A New Challenge for Information Mining

Roberto Paiano
Department of Engineering for Innovation, University of Salento, 73100 Lecce, Italy
roberto.paiano@unisalento.it

Stefania Pasanisi
Department of Engineering for Innovation, University of Salento, 73100 Lecce, Italy
stefania.pasanisi@unisalento.it

Abstract

In the field of "Data Exploration" many approaches have been developed to solve the problem of management of big data that are also semantically rich. Nowadays, there is a strong need to support the discovery-oriented applications where data discovery is a highly ad hoc interactive process to support the users by assisting the navigation in the data to find interesting objects. In this work starting by a theoretical data exploration system, where we identified the main features that a data exploration system must have to an efficient exploratory experience, we propose a combination of two data exploration techniques faceted navigation and data mining with the aim to improve the discovery information during exploration. This approach is contextualized better in Information Mining. Information mining, in fact, aims at discovering knowledge, i.e. more general patterns within objects or collections of objects.

Keywords: Data Exploration, Data Mining, Faceted Search, Rich Data Set, Information Mining

1. Introduction

The continued growth in data volume, velocity, variety, complexity and the increased importance of information for companies, needing a system of management of different knowledge from the past, forces us to adopt strategies and develop methods to explore and interpret data.
Today the world of technologies and services evolves according to four main drivers: Big Data, Mobile, Social and Cloud. You must govern the drivers of this change through advanced exploration technologies (Semantic Engine, Predictive Analytics, Social Listening, Sentiment Analysis, Data Mining, Exploratory Search, Exploratory Data Analysis, Faceted Search, etc.). Today's organizations need effective methods and tools to harness the wealth of data available to facilitate the availability, scope and knowledge sharing as well as for the chance to perform predictive analyzes useful for decision-making purposes. Organizations that invest in this will have a better chance of survival and, for this reason; the information itself will become a very important factor in production.

Big data and big data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies. A definition of big data is given below. Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies and techniques to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes (Mills et al., 2012; Sicular, 2013).

By itself, stored data does not generate business value, and this is true of traditional databases, data warehouses, and the new technologies such as Hadoop for storing big data. Once the data is appropriately stored, however, it can be analyzed, which can create tremendous value.

Data analytics refers to the BI technologies that are grounded mostly in data mining and statistical analysis. As mentioned previously, most of these techniques rely on the mature commercial technologies of relational DBMS, data warehousing, ETL, OLAP, and BPM. We can

distinguish in Text Analytics, Web Analytics, Network Analytics, Mobile Analytics (Chaudhuri *et al*., 2011).

There are three kinds of analytics (Chen et al., 2012):

- Descriptive analytics, such as reporting/OLAP, dashboards/scorecards, and data visualization, are backward looking (like a car's rear view mirror) and reveal what has occurred.
- Predictive analytics suggest what will occur in the future. The methods and algorithms for predictive analytics such as regression analysis, machine learning, and neural networks have existed for some time. Marketing is the target for many predictive analytics applications; here the goal is to better understand customers and their needs and preferences.
- Exploratory or discovery analytics (although these are just other names for predictive analytics): they normally refer to finding relationships in big data that were not previously known. The ability to analyze new data sources—that is, big data—creates additional opportunities for insights and is especially important for firms with massive amounts of customer data.

The managed information turned from analytics in the early period to qualitative in these last years. Qualitative research is a broad methodological approach that encompasses many research methods. The aim of qualitative research may vary with the disciplinary background, such as a psychologist seeking to gather an in-depth understanding of human behavior and the reasons that govern such behavior. Qualitative methods examine the *why* and *how* of decision making, not just *what*, *where*, *when*, or *who* (Alasuutari,2010). Maxwell (2005) suggests that qualitative research questions tend to fall into three categories: questions about meaning, or how people make sense of the world; questions that illuminate context; and questions that investigate processes (Maxwell, 2005). Marshall & Rossman (2006), in turn, separate qualitative research questions into exploratory questions, which investigate a phenomenon that is little understood, explanatory questions, which explain a phenomenon, descriptive questions, which seek to describe a phenomenon, and emancipatory question, which are meant to engage in social action around a phenomenon. In addition, the dataset turned from numeric dataset to rich data set.

When facing the challenge of data abundance, we should first distinguish between two ample categories of Big Data: those that are semantically poor (henceforth "poor"), for instance sensor readings, and those that are more complex, i.e., multi-faceted, hierarchical, etc., in a word, semantically rich (hence - forth "rich"). It is possible a characterization of the data on the basis of the semantic concepts and size:

- Small amounts of semantically rich data, where Faceted Search systems or traditional systems of artificial intelligence are very effective;
- Large amounts of data semantically poor, faced with NoSQL database systems that support queries to data arranged in simple data models of this type;
- Large amounts of semantically rich data: this set is all the traditional challenges of Database Research and Data Exploration, powerful computing tools, both mathematical and computational resources are needed to make effective exploration.

When we talk of "rich data set" we intend datasets where objects are classified according to powerful taxonomies. Examples of this second kind are business data, data about health, and in fact most of the data that must be directly examined by users to the purpose, for instance, of taking a decision (Di Blas *et al.*, 2014). Let us examine the main Data Exploration techniques into analyzing a rich data set.

Because of the complexity of these data, very important is the concept of data exploration to transform the data into information we need. A definition of Data exploration is the following. Data exploration is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for (Idreos *et al.*, 2015). Nowadays the user need more than a simple data exploration but need to explore it in interactive way and being able to find her way through large

amounts of data in order to gather the necessary information (Guido et al., 2015). Information mining is distinguished from traditional approaches to data analysis such as query and reporting by the fact that it is aimed at the discovery of information and knowledge, without a previously formulated hypothesis.

Starting from a theoretical data exploration system, where we identified the main features that a data exploration system must own in order to have an effective exploratory experience, we propose an innovative combination of two data exploration techniques: faceted navigation and data mining improving the discovery information during exploration. The paper is organized as follow: in section II a background of data exploration techniques. In the section III we describe a theoretical Data Exploration System to meet the information needs. In the section IV a different challenge of Information Mining: Combining Data Mining and Faceted search is presented. In the section V we present an evaluation of combination Facet Navigation and Data Mining with a case study on EDOC project experience. Finally, in the section VI we conclude the paper with some considerations about results.

2. Background

Traditional data management systems assume that when users pose a query a) they have good knowledge of the schema, meaning and contents of the database and b) they are certain that this particular query is the one they wanted to pose. In short, we assume that users know what they are looking for. In response, the system always tries to produce correct and complete results. Traditional DBMSs are designed for static scenarios with numerous assumptions about the workload (Idreos *et al*., 2015).

The increasing amount of data has led to the build more dynamic data-driven applications that, often, have different requirements than common database systems. Indeed, managing an employee or an inventory database is a drastically different setting than looking for interesting patterns over a scientific database. Consider an astronomer looking for interesting parts in a continuous stream of data (possibly several TBs per day): they do not know what they are looking for, they only wish to find interesting patterns; they will know that something is interesting only after they find it. In this setting, there are no clear indications about how to tune a database system or how the astronomer should formulate their queries. Typically, an exploration session will include several queries where the results of each query trigger the formulation of the next one. This data exploration paradigm is the key ingredient for a number of discovery-oriented applications, e.g., in the medical domain, genomics and financial analysis (Idreos *et al.*, 2015). Such novel requirements of modern exploration driven interfaces have led to rethinking of database systems across the whole stack, from storage to user interaction.

The research in this ambit can be subdivided in these sectors: a) *Visualization tools* for data exploration are receiving growing interest (A.Parameswaran *et al.*, 2013),(E. Wu *et al*, 2014); b) *New exploration interfaces* emerged aiming to facilitate the user's interactions with the underlying database (K. Dimitriadou *et al*, 2014), (S. Idreos *et al*, 2013),(A. Nandi *et al*, 2013); c) Numerous *novel optimizations* have been proposed for offering interactive exploration times (S. Agarwal *et al*, 2014), (N. Kamat *et al*, 2014), (A. Kalinin *et al*, 2014); d) Database architecture has been re-examined to match the characteristics of the new exploration workloads (I. Alagiannis *et al*, 2012), (S. Idreos *et al*, 2011), (S. Idreos *et al*, 2013), (M.Kersten *et al*, 2011). Together, these pieces of work contribute towards providing data exploration capabilities that enable users to extract knowledge out of data with ease and efficiently.

The main techniques for data elaboration and data exploration are Faceted Search and Data Mining.

Faceted Search, also called faceted navigation or faceted browsing, is an exploratory search mechanism. Interesting definition of Faceted Search is the following "Faceted search is an exploratory approach, which provides an iterative way of refining search results by facets."

(BifanWei et al, 2013). The introduction of the faceted concept comes from the Ranganatan that in 1991 describes the multidimensional aspects of a document by defining 5 faceted (Ranganatan, 1991). Starting from the Ranganatan idea there are several other definitions of faceted and a very interesting one is one where faceted are a set of terms related to a specific aspect of a topic (Spiteri, 2008). Each term in a facet is an attribute or a category. Starting from the facet definition comes the faceted search definition meant as the navigation (or faceted browsing) that is a navigation paradigm interactive, heuristic and based on progressive refinement that enable the user to analyze an iteratively select faceted in order to obtain the desired result (Ben-Yitzhak *et al.*, 2008), (Dachset *et al*., 2008). The category definition is the starting point for the facet paradigm and in this research area the main effort was in the defining techniques useful to extract in automatic or semi-automatic way faceted starting from the text (Stoica *et al.*, 2008), (Ling *et al*., 2008).

Data Mining is an interdisciplinary subfield of computer science and it is the process of discovering interesting and useful patterns and relationships in large volumes of data (big data). The fields of Data Mining combine tools from statistics and artificial intelligence (such as neural networks) with database management to analyze large digital collections, known as data sets. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data Mining is widely used in business (insurance, banking, retail), science research and government security. Data mining tasks can broadly be classified into two categories: predictive or supervised and descriptive or unsupervised. The predictive techniques learn from the current data in order to make predictions about the behavior of new datasets. On the other hand, the descriptive techniques provide a summary of the data (Mukhopadhyay, *et al.*, 2014). A possible list of Data Mining Techniques is (Srivastava *et al.*,2002): Classification, Clustering, Association Rules, Sequential Patterns, Regression, Deviation Detection. The four areas that contributed to the growth of data mining in its current form are Artificial Intelligence, Machine Learning, Statistics Databases (Ramzan *et al.*,2014). Data Mining is being used for a wide variety of applications. Below a list of Data Mining current trends and applications (Gupta *et al.*,2014). Prediction and Description (e.g., Election Campaign), Relationship Marketing, Customer Profiling, Customer Segmentation, Outliers Identification and Detecting Fraud, Website Design and Promotion, Web Content Mining, Social Media, Surveillance. Data mining allows you to do many types of data processing and to provide a solution to several classes of problems.

Exploratory Data Analysis, or EDA for short, is a term coined by John W.Tukey in the book "Exploratory Data Analysis" in 1977 (Tukey, 1977). In contrast to statistical approaches aimed at testing specific hypotheses, Exploratory Data Analysis (EDA) is a quantitative tradition that seeks to help researchers understand data when little or no statistical hypotheses exist, or when specific hypotheses exist but supplemental representations are needed to ensure the interpretability of statistical results. In this way, EDA seeks to answer the broad scientific questions of "what is going on here" and "how might I be fooled by my statistical results" (Beherens *et al*.,2003).

In 2006, Marchionini (G. Marchionini,2006) postulates the idea of Exploratory Search as a model in which the user learns and investigates information after a first step of Lookup. Exploratory Search, as Marchionini states, is similar to learn search activity and social searching where people use the same strategy for locating, comparing and assessing results. In exploratory search people usually submit a tentative query to get them near relevant documents then explore the environment to better understand how to exploit it, selectively seeking and passively obtaining cues about where their next steps lie. Exploratory search can be considered a specialization of information exploration, a broader class of activities where new information is sought in a defined conceptual area; exploratory data analysis is another example of information exploration activity. Exploratory search systems (ESSs) capitalize on new technological capabilities and interface paradigms that facilitate an increased level of interaction with search systems. Examples of ESSs include information visualization systems,

document clustering and browsing systems, and intelligent content summarization systems. ESSs go beyond returning a single document or answer in response to a query, and instead aim to instigate significant cognitive change through learning and improved understanding (White *et al*., 2006).

More recently, the research comes back with a new paradigm for access to rich data set, Exploratory Computing. Using this new paradigm, some Exploratory Portal have been developed in several fields of interest (archeology, tourism, education, etc. (N. Di Blas *et al*.,2014), (N. Di Blas *et al*.,2012), (L.Spagnolo *et al*.,2010)). The Exploratory Computing approach as explained in (Paolini *et al*., 2014), and in its manifesto (N. Di Blas *et al*.,2014), allows users to investigate complex dataset composed of rich information. The user can interact with the data and can discover information features that he/she did not see at a first lookup. The innovation of the Exploratory Computing has several features such as serendipitous discovery, at-a-glance understanding, niche finding, raise of interest, sense-making.

Information mining represents a further way to the strategic knowledge. In 1998, IBM calls Information Mining to the process of extracting previously unknown, comprehensible, and actionable information from any source including transactions, documents, e-mail, web pages, and other, and using it to make crucial business decisions (Tkach & Daniel, 1998). Another definition is the following "Information mining is the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in heterogeneous information sources", that is Information Mining tries to combine the analysis of heterogeneous information sources with the prominent aim of producing comprehensible results (Kruse & Borgelt 2003). More recently, the term has been used to indicate the process to acquire knowledge from the interesting patterns discovered by mining from data or information granules, and it is a post-process of the mining processes. Consistent verification, information abstraction, hypothesis generation, hypothesis verification, and information deduction are activities of information mining (Goto, 2015).

3. The theoretical Data Exploration System to meet the information needs

In general, we can classify information needs into two very broad categories: a) precision-oriented ones (e.g. find the telephone of a store) and b) recall-oriented ones (e.g. decide which car to buy). Only some prototype information systems provide means for supporting recall-oriented information needs. Recall-oriented needs frequently aim at decision making, over one or more criteria, and have an exploratory nature, like search tasks in the medical, legal, patent, and academic field, consumer related tasks like car buying (Tzitzikas et al., 2016). Wildemuth and Freund (Wildemuth and Freund,2012) have identified the following as key attributes for exploratory tasks:

1. they are associated with the goals of learning and/or investigation
2. they are general rather than specific
3. they are open-ended
4. they target multiple items
5. they involve uncertainty
6. they elicit through ill-structured information problems
7. they are dynamic
8. they are lengthy
9. they are multi-faceted
10. they are complex
11. they are accompanied by other information and cognitive behaviors, like sense making

The taxonomy of tasks, related to the two different kinds of information needs is illustrated in Figure 1:

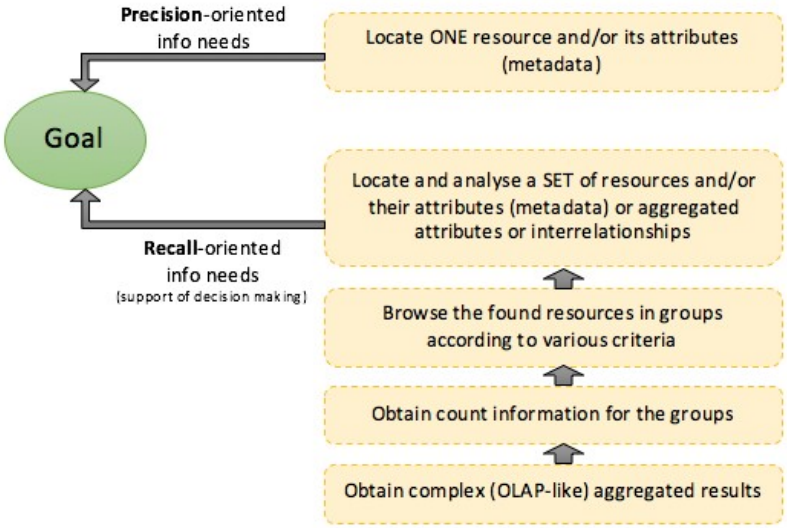


Figure 1. kinds of information needs (Wildemuth and Freund,2012)

In the precision-oriented information needs category, the task's goal is to locate one resource and get information about its attributes or metadata while in the recall-oriented information needs category the task's goal is to locate (and get information about) a set of resources. In this category we can distinguish goals that require accessing sets of resources just in groups, or in groups accompanied by count information for getting an overview of a set of resources, e.g. as in *Faceted Dynamic Taxonomies* (FDT). Furthermore, we may have goals that require more complex aggregated results like those provided by data warehouses. For instance, aggregations of arithmetic (min, max, average) and Boolean functions over the numeric attributes of the documents in the answers of free-text queries. Moreover, counts are computed and displayed over combinations (pairs, triples, quadruplets, etc.) of attributes (of grouping criteria in general). In comparison to *OnLine Analytical Processing* (OLAP) queries, in exploratory search the information demand in unknown a priori (in OLAP it is known and the schema is fixed) and the objective is not only to compute and see various aggregate values (e.g. sales per month and department), but also to support a flexible process for finding the desired individual resources (Tzitzikas et al., 2016).

In our previous paper (Guido et al., 2015) we have identified the main features for an ideal data exploration system that allows the user to have a new and more interesting navigational experience and we have highlighted what techniques meet these main features to obtain better results from a data exploration. The main features are derived by the common needs of users that have to explore and understand large and rich data set with or without a specific goal:

- *Investigation and inspiration seeking:* the user who has an ill-defined idea of what to look for and through the exploration of the dataset moves on, refines, focuses, expands or changes her initial attitude;
- *Researching:* the user who wants to refine or verify some research hypothesis, or who is looking for research hypothesis;
- *Leisure browsing and learning:* the user who wants to stroll around to augment her knowledge about the dataset and can to do a serendipitous discovery;
- *Supervision and decision-making*: the user who needs to understand "how things are going" to decide about something;
- *Set comparison*: the user needs to compare two phenomena, under various perspectives.

- *Categories search*: it is necessary define a coherent set of categories and provide analytic values about distribution of the categories (the feedback is useful for the user and a simple absolute value of values may not address this requirement);
- *Set Exploration*: in order to explore a dataset it is necessary to have the possibility to combine several categories to create a complex set, to create a new set starting from the current one, to combine dataset using logical operators;
- *Interactivity*: an interactive process that implements mechanisms advanced of Human-Computer Interaction is necessary to support sophisticated exploration activities. These mechanisms must be allowed to quickly query the system in order to have new dataset to explore, to create subset starting from the current set in interactive way and using also logical operator, to query the system considering more than two categories in a single query. Thus, just like in a human dialog, a flow of interactions (as opposed to one very powerful interaction) is needed, since users build upon what they discover through the exploration;
- *Correlation between categories*: strong correlation between the categories (the result of a search of a category affects the result of another category even though not expressly stated in the research);
- *Complex answer to simple query:* the ideal data exploration system must be able to provide complex answer to simple query.

The first 5 features characterize the different approaches to the exploration of a user that the system must be able to meet, while the last 5 features express the functionality that the system must possess for effective exploration. Downstream of this critical analysis of the main features of an ideal data exploration system, they have been compared frequently used techniques of exploration, Faceted Search and Data Mining, to discover differences and similarities on the basis of the satisfied characteristics.

Another determinant property is the Visualization: the results determined by the system should be shown to users in a comprehensive way. Thus, efficient and effective visualizations are needed. Research on visualization carried out in the area of Exploratory Data Analysis can come to the rescue in this task.

It is clear that each technique has many features, but not all, and that therefore for obtaining an effective exploration it is necessary to use more techniques together through their skillful combination. We think that this idea open the way to a theoretical data exploration system; we are walking along this road, step by step, to reach the goal of an ideal data exploration system.

4. A new challenge of Information Mining: Combining Data Mining and Faceted search

Case study for this analysis was the Exploratory Portal learning4all, for EDOC@Work3.0 project. By Exploratory Portal, we mean a highly interactive delivery environment, where the exploration can take place through a number of strongly interconnected (and interdependent) interactions.

An exploratory portal takes advantage of the principles and the aims of exploratory computing technique: in this context “exploration” is not search, nor faceted search, nor data mining, nor logic reasoning, nor data visualization: it is a combination of all these approaches, and something more. The exploratory portal L4ALL is characterized by a "repository" shared meaningful learning experiences that have made significant use of technology to innovate and improve teaching methods: several hundred experiences to represent, as appropriate, the diversity and the variety of situations in Italian school through experiences, formats and different pedagogical approaches with a wide variety of technologies used, the school realities examined (level of school, location, socio-economic, environmental and cultural conditions, etc.) and also with an analysis of experiences thorough and methodologically valid. This represented the rich data set of the study: a number of educational experiences carried on at school with a strong support by ICT. Each

experience has some formatted data (location, school level, etc.) some multimedia data (various text files, audio files, video files, etc.) and is classified according to nearly 60 facets. All the objects were classified by pedagogy experts according to a complex taxonomy consisting of 28 attributes' categories and more than 300 attributes. Categories and attributes are organized into widgets supporting both selection and exploration. Each widget shows the value of the attributes for the current state of the dataset; different visualization strategies can be chosen by the user: absolute value, percentage, word-cloud, histogram, etc. The current set of objects is shown on a "canvas". The properties determined by the EC system should be shown to users in a comprehensive way. Thus, efficient and effective visualizations are needed (Di Blas et al., 2014)
Starting from the limitations of exploratory computing, in this work we want to identify innovative methods that combine techniques that have proved successful in other contexts (Data Mining) to enhance the 'information discovery. Thus we try to integrate Data Mining techniques in exploratory portal to support the information discovery with the main aim to identify, through the use of Data Mining models, the patterns of knowledge useful to exploratory experience of a user inside of the educational experiences repository that represents the rich data set.

We have chosen to implement, between the different existing data mining techniques, the Cluster Analysis and we have used for this aim the WEKA as tool of development. Below we explain the reasons for both choices.

4.1. Cluster Analysis

Clustering is a machine learning technique used for discovering groups or pattern in a dataset. These groups or sets of similar data are known as clusters.

The Clustering algorithms allow performing segmentation operations on the data, that is to identify homogeneous patterns, which have regularities in them able to characterize and differentiate from the other patterns.

There are a large number of clustering algorithms. The main reason for having many clustering methods is the fact that the notion of "cluster" is not precisely defined (Estivill-Castro, 2000). Consequently, many clustering methods have been developed, each of which uses a different induction principle. Farley and Raftery (1998) suggest dividing the clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) suggest categorizing the methods into additional three main categories: density-based methods, model-based clustering and grid-based methods. An alternative categorization based on the induction principle of the various clustering methods is presented in Estivill-Castro (2000). The algorithm chosen to be used in a given context depends on the type of data available, the particular purpose and application. If the cluster analysis is used as a descriptive or exploratory tool, you can try different algorithms on the same data to see what each of them can do.

In this work we have chosen to implement the partitional clustering. Partitional clustering algorithms generate various partitions and then evaluate them by some criteria. They are also referred to as nonhierarchical as each instance is placed in exactly one of k mutually exclusive clusters. Because only one set of clusters is the output of a typical partitional clustering algorithm, the user is required to input the desired number of clusters (usually called k). One of the most commonly used partitional clustering algorithms is the k-means clustering algorithm. The user is required to provide the number of clusters (k) before starting and the algorithm first initiates the centers (or centroids) of the k partitions. In a nutshell, k-means clustering algorithm then assigns members based on the current centers and re-estimates centers based on the current members. These two steps are repeated until a certain intra-cluster similarity objective function and inter-cluster dissimilarity objective function are optimized. Therefore, sensible initialization of centers is a very important factor in obtaining quality results from partitional clustering algorithms.

The most well-known and commonly used partitioning algorithms include: K-means clustering (MacQueen, 1967), in which, each cluster is represented by the center or means of the data points belonging to the cluster; K-medoids clustering or PAM (Partitioning Around Medoids), (Kaufman & Rousseeuw, 1990), in which, each cluster is represented by one of the objects in the cluster. A variant of PAM is named CLARA (Clustering Large Applications) which is used for analyzing large data sets.

4.1.1. K-means algorithm

In k-means clustering, each cluster is represented by its center (i.e., centroid) which corresponds to the mean of points assigned to the cluster. Recall that, k-means algorithm requires the user to choose the number of clusters (i.e., *k*) to be generated.

The algorithm starts by randomly selecting *k* objects from the dataset as the initial cluster means.

Next, each of the remaining objects is assigned to its closest centroid, where closest is defined using the Euclidean distance between the object and the cluster means. This step is called cluster assignment step. After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster centroid update is used to design this step. All the objects are reassigned again using the updated cluster means The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e. until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration. The algorithm can be summarizing as follow:

1. Specify the number of clusters (K) to be created (by the analyst)
2. Select randomly *k* objects from the dataset as the initial cluster centers or means
3. Assigns each observation to their closest centroid, based on the Euclidean distance between the object and the centroid
4. For each of the *k* clusters update the cluster centroid by calculating the new mean values of all the data points in the cluster. The centroid of a K_{th} cluster is a vector of length *p* containing the means of all variables for the observations in the K_{th} cluster; *p* is the number of variables.
5. Iteratively minimize the total within sum of square. That is, iterate steps 3 and 4 until the cluster assignments stop changing or the maximum number of iterations is reached. Usually 10 as the default value for the maximum number of iterations.

K-means clustering is very simple and efficient algorithm. However, there are some weaknesses, including:

- It assumes prior knowledge of the data and requires the analyst to choose the appropriate *k* in advance
- The final results obtained are sensitive to the initial random selection of cluster centers.

To overcome these difficulties there are some solutions that briefly are: in respect to the first problem: compute k-means for a range of *k* values, for example by varying *k* between 2 and 20 and then, choose the best *k* by comparing the clustering results obtained for the different *k* values. The solution in respect to the second problem: compute K-means algorithm several times with different initial cluster centers. The run with the lowest total within-cluster sum of square is selected as the final clustering solution.

4.1.2. Partitioning Around Medoids (PAM) algorithm

The use of means implies that k-means clustering is highly sensitive to outliers. This can severely affects the assignment of observations to clusters. A more robust algorithm is provided by PAM algorithm which is also known as k-medoids clustering.

The pam algorithm is based on the search for *k* representative objects or medoids among the observations of the dataset. These observations should represent the structure of the data. After finding a set of *k* medoids, *k* clusters are constructed by assigning each observation to the nearest

medoid. The goal is to find *k* representative objects which minimize the sum of the dissimilarities of the observations to their closest representative object. For a given cluster, the sum of the dissimilarities is calculated using Manhattan distance.

PAM works efficiently for small data sets but is not very scalable. To treat big datasets it is possible to use a sampling based method, called CLARA. The idea behind CLARA is the following: instead of taking into account the entire set of data, a small subset of the actual data is chosen assuming that it is representative of all the data. The medoids are, therefore, chosen from this sample using PAM. If the samples are selected rather randomly, they should represent quite closely together of the original data and the identified representative medoids should be similar to those that would have been constructed using the entire set of data.

Data mining software

Today, many kinds of Data mining software are available on the internet. Each tool has different methods of analyzing and interpreting the information from a grouped data. Data mining can be difficult, especially if you do not know what some of the best free data mining tools are.

RapidMiner, RapidAnalytics, WEKA, PSPP, KNIME, Orange, Apache Mahout, jHepWork, Rattle, GhostMiner, XENO, SAS Enterprise Miner, Polyanalyst and IBM SPSS modeler are the most common Data mining tools used. In our work we have evaluated the following software: Orange and WEKA.

Orange is a machine learning and data mining suite for data analysis through Python scripting and visual programming. It focuses on simplicity, interactivity through scripting, and component-based design. Orange library is a hierarchically-organized toolbox of data mining components. The main branches of the component hierarchy are: *data management and preprocessing* for data input and output, *classification*, *regression*, *association* for association rules and frequent item sets mining, *clustering*, which includes k-means and hierarchical clustering approaches, *evaluation* with cross-validation and other sampling-based procedures, *projections* with implementations of principal component analysis, multi-dimensional scaling and self-organizing maps.

The library is designed to simplify the assembly of data analysis workflows and crafting of data mining approaches from a combination of existing components. Orange scripting library is also a foundation for its visual programming platform with graphical user interface components for interactive data visualization (Janez et al., 2013).

Below the focus on the Clustering algorithms implemented by Orange.

- *Hierarchical Clustering*: computes hierarchical clustering of arbitrary types of objects from the matrix of distances between them and shows the corresponding dendrogram supports three kinds of linkages. In Single linkage clustering, the distance between two clusters is defined as the distance between the closest elements of the two clusters. Average linkage clustering computes the average distance between elements of the two clusters, and complete linkage defines the distance between two clusters as the distance between their most distant elements (Hierarchical Clustering. Documentation for Orange v2.7, 2014).
- *K-Means Clustering*: applies the K-means clustering algorithm to the data from the input and outputs a new data set in which the cluster index is used for the class attribute. The original class attribute, if it existed, is moved to meta attributes (K-Means Clustering. Documentation for Orange v2.7, 2014).

Weka is a suite of machine learning software applications written in the Java programming language. Weka is Waikato Environment for Knowledge Analysis. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization (Ian et al.,2011). Weka provides access to

SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka (Reutemann et al., 2004). Weka provides comprehensive sets of data pre-processing tools, learning algorithms and evaluation methods, graphical user interfaces and an environment for comparing learning algorithms.

Weka contains "clusters" for finding groups of similar instances in a dataset. Some implemented schemes are: K-means, EM, Cobweb, X-means, FarthestFirst. Another feature is the panel Experimenter that makes it easy to compare the performance of different learning schemes. The evaluation options present are: cross-validation, learning curve, hold-out and it is possible also to iterate over different parameter settings (Witten et al., 2016)

For this work we have used at first the Orange software. Orange resulted interesting for the capability to design the data analysis process through the visual programming, but we met customization issues during the development and, furthermore, the software implements few partitional clustering algorithms. For this reasons our choice was changed and we are now orienting to the WEKA software.

5. Evaluation of combination Facet Navigation and Data Mining

In order to evaluate effort and performance obtained in used traditional and clustering analysis based approaches we refer to a case study related to the L4All portal in EDOC project experience.

5.1. Traditional Approach in L4All Exploratory Portal

L4All (Fig. 1) hosts nearly 300 objects describing educational experiences in which the use of technology was relevant. Each object entails several information items: an abstract, some structured data, one or more reports, interviews, documents produced within the experiment, etc. All the objects are classified by pedagogy experts according to a complex taxonomy consisting of 39 attributes' categories and more than 300 attributes. Categories and attributes are organized into widgets (see Figure 2 – left hand side) supporting both selection and exploration. Simple selection or complex selection operations, with boolean operators, are possible. Each widget shows the value of the attributes for the current state of the dataset with different visualization. The current set of objects is shown on a "canvas" (see Figure 2 - right side of the interface). Thanks to advanced Human-Computer Interaction mechanisms, the portal can support sophisticated exploration activities in the cycle <selection, feedback, selection>. Based on L4All, a number of scientific investigations by different research groups took place: on the relation between different forms of group-work and inclusion, on digital storytelling and related benefits, etc. (Di Blas, Paolini, 2013; Falcinelli, 2012; Falcinelli, Laici, 2012). Let us see an example of investigation. In the case of the research on "Expertise with technology" of a teacher and "Student's performance ", the main point was investigating whether there was any relation between the two. In order to answer this question, the value "Excellent" was selected within the facet "Expertise with technology"; taking a look at the values related to level of performance (average, high or low) within the facet "student's performance" and comparing them with the Universe (the initial set) it appeared that a relation was there: most of the values are average high. Thus it was clear that the expertise with technology of a teacher is an important factor to student's performance. It is important to note that the exploration, in the exploratory portal, is aimed at experts in the domain that, on base of their knowledge of domain, are able to discover the information.

5.2. Cluster Analysis on Facets

Our aim was to identify, through the use of data mining models, patterns knowledge inside of the facets of the exploratory portal. To achieve this purpose, we have applied the clustering

algorithm on experiences of the exploratory portal. The portal is schema-driven through a modeling of taxonomy, the data and the portal layout on Excel. After the first phases of information retrieval and pre-processing of dataset (cleaning, enrichment, coding) we uploaded a dataset of the facets of the experiences in csv format on tool Weka.

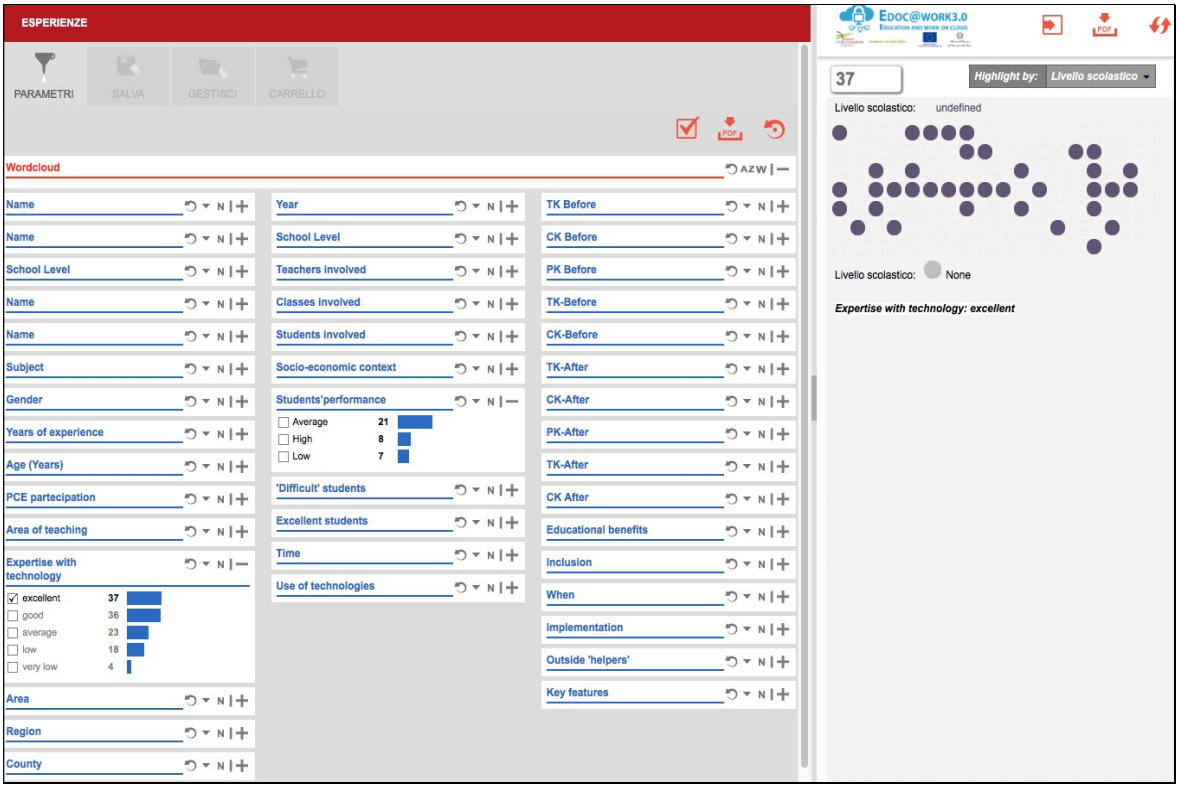


Figure 2. A ScreenShot of L4All Portal

The starting dataset consists of the general scheme of the complex taxonomy on which the modeling of experiences is based. The proposed general scheme consists of two excel files: one related to the data and one related to the annexes of the experiences.

The data file consists of the following types of sheets:

- Widget: only one sheet, defines the overall layout and the number of columns in which subdivide the widgets in the interface;
- Define Widget: one sheet for each facet, defines the structure of each widget, the labels displayed for each widget;
- Widget label: one sheet for each facet, defines the data of the experience.

The connection between the sheets is through the widget id. The schema presented defines all aspects of the data for our case study.

From this starting dataset we have extracted and built the dataset on which to apply the data mining clustering technique.

We selected the relevant facets for our purpose (for example: the facets related to the municipality and province are not relevant in looking for similar relationships in the experiences and they were not taken into consideration, instead the facet "macro region" - with attributes north, center, south, islands - are useful to indicate the geographic area). So starting with the 39 initial

facets we extracted 23 facets for a total of 42 types of attribute and, after the operations of cleaning, enrichment and coding, we have a total of 118 instances.

Then we uploaded a dataset of the facets of the experiences in csv format on the Weka tool that implements several clustering's algorithms: we tested *SimpleKmeans* on our dataset, described in paragraph 4.1. We tested the algorithm with different values of K, to find the optimal centroids.
In general, as you know, there is no method for determining the exact value of K, but an accurate estimate can be obtained, for example, monitoring the value of the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid. Mathematically, we can write (1):

$$SSE = \sum_{i=1}^{K} \sum_{x \in c_i} dist\,(x, c_i)^2 \qquad (1)$$

In our case we estimated in k = 8 the best number of cluster. We obtained the following clustered instances:

```
=== Model and evaluation on training set ===

Clustered Instances

0       17 ( 14%)
1        8 (  7%)
2       23 ( 19%)
3       23 ( 19%)
4       14 ( 12%)
5       15 ( 13%)
6        9 (  8%)
7        9 (  8%)
```

Figure 3: Clustered Instances

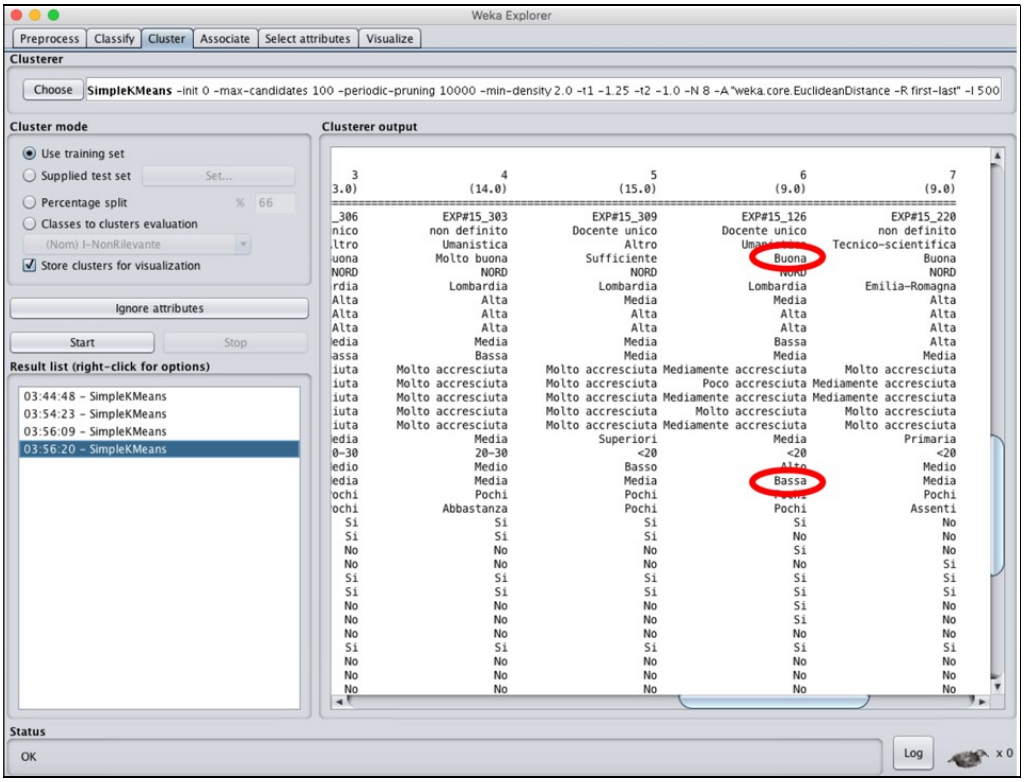


Figure 4. Results of Weka's Clustering

In the Figure 4 we show that, in a particular cluster, attributes are grouped in the "good" attribute of "Expertise with technology" with the "Low" attribute of "Student's performance" together.

Thus, we can deduce that the level of Student's Performance is influenced by other factors over "Expertise with technology" of teacher. These factors can be searched inside the cluster, providing useful information to a significant exploration. These aspects are not deducible only by exploration through the portal and, for this reason, the clustering technique allows to user to navigate better during the search.

In the Table 1 a characterization of the identified clusters is presented, representing patterns of knowledge where a significant exploration is possible.

Table 1. Patterns of knowledge

Clust	Characterization
C0	Teaching area: **Humanistic**, expertise with technology of teacher: **Very good**, Macro-region: **South**, school level: **Primary**, social-economic context: **Average**, class performance: **Average**
C1	Teaching area: **Technical-scientific**, expertise with technology of teacher: **Very good**, Macro-region: **South**, school level: **Primary**, social-economic context: **Low**, class performance: **Average**
C2	Teaching area: **Humanistic**, expertise with technology of teacher: **Sufficient**, Macro-region: **South**, school level: **College**, social-economic context: **Average**, class performance: **Average**
C3	Teaching area: **Other**, expertise with technology of teacher: **Good**, Macro-region: **North**, school level: **Secondary**, social-economic context: **Average**, class performance: **Average**
C4	Teaching area: **Humanistic**, expertise with technology of teacher: **Very good**, Macro-region: **North**, school level: **Secondary**, social-economic context: **Average**, class performance: **Average**
C5	Teaching area: **Other**, expertise with technology of teacher: **Sufficient**, Macro-region: **North**, school level: **College**, social-economic context: **Low**, class performance: **Average**
C6	Teaching area: **Humanistic**, expertise with technology of teacher: **Good**, Macro-region: **North**, school level: **Secondary**, social-economic context: **High**, class performance: **Low**
C7	Teaching area: **Technical-scientific**, expertise with technology of teacher: **Good**, Macro-region: **North**, school level: **Primary**, social-economic context: **Average**, class performance: **Average**

Through patterns of knowledge, the user can explore the information within the more interesting cluster, facilitating the correct interpretation of the results of the exploration and, furthermore, can use the relevant properties of each cluster to refine the information search on the entire dataset in order to conduct a more effective general exploration.

6. Results and Discussion

The clustering of rich data set discovers new properties (semantic relationship between attributes) compared to the results of exploration conducted on the portal. This has led to consider the introduction of cluster analysis of the facet very useful to improve exploratory experience. It is obtained in this way by the combination of two different paradigms: Faceted Search, with its fast interaction for the creation of subsets, and Data Mining, with its ability to understand the properties of the datasets. This combination leads to develop a series of new features and opens up new challenges and opportunities not previously available.

Therefore, among the results of this work there are:

- the identification of patterns of knowledge, by the application of Data Mining tools;
- the patterns of knowledge allow the user to explore the information within the more interesting cluster facilitating the correct interpretation of exploration results;
- the relevant properties that allowed the tool to build clusters can be used by the user as a guide or indicator to conduct a more effective general exploration on all rich data set;

- This approach, also, facilitates the exploration to a not-expert user of domain and increases the "awareness exploratory" to an expert user of domain.

In the following figure (see Figure 5) it has represented a scheme of the new approach proposed to Rich Data Set's Exploration:

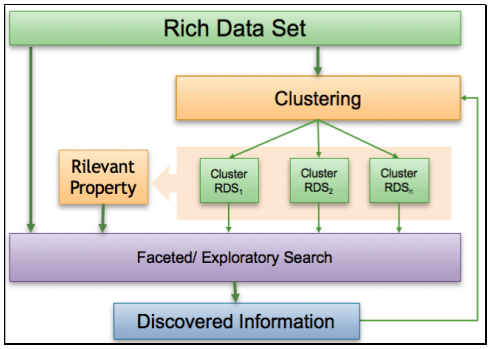


Figure 5: New Approach Rich Data Exploration

Other experiments are running in order to validate our idea, both in order to optimize this clustering model by applying new algorithms and distance measures to the datasets presented here, and both applying these techniques to a different domain from the didactic one. Other experiments are also conducted to improve user exploration by skillfully combining multiple methods and exploration techniques through the application of a variety of models such as the Association Rule to extract hidden relationships and association rules between data and Artificial Neural Network mechanisms of learning applicable to classification and forecasting problems.

6. Conclusions

The present paper aims to make a combination between two Data Exploration Techniques: Facet Search and Data Mining, in order to evaluate the improvement in terms of performance and effort that is possible to obtain during an exploratory experience. The combination is a new approach to the discovery and management of information by improving the exploratory experience of a user. The results obtained are encouraging because compared to the previous approach, where exploration is aimed at domain experts, who are able to make a user exploratory research based on their knowledge, we think it is useful to investigate this scientific research context with the aim of supporting a non-domain expert user in finding its way through an exploration. This approach introduces us into the field of Information Mining that aims at discovering knowledge, i.e. more general patterns within objects or collections of objects.
In summary, the results obtained in terms of performance and effort during the case study we have conducted to perform the evaluation can confirm our expectations.

References

Agarwal, S., Milner, H., Kleiner, A., Talwalkar, A., Jordan, M., Madden, S. Mozafari, S., & Stoica, I. (2014). Knowing when you're wrong: Building fast and reliable approximate query processing systems. In Proceedings of the ACM SIGMOD Conference on Management of Data, 2014.

Alagiannis, I., Borovica, R., Branco, M., Idreos. S., & Ailamaki, A. N. (2012). Efficient query execution on raw data files. In Proceedings of the ACM SIGMOD Conference on Management of Data, 2012 (pp. 241–252).

Alasuutari, P. (2010). *The rise and relevance of qualitative research.* International journal of social research methodology 13(2), 139-155.

Anirban, M. et. al. (2014). *A survey of multiobjective evolutionary algorithms for data mining: Part I.* Evolutionary Computation, IEEE Transactionson 18.1: 4-19, 20-35.

Behrens J. T. & Chong H. Y. (2003). *Exploratory data analysis*. Handbook of psychology.

Ben-Yitzhak, O. et. al. (2008). Beyond basic faceted search. In Proceedings of the International Conference on Web search and web data mining, 2008 (p. 33 – 44). Palo Alto, California, USA.

Bifan W., Jun L., Qinghua Z., Wei Z., Xiaoyu F., & Boqin F. (2013). A survey of faceted search. J. Web Eng. 12, 1-2 (February 2013), 41-64.

Dachselt, R., Frisch, M., & Weiland, M. (2008). FacetZoom: a continuous multi-scale widget for navigating hierarchical metadata. In Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (p. 1353-1356). Florence, Italy.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Stajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: data mining toolbox in Python. JMLR. 14(1): 2349–2353.

Di Blas, N. & Paolini, P. (2013). Technology and Group Work: Inclusion or Diversification of Talents?. In Parmigiani, D., Pennazio, V., & Traverso, A. (Eds.). Learning & Teaching with Media & Technology. ATEE-SIREM Winter Conference Proceedings, 2013, March 7-9 (pp. 218-231). Genoa, Italy. Brussels, ATEE aisbl.

Di Blas, N., Fiore, A., Mainetti, L., Paolini, P., & Vergallo, R. (2014). A Portal of Educational Resources: Providing Evidence for Matching Pedagogy with Technology. In Research in Learning Technology, vol. 22, 2014, May 2014, p. 1-26, ISSN: 2156-7069. UK: Co-Action Publishing.

Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., & Tanca, L. (2014, October). Exploratory computing: a draft Manifesto. In Data Science and Advanced Analytics (DSAA), 2014 International Conference (pp. 577-580). IEEE.

Di Blas, N., Paolini, P., & Spagnolo, L. (2012). Policultura Portal: 15.000 Students Tell their Stories about Cultural Heritage. In N. Proctor and R. Cherry (Eds.), Museums and the Web 2012. Selected Papers from an International Conference. Archives & Museum Informatics.

Dimitriadou, K., Papaemmanouil, O., & Diao, Y. (2014). Explore-by-Example: An Automatic Query Steering Framework for Interactive Data Exploration. In Proceedings of the ACM SIGMOD Conference on Management of Data.

Estivill-Castro, V. & Yang, J. A. (2000). Fast and robust general purpose clustering algorithm. Pacific Rim International Conference on Artificial Intelligence (pp. 208-218).

Falcinelli, F. & Laici, C. (2012). Teaching with ICT: The Policultura and Moodle Didactic Format Experimented in Schools, IJCEE, January-March 2012, Vol. 2, No. 1.

Falcinelli, F. (2012, November 20-30). Evidence-Based Research About the Impact of ICT on Italian Schools: The Cl@ssi2.0 Project. Online Educa Berlin 2012. Berlin.

Fraley C. & Raftery A. E. (1998). *How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis.* Technical Report No. 329. Department of Statistics University of Washington.

Goto, Y. (2015). *Information Mining for Big Information*. Information Granularity, Big Data, and Computational Intelligence. Springer International Publishing, 23-38.

Guido A. L., Paiano R., Pandurino A., & Pasanisi S. (2015). Searching issues: a survey on data exploration techniques. International Journal of Emerging Trends and Technology in Computer Science, vol. 4, p. 183-188, ISSN: 2278-68.

Gupta, G. K. (2001). Introduction to data mining with case studies. PHI LearningPvt. Ltd., 2014.

Han, J. and Kamber, M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Hierarchical Clustering. (2014). Documentation for Orange v2.7. Retrieved from https://docs.orange.biolab.si/2/widgets/rst/unsupervized/hierarchicalclustering.html#hierarchical-clustering.

Ian, H. W., Eibe, F., & Mark, A. (2011). Data Mining: Practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, San Francisco.

Idreos, S. & Liarou, E. (2013). dbTouch: Analytics at your fingertips. In Proceedings of the biennial Conference on Innovative Data Systems Research (CIDR), 2013.

Idreos, S. (2013). Big Data Exploration,Taylor and Francis.

Idreos, S., Alagiannis, I., Johnson, R., & Ailamaki, A. (2011). Here are my Data Files. Here are my Queries. Where are my Results? In Proceedings of the biennial Conference on Innovative Data Systems Research (CIDR), 2011.

Idreos, S., Papaemmanouil, O., & Chaudhuri, S. (2015). *Overview of data exploration techniques*. Proceedings of the 2015 ACM SIGMOD. International Conference on Management of Data. ACM, 2015.

Kalinin, A., Cetintemel, U., & Zdonik, S. (2014). Interactive Data Exploration using Semantic Windows. In Proceedings of the ACM SIGMOD Conference on Management of Data, 2014.

Kamat, N., Jayachandran, P., Tunga, K., & Nandi, A. (2014). Distributed Interactive Cube Exploration. In Proceedings of the International Conference on Data Endineering (ICDE).

Kaufman, L., & Rousseeuw, P. J. (1990). *Partitioning around medoids (program pam).* Finding groups in data: an introduction to cluster analysis: 68-125.

Kersten, M., Idreos, S., Manegold, S. & Liarou, E. (2011). The Researcher's Guide to the Data Deluge: Querying a Scientific Database in Just a Few Seconds. Proceedings of the Very Large Data Bases Endowment (PVLDB), 4(12):1474–1477.

K-means Clustering. (2014). Documentation for Orange v2.7. Retrieved from https://docs.orange.biolab.si/2/widgets/rst/unsupervized/kmeansclustering.html#k-means-clustering.

Kruse, R., & Borgelt, C. (2003). *Information mining*. International Journal of Approximate Reasoning (IJAR), Vol.32(2), pp. 63-66.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14.

Marchionini, G. (2006). Exploratory search: from finding to understanding. Communication of the ACM, vol. 49, no. 4, p. 41.

Nandi, A. (2013). Querying Without Keyboards. In Proceedings of the biennial Conference on Innovative Data Systems Research (CIDR).

Paolini, P., & Di Blas, N. (2014, October). Exploratory portals: The need for a new generation. In Data Science and Advanced Analytics (DSAA), 2014 International Conference on (pp. 581-586). IEEE.

Parameswaran, N. P. & Garcia-Molina, H. (2013). SeeDB: Visualizing Database Queries Efficiently. Proceedings of the Very Large Data Bases Endowment (PVLDB), 7(4):325–328.

Ramzan, M. & Majid, A. (2014). *Evolution of data mining: Anoverview.* IT in Business, Industry and Government (CSIBIG), 2014 Conferenceon. IEEE.

Ranganatan, S. R. (1991). Elements of library classification (1st ed). Bombay, New York: South Asia Books. 168p.

Reutemann, P., Pfahringer, B., Frank, E., (2004). *Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners.* 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag.

Spagnolo, L., Bolchini, D., Paolini, P., & Di Blas, N. (2010). Beyond Findability: Search-Enhanced Information Architecture for Content-Intensive RIAs. Journal of Information Architecture, 2(1), 19-36.

Spiteri, L. (2008). A simplified Model for Facet Analysis. Canadian Journal of Information and Library Science, 23(1-2) p.1-30.

Srivastava, Jaideep, Prasanna, D., & Kumar, V. (2002). *Web mining: Accomplishments and future directions*. National Science Foundation Workshop on Next Generation Data Mining (NGDM'02).

Tkach, Daniel, S. (1998). *Information mining with the IBM intelligent miner family.* An IBM Software Solutions White Paper: 1-29.

Tukey & John, W. (1977). Exploratory data analysis: 2-3.

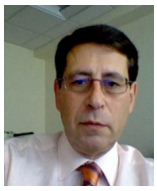
Tzitzikas, Yannis, Nikos, M., & Papadakos, P. (2016). *Faceted exploration of RDF/S datasets: a survey.* Journal of Intelligent Information Systems: 1-36.

White, R. W., Muresan, G., & Marchionini, G. (2006, December). Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems. In ACM SIGIR Forum, Vol. 40, No. 2, pp. 52-60. ACM.

Wildemuth, B. M. & Freund., L. (2012). Assigning Search Tasks Designed to Elicit Exploratory Search Behaviors. In Procs of the Symposium on Human-Computer Interaction and Information Retrieval, HCIR '12 (p 4:1–4:10). New York, NY, USA, ACM.

Witten, Ian, H., et al. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, E., Battle, L., & Madden, S. (2014). The Case for Data Visualization Management Systems. Proceedings of the Very Large Data Bases Endowment (PVLDB), 7(10), pp. 903–906.



Roberto Paiano (b. June 14, 1958) graduated in Electronic Engineering at the University of Bologna. He worked in IBM for 10 years. He was team leader at IBM RNSL and Project Manager at the CORINTO Consortium (National Research Consortium about Object-Oriented Technology). He was member of the IEEE. Currently, he is assistant professor at University of Salento (Italy). He has authored papers about information systems, Web modeling and design, metrics for the Web development. His current research interests are: the methodology of design of Web information systems, the automatic code generation using Open-Source Frameworks and Information Systems modeling.

Stefania Pasanisi (b. October 6, 1978) graduated in Automation Engineering at the University of Salento (Italy) in April 2009. After the degree she worked for five years in the company (Lecce) on projects for observational study and experimental project in the medical field and for design and development software and web applications. Since November 2014 she is a PhD student in Engineering of Complex Systems at the University of Salento (Italy). Her main research areas include advanced semantics exploration techniques on dynamic and complex information spaces, Exploratory Computing Technique and Data Mining. She participates tpo several research projects and she is (co-) author of several scientific papers.