occidentale	67/67
Europa orientale	43/43
Europa settentrionale	52/52
Italia centrale	125/125
Italia meridionale e insulare	148/148
Italia settentrionale	108/108
Oceania	29/29

Selected 157/157

FIGURE 4.12: An example of query preview along Personnel, Company type, and Current Market attributes

Let us suppose that the user selects the attributes 'Personnel', 'Current Market', and 'Company type'. After having pressed the button 'Next', the interface updates and shows the distribution of the number of exhibitors along the selected attributes (Figure 4.12). Users can see that the majority of exhibitors are Producer ('Produttore') and that most have less than 50 employees.

The organizer is probably interested in small companies (he may select small companies taking those that have a few employees) that are producers and comes from the North of Italy and wants to promote northern Italy producers to resellers; if they are interested there are more possibilities that they will attend next trade fair. So the organizer wants to choose among the exhibitors displayed on the screen (that are all the exhibitors, probably of a specific edition, that are present in the organizer database) those corresponding to the desired characteristics.

So the organizer selects in personnel the range 1-10, then in the section 'company type' selects the label corresponding to producer ('produttore') and finally in the 'current market' section he clicks on the North of Italy ('Italia settentrionale'), as shown in the Figure 4.13.

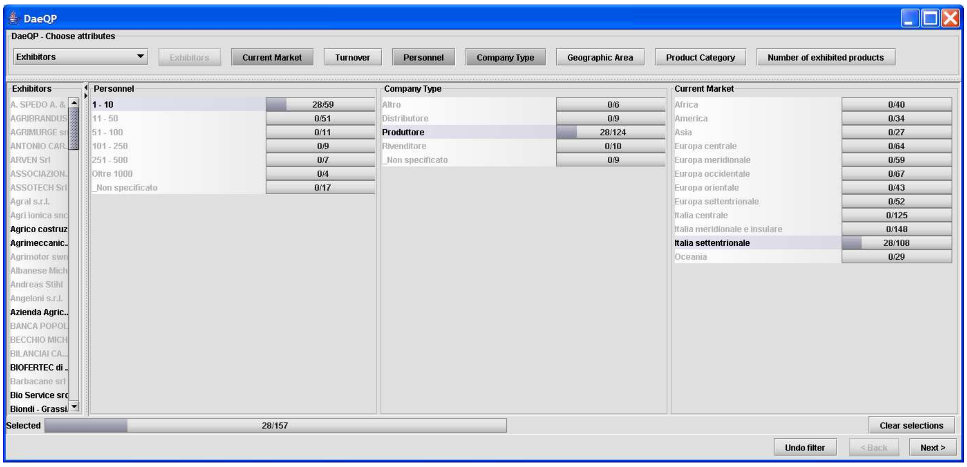
The organizer see first the overall distribution of the exhibitors in his database, then he clicks on the bars that have some exhibitor inside. In Figure 4.13, for some attributes zero exhibitors are present one or more items, this is an important information for the user because when using a classical form-based interface (still the most used in the web) the user may obtain the so called zero hit queries. It is well known that zero hit queries produce frustration for users who do not know what parameters are allowed and when the data satisfy a query. With the query preview technique zero hit queries are avoided, because the user can see what combination of parameters will produce acceptable queries.

All the actions performed with DaeQP are immediate and reversible, so users can see the results in few milliseconds and quickly perform many queries. If the query preview shows still too many records, the user may need to further reduce the selected data set. To address this need, the tool allows query refinement by clicking on the button 'Next'. The query refinement phase supports dynamic queries over other relevant attributes of the database. In this way, the user can see more details of the dataset retrieved in the first phase, and get the reduced set of data he is actually interested in. For example, the user can specify further attributes, such as the company income, or he can zoom on the selected value of an attribute shown in the first preview.

On the side bar shown in the the Figure 4.13 it is possible to see the details of the selected exhibitors. Those are shown in light gray, as the user apply or remove the selections the status of the exhibitors changes.

4.10 TimeSearcher: Time series analysis

Time series are widely used in applications such as electrocardiograms (EKGs), seismographs, industrial processes, meteorology, and sound recordings. They consist of sequences of real numbers, representing the measurements or observations of a real variable at equal time intervals. From an algorithmic perspective, there is a long history on time se-



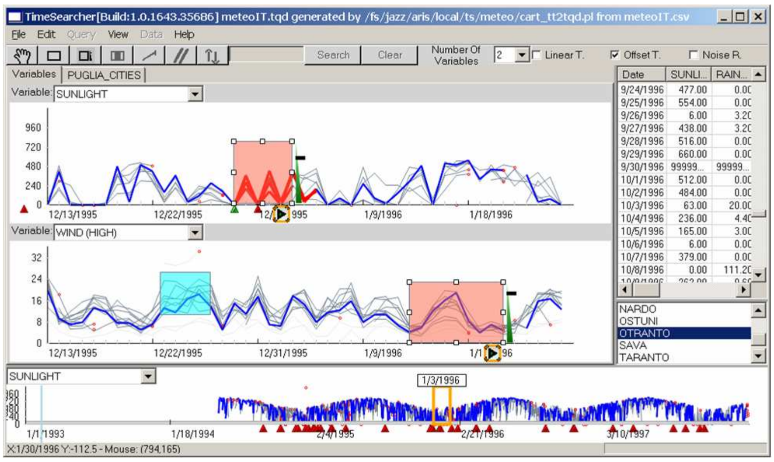
Personnel	
1 - 10	28/59
11 - 50	0/51
51 - 100	0/11
101 - 250	0/9
251 - 500	0/7
Oltre 1000	0/4
_Non specificato	0/17

Company Type	
Altro	0/6
Distributore	0/9
Produttore	28/124
Rivenditore	0/10
_Non specificato	0/9

Current Market	
Africa	0/40
America	0/34
Asia	0/27
Europa centrale	0/64
Europa meridionale	0/59
Europa occidentale	0/67
Europa orientale	0/43
Europa settentrionale	0/52
Italia centrale	0/125
Italia meridionale e insulare	0/148
Italia settentrionale	28/108
Oceania	0/29

Figure 4.13: DaeQP after the selection of personnel, company type and current market. The selected exhibitors are displayed in the sidebar (Exhibitors)

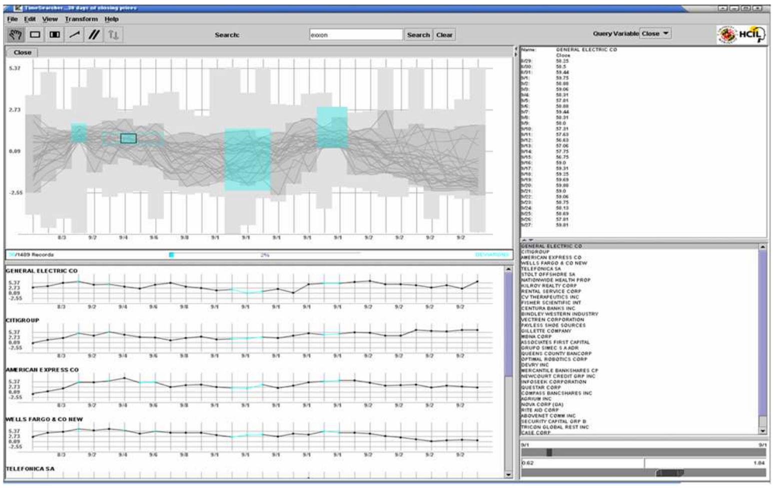
ries, originally grounded on statistical analysis. Today, with the aid of computers, users can analyze time series data using classical statistical models, and also explore data using visualization tools. These interfaces enable users to see the data and apply their powerful perceptual abilities to identify trends or spot anomalies.



Date	SUNLI...	RAIN...
9/24/1996	477.00	0.0C
9/25/1996	554.00	0.0C
9/26/1996	6.00	3.2C
9/27/1996	438.00	3.2C
9/28/1996	516.00	0.0C
9/29/1996	660.00	0.0C
9/30/1996	99999...	99999...
10/1/1996	512.00	0.0C
10/2/1996	484.00	0.0C
10/3/1996	63.00	20.0C
10/4/1996	236.00	4.4C
10/5/1996	165.00	3.0C
10/6/1996	6.00	0.0C
10/7/1996	379.00	0.0C
10/8/1996	0.00	111.2C

Figure 4.14: TimeSearcher 2. The top left area shows a detail view of two months for two variables. An overview for the entire five years of data is shown at the bottom. On the detail view patterns can be selected for searching

TimeSearcher 2 was developed during my visit at the Human-Computer interaction laboratory at the University of Maryland, in collaboration with Aleks Aris, with the participation of Amir Khella, and under the direction of Catherine Plaisant and Ben Shneiderman. The work presented in this section (Figure 4.14) builds on previous work at the University of Maryland that explored the use of timeboxes to query



Name:	GENERAL ELECTRIC CO
	Close
8/29:	58.25
8/30:	58.5
8/31:	59.44
9/1:	59.75
9/2:	58.88
9/3:	59.06
9/4:	58.31
9/5:	57.81
9/6:	58.88
9/7:	59.44
9/8:	58.31
9/9:	58.0
9/10:	57.31
9/11:	57.63
9/12:	56.63
9/13:	57.06
9/14:	57.75
9/15:	56.75
9/16:	59.0
9/17:	59.31
9/18:	59.25
9/19:	59.69
9/20:	59.88
9/21:	59.0
9/22:	59.06
9/23:	58.75
9/24:	58.13
9/25:	58.69
9/26:	57.81
9/27:	58.81

GENERAL ELECTRIC CO
CITIGROUP
AMERICAN EXPRESS CO
WELLS FARGO & CO NEW
TELEFONICA SA
STOLT OFFSHORE SA
NATIONWIDE HEALTH PROP
KILROY REALTY CORP
RENTAL SERVICE CORP
CV THERAPEUTICS INC
FISHER SCIENTIFIC INT
CENTURA BANKS INC
BINDLEY WESTERN INDUSTRY
VECTREN CORPORATION
PAYLESS SHOE SOURCES
GILLETTE COMPANY
MDNA CORP
ASSOCIATES FIRST CAPITAL
GRUPO SIMEC S A ADR
QUEENS COUNTY BANCORP
OPTIMAL ROBOTICS CORP
DEVRY INC
MERCANTILE BANKSHARES CP
NEWCOURT CREDIT GRP INC
INFOSEEK CORPORATION
QUESTAR CORP
COMPASS BANCSHARES INC
AGRIUM INC
NOVA CORP (GA)
RITE AID CORP
ABOVENET COMM INC
SECURITY CAPITAL GRP B
TRICON GLOBAL REST INC
CASE CORP

Figure 4.15: TimeSearcher 1. As users draw or drag the timeboxes the lines are filtered to show only those that pass through the boxes

time series data [98] (Figure 4.15). Timeboxes are rectangular regions that are selected and directly manipulated on a timeline overview of the data. The boundary values of the timeboxes specify the relevant parameters of the query. This work enhances the concept of timeboxes by differentiating two types of timeboxes. The first type (the original) is used to filter the data and reduce the scope of the search, whereas the second type is used to perform a specific pattern search anywhere in the remaining data. Also the general browsing interface of TimeSearcher has been enhanced, allowing users to deal with long time series of multiple heterogeneous variables. The target users may be expert or intermittent users; the interfaces do not require any specialized analysis skills such as statistical knowledge. The aim is to provide interfaces that extend the initial exploratory analysis. Users first gain an overview of the data, then filter and zoom on the data to spot patterns of interest. Next, a search can be performed to find similar patterns at other times and in other series helping users to make hypothesis about phenomena in the data (then experts may use external tools that validate the hypothesis).

4.10.1 Time series related work

The characteristics of time series vary widely and are highly dependent on the application domain. Also data that originally are not conceived as time series can be transformed into time series. Examples include the mapping of a video stream as a time series, or the mapping of the human movement in 3D space to a time series (each coordinate becomes a single time series) [48]. There is an abundant literature focused on time series data. Some tools deal with a single time series and may look for patterns within that series [105], others deal with multiple time series and allow users to find the time series of interest [105]. Very few allow multiple variables to be browsed and searched [194]. Several papers propose the visualizations of time series but allow only limited interaction with the data, such as altering the order or accessing details. Interactive search facilities are rare and offer limited features [122, 123, 75]. Another direction is to provide predictions with visual results but the interaction is limited [171]. This section focuses on tools that include a graphical user interface allowing users to interact with the data by using simple widgets and without the need of extensive training or specialized skills such as statistical analysis expertise. One of the first attempts to visualize and interact with time series was Diamond Fast [194]. Diamond Fast visualizes time series on the screen and permits users to move and resize them, hence compare more than two time series. It also includes some management of missing values. Diamond Fast was only capable of managing short time series which is not enough today, but remains an important inspiration for new tools and a good application in several domains. Other newer tools like in ILOG [110] and in Personal Stock Monitor [157] permit a high level of interaction to browse the data with enhanced zoom features, but they lack search capabilities and are limited to the visualization of a single time series. Brodbeck and Girardin [26] offer a semantic zoom implementation and the possibility to visualize very long time series. Tools that allow users to search for patterns vary in the way they let users specify the pattern, adjust the search algorithm, and browse the results. Some tools allow users to discover patterns interactively. For example, Carlis and Konstan [48] show that by using a simple interaction technique (tightening or relaxing a spiral view) users can visually reveal periodic patterns in serial periodic data without the need of running specific search algorithms. The first version of Timesearcher [26, 104] allows users to specify patterns by concatenating multiple boxes together to form a pattern at a particular time in the series. The series are filtered to show only those that pass through that set of time-specific timeboxes. It includes useful tools to interactively search by slope, or permit the specification of queries that allow a range specification in the time axis, but there is no mechanism to search for a pattern occurring at different times in the dataset. Timesearcher also started to explore the specification of queries on multiple variables and here there is the continue of this work. Other tools permit users to specify a pattern of interest and then to see search results with similar patterns. Chortaras [58] requires users to specify numerical parameters of the pattern, VizTree [131] asks users to specify the shape of the pattern by dividing the pattern into segments and specifying whether each segment belongs to a specific range. QuerySketch [205] allows users to directly sketch the shape of the pattern. An interesting contribution is IPBC [57], which presents a 3D tool that allows users to select a pattern of interest in the data itself, and then initiate a search for similar patterns. Results are highlighted showing where the similar patterns are in the dataset. This tool is customized to display time series that have periodic characteristics (e.g. hours, days, weeks, months, etc.). Visualization combined with interaction may be useful to generate hypotheses and to confirm analysis done using statistical algorithms. We believe that allowing users to select a pattern from the data itself is particularly useful in the exploratory phase of the analysis. During early data exploration, users do not know what pattern they may want to look for. They browse the data and when they see some anomalies or surprising shapes in the data, they can zoom in, select the pattern, and then start the search for similar patterns, enabling them to see where and when similar behavior occurred. Of course, allowing multiple pattern specification methods is an even more powerful approach (see section 4.10.3). No tool that we know of other than TimeSearcher allows the specification of multiple patterns on multiple variables.

4.10.2 The application

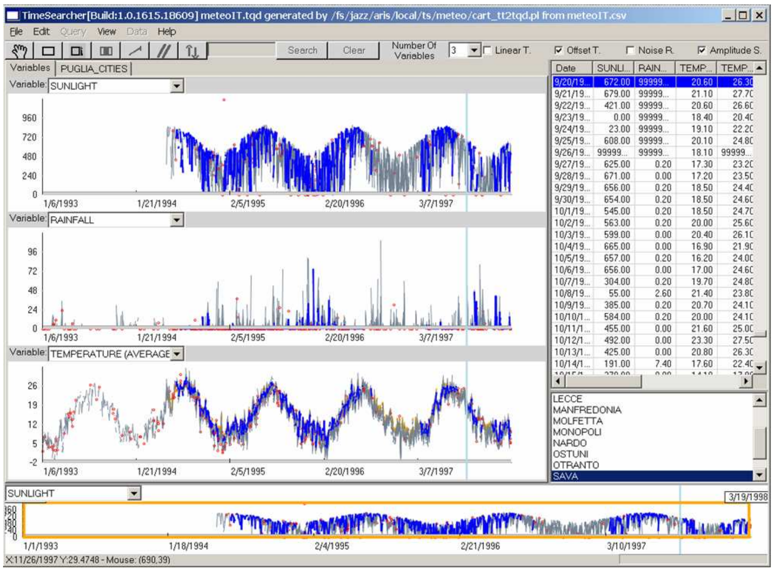
TimeSearcher 1's basic browsing capability was extended to include multiple heterogeneous variables and handles tens of thousands of time points. In addition, TimeSearcher 2's new search interface combines both filter and pattern search capability, implementing a three-step approach that can be extended to a variety of time series search interfaces. The work was primarily guided by petroleum industry production data (oil and gas well data) and meteorology data. The screens shown in the paper mostly use the meteorology data, as it is more familiar.

TimeSearcher interface

For exploratory data analysis users need an application that offers an overview of the time series and the possibility to zoom in and out. We found that in some cases the data had never been represented visually before and users were eager to explore their data first to see patterns, make hypotheses, understand outliers, and recognize that some data were missing. TimeSearcher 2 allows typically up to eight multiple heterogeneous variables to be shown at once, this limitation may be overcome using higher resolution screen. Figure 4.16 shows an example using the meteorology data that refers to locations in Puglia, a region of Southern Italy (names have been changed for this example). Three variables are shown: amount of sunlight, rainfall and average temperature, for a set of items - in this case Italian cities. On the right hand side, a scrollable table shows the numerical values and the list of locations. Approximately 5 years of data is shown and the seasonal patterns of most variables are visible. Users can highlight a specific time point in all variables by clicking on the background, which draws a light blue vertical line and highlights values at the corresponding time in the table (upper right). Figure 4.17 shows all eight variables for all the items - which are drawn overlapped. Figure 4.18 provides an alternate view showing separated views of individual items (here locations) for a single variable.

The use of a detail+overview browser to provide access to details is shown in Figure 4.19. Users may want to look at the second half of 1997, which they can achieve by moving and resizing the orange field of view box on the bottom (also delimited by dates), as it determines the range of time interval displayed on the display. Users often need to compare different time periods, so the detail view can be split into 2 panes each showing a different time period, controlled by independent field of view boxes in the overview [118]. Users can use the browsing capabilities of Timesearcher 2 to explore data. They can visually spot interesting patterns, and quickly browse recent data in search of other instances of the pattern. Nevertheless, this process becomes cumbersome when looking at extensive archives of data, and search capabilities become necessary. They are described in the next section.

Three-step interactive search One of the innovations in Timesearcher 2 is the capability to perform pattern search. Our goal was to create a



TimeSearcher[Build:1.0.1615.18609] meteoIT.tqd generated by /fs/jazz/aris/local/ts/meteo/cart_tt2tqd.pl from meteoIT.csv

Number Of Variables 3 □ Linear T. ☑ Offset T. □ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

Variable: RAINFALL

Variable: TEMPERATURE (AVERAGE

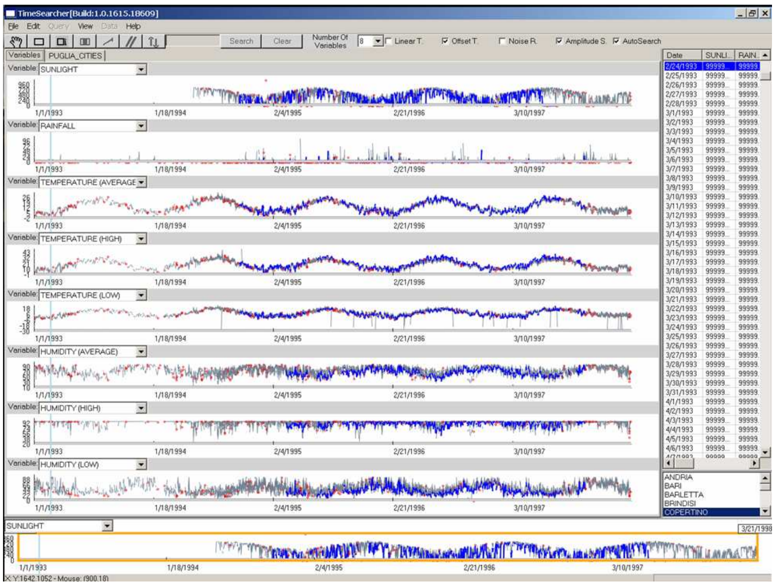
Date	SUNLI...	RAIN...	TEMP...	TEMP...
9/20/19...	672.00	99999...	20.60	26.30
9/21/19...	679.00	99999...	21.10	27.70
9/22/19...	421.00	99999...	20.60	26.60
9/23/19...	0.00	99999...	18.40	20.40
9/24/19...	23.00	99999...	19.10	22.20
9/25/19...	608.00	99999...	20.10	24.80
9/26/19...	99999...	99999...	18.10	99999...
9/27/19...	625.00	0.20	17.30	23.20
9/28/19...	671.00	0.00	17.20	23.50
9/29/19...	656.00	0.20	18.50	24.40
9/30/19...	654.00	0.20	18.50	24.60
10/1/19...	545.00	0.20	18.50	24.70
10/2/19...	563.00	0.20	20.00	25.60
10/3/19...	599.00	0.00	20.40	26.10
10/4/19...	665.00	0.00	16.90	21.90
10/5/19...	657.00	0.20	16.20	24.00
10/6/19...	656.00	0.00	17.00	24.60
10/7/19...	304.00	0.20	19.70	24.80
10/8/19...	55.00	2.60	21.40	23.80
10/9/19...	385.00	0.20	20.70	24.10
10/10/1...	584.00	0.20	20.00	24.10
10/11/1...	455.00	0.00	21.60	25.00
10/12/1...	492.00	0.00	23.30	27.50
10/13/1...	425.00	0.00	20.80	26.30
10/14/1...	191.00	7.40	17.60	22.40

LECCE
MANFREDONIA
MOLFETTA
MONOPOLI
NARDO
OSTUNI
OTRANTO
SAVA

SUNLIGHT 3/19/1998

X:11/26/1997 Y:29.4748 - Mouse: (690,39)

Figure 4.16: Three variables (amount of sunlight, rainfall and average temperature are shown. A location has been selected, highlighting its data in dark blue. A date is also selected and highlighted in the table and timeline



TimeSearcher[Build:1.0.1615.18609]

Number Of Variables 8 □ Linear T. ☑ Offset T. □ Noise R. ☑ Amplitude S. ☑ AutoSearch

Variable: SUNLIGHT

Variable: RAINFALL

Variable: TEMPERATURE (AVERAGE

Variable: TEMPERATURE (HIGH)

Variable: TEMPERATURE (LOW)

Variable: HUMIDITY (AVERAGE)

Variable: HUMIDITY (HIGH)

Variable: HUMIDITY (LOW)

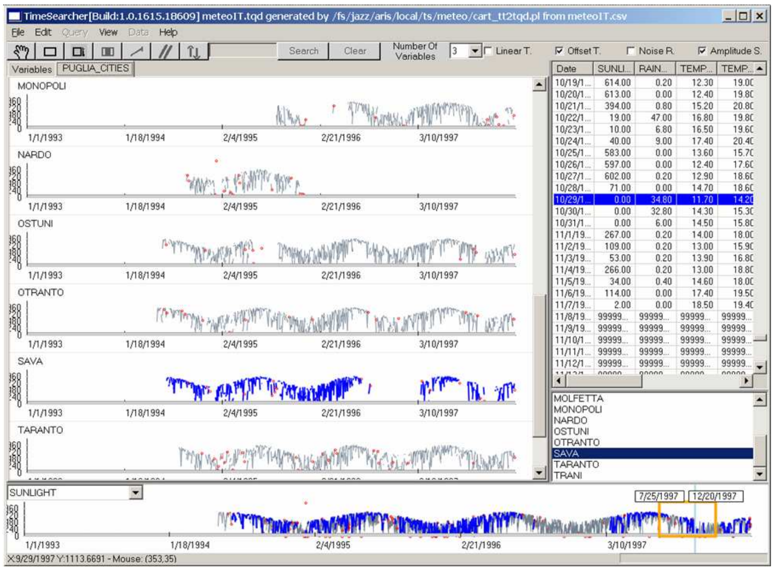
Date	SUNLI...	RAIN...
2/24/1993	99999...	99999.
2/25/1993	99999...	99999.
2/26/1993	99999...	99999.
2/27/1993	99999...	99999.
2/28/1993	99999...	99999.
3/1/1993	99999...	99999.
3/2/1993	99999...	99999.
3/3/1993	99999...	99999.
3/4/1993	99999...	99999.
3/5/1993	99999...	99999.
3/6/1993	99999...	99999.
3/7/1993	99999...	99999.
3/8/1993	99999...	99999.
3/9/1993	99999...	99999.
3/10/1993	99999...	99999.
3/11/1993	99999...	99999.
3/12/1993	99999...	99999.
3/13/1993	99999...	99999.
3/14/1993	99999...	99999.
3/15/1993	99999...	99999.
3/16/1993	99999...	99999.
3/17/1993	99999...	99999.
3/18/1993	99999...	99999.
3/19/1993	99999...	99999.
3/20/1993	99999...	99999.
3/21/1993	99999...	99999.
3/22/1993	99999...	99999.
3/23/1993	99999...	99999.
3/24/1993	99999...	99999.
3/25/1993	99999...	99999.
3/26/1993	99999...	99999.
3/27/1993	99999...	99999.
3/28/1993	99999...	99999.
3/29/1993	99999...	99999.
3/30/1993	99999...	99999.
3/31/1993	99999...	99999.
4/1/1993	99999...	99999.
4/2/1993	99999...	99999.
4/3/1993	99999...	99999.
4/4/1993	99999...	99999.
4/5/1993	99999...	99999.
4/6/1993	99999...	99999.

ANDRIA
BARI
BARLETTA
BRINDISI
COPERTINO

SUNLIGHT 3/21/1998

X: Y:1642.1052 - Mouse: (900,18)

Figure 4.17: Many variables can be shown at once, here 8 variables over 5 years, for a complete overview of the data available for the location



TimeSearcher[Build:1.0.1615.18609] meteoIT.tqd generated by /fs/jazz/aris/local/ts/meteo/cart_tt2tqd.pl from meteoIT.csv

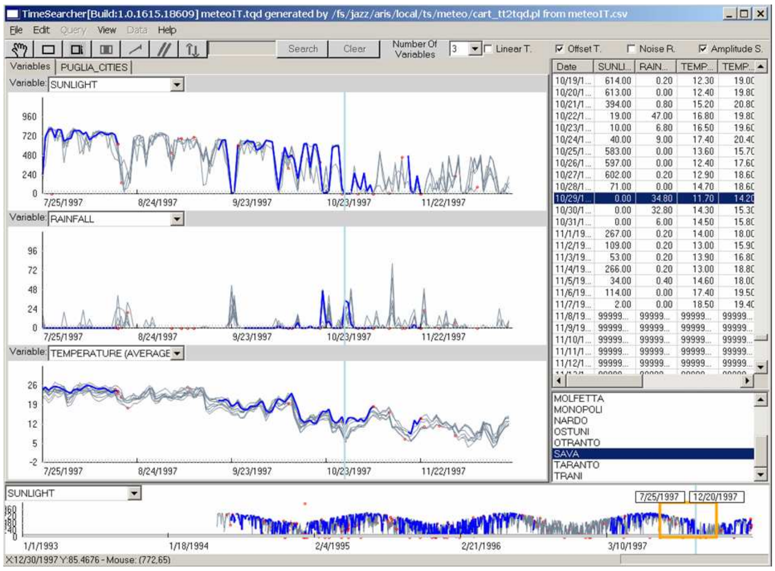
Number Of Variables: 3 ☐ Linear T. ☑ Offset T. ☐ Noise R. ☑ Amplitude S.

Date	SUNLI...	RAIN...	TEMP...	TEMP...
10/19/1...	614.00	0.20	12.30	19.0
10/20/1...	613.00	0.00	12.40	19.8
10/21/1...	394.00	0.80	15.20	20.8
10/22/1...	19.00	47.00	16.80	19.8
10/23/1...	10.00	6.80	16.50	19.6
10/24/1...	40.00	9.00	17.40	20.4
10/25/1...	583.00	0.00	13.60	15.7
10/26/1...	597.00	0.00	12.40	17.6
10/27/1...	602.00	0.20	12.90	18.6
10/28/1...	71.00	0.00	14.70	18.6
10/29/1...	0.00	34.80	11.70	14.2
10/30/1...	0.00	32.80	14.30	15.3
10/31/1...	0.00	6.00	14.50	15.8
11/1/19...	267.00	0.20	14.00	18.0
11/2/19...	109.00	0.20	13.00	15.9
11/3/19...	53.00	0.20	13.90	16.8
11/4/19...	266.00	0.20	13.00	18.8
11/5/19...	34.00	0.40	14.60	18.0
11/6/19...	114.00	0.00	17.40	19.5
11/7/19...	2.00	0.00	18.50	19.4
11/8/19...	99999...	99999...	99999...	99999...
11/9/19...	99999...	99999...	99999...	99999...
11/10/1...	99999...	99999...	99999...	99999...
11/11/1...	99999...	99999...	99999...	99999...
11/12/1...	99999...	99999...	99999...	99999...

MOLFETTA
MONOPOLI
NARDO
OSTUNI
OTRANTO
SAVA
TARANTO
TRANI

X:9/29/1997 Y:1113.6691 - Mouse: (353,35)

Figure 4.18: This alternate view shows only one variable but each location is drawn separately (opposed to overlapping). It clearly shows the periods where data is missing at each location. The start of a missing data period is indicated with a small red circle



TimeSearcher[Build:1.0.1615.18609] meteoIT.tqd generated by /fs/jazz/aris/local/ts/meteo/cart_tt2tqd.pl from meteoIT.csv

Number Of Variables: 3 ☐ Linear T. ☑ Offset T. ☐ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

Variable: RAINFALL

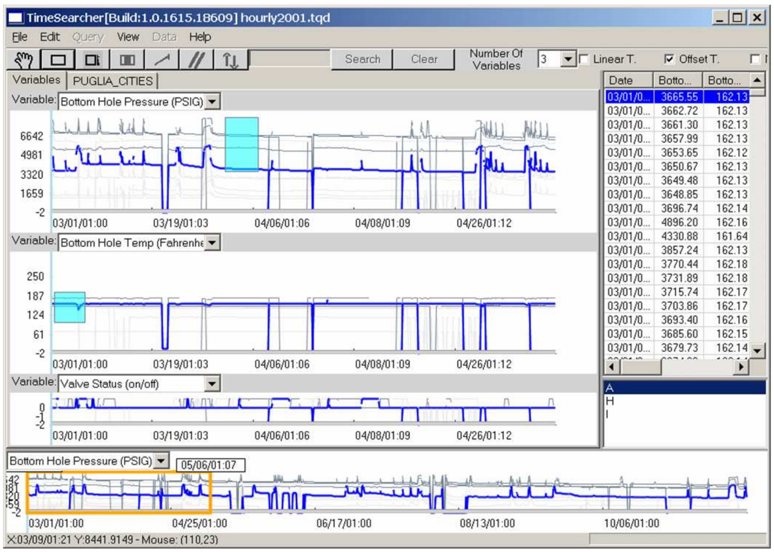
Variable: TEMPERATURE (AVERAGE

Date	SUNLI...	RAIN...	TEMP...	TEMP...
10/19/1...	614.00	0.20	12.30	19.0
10/20/1...	613.00	0.00	12.40	19.8
10/21/1...	394.00	0.80	15.20	20.8
10/22/1...	19.00	47.00	16.80	19.8
10/23/1...	10.00	6.80	16.50	19.6
10/24/1...	40.00	9.00	17.40	20.4
10/25/1...	583.00	0.00	13.60	15.7
10/26/1...	597.00	0.00	12.40	17.6
10/27/1...	602.00	0.20	12.90	18.6
10/28/1...	71.00	0.00	14.70	18.6
10/29/1...	0.00	34.80	11.70	14.2
10/30/1...	0.00	32.80	14.30	15.3
10/31/1...	0.00	6.00	14.50	15.8
11/1/19...	267.00	0.20	14.00	18.0
11/2/19...	109.00	0.20	13.00	15.9
11/3/19...	53.00	0.20	13.90	16.8
11/4/19...	266.00	0.20	13.00	18.8
11/5/19...	34.00	0.40	14.60	18.0
11/6/19...	114.00	0.00	17.40	19.5
11/7/19...	2.00	0.00	18.50	19.4
11/8/19...	99999...	99999...	99999...	99999...
11/9/19...	99999...	99999...	99999...	99999...
11/10/1...	99999...	99999...	99999...	99999...
11/11/1...	99999...	99999...	99999...	99999...
11/12/1...	99999...	99999...	99999...	99999...

MOLFETTA
MONOPOLI
NARDO
OSTUNI
OTRANTO
SAVA
TARANTO
TRANI

X:12/30/1997 Y:85.4676 - Mouse: (772,65)

Figure 4.19: To see more details users can zoom on the timeline by narrowing the orange field-of-view box drawn on the overview shown at the bottom of the screen. The overview represents the complete 5 years of data for one variable, while the detail view is zoomed on the second half of 1997


TimeSearcher[Build:1.0.1615.18609] hourly2001.tqd
File Edit Query View Data Help
Search
Clear
Number Of Variables
3
□ Linear T.
☑ Offset T.
Variables
PUGLIA_CITIES
Variable: Bottom Hole Pressure (PSIG)
6642
4981
3320
1659
-2
03/01/01:00
03/19/01:03
04/06/01:06
04/08/01:09
04/26/01:12
Variable: Bottom Hole Temp (Fahrenhe
250
187
124
61
-2
Variable: Valve Status (on/off)
0
-1
-2
Bottom Hole Pressure (PSIG)
05/06/01:07
03/01/01:00
04/25/01:00
06/17/01:00
08/13/01:00
10/06/01:00
X:03/09/01:21 Y:8441.9149 - Mouse: (110,23)

Date	Botto...	Botto...
03/01/0...	3665.55	162.13
03/01/0...	3662.72	162.13
03/01/0...	3661.30	162.13
03/01/0...	3657.99	162.13
03/01/0...	3653.65	162.12
03/01/0...	3650.67	162.13
03/01/0...	3649.48	162.13
03/01/0...	3648.85	162.13
03/01/0...	3696.74	162.14
03/01/0...	4896.20	162.16
03/01/0...	4330.88	161.64
03/01/0...	3857.24	162.13
03/01/0...	3770.44	162.18
03/01/0...	3731.89	162.18
03/01/0...	3715.74	162.17
03/01/0...	3703.86	162.17
03/01/0...	3693.40	162.16
03/01/0...	3685.60	162.15
03/01/0...	3679.73	162.14

A
H
I

Figure 4.20: Example of filtering over multiple variables, using an oil production dataset. Each item is a well, for which hourly data is available (pressure, temperature, etc.). Users have narrowed down the list of wells down to three (A, H, and I) by filtering the series to keep only the wells with high b̈ottom hole pressureëarly April and high b̈ottom hole temperatureïn early March

simple interaction for searching for a pattern in the time series. Once the pattern is specified, the search for similar patterns can be started. Pattern matching may be slow so to address this problem, we propose a three-step framework to interactively search the data. In Step 1, users reduce the scope of the query by drawing timeboxes in one or more detail views. In Step 2, users specify a pattern and get a large number of results with an approximate search. In step 3, users refine the query to narrow down the result set by dynamically manipulating the parameters of the search.

In Step 1, the goal is to reduce the size of the data to be searched. User blue time boxes on the display. A boolean AND operation is performed among all time boxes so that the remaining lines are only those that go through all the boxes. Filtered out items appear light grey on the screen (and can be removed if needed). The immediate feedback given to users clarifies the effect of users' actions and reveals the logic of the combinatory effect of timeboxes. A smaller number of items will allow the next search steps to be performed more rapidly.

Step 2 corresponds to the selection of a specific pattern and the initial search for similar sequences within the scope specified in Step 1. Users select the pattern in the dataset itself by selecting a line - therefore highlighting it - then drawing a box enclosing the pattern. Figures 9 to 12 show an example with only one variable to keep the

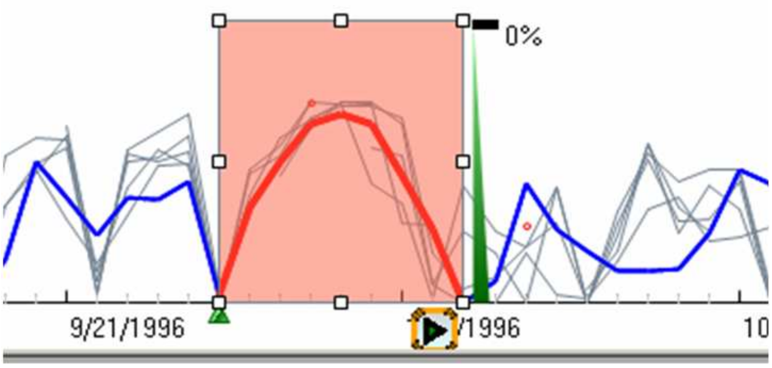
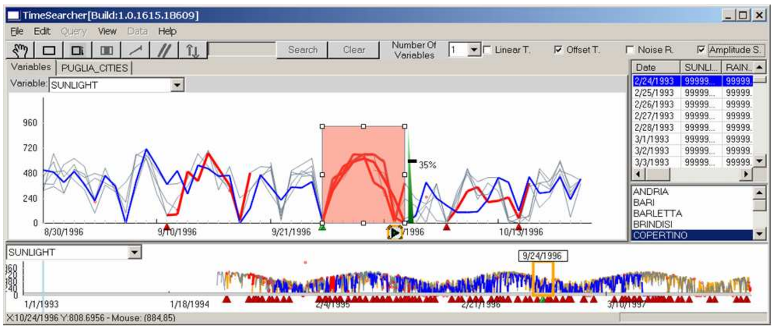


Figure 4.21: Users select a pattern by first selecting a line (i.e. a time series) and drawing a box around the pattern of interest. The default tolerance is set to zero. The reference pattern is tagged by a green triangle on the detail and on the overview

figures small. After drawing the box (Figure 8), an adjustable tolerance slider appears, allowing users to roughly set the tolerance of the match between the selected pattern and the search results.

The initial search is triggered explicitly by clicking on the arrow-shaped button located below the box. The result of the search is displayed with red triangle markers under the horizontal axes on the overview - and in the detail view(s) when applicable (Figure 4.22). Together with the triangles, the matched patterns appear with a different color (red) in the detail view and in the overview, when applicable. This step might become slow when the dataset is large but it may need to be performed only once (see discussion section). Finally, in Step 3, users reduce the tolerance and adjust search parameters, incrementally. In order to rapidly adjust the size of the result set, the users benefit from having immediate feedback about the effect of their action on the results. Users can adjust the tolerance level and immediately see the effect on the cardinality of the result set (Figure 4.23). Finally, they can use all the basic features presented in the previous subsection to browse the results (figure 4.24): overview and details, synchronized table to view numerical values, multiple detail views to compare patterns, etc. When inadequate matches are found users can readjust the tolerance or modify parameters of the search (Figure 4.25). Users can repeat the operation by selecting patterns in multiple variables (Figure 4.14). At this stage our implementation handles the variables independently, i.e. results of the pattern matching in one variable are highlighted in that variable view only and no boolean operation is performed. The overview shows the positions of all patterns matched over the entire



Number Of Variables: 1 □ Linear T. ☑ Offset T. □ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

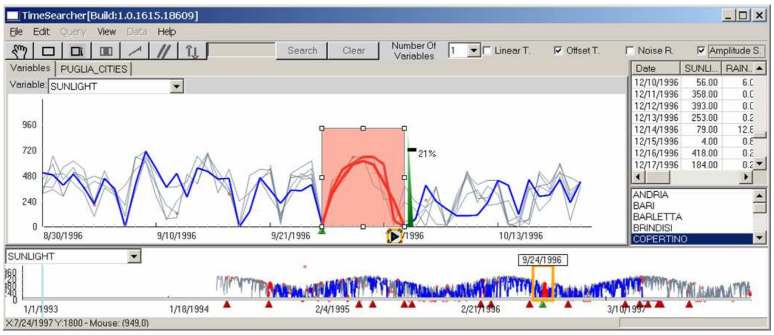
Date	SUNLI...	RAIN..
2/24/1993	99999...	99999.
2/25/1993	99999...	99999.
2/26/1993	99999...	99999.
2/27/1993	99999...	99999.
2/28/1993	99999...	99999.
3/1/1993	99999...	99999.
3/2/1993	99999...	99999.
3/3/1993	99999...	99999.

ANDRIA
BARI
BARLETTA
BRINDISI
COPERTINO

SUNLIGHT

X:10/24/1996 Y:808.6956 - Mouse: (884,85)

Figure 4.22: The slider is first adjusted to an arbitrary high level of tolerance and the search is initiated by clicking on the arrow shaped button below the pattern. The starting point of each matching sequence is indicated by a red triangle. Here there are too many to examine



Number Of Variables: 1 □ Linear T. ☑ Offset T. □ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

Date	SUNLI...	RAIN..
12/10/1996	56.00	6.0
12/11/1996	358.00	0.0
12/12/1996	393.00	0.0
12/13/1996	253.00	0.2
12/14/1996	79.00	12.8
12/15/1996	4.00	0.8
12/16/1996	418.00	0.2
12/17/1996	184.00	0.2

ANDRIA
BARI
BARLETTA
BRINDISI
COPERTINO

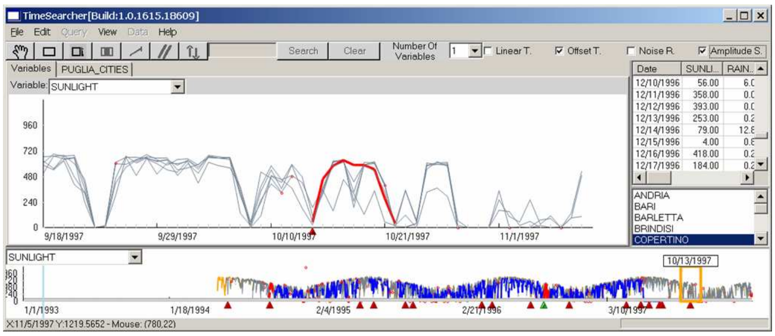
SUNLIGHT

X:7/24/1997 Y:1800 - Mouse: (949,0)

Figure 4.23: The tolerance level is reduced iteratively until the result set is small enough to be examined

time series. Users are then able to browse the results to see how patterns found in separate variables match in time. Even though the standard setup shows an overview of the entire dataset at the bottom of the screen on a single variable, users can stretch the field-of-view box (the box in the overview panel) to cover the entire time span of the dataset to see an overview of all variables at once in the area generally reserved for details.

Search algorithm used The search consists of two sets of algorithms. Transformation algorithms are applied to the search pattern and to sequences of the times series. Comparison algorithms compare the transformed search pattern and the transformed sequences. Transformation: We implemented four transformations. They are offset translation, magnitude scaling, linear trend removal and noise reduction. Offset translation and magnitude scaling are applied by default but users can specify which transformations are used using the check boxes available in the top-right area of the screen. Comparison: Let's first define the tolerance percentage and tolerance value. The tolerance percentage is a value chosen by the user with the search widget. Currently it is a percentage value ranging from 0% to 100% selected with a slider. 0% tolerance corresponds to exact matches only, and 100% returns a very large number of matches. The tolerance value is used to determine whether the sequence under comparison matches the pattern. We defined the tolerance value to be the range (Max - Min) of the transformed pattern multiplied by the tolerance level. To illustrate, suppose the range of the transformed pattern is 10 units, and the tolerance level chosen by the user is 20%. The resulting tolerance value will be 2 units.



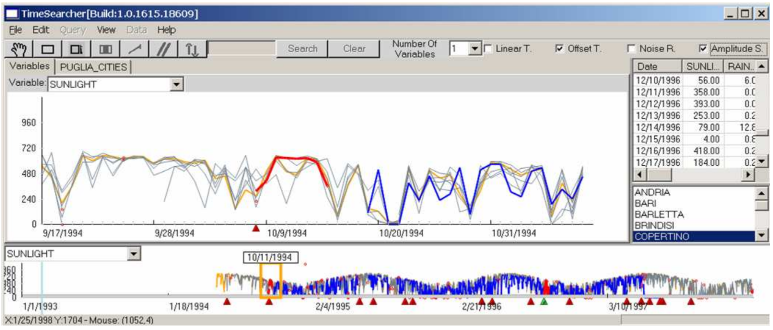
Number Of Variables: 1 ☐ Linear T. ☑ Offset T. ☐ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

Date	SUNLI...	RAIN..
12/10/1996	56.00	6.0
12/11/1996	358.00	0.0
12/12/1996	393.00	0.0
12/13/1996	253.00	0.2
12/14/1996	79.00	12.8
12/15/1996	4.00	0.8
12/16/1996	418.00	0.2
12/17/1996	184.00	0.2

X:11/5/1997 Y:1219.5652 - Mouse: (780,22)

FIGURE 4.24: Moving the field-of-view box to the red arrows reveals a sequence matching the search pattern



Number Of Variables: 1 ☐ Linear T. ☑ Offset T. ☐ Noise R. ☑ Amplitude S.

Variable: SUNLIGHT

Date	SUNLI...	RAIN..
12/10/1996	56.00	6.0
12/11/1996	358.00	0.0
12/12/1996	393.00	0.0
12/13/1996	253.00	0.2
12/14/1996	79.00	12.8
12/15/1996	4.00	0.8
12/16/1996	418.00	0.2
12/17/1996	184.00	0.2

X:1/25/1998 Y:1704 - Mouse: (1052,4)

FIGURE 4.25: When a search result is found unacceptable (e.g. here users realize that they care about the amplitude of the pattern) algorithm parameters can be adjusted (here by turning off the "amplitude scaling" option at the top right of the screen)

We implemented two comparison algorithms, both using a sliding window on the time series having the same size as the pattern. The first algorithm calculates the Euclidean distance between the window and the pattern, and determines the sequence as a match only if this distance is less than or equal to the tolerance value multiplied by the length of the pattern. The Euclidean distance is the following formula: $D(Q,C) \equiv \sqrt{\sum_{i=1}^{n}(Q_i - C_i)^2}$ where Q and C are two time series with n points and D is the distance function for two time series but we remove the square root operation to speed-up the calculation (which is acceptable as Euclidean distance is a monotone function) The second algorithm calculates the difference between corresponding point values. If the difference for every point is within the tolerance value, the sequence is determined to be a match; otherwise, it is excluded from the result set. Those simple algorithms were chosen because the data provided by our users were not extremely large, had no periodicity, and because our users could not predict the size of the pattern to be searched, which meant that existing advanced search algorithms could not be exploited profitably (see discussion in section 4.10.3).

The tolerance value depends on the transformations selected by the user. Hence, a 10% tolerance in one transformation corresponds to a different tolerance value into another transformation. As the user selects or deselects transformations to be applied, the tolerance value changes although the tolerance percentage remains the same. Overall we can see that the tolerance value remains an arbitrary measure that has limited meaning for users therefore we plan to change the display to show + and - signs instead of displaying a number for the tolerance. The meteorological dataset contains 2000 time points (daily means of temperatures, humidity, etc.). Each time point has 8 variables and it is possible to perform the search on 15 items (locations in our case). All queries in this reasonably sized dataset can be performed in real time because we can load all the data in the high speed store, therefore achieving dynamic queries [177]. Dynamic queries imply rapid, incremental, and reversible actions; and the immediate display of feedback (less than 100 milliseconds). The oil production data sample has approximately 10,000 time points and 7 variables for 10 wells, and dynamic queries were not achieved in step 3 at this point.

4.10.3 Discussions

The tool still has many opportunities for improvement. Main issues include dealing with larger datasets, improving the interface, and dealing with missing data. Dealing with larger datasets: Our algorithms can be optimized but for much larger datasets further development is needed to handle each step separately so that slower steps can be triggered explicitly (by pressing a button) while others are triggered implicitly to achieve dynamic queries. One technique consists of using Step 2 as another scope reducing step. Once users have run the search once, with a high tolerance value of X, the result set R can be used as the new reduced scope of refinement queries using tolerance Y<X (i.e. when dragging the slider up). This makes the operation of reducing the tolerance faster as the tolerance level is reduced). Increasing the tolerance would require running a search on the entire data again, unless the algorithm keeps track of the value for which a pattern was dropped off the result list. And if the value is greater than X, or when transformations are checked or unchecked, then a complete search on the entire scope defined in Step 1 has to be performed again. Another possibility is to index the time series in step 2 so the dynamic queries may be applied also for a bigger data set. When the data (indexed or not) is larger than the high-speed storage capacity, the three-step framework can be applied as well. The initial search of Step 2 may be slow, but Step 3 can become interactive when the Step 2 result set is small enough to fit into the high-speed storage.

When the size of the pattern is identified in advance, simple search can be replaced by faster indexed searches that find a small set of results then apply the Euclidian distance to quickly refine this set. This is the case when the data are cyclic and patterns of interest match that cycle. For example, for EKGs the cycle is on the order of a second. Similarly, weekly cycles are likely to be present in time series of human work activities. Those cycles are important to index the data for pattern search because very efficient indexing techniques are available when the duration of the patterns to be searched is known in advance [159, 126]. When the length of the search pattern cannot be known in advance but only approximated to fall within a fixed range, multiple indices can be created for a reasonable set of different pattern lengths, having an index for each length.

Improving interaction: Our early feedback highlighted the benefits of good browsing capabilities. Exploring the data visually is extremely important, and providing access to the numerical values needs to be supported with synchronized views of the graphs, tables and lists. Multiple presentations of the data are useful (for example overlapping versus sequential views of the items). User suggestions included providing multiple methods to specify the pattern and the options of the search algorithm. For example, one alternative to specifying tolerance would be to allow users to interactively modify the selected pattern or to draw boundaries of tolerance on the pattern by interacting with the line using direct manipulation. When the users know the pattern they are looking for and it is not easy to find an existing similar one, pure sketching may help [205]. The selection of patterns in existing data would also be complemented by a pure sketching option. This could easily be added to the current software. Alternatives can be offered as well for the selection of search parameters. Our early user feedback indicates that some users are confused by the tolerance slider label, because the tolerance measure has no real meaning. We initially chose to display a number to allow users to return to a previous value that had been found useful. Hopefully, the dynamic query behavior will help users quickly understand how to use the tolerance slider. An option might be to avoid displaying a numerical value and only provide + and - controls. Another natural suggestion is to extend the scope-setting Step 1 by allowing users to limit the time range where the search should be performed. Note that currently, nothing requires users to start with Step 1, and they can start with Step 2 and 3. Nevertheless it might be good to encourage users to reduce the scope of the search with the filter boxes when the size of the data does not allow the use of dynamic queries.

Dealing with missing data: Missing data is another common problem that needs to be carefully addressed. Time series make it easy to inform users of the fact that data is missing [73] and TimeSearcher 2 shows the location of the beginning of missing data (Figure 5). Nevertheless, standard annotation mechanisms are needed to inform users about the default method used to handle missing data in the search (is the missing data ignored? Is it considered a perfect match? Is it replaced by an estimated value?), and users should be able to specify what method is to be used. Our experience suggests that providing annotation mechanisms will also be important to document the reasons for the absence of the data and that exploratory tools such as Timesearcher are a natural environment to gather anecdotal knowledge about the missing data. Evaluation: The formal evaluation of exploratory tools such as TimeSearcher remains a challenges[50]. Formative usability studies will allow us to improve the interface but, more importantly, we will continue working with users to identify and report on case studies, that will help us understand the range of data type and application for which TimeSearcher 2's pattern search can be effective.

Chapter 5

Design for usability

5.1 Introduction

The framework presented in this thesis has been developed with the aim of making it useful and usable by its intended users. In this chapter we will first describe what is usability, then the usability principles adopted for the design of the framework and how we did get user requirements. Section 5.5 is dedicated to the methods for usability evaluation and the following two sections describe the evaluation of the framework.

5.2 Usability of interactive systems

It is now widely acknowledged that usability is a crucial factor of the overall quality of interactive applications. Several definitions of usability have been proposed. Nielsen defined a model in which usability is presented as one of the aspects that characterizes a global feature of a system that is acceptability by the users, reflecting whether the system is good enough to satisfy needs and requirements of the users.

The acceptability of a computer system is a combination of its social acceptability and its practical acceptability [148], as shown in Figure 5.2 its practical acceptability is analyzed within various categories, including traditional categories such as cost, support, reliability, compatibility with existing systems, etc., as well as the category of usefulness. Usefulness can be broken down into the two categories of utility and usability, where utility is the question of whether the functionality of the system in principle can do what is needed, and usability is the question of how well users can use the functionality.

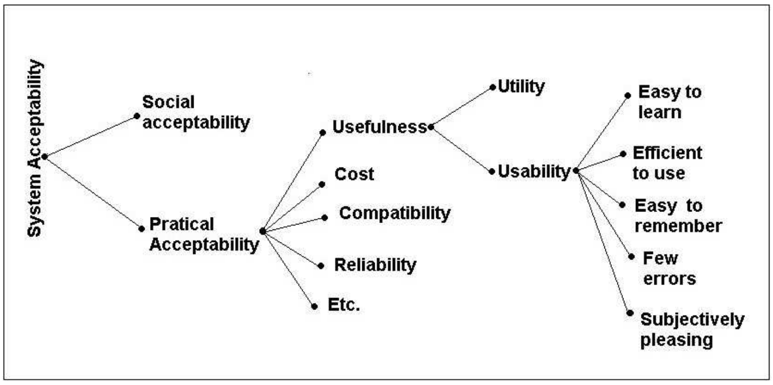


Figure 5.1: Usability Definition about Nielsen

In Nielsen model, usability is not a one-dimensional property of a system, rather it has multiple components. It can be decomposed into five attributes:

- Learnability: the system should be easy to learn so that the user can rapidly start getting some work done with the system;
- Efficiency: the system should be efficient to use, so that once the user has learned the system, a high level of productivity is possible;
- Memorability: the system should be easy to remember, so that the casual user is able to return to the system after some period of not having used it, without having to learn everything all over again;
- Few Errors: the system should have a low error rate, so that users make few errors during the use of the system, and so that if they do make errors they can easily recover from them;
- Satisfaction: the system should be pleasant to use, so that users are subjectively satisfied when using it; they like it.

Only by defining the abstract concept of "usability" in terms of these more precise and measurable components, we can arrive at an engineering discipline where usability is not just argued about but is systematically approached, improved, and evaluated.Different authors in the Human-Computer Interaction (HCI) literature have proposed different usability principles for interactive applications. Nielsen [148] reports the following usability principles:

1. Simple and natural dialogue: Dialogue should not contain information that is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility. All information should appear in a natural and logic order.
2. Speak the user's language: Dialogue should be expressed clearly in words, phrases and concepts familiar to the user.
3. Minimize use memory load: The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be always visible or simply retrievable when it is necessary.
4. Consistency: Users should not have to wonder whether different words, situations, or actions should have always the same meaning.

5. Feedback: The system should inform the user about what he is doing by means appropriate, effective and efficiency feedback.
6. Clearly marked exits: Users often choose functions mistakenly and then they need simple to leave the unintentional status.
7. Shortcut: Accelerator, unseen by the novice user, may often speed up the interaction for expert user such that the system caters to both inexperienced and experienced users.
8. Good error messages: Messages should be expressed in a easy language (no codes), should indicate the problem and should suggest the solution in a constructive way.
9. Prevent errors: Even better than good error messages is a careful design that prevents a problem from occurring in the first page.
10. Help and documentation: Even better it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the users´ task, and it should specify the concrete steps to execute.

5.3 Designing usable systems

FairsNet has been developed by using a User-Centered Design methodology in order to build a system that satisfies clear usability objectives. As a consequence, the main key points have been the analysis of all possible users of the system and, in general, of all the stakeholders, as well as the evaluation process that involved all phases of the system design and development. Moreover, during the system development, users interacted with project team according to a Participatory Design approach [143], as it is described in this document. In this section, we briefly illustrate the user-centred methodology, and then we describe our approach to the FairsNet design.

5.3.1 User-Centered Design

User-Centered Design implies that final users are involved from the very beginning of the planning stage, and identifying user requirements becomes a fundamental and crucial phase to any development project.

Early involvement of users has the potential for preventing serious mistakes and for identifying what is effectively needed for them [161]. Poor or inadequate requirement specifications can result in interaction difficulties, including lack of facilities and usability problems. The basic principles of User-Centered Design are:

1. analyze users and tasks;
2. design and implement the system iteratively through prototypes of increasing complexity;
3. evaluate design choices and prototypes with users.

User-Centered approach requires understanding reality: who will use the system, where, how, and to do what. Then, the system is developed iterating a design-implementation-evaluation cycle. In this way it is possible to avoid serious mistakes and to save re-implementation time, since the first design is based on empirical knowledge of user behaviour, needs, and expectations. In the traditional software life cycle, the standard waterfall model, which is system-centered, usability is not adequately addressed. Moreover, there are some significant drawbacks. For instance, the system is tested only at the end of the cycle, when unfortunately is too late for going through radical design modifications to cope with possible discrepancies with the requirements. Another problem is that these requirements are collected with customers, who often are different from the people who will use the system. Customers are the people who negotiate with designers the features of the intended system, while users or end users are those people that will actually use the designed systems [63]. A direct consequence of the restricted nature of the requirement specifications is that, usually, system testing is not only performed late in the development cycle, but is also limited to some of its functional aspects, thus neglecting system usability.

In order to create usable interactive systems, it is therefore necessary to augment the standard life cycle to explicitly address usability issues: the User-Centered Design methodology stresses the iteration of the design-implementation-evaluation cycle.

The key principles of the User Centered Design methodology have been captured in the standard ISO 13407 standard (Human-Centered Design process for interactive systems), that is shown in Figure 5.2. The key points of user-centred design are: 1) analyse users, task, as well as the context in which users operate (as indicated in block 2 and 3 of

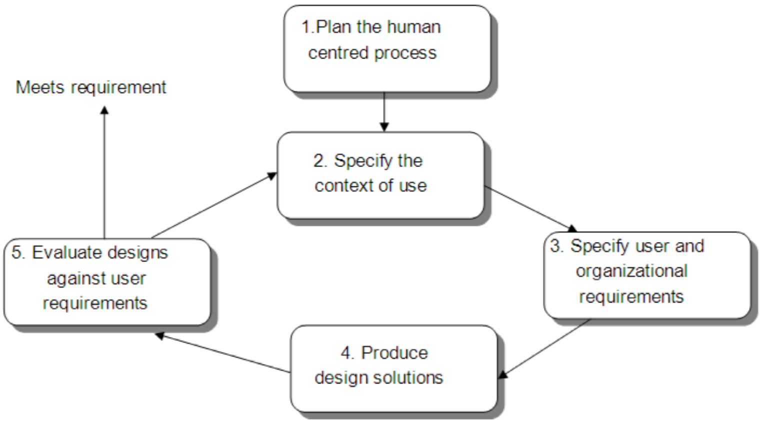

1.Plan the human centred process
2. Specify the context of use
3. Specify user and organizational requirements
4. Produce design solutions
5. Evaluate designs against user requirements
Meets requirement

Figure 5.2: ISO 13407: Human-centred design process for interactive systems

Figure 5.2); 2) design and implement the system through prototypes of increasing complexity (block 4 of Figure 5.2); 3) evaluate design choices and prototypes with user requirements and possibly real users (block 5 of Figure 5.2). Iterate this cycle and stop it when design meets requirement. The iterative process is stopped when requirements are met. From this it follows that evaluation represents the central phase in the development cycle. For this reason, within the HCI community, Hartson and Hix have developed the star life cycle model shown in Figure 5.3.1 [101].

The star model recognizes that this approach needs to be complemented by a bottom-up (synthetic) approach, and can start from any point in the star (as shown by the entry arrows), and followed by any other stage (as shown by the double arrows).

Design and implementation of FairsNet follows a User-Centred Design: users have an active role, they have to understand problems and also to propose solutions. This has been possible thanks to the presence in the Consortium of members that are real users of the system that will be developed. Other users have also been involved by the various partners. The system is designed iterating a design-implementation-evaluation cycle. In this way it is possible to avoid serious mistakes and to save re-implementation time since the first design is based on empirical knowledge of user behaviour, needs, and expectations.

5.3.2 Participatory Design

Participatory Design (known as the Scandinavian approach) [3] acknowledges the importance of involving users in the design process and,

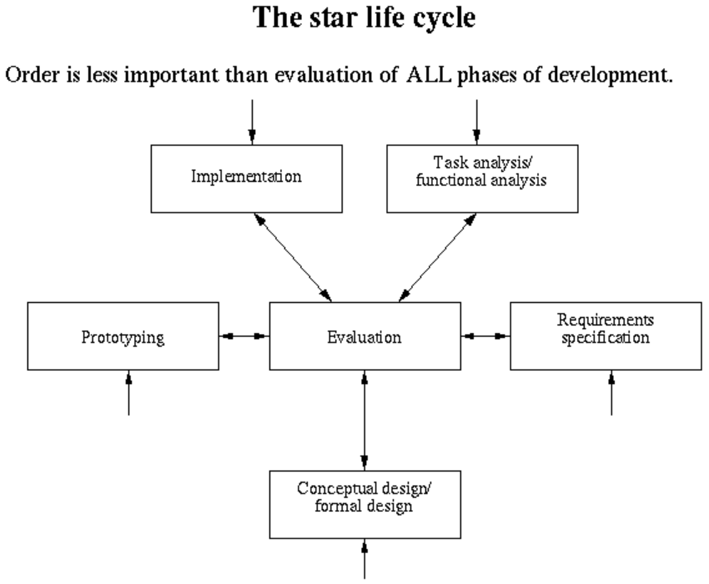

The star life cycle
Order is less important than evaluation of ALL phases of development.
Implementation
Task analysis/ functional analysis
Prototyping
Evaluation
Requirements specification
Conceptual design/ formal design

FIGURE 5.3: The Star Life Cycle Model

indeed, argues that they have a right to be involved in the design of the systems which they will subsequently use. Users participate by analyzing organizational requirements and considering appropriate social and technical structures to support both individual and organizational needs.

Participatory Design is a philosophy, which encompasses the whole design cycle, and incorporates the user not only as an experimental subject but also as a member of the design team. Users are therefore active collaborators in the design process, rather than passive participants whose involvement is entirely governed by the designer. The argument is that the users are experts in the work context and a design can only be effective within that context if these experts are allowed to contribute actively to the design.

Participatory Design therefore aims to refine system requirements iteratively through a design process in which the user is actively involved, being included in the design team, in order that they contribute to every stage of the design process. Participatory Design encompasses several methods to facilitate data exchange between the users and the designers. In FairsNet, we have adopted the following:

- Brainstorming

- Storyboarding

- Workshops

- Pencil and paper exercises

Brainstorming

Brainstorming is a technique designed to help creative thinking in initial product development. A large number of ideas are generated, many of which are discarded. In the process it is hoped that innovative ideas will arise that can then be followed up in more detail. The number of people in a brainstorming group can range from 2 - 12 people. The members of the group should have a range of experiences and stakes in the problem to be solved. Hence the group should not only consist of experts, but also lay people and a range of people in between. They should all have some familiarity of the problem preferably from different viewpoints. Brainstorming sessions are relatively easy to run, but do need to be handled carefully. The group needs to be carefully managed so that all participants can contribute without being criticized, but it is necessary to have one leader or chairperson to present the problem. It is also important to keep the conversation from straying too far and to prevent any criticism between the group members, as this will halt the flow of creativity.

Storyboarding

Storyboarding can be used as a means of describing the user's day-to-day activities as well as the potential designs and the impact they will have.

Workshops

Workshop can be used to fill in the missing knowledge of participants and provide a more focused view of the design. They may involve mutual enquiry in which all parties attempt to understand the context of the design from each other's point of view. The designer questions the user about the work environment in which the design is to be used, and the user can query the designer on the technology and capabilities that may be available. This establishes common ground between the user and designer and sets the foundation for the design that is to be produced.

Pencil and paper exercises

Pencil and paper exercises allow designs to be talked through and evaluated with very little commitment in terms of resources. Users can walk through typical tasks using paper mock-ups of the system design. This is intended to show up discrepancies between the user's requirements and the actual design as proposed. Such exercises provide a simple and cheap technique for early assessment of models.

The predominant activity in designing systems like FairsNet is that the participants in the design team teach and instruct each other; since domain experts (in this case fair experts) understand the practice and system designers know the technology. The knowledge relevant to the problem is distributed and can be opinionated - it must be acquired and formalized with a careful user requirements analysis.

5.4 Gathering user requirements

In this sectin is briefly reported the approach for gathering data about users and general system requirements taken by the FairsNet partners. As described above, the approach of FairsNet partners is consistent with the user-centred methodology we decided to adopt, even if the specific methods used by each partner may slightly vary depending on the partner's specific experience and expertise and circumstances of their users.

In the United Kingdom the adopted approach has been the combination of the retrieval of published material, including periodicals, literature and web sites with interviews and information-gathering visits. The findings of these activities have been combined with the experiences of the system designer, gained by attending, and in some cases exhibiting at, several trade fairs, exhibitions and related events over recent years. The analysis and information gathering has suggested that the key users and beneficiaries for the system are the Organisers (principally), the Exhibitors, the Service Providers/Contractors and the event Visitors/Delegates themselves (a description of user types is given in Section 1.2.3). The selection of users has been targeted to get the representative views of the above categories of event participants and to cover a range of events. It was also felt important to capture input regarding events that combine traditional fair activity (exhibiting and selling) with complimentary activities, such as conferences and seminars. Several specific venues and events were considered, and participants were interviewed/consulted.

The aim of the FairsNet user requirements analysis performed by INMARK partner in Spain has been to identify the needs and requirements of those users who are the main actors for the FairsNet system, namely trade fairs organisers, exhibitors and professional visitors.

Fair organiser's requirements and opportunities of the FairsNet system were discussed with Senior Executives from the Fundación Semana Verde de Galicia (FSVG), in a series of working meetings in the FSVG headquarters in Silleda (Galicia) during April and May 2000. On the other hand, INMARK has interviewed exhibitors and professional visitors attending the Semana Verde de Galicia International Fair held in Silleda between 3-7 May 2000. A total of 60 direct face to face interviews were performed during the exhibition. The main objective of these in depth interviews was to obtain further information on the current business processes in the fair industry, as well as to know the vision of exhibitors and professional visitors on virtual fairs and what they would like to get from FairsNet.

The user requirement analysis has been carried out in Italy by the partners University of Bari (Uniba for short) and Luiss. To this purpose, Uniba identified Agrilevante as a pilot site. Agrilevante is a fair on Agriculture, held every year in September at the venue Fiera del Levante in Bari. The Uniba approach to data gathering has involved on-site visits at the Fiera del Levante venue, interviews to people belonging to different user types involved with Agrilevante, but also with other fairs, and questionnaires submitted to various users. Fair organisers have been interviewed, and also people of the technical and maintenance office who, at Fiera del Levante, offer technical support to exhibitors. Such a support is the one due by fair organiser in order to gather exhibitors' requirements, identify their business process, and discuss opportunities for a fair support system such as FairsNet.

Fair exhibitors have been interviewed in order to obtain information on the business process related to their participation to fair events, and to understand their attitude toward systems like FairsNet. In order to gather more information on a larger sample of exhibitors, a questionnaire has been designed and submitted to exhibitors participating at the Tecnorama Fair held in Bari at Fiera del Levante during 4-7 May 2000. Uniba also interviewed generic visitors of that fair (in particular a group of Computer Science students that visited the fair and stopped at the Uniba stand where a poster on the FAIRWIS Project was displayed), people working at the fair, and also the people responsible of the Fidanzia Sistemi srl, a company of about 60 employees, which is an official service provider of Fiera del Levante and of many Italian fairs as well (eg. Fiera di Milano, Fiera di Bologna).

The approach of LUISS has been to identify a selection of venues, organisers, exhibitors and professional visitors, and to submit them e-mail questionnaires. Their answers have been then the basis for a series of interviews, tailored on the individual activity and experience of the person. More information has been gathered from the involved people in form of papers, web sites, working documents, checklists and so on. The requirement gathering was performed by involving primarily the Fiera di Roma venue, located in Roma, where several fairs take place all year around, and the IGStudents Foundation, that is promoted by the Società per l'Imprenditorialità Giovanile S.p.A. (an Italian company for the promotion of young entrepreneurs). The IGStudents mission consists in the promotion of the country's development, using enterprise as a means to improve the link between the world of education and the working world, to aid the emergence of vocation and to promote professional experiences suitable to the growth of transversal competencies among young people. Other organisations have been involved too.

The interviewed persons gave often their experiences in more than one role, i.e. exhibitors and organisers answered also as business visitors and press. Their answers have been integrated with the LUISS team's own experiences in several Italian and foreign trade fairs, as professional visitors and sometime exhibitors and related event organisers. The interviewed users have been involved in checking LUISS drafts while the work was in progress, to be sure that they would reflected the experiences they were based on.

5.5 Usability Evaluation in the Software Life Cycle

In order to design usable systems, we have seen that in User-Centered design usability evaluation plays a fundamental role. The HCI research has provided several methods that can help the designers in taking their decisions during the different stage of the development of a usable system.

Many different techniques can be applied for collecting user information, among them direct and indirect observation, interviews and questionnaires [94, 70, 161]. Direct and indirect observation means observing the users while they carry out their tasks at their workplace. It is the most reliable and precise method for collecting data about users, especially valuable for identifying user classes and related tasks. Moreover, it allows identifying critical factors, like social pressure, that can have a strong effect on user behaviour when the system will be used in the field. Unfortunately, direct observation is very expensive because it requires experimenters to observe each user individually. It can also lead to a level of 'artificial' behaviour, as those being observed react to the observation. For these this reasons, it is most useful when a reduced number of observations is enough to generalize behavioural predictions or when hypotheses have to be tested rather than generated. Interviews collect self-reported experience, opinion, and behavioural motivations. They are essential to gaining finding out procedural knowledge as well as problems with currently used tools. Interviews cost a bit less than direct observations, because they can be shorter, easier, and quicker to document to code. However, they still require skilled experimenters interviewers to be effective. By contrast, self-administered questionnaires can be handed out and collected by untrained personnel allowing to the gathering from various users of a large huge quantity of data at relatively low cost. They allow statistical analyses and stronger generalizations than interviews but also lack the flexibility of questioning. Questionnaires provide an overview on the current situation as well as specific answers. They can be readily produced and distributed to reach as wide an audience as possible bur need to be carefully constructed as not to predetermine the answers given to them.

Which combination of these methods is best worth to applying depends both on requirements and budget. By elaborating the outcome of the knowledge phase, designers define a first version of the system. At this stage, design techniques (e.g., task-centered [154] or scenario-based [161]) provide satisfying solutions. The goal is to explore different design alternatives before settling on a single proposal to be further developed. Possibly, in this way designers will propose different solutions and different interaction strategies. Techniques like such as paper mock-ups and prototyping can be applied.

Paper mock-ups are the cheapest: pieces of the system interface are drawn on paper and an interviewer experimenter simulates the interaction with a user. Despite its simple trivial appearance, this technique allows for the collecting of reliable data, which can be used for parallel reviewing.

Prototyping allows testing some functionality in depth (vertical prototyping) or the whole interface (horizontal prototyping). Then, one or more solutions can be evaluated with or without users. This step, called formative evaluation, aims at checking some choices and getting hints for revising the design. Different methods can be used for evaluating systems at the different phases of their development: the most commonly adopted are user-based methods and inspection methods.

User-based methods mainly consist of user testing, in which usability properties are assessed by observing how the system, or a prototype of the system, is actually used by some representatives of real users performing real tasks [161, 70, 114]. Usability inspection methods involve expert evaluators only, who inspect the user interface in order to find out possible usability problems, provide judgements based on their knowledge, and make recommendations for fixing the problems and improving the usability of the application.

User-based evaluation provides a sounder trusty evaluation from the user perspective, because it assesses usability through samples of real users. However, it has a number of drawbacks, such as the difficulty to properly select a correct sample of the user community, and to train it to manage not only the main application features but also the most sophisticated and advanced facilities of an interactive system.

With respect to user-based evaluation, usability inspection methods are more subjective, having heavy dependence upon the inspector skills and preconceptions. Among the inspection methods, we may include [150]:

- heuristic evaluation;
- cognitive walkthrough;
- formal usability inspection;
- guidelines reviews.

Heuristic evaluation is the most informal method; it involves a usability expert who analyses the dialogue elements of the user interface to check if they conform to usability principles, usually referred as heuristics, hence the name of this method. In a cognitive walkthrough, the expert uses some detailed procedures to simulate users' problem solving processes during the user-computer dialogue, in order to see if the functionalities provided by the system are efficient for users and lead the to correct actions.

Formal usability inspection is a review of users' potential task performance with a product. It was designed to help engineers to review a product and identify any find a large number of usability defects. It is very similar to the traditional 'code inspection' methods with which software developers have long been are familiar. It is carried out by the engineer designing the product and a team of peers, looking for defects. Finally, in a guidelines review, the experts inspects the interface to check if it is conforms to a list set of usability guidelines. The method can be considered as a cross between heuristic evaluation and standard inspection, the latter is another kind of inspection to check the compliance of the interface to some interface standards. A detailed description of these and other inspection methods can be found in [150].

The main advantage of inspection methods is the cost saving: they do not involve users nor require any special equipment or lab facilities [148, 150]. In addition, experts can detect a wide range of problems and possible faults of a complex system in a limited amount of time. For these reasons, inspection methods have achieved widespread use in recent the last years, especially in the industrial environments [149], since industry is very much interested in effective and formalised methods, that can provide good results whilst being still cost-effective and easily operated.

Inspection methods aim at finding usability problems in an existing user interface, and to then make recommendations for fixing such these problems. Hence, they can be applied at various steps of the software development, and are appropriate certainly used for evaluating the design of the system in a prototype form, even a paper prototype, so that possible defects can be fixed as soon as possible.

When a system implementation is available, user-based evaluation is often recommended. It includes experimental methods, observational methods, and survey techniques. Among experimental methods, controlled experiments are very valuable; they provide empirical evidence to support specific hypotheses. They allow a comparative evaluation, which is very useful when alternative prototypes or versions of the same system are available. An experiment consists of the following steps: formulation of the hypotheses to be tested, definition of the experimental conditions that differ only in the values of some controlled variables, execution of the experiment, and analysis of collected data.

In order to verify the usability of a single prototype, we can also observe users working with it. A valid technique is the thinking aloud, in which users are asked to think aloud when they use the system or prototype. In this way, evaluators can detect users' misconceptions and the system elements that cause them.

Both experimental and observational methods are used for collecting data about system and user performance; they do not provide data about users' satisfaction that is a subjective measure that can be obtained by survey techniques, such as interviews and questionnaires [161, 70].

By considering the industry's interest for cost effective heap but effective methods, heuristic evaluation plays an important role. It prescribes having a small set of experts analyzing the system, and evaluating its interface against a list of recognized usability principles, the heuristics. Some researches have shown that heuristic evaluation is a very efficient usability engineering technique [117], with a high benefit cost-ratio [149], and therefore it falls within the so-called discount usability methods.

In principle, only one evaluator can conduct heuristic evaluation. However, in an analysis of various studies, it has been assessed that single evaluators are able to find only the 35% of the total number of the existent usability problems [148, 150]. Different evaluators tend to find different problems. Therefore, the more experts that are involved in the evaluation, the more problems it is possible to find. The mathematical model defined in [149] shows that reasonable results can be obtained by having only five evaluators.

5.6 Evaluation Process of FairsNet

Evaluation is not a single phase of the development process of an application, but is an iterative set of processes deployed across a set of prototypes. This has the advantages that problems can be identified as early as possible and can then be corrected easier and cost-effectively. FairsNet has adopted evaluation techniques that are comprehensive and cost- effective. Specifically, the evaluation process is based on the use of the following techniques:

- user and task observation;

- scenarios;

- simplified thinking aloud;

- heuristic evaluation.

User and task observation

The first step for designing a usable system is to know who will use it. There are different users; so it is important to analyse them. We can use different query techniques: direct and indirect observation, interviews, questionnaires, which have been described in the previous section.

Scenarios

Scenarios is a "story about use". Stories can be of different lengths and different levels of detail, and, indeed, the word "scenario" is used in many different ways in the literature on user and task analysis.

Scenarios can be about users, their work, their environments, how they do tasks, the tasks they need to do, and all combinations of these elements. Scenarios can focus on the primary users- the people who will actually use what you develop- or the secondary users- the people who benefit by what the primary users do. We distinguish four types of scenarios that vary in their level of detail and the use the development team might make of them. We include:

a. Brief scenarios are very brief stories that give just the facts of a real situation the primary user had to deal with, but that don't go into detail on how the user does the task.

b. Vignettes Brief narratives, sometimes with figures that give readers a high-level, broad brush view of a user, the user's environment, and the user's current way of doing something.

c. Elaborated scenarios are narratives with more details. Which details you focus on depend on what you want the team to take from the story.

d. Complete task scenarios are narratives that carry the story from the beginning to the end of a task or sequence of tasks.

Simplified thinking aloud

The thinking aloud method involves having one test user at a time use the system (or a prototype) for a given set of tasks while being asked to "think out loud". By verbalizing their thoughts, users allow an observer to determine not just what they are doing with the interface, but also why they are doing it. Traditionally, thinking aloud studies are conducted with psychologists or user interface experts as experimenters who videotape the subjects and perform detailed protocol analysis. A major difference between simplified and traditional thinking aloud is that data analysis can be done on the basis of the notes taken by the experimenter instead of by videotapes. Recording, watching, and analysing the videotapes is expensive and takes a lot of time.

5.7 Evaluation in FairsNet life cycle

Evaluation is not just something that happens in the delivery phase. As described above, it occurs in some form at all stages in the software cycle: analysis, design, development and delivery. A specific activity we have performed in FairsNet was related to designing and implementing a monitoring system that, through the evaluation techniques described in Section 5.5 and applied during the whole life-cycle, could provide information about the quality of the developed system. The main focus has been placed upon the evaluation at an operational level, namely on issue concerning the quality of the modules developed with the FairsNet, their usability and ease of use, the drawbacks and any problems, as well as the strong points and benefits of each module.

FairsNet system has been design according to user-centered design, as we have described above, in order to develop a system that was effective, efficient, and used by the end user with satisfaction. So, in each phase of the software process the end user has been involved in various ways and for different purposes. In analysis phase, questionnaires, interviews, user observations, study of existing literature and documents have been used to collect information in a systematic way. This activity has been extensively documented in [79].

During the initial phases of the project, prototypes of various types and complexity have been evaluated in order to explore different design alternatives. The goal was to individuate the best interface to be further developed. Various techniques have been used: scenarios, heuristic evaluation, interviews, and simplified thinking aloud.

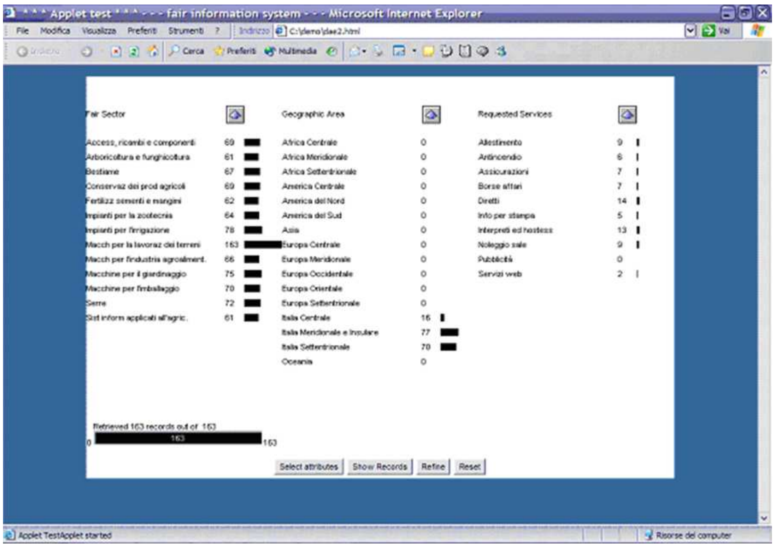
Some meetings have been organized among the FairsNet partners in order to evaluate prototypes and solve problems that came out in developing the FairsNet system. In particular, the members of the University of Bari, FSGV, and University of Rome LUISS met on 3-4 February 2003 in Rome to examine some FairsNet prototypes together with user.

As we have described, DAE offers various tools for analysing data. During the FairsNet project we have tested the tools with data related to different fairs. We have also evaluated these tools with users and we report here the results.

One of the first data set provided by the organizer of Agrilevante, a fair on agriculture organised at Fiera del Levante in Bari, Italy, every September. We then showed to the users the possibility they have to analyse these data through the DAE tools. Users were two people of the management of Agrilevante, involved in the fair organization. As an example of use of the tool DAEQP, we presented to the users the following scenario: the organiser of Agrilevante wants to perform a segmentation of the exhibitors of the last edition of the fair. Let us suppose that the organiser wants to find out which were the most requested services at the fair, which exhibitors requested them, and so on. The objective of the organizer is to increase the fair income by selling more services. Therefore s/he is interested in selecting exhibitor segments for starting appropriate marketing campaign promoting the fair services. In order to help the user (in this case the organiser) in his/her analysis, DAEQP is able to visualize an overview in which data are visualized along some major attributes. After accessing the system, the user is first asked to select three major attributes among those attributes considered in the database and shown on the screen. The user selected Fair sector, Geographic area, and Requested services; the resulting overview is shown in Figure 5.4.

The users immediately were pleased to see that the module provides very simply a lot of information and the interaction is easy and understandable after a few minutes of work.

The other tool we tested in that trial was DaeTL, described in Section 4.8. It is also useful for identifying important correlations among data. During a meeting with Agrilevante organisers, after showing for a few minutes the use of the DAE modules, we let them interact with the modules (prototypes at that time), and observed them working with the system asking them to think aloud, according to one of the methods illustrated in Section 5.5. This user observation activity provided useful indications for improving that version of the prototypes. Beside, we were very pleased to hear one of the fair organisers involved in that trial to say: “I wish I would have tools like these for directly analyse


* * * Applet test * * * - - - fair information system - - - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indirizzo C:\demo\dae2.html
Vai
Indietro Cerca Preferiti Multimedia
Select attributes
Show Records
Refine
Reset
Applet TestApplet started
Risorse del computer

Fair Sector		Geographic Area		Requested Services	
Access, ricambi e componenti	69	Africa Centrale	0	Allestimento	9
Arboricoltura e funghicoltura	61	Africa Meridionale	0	Antincendio	6
Bestiame	67	Africa Settentrionale	0	Assicurazioni	7
Conservaz dei prod agricoli	69	America Centrale	0	Borse affari	7
Fertilizz sementi e mangimi	62	America del Nord	0	Diretti	14
Impianti per la zootecnia	64	America del Sud	0	Info per stampa	5
Impianti per l'irrigazione	78	Asia	0	Interpreti ed hostess	13
Macch per la lavoraz dei terreni	163	Europa Centrale	0	Noleggio sale	9
Macch per l'industria agroaliment.	66	Europa Meridionale	0	Pubblicità	0
Macchine per il giardinaggio	75	Europa Occidentale	0	Servizi web	2
Macchine per l'imballaggio	70	Europa Orientale	0		
Serre	72	Europa Settentrionale	0		
Sist inform applicati all'agric.	61	Italia Centrale	16		
		Italia Meridionale e Insulare	77		
		Italia Settentrionale	70		
		Oceania	0		

Retrieved 163 records out of 163

0 163 163

Figure 5.4: Agrilevante trade fair: an example of query preview along Fair sector, Geographic area, and Requested services attributes

my data". This and other comments we got emphasize the importance and utility of the tools provided by DAE, that give users the feeling of "putting their hands" on the data and allow them to explore and manipulate data as they wish.

Another data set was provided by the project partner FSVG. They were data related to the fair Semana Verde 2002. Another trial was then set up for the meeting that the FairsNet Consortium organised in Rome, at LUISS during 3-4 February 2003. We showed to the users the possibilities they have to analyse their data through the DAE tools. Users were two people of the management to FSVG, involved in fair organization. We used the same protocol of the Agrilevante trial. We first showed for a few minutes the use of two of the DAE modules, namely DaeQP and DaeTL, then we let users interact with the modules, and observed them using the thinking aloud evaluation method. As it usually occurs with user evaluation, very useful indications came from users for improving the prototypes, as well as the confirmation of the validity of such analysis tool in user daily activities.

Among the various DAE tools, ARVIS is the most original tool developed in this thesis. It has been described in Section 4.5. Differently from DaeQP and DaeTL, ARVIS is designed to be used primarily by a data miner rather than directly by a company manager.

The usability evaluation of ARVIS has been performed by heuristic evaluation at various stages of the development cycle, starting from a paper mock-up and also by user testing.

The initial mock-up interface was presented to the team of Data Mining experts participating in the FairsNet project carrying out a simulation of the some typical tasks. Encouraging results from the tests were provided and even suggestion on how to improve the interface.

Among the heuristic evaluations performed during ARVIS design and development, we describe here the evaluation of a running prototype, that included almost all the planned functionalities. Three inspectors with a good data mining knowledge performed heuristic evaluation. Following the Nielsen's Usability Heuristic, they explored the system interface to discover usability problems, and collected their inspection results in a detailed report. At the end of individual inspections, the three inspectors met to discuss their findings and produced a final report. The detected usability problems have been classified in two main categories:

1. Category 1: presentation problems (e.g. widgets meaning not clear, tooltip absence)

2. Category 2: problems that can reduce the tool effectiveness (e.g. it is not clear how to perform a task using the system functionality)

Analysing the results of the heuristic evaluation, we can say that 40% of the problems are in Category 1 and 60% are in Category 2. An example of detected problems is shown in Figure 5.5. A first problem is that a same command is duplicated in the interface, thus violating a principle of minimalist design. Another problem is related to the interface consistency, since when checking a box such as "Support" in the tool bar, the corresponding box in the control panel at the bottom left of the Figure 5.5 is not checked, and vice versa.

In line with the results obtained from the heuristic evaluation, the prototype was modified/improved trying to solve the usability problems. Finally, a user test on the new prototype has been performed. Four selected users from the University of Bari have participated to the user test. Two of the users were PhD students with deep data mining knowledge. Other two users were data mining researchers. The test involved: the prototype; a case study, user tasks corresponding to the case study; a questionnaire to assess users satisfaction and easy of use of the tool. Each users was required to run and interact with the prototype by executing the list of defined tasks while the evaluator

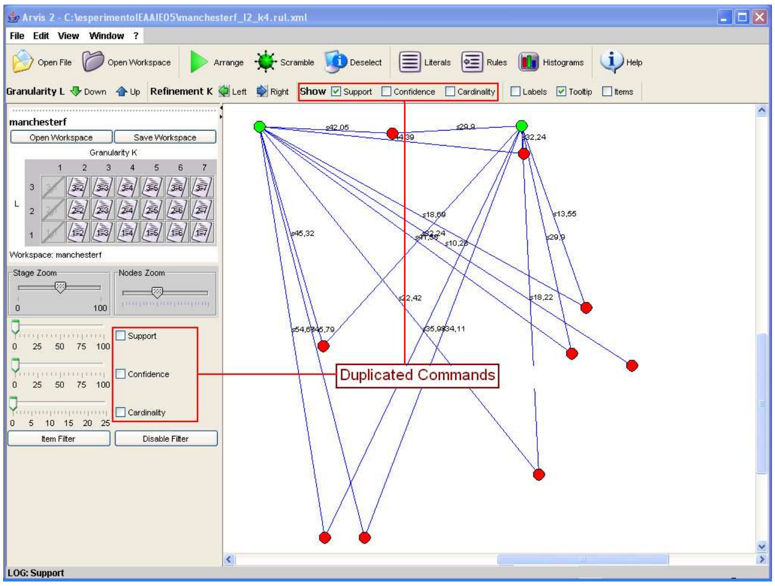


Figure 5.5: ARVis interface, the red box indicate the usability problems detected

performed thinking aloud. The evaluator annotated her comment on pre-formatted sheets, that were successively carefully analysed.

At the end of the interactions session, each user was asked to fill the questionnaire.

The first part of the questionnaire had closed questions pertaining to the simplicity or complexity of carrying out user tasks. From the questionnaires analysis it can be deduced that at least three of the users found each of the supported tasks fairly easy to carry out. Three users found it fairly easy to understand and interact with the graph visualization and to perform a mining task in general.

The second part of the questionnaire also had closed questions. The part aimed at assessing interface designing aspects. At least three of the users found each of a tested design aspects reasonably well-adhered to. All the users observed that a consistency was very well-applied in the interface design and found the interface perceptiveness fairly well-adhered to. Three users noted that it was fairly easy to remember or acquire the aspects relevant to a particular mining task. Moreover, the same percentage observed that the system responded to user operations in a reasonably valuable way. Three of the users found the interface elements fairly well-organized.

The last part of the questionnaire had open questions pertaining to strengths, weaknesses, and capability of the system/interface. It is interesting to realize that the subjects highlighted the same features as the most-liked. Three users mentioned consistency as one of the main features they liked most about the interface. Two users mentioned good layout/organization. As for the supported system functionalities, three users were very satisfied.

Conclusions

In this thesis it has been presented the results of about five years of work done in the fields of information visualization, databases, web-based systems, in particular in the trade fair domain, have been presented. The framework was conceived in that domain then the various tools composing it have been refined and used in other domains, because ideas in the tools are general enough.

DAE is modular and every module can work stand-alone. The rationale of a unique framework is to give the user the possibility to work with a unique system so that on the same data set the user may perform different analyses using different tools.

In the future work it is planned to extend the use of these tools to multiple devices. This is a challenging task, since the problems with small devices are well known.

The review of the literature, of some commercial systems and the work presented here suggest that there are many fields in which the framework can be applied and further developed. The field of visual data mining is currently a lively field and users will benefit greatly from systems that combine both improved data mining algorithms and empowering interactive interfaces.

This work illustrates how traditional features such as overview and details, combined with an interactive search interface, can help users to perform exploratory analysis on time series. This is another area that may have promising developments.

Our interaction with users have shown that the work presented here is useful for several types of users, from the data mining expert that can benefit from tools that may relieve some of his activities to users who are not computer science experts, like a company manager, that will have support in their decision making activity.

Bibliography

[1] H. A. Abbass. Introducing data mining. In H. A. Abbass, editor, *Short Lectures in Data Mining and Heuristics*, pages 1–9. UNSW, ADFA, Canberra, 2000.

[2] H. A. Abbass, M. Towsey, and G. Finn. C-net: A method for generating nondeterministic and dynamic multivariate decision trees. *Knowledge and Information Systems: An International Journal*, 5(2), 2001.

[3] P. S. Adler and T. A. Winograd, editors. *Usability: Turning technologies into tools.* New York: Oxford University Press, 1992.

[4] Advanced Visual Systems, "AVS/Express". http://www.avs.com, accessed jan 2005.

[5] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, page 94–105, 1998.

[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press/The MIT Press, 1996.

[7] C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *CHI '94: Proceedings of the SIGCHI Conference on Human factors in computing systems*, pages 313–321, 1994.

[8] N. Andrienko, G. Andrienko, and P. Gatalsky. Exploratory spatio-temporal visualization: An analytical review. *Journal of Visual Languages and Computing*, 14(6):503–541, 2003.

[9] M. Ankerst. *Visual Data Mining*. PhD dissertation, Ludwig-Maximilians-Universität München, Fakultät für Mathematik und Informatik, 2000.

[10] M. Ankerst. Visual data mining with pixel-oriented visualization techniques. In *Proceedings of the KDD Workshop on Visual Data Mining*, 2001.

[11] M. Ankerst, M. Breunig, H. P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 49–60. ACM Press, 1999.

[12] A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba. Discovery of spatial association rules in georeferenced census data: A relational mining approach, intelligent data analysis. *Intelligent Data Analysis*, 7(6):541–566, 2003.

[13] Artificial Intelligence Software, “VisualMine”. http://www.visualmine.com/ , 2001.

[14] A. M. Bagirov, A. M. Rubinov, and J. Yearwood. *Heuristic and Optimization for Knowledge Discovery*, chapter 2. Idea Group Publishing, USA, 2001.

[15] S. Balkin and J. Ord. Automatic neural network modeling for univariate times series. *International Journal of Forecasting*, 16:509–515, 2000.

[16] F. M. Barbini, P. Buono, M. F. Costabile, A. D’Atri, E. Pauselli, S. Swift, and Y. Ursa. Requirement analysis for on-line trade fairs. In *International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations*, pages 142–146, Rome, 14-15 September 2001.

[17] B. B. Bederson, J. Meyer, and L. Good. Jazz: An extensible zoomable user interface graphics toolkit in java. In *UIST - ACM Symposium on User Interface Software and Technology*, volume 2, pages 171–180. CHI Letters, 2000.

[18] B. B. Bederson and B. Shneiderman. *The Craft of Information Visualization: Readings and Reflections.* Morgan Kaufmann Publisher, San Francisco, California, 2003.

[19] B. B. Bederson, B. Shneiderman, and M. Wattemberg. Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *ACM Transactions on Graphics*, 21(4):833–854, October 2002.

[20] Belmont Research Inc, “Cross Graphs”. http://www.belmont.com/cg.html , 2001.

[21] J. Bertin. *Semiology of Graphics.* University of Wisconsin Press, Madison, WI, 1983.

[22] J. Blanchard, F. Guillet, and H. Briand. Exploratory visualization for association rule rummaging. In *Proceedings of the fourth International Workshop on Multimedia Data Mining MDM/KDD2003*, pages 107–114, 2003.

[23] D. L. Boley. Principal direction divisive partitioning. *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

[24] B. Borner and C. Chen editors. In *Visual Interfaces to Digital Libraries: Motivation, Utilization, and Socio-Technical Challenges*, volume 2539 of *Lecture Notes in Computer Science.* Springer-Verlag, London, 2003.

[25] P. S. Bradley, U. M. Fayyad, and O. L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.

[26] D. Brodbeck and L. Girardin. Trend analysis in large time series of high-throughput screening data using a distortion-oriented lens with semantic zooming (poster). In *InfoVis 2003: Adjunct Proceedings of the IEEE Symposium on Information Visualization*, pages 74–76. IEEE Press, 2003.

[27] C. M. Brown. *Human-Computer interface design guidelines.* Ablex, Norwood, NJ, 1988.

[28] D. Bruzzese and P. Buono. Combining visual techniques for association rules exploration. In M. F. Costabile, editor, *Proceedings of the Working Conference on Advanced Visual Interfaces AVI 2004*, pages 381–384. ACM Press, 2004.

[29] D. Bruzzese and C. Davino. Statistical pruning of discovered association rules. *Computational Statistics*, 16:387–389, 2001.

[30] D. Bruzzese and C. Davino. Visual post analysis of association rules. *Journal of Visual Languages and Computing*, 14(6):621–635, 2003.

[31] D. Bruzzese and C. Davino. Visualizing association rules. In S. J. Simoff, M. Noirhomme, and M. Boehlen, editors, *Visual Data Mining: Theory and Applications*, volume 1 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2004.

[32] P. Buono. Analysing association rules with an interactive graph-based technique. In *HCI International 2003*, pages 675–679, Crete, Greece, 22-27 June 2003.

[33] P. Buono. Visual data analysis: the case of trade fairs. In *AICA 2004*, pages 733–743, Benevento, Italy, 28-30 September 2004.

[34] P. Buono, M. F. Costabile, A. D'Atri, M. Hemmje, G. Jaeschke, C. Muscogiuri, E. Pauselli, and F. Barbini. FAIRWIS: A system for improving on-line trade fair services. In *E-work and E-commerce E2001*, pages 519–525, Venice, October 17-19 2001. IOS Press.

[35] P. Buono, M. F. Costabile, D. Grilli, S. P. Guida, P. Lops, and G. Semeraro. Integrating machine learning and filtering techniques to improve recommendations. In *CHI '03 Workshop on Designing Personalized User Experiences for eCommerce: Theory, Methods, and Research*, Fort Lauderdale, USA, 6-7 April 2003.

[36] P. Buono, M. F. Costabile, S. P. Guida, R. Lanzilotti, and A. Piccinno. Improving web interaction trough personalization. In *HCI International 2003*, pages 522–526, Crete, Greece, 22-27 June 2003.

[37] P. Buono, M. F. Costabile, S. P. Guida, and A. Piccinno. Integrating user data and collaborative filtering in a web recommendation system. In S. Reich, M. M. Tzagarakis, and P. M. E. D. Bra, editors, *Hypermedia: Openness, Structural Awareness, and Adaptivity*, volume 2266 of *Lecture Notes in Computer Science*. Springer, 2002.

[38] P. Buono, M. F. Costabile, S. P. Guida, A. Piccinno, and G. Tesoro. Integrating user data and collaborative filtering in

a web recommendation system. In *Third Workshop on Adaptive Hyper-text and Hypermedia*, pages 129–140, Sonthofen, Germany, July 2001.

[39] P. Buono, M. F. Costabile, M. Hemmje, G. Jaeschke, and C. Muscogiuri. Providing on-line trade fair services with fairwis. In *International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations*, pages 147–152, Rome, 14-15 September 2001.

[40] P. Buono, M. F. Costabile, A. Piccinno, and G. Minardi. Upe: The fairwis personalisation component. In *International Workshop on Open Enterprise Solutions: Systems, Experiences, and Organizations*, pages 153–154, Rome, 14-15 September 2001.

[41] P. Buono, M. F. Costabile, A. Piccinno, and T. Roselli. Web recommendation systems: The case of on-line trade fairs. In *PC-HCI 2001*, pages 405–406, Patras, Greece, December 7-9 2001.

[42] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2), 1998.

[43] C. Ahlberg, IVEE Development AB, “Spotfire”. http://www.ivee.com, 2001.

[44] S. K. Card. Information visualisation. In J. Jacko and A. Sears, editors, *The Human-Computer Interaction Handbook*, pages 544–582. Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

[45] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization.* Morgan Kaufmann Publisher, San Francisco, California, 1999.

[46] S. K. Card and D. Nation. Degree-of-interest trees: a component of attention-reactive user interface. In *Proceedings of Advanced Visual Interface '02.* Trento, Italy, 2002.

[47] S. K. Card, G. G. Robertson, and W. York. The webbook and the webforager: An information workspace for the world-wide web. In *CHI '96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 111–117. ACM New York, NY, 1996.

[48] J. V. Carlis and J. A. Konstan. Interactive visualization of serial periodic data. In *ACM Symposium on User Interface Software and Technology*, pages 29–38. ACM Press, New York, 1998.

[49] T. Catarci, M. F. Costabile, S. Levialdi, and C. Batini. Visual query systems for databases: a survey. *Journal of Visual Languages and Computing*, 8:215–260, 1997.

[50] K. Chakrabarti, E. Keogh, S. Mehrotra, and M. Pazzani. Locally adaptive dimensionality reduction for indexing large time series databases. *ACM Trans. Database Syst.*, 27(2):188–228, 2002.

[51] N. Chandler. Are you winning the performance management race? *DM Direct Newsletter Archives*, Jan 18, 2002.

[52] P. Cheeseman and J. Stutz. Bayesian classification (autoclass): Theory and results. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining.* AAAI Press/MIT Press, 1996.

[53] C. Chen. *Information Visualisation and Virtual Environments.* Springer-Verlag, London, 1999.

[54] C. Chen. *Mapping Scientific Frontiers: The Quest for Knowledge Visualisation.* Springer-Verlag, Berlin, 2003.

[55] C. Chen and R. J. Paul. Visualizing a knowledge domain's intellectual structure. *IEEE Computer*, 34(3):65–71, March 2001.

[56] V. Cherkassky and F. Mulier. *Learning from Data – Concepts, Theory and Methods.* John Wiley and Sons, New York, 1998.

[57] L. Chittaro, C. Combi, and G. Trapasso. Data mining on temporal data: a visual approach and its clinical application to hemodialysis. *Journal of Visual Languages and Computing*, 14:591–620, 2003.

[58] A. Chortaras. *Efficient Storage Retrieval and Indexing of Time Series Data.* Msc thesis, Imperial College, London, 2002.

[59] M. Clemens. http://www.idiagram.com/kv_venn.html.

[60] W. Cleveland. *Visualizing Data.* Hobart Press, Summit, NJ, 1993.

[61] A. Cliff and P. Haggett. *Atlas of Disease Distributions.* Blackwell, Oxford, 1988.

[62] Comshare Commander Decision. http://www.performance.geac.com/, accessed jan 13, 2005.

[63] M. F. Costabile. Usability in the software life cycle. In S. Chang, editor, *Handbook of Software Engineering and Knowledge Engineering*, volume 1, pages 179–192. World Scientific, December 2001.

[64] M. F. Costabile and D. Malerba. Special issue on visual data mining, editor's foreword. *Journal of Visual Languages & Computing*, 14(6):499–501, 2003.

[65] Cyber-Geography Research. http://www.cybergeography.org/atlas/.

[66] Cygron Research & Development Ltd. , "DataScope". http://www.cygron.com/ , 2001.

[67] Data Description, "Data Desk". http://www.datadesk.com/datadesk/ , 2001.

[68] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge mining with VxInsight: Discovery through interaction. *Journal on Intelligent Information Systems*, 11(3):259–285, March 1998.

[69] T. D'Hers and S. Vickery. Comparing business intelligence platforms - sql server 2000 analysis services compared with ibm db2 olap server 8.1 and hyperion essbase 6.5.

[70] G. A. Dix, J. Finlay and R. Beale. *Human Computer Interaction.* Prentice Hall, 1998.

[71] K. Doan, C. Plaisant, B. Shneiderman, and T. Burns. Interface and data architecture for query preview in networked information systems. *ACM Transactions on Information Systems*, 17(3):320–341, 1999.

[72] S. Džeroski and N. Lavrač, editors. *Relational Data Mining.* Springer, 2001.

[73] C. Eaton, C. Plaisant, and T. Drizd. The challenge of missing and uncertain data. In *Vis 2003: Adjunct Proceedings of the IEEE Visualization Conference*, pages 40–42. IEEE Press, 2003.

[74] S. G. Eick, J. L. Steffen, and E. E. J. Sumner. Seesoft: A tool for visualizing line-oriented software statistics. *IEEE Transactions on Software Engineering*, 18(11):957–968, 1992.

[75] Embedded Component History Object. http://www.echohistorian.com, accessed october 25, 2004.

[76] L. Ertöz, M. Steinbach, and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In D. Barbara and C. Kamath, editors, *Proceedings of the SIAM International Conference on Data Mining*, 2003.

[77] M. Ester, H. P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.

[78] B. S. Everitt. *Cluster Analysis, 3rd ed.* Edward Arnold, 1993.

[79] FAIRWIS Consortium. User requirements report. Preliminary specifications for the production of FAIRWIS. General description and needs for valuable validation sites. Technical Report IST-1999-12641, D1 2002.06.30, 2000.

[80] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 1– 36. AAI/MIT press, 1996.

[81] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* AAAI Press/The MIT Press, 1996.

[82] U. Fayyad, Piatetsky-Shaprio, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* MIT Press, Cambridge, MA, 1996.

[83] Forest & Trees. http://www.ca.com, accessed jan 13, 2005.

[84] C. Fraley and A. Raftery. Mclust: Software for model-based cluster and discriminant analysis, 1999.

[85] G. W. Furnas. Generalized fisheye views. In *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 16–23. ACM Press, 1986.

[86] M. Ganesh, E. Han, V. Kumar, S. Shekhar, and J. Srivastava. Visual data mining: Framework and algorithm development, TR-96-021, Department of Computer Science, University of Minnesota, Minneapolis, 1996.

[87] V. Ganti, J. Gehrke, and R. Ramakrishnan. Cactus: Clustering categorical data using summaries. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 73–83, 1999.

[88] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[89] G.L.Andrienko, "Descartes". http://allanon.gmd.de/and/and.html , 2001.

[90] S. Goil, H. Nagesh, and A. Choudhary. Mafia: Efficient and scalable subspace clustering for very large data sets, 1999.

[91] M. Gross. *Visual computing: The integration of computer graphics, visual perception and imaging.* Springer-Verlag, Heidelberg, 1994.

[92] S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, page 73–84, 1998.

[93] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proceedings of the International Conference on Data Engineering (ICDE)*, page 512–521, 1999.

[94] J. Hackos and J. C. Redish. *User and Task Analysis for Interface Design.* John Wiley & Sons, 1998.

[95] J. Han. OLAP Mining: Integration of OLAP with Data Mining. In S. Spaccapietra and F. Maryanski, editors, *Data Mining and Reverse Engineering: Searching for Semantics*, volume 124 of *IFIP Conference Proceedings.* Chapman & Hall, 1998.

[96] J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, and O. R. Zaiane. DBMiner: A system for mining knowledge in large relational databases. In A. Press, editor, *Proc. 1996 Int'l Conf. on Data Mining and Knowledge Discovery (KDD'96)*, pages 250–255, Portland, Oregon, 1996.

[97] M. C. Hao, U. Dayal, M. Hsu, T. Sprenger, and M. H. Gross. Visualization of directed associations in e-commerce transaction data. Hewlett Packard Research Laboratories, 2001.

[98] B. S. Harry Hochheiser. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[99] J. A. Hartigan. *Clustering Algorithms.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1975.

[100] J. A. Hartigan and M. Wong. Algorithm as136: A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[101] H. R. Hartson and D. Hix. *Developing User Interfaces.* John Wiley, New York, 1993.

[102] F. Hillier and G. Lieberman. *Introduction to Operations Research.* McGrawHill, Boston, 2001.

[103] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 58–65, 1998.

[104] H. Hochheiser. *Interactive Graphical Querying of Time Series and Linear Sequence Data Sets.* Ph.D. dissertation, University of Maryland Dept. of Computer Science, 2003.

[105] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization Journal*, 3(1):1–18, 2004.

[106] H. Hofmann, A. P. J. M. Siebes, and A. F. X. Wilhelm. Visualizing association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 227–235. ACM Press, 2000.

[107] http://java.sun.com/webservices/.

[108] S. Hui and G. Jha. Data mining for customer service support. *Information & Management*, 38:1–13, 2000.

[109] IBM, "Open Visualization Data Explorer (DX)". http://www.research.ibm.com/dx/ , 2001.

[110] ILOG - JViews. http://www.ilog.com, accessed october 25, 2004.

[111] W. H. Inmon. The data warehouse and data mining. *Communications of the ACM*, 39(11):49–50, 1996.

[112] A. Inselberg. N-dimensional graphics, part i - lines and hyperplanes, 1981.

[113] Intelligent Miner. http://www-306.ibm.com/software/data/iminer/fordata/, accessed jan 2005.

[114] J. B. J. Whiteside and K. Holtzblatt. Usability enginnering our experience and evolution. In M. Helander, editor, *Handbook of Human-Computer Interaction*, pages 791–817. Elsevier Science, 1988.

[115] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data.* Prentice-Hall Advanced Reference Series. Prentice-Hall, 1988.

[116] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[117] R. Jeffries and H. Desurvire. Usability testing vs. heuristic evaluation: Was there a context? *ACM SIGCHI Bulletin*, 24(4):39–41, 1992.

[118] D. F. Jerding and J. T. Stasko. The information mural: a technique for displaying and navigating large information spaces. In *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization*, page 43. IEEE Computer Society, 1995.

[119] G. Karypis, E. H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.

[120] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data - Introduction to Cluster Analysis.* John Wiley & Sons, 1990.

[121] D. Keim and M. Ankerst. Visual data mining and exploration of large databases, a Tutorial. In *Proceedings of the International Workshop on Visual Data Mining, in conjunction with ECML/PKDD2001.* 4 September 2001. Freiburg, Germany.

[122] D. A. Keim. Pixel-oriented visualization techniques for exploring very large databases. *Journal of Computational and Graphical Statistics*, March:58–77, 1999.

[123] P. R. Keller and M. M. Keller. *Visual Cues: Practical Data Visualization.* IEEE Computer Society Press, 1994.

[124] S. Kelly. *Data Warehousing in Action.* John Wiley & Sons, 1997.

[125] H.-G. Kemper and P.-L. Lee. The customer-centric data warehouse–an architectural approach to meet the challenges of customer orientation. In *HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 8*, page 231.3. IEEE Computer Society, 2003.

[126] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 102–111. ACM Press, 2002.

[127] S. Kimani. *Visual Information Discovery.* Ph.D. dissertation, Università degli Studi di Roma "La Sapienza", 2003.

[128] S. Kimani, S. Lodi, T. Catarci, G. Santucci, and C. Sartori. Vidamine: a visual data mining environment. *Journal Visual Languages Computing*, 15(1):37–67, 2004.

[129] I. Kopanakis and B. Theodoulidis. Visual data mining modeling techniques for the visualization of mining outcomes. *Journal of Visual Languages and Computing*, 14(6), 2003.

[130] J. Lamping, R. Rao, and P. Pirolli. A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408. ACM, New York, NY, 1995.

[131] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom. Visually mining and monitoring massive time series. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 460–469. ACM Press, 2004.

[132] F. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. *Machine Learning*, 55:175–210, 2004.

[133] M. Ludl and G. Widmer. Relative unsupervised discretization for association rule mining. In D. Zighed, H. Komorowski, and J. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *LNAI*, pages 148–158. Springer-Verlag, 2000.

[134] J. D. Mackinlay, G. G. Robertson, and S. K. Card. The perspective wall: detail and context smoothly integrated. In *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 173–176. ACM Press, 1991.

[135] O. L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 1(2):183–201, 1997.

[136] M. May. Spatial knowledge discovery: The SPIN! system. In K. Fullerton, editor, *Proceedings of the EC-GIS Workshop*, 2000.

[137] B. H. McCormick, T. A. DeFanti, , and M. D. Brown. Visualization in scientific computing. *Computer Graphics*, 6(21), 1987.

[138] R. S. Michalski, I. Bratko, and M. Kubat, editors. *Machine learning and data mining.* John Wiley and Sons, Chichester, England, 1999.

[139] MicroStrategy, Inc. *The 5 Styles of Business Intelligence: Industrial-strength business intelligence, white paper.* http://www.microstrategy.com, accessed Jan 2005.

[140] Miron Livny, Raghu Ramakrishnan e Kent Wenger, "DEVise". http://www.cs.wisc.edu/ devise/, 2001.

[141] H. Morris. Analytic applications: Beyond business intelligence. *DM Review Magazine*, 4, 2002.

[142] B. Moxon. Defining data mining, 1996.

[143] C. Muscogiuri, G. Jäschke, A. Paradiso, and M. Hemmje. FAIR-WIS: An integrated system offering trade fair web-based information services - a r&d case study. In R. Sprague, editor, *Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35)*, page 307. Los Alimatos: IEEE Press, 2002.

[144] H. Nagesh, S. Goil, and A. Choudhary. Adaptive grids for clustering massive data sets. In *Proceedings of the SIAM International Conference on Data Mining*, 2001.

[145] H. A. D. D. Nascimiento. Interactive graph clustering based on user hints. In *Pan-Sydney area Workshop On Visual Information Processing*, 2000.

[146] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, page 144–155, 1994.

[147] C. Niederée, C. Muscogiuri, and M. Hemmje. Taxonomies in operation, design, and meta-design. In *WISEW '02: Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops) - (WISEw'02)*, page 140. IEEE Computer Society, 2002.

[148] J. Nielsen. *Usability Engineering.* Academic Press, Cambridge, 1993.

[149] J. Nielsen. Guerrilla hci: Using discount usability engineering to penetrate intimidation barrier. In R. Bias and D. Mayhew, editors, *Cost-Justifying Usability.* Academic Press, 1994.

[150] J. Nielsen and R. Mack. *Usability Inspection Methods.* John Wiley and Sons, New York, 1994.

[151] G. M. Nielson, H. Hagen, and H. Muller. Scientific visualization : overviews, methodologies, and techniques. *IEEE Computer Society*, 1997.

[152] O. Niggemann. *Visual Data Mining of Graph-Based Data.* PhD dissertation, University of Paderborn, Germany, Department of Mathematics and Computer Science, 2001.

[153] K.-H. Ong, K.-L. Ong, W. K. Ng, and E.-P. Lim. Crystal clear: Active visualization of association rules. Nanyang Technological University.

[154] F. Paterno'. Task models for interactive software systems. In S. Chang, editor, *Handbook of Software Engineering and Knowledge Engineering*, volume 1, pages 817–836. World Scientific, December 2001.

[155] G. Paul. Kant, Immanuel. In E. Craig, editor, *Routledge Encyclopedia of Philosophy*, http://www.rep.routledge.com/article/DB047SECT4, (1998, 2004). London: Routledge.

[156] N. Pendse. FASMI, http://www.olapreport.com/fasmi.htm.

[157] Personal StockMonitor. http://www.personalstockmonitor.com, accessed october 25, 2004.

[158] Pilot Software. http://www.pilotsoftware.com/, accessed jan 13, 2005.

[159] C. Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM Press, 2004.

[160] C. Plaisant, J. Grosjean, and B. B. Bederson. Spacetree: Supporting exploration in large node link tree. In *Design Evolution and Empirical Evaluation IEEE Symposium on Information Visualization*, pages 57–64, 2002.

[161] J. Preece. *Human-Computer Interaction.* Addison Wesley, 1994.

[162] C. T. Ragsdale. *Spreadsheet Modeling and Decision Analysis.* South-Western College Publishing, USA, 2001.

[163] C. P. Rainsford and J. F. Roddick. Visualisation of temporal interval association rules. In *2nd International Conference on Intelligent Data Engineering and Automated Learning, (IDEAL 2000)*, volume 1983, pages 91–96, Shatin, N.T., Hong Kong, 2000. Springer.

[164] R. Rao and S. Card. The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of Conference Human Factors in Computing Systems*, 1994.

[165] R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masinter, P. Halvorsen, and G. G. Robertson. Rich

interaction in the digital library. *Communications of the ACM*, 38(4):29–39, 1995.

[166] M. Rasmussen. gcluto – an interactive clustering, visualization, and analysis system. Technical Report TR# 04–020, Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN, 2004.

[167] B. D. Ripley. Neural networks and related methods for classification. *Journal of the Royal Statistics Society*, 53(3):409–437, 1994.

[168] G. Robertson, S. Card, and J. Mackinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4):57–71, April 1993.

[169] M. Sarkar and M. H. Brown. Graphical fisheye views. *Commun. ACM*, 37(12):73–83, 1994.

[170] R. Sarker, H. Abbass, and C. Newton. Solving two multi-objective optimization problems using evolutionary algorithms. In M. Mohammadian, R. Sarker, and X. Yao, editors, *Intelligent Systems in Control*, pages 218–232. Idea Group Publishing, USA, 2002.

[171] SAS Institute. http://www.sas.com, accessed october 25, 2004.

[172] W. R. Saul. *Information Anxiety.* Doubleday, New York, 1989.

[173] E. Schikuta and M. Erhart. The bang-clustering system: Grid-based data analysis. In *Advances in Intelligent Data Analysis*, volume 1280 of *Lecture Notes in Computer Science*, pages 513–526. Springer, 1997.

[174] D. W. Scott. *Mutivariate Density Estimation.* John Wiley and Sons, New York, 1992.

[175] G. Sheikholeslami, S. Chatterjee, and A. Zhang. A multi-resolution clustering approach for very large spatial databases. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, pages 428–439, 1998.

[176] T. Shin and I. Han. Optimal single multi-resolution by genetic algorithms to support artificial neural networks for exchange-rate forecasting. *Expert Systems with Applications*, 18:257–269, 2000.

[177] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, 1994.

[178] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. Technical Report UMCP-CSD CS-TR-3665, College Park, Maryland 20742, U.S.A., 1996.

[179] B. Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[180] B. Shneiderman. Dynamic queries: For visual information seeking. *IEEE Software*, 11(6):70–77, November 1994.

[181] B. Shneiderman and C. Plaisant. *Designing user interfaces.* Addison Wesley, Washington, D.C., 2004.

[182] B. Shneiderman, C. Plaisant, K. Doan, and T. Bruns. Interface and Data Architecture for Query Preview in Networked Information Systems. *ACM Transaction on Information System*, 17:320–341, 1999.

[183] B. Shneiderman, C. Williamson, and C. Ahlberg. Dynamic queries: database searching by direct manipulation. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 669–670, 1992.

[184] S. F. Silva and T. Catarci. Visualization of linear time-oriented data: A survey. In *WISE '00: Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1*, page 310. IEEE Computer Society, 2000.

[185] S. J. Simoff. Towards the development of environments for designing visualisation support for visual data mining. In *Proceedings of the International Workshop on Visual Data Mining, in conjunction with ECML/PKDD2001.* September 2001. Freiburg, Germany.

[186] P. Smith, D. Pregibon, and C. Faloutsos. Data-driven evolution of data mining algorithms. *Communication of the ACM*, 45(8):33–37, 2002.

[187] R. Spence. *Information Visualization.* Addison-Wesley Publ. Co., Reading, MA, 2001.

[188] Spotfire. http://www.spotfire.com, accessed jan 2005.

[189] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *Proceedings of the ACM SIGKDD Workshop on Text Mining*, 2000.

[190] E. Tanin, A. Lotem, I. Haddadin, B. Shneiderman, C. Plaisant, and L. Slaughter. Facilitating network data exploration with query previews: A study of user performance and preference. *Behaviour & Information Technology*, 19(6):393–403, 2000.

[191] S. T. Teoh and K.-L. Ma. Starclass: Interactive visual classification using star coordinates. In *Proceedings of SIAM International Conference on Data Mining.* San Francisco, 2003.

[192] E. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT, 1983.

[193] A. Unwin, H. Hofmann, and K. Bernt. The twokey plot for multiple association rules control. AT&T Florham Park.

[194] A. Unwin and G. Wills. Exploring time series graphically. *Statistical Computing and Graphics Newsletter*, 2:13–15, 1999.

[195] J. van Wijkvan and E. van Selow. Cluster and calendar-based visualization of time series data. In G. Wills and D. Keim, editors, *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 4–9. IEEE Computer Society, October 25-26, 1999.

[196] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, Second Edition, 2000.

[197] Visual Analytics. http://www.visualanalytics.com, accessed jan 2005.

[198] Visual Insights, "ADVIZOR/2000". http://www.vdi.com, 2001.

[199] Visual Numerics, "JWAVE". http://www.vni.com/vnihome.html, 2001.

[200] C. S. Wallace and D. L. Dowe. Intrinsic classification by mml - the snob program. In C. Zhang, J. Debenham, and D. Lukose, editors, *Proceedings of the Australian Joint Conference on Artificial Intelligence*, pages 37–44. World scientific, 1994.

[201] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei. Content-based image indexing and searching using daubechies' wavelets. *International Journal of Digital Libraries (IJODL)*, 1(4):311–328, 1998.

[202] W. Wang, J. Yang, and R. R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, page 186–195, 1997.

[203] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1999.

[204] C. Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann Publisher, San Francisco, California, 2000.

[205] M. Wattenberg. Sketching a graph to query a time-series database. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 381–382. ACM Press, 2001.

[206] C. White. Intelligent business strategies: Understanding performance management. *DM Direct Newsletter Archives*, February 2003.

[207] O. G. Wilbert. *The essential guide to user interface design: an introduction to GUI design principles and techniques, second edition.* John Whiley & Sons, New York, 2002.

[208] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing).* Springer Verlag, 1999.

[209] P. C. Wong, P. Whitney, and J. Thomas. Visualizing association rules for text mining. Pacific Northwest National Laboratory, 1999.

[210] XForms. http://www.w3.org/tr/xforms11/, accessed jan 10, 2005.

[211] X. Xu, M. Ester, H. P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings of the International Conference on Data Engineering*, pages 324–331, 1998.

[212] X. Yao. Evolving artificial neural networks. In *Proceedings of the IEEE*, volume 87, pages 1423–1447. 1999.

[213] R. Zembowicz and J. M. Zytkov. From contingency tables to various forms of knowledge in databases. In U. Fayyad, G. Piatetsky-Shaprio, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 329–349. MIT Press, Cambridge, MA, 1996.

[214] G. Zhang and B. P. M. Hu. A simulation study of artificial neural networks for nonlinear time-series forecasting. *Computers and Operations Research*, 28:381–396, 2001.