International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 4, Issue 6, November - December 2015 ISSN 2278-6856

Searching issues: a survey on data exploration techniques

Anna Lisa Guido, Roberto Paiano, Andrea Pandurino, Stefania Pasanisi[1]

[1]Department of Engineering for Innovation Salento University Lecce, Italy

Abstract

In the field of "Data Exploration" many approaches have been developed to resolve the problem of management of big data that are also semantically rich. 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. In this paper, we first introduce the main features of an idealdata exploration system.Next, we present an overview of the main techniques of data exploration introducing related methods and techniques such as Faceted Search and Data Mining and then we conduct a comparative study of Faceted Search, Data Mining. The aim of our work is, through a critical analysis of the main features of an ideal data exploration system, discover the differences and similarities between the techniques analyzed. Some directions for future research are finally presented.

Keywords: Data Exploration, Faceted Search, Data Mining, large and rich data set.

1. INTRODUCTION

Since the beginning of the ICT, the management and the visualization of the information are considered main research task. The managed information turned from analytics in the early period to qualitative in these last years. They turned from numeric dataset to rich dataset. Moreover, nowadays the users need more than a simple data elaboration but need to explore it in interactive way: for example, the user to choose what to see in the TV channels broadcast could do an exploration on the rich dataset of broadcast program classified by genre, time band, channel type, etc. Thus, the ability to explore quickly rich and large data set for discover patterns, is becoming a crucial competitive advantage and consequently a very interesting research field. For this aim, several attempts to introduce new techniques were born in the last ten years. The term “data exploration” is an informative search and generally refers to data consumers being able to gather the necessary information through large amounts of data. Data exploration is used to analyze the data and information from the data in order to take a further analysis.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:is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters and to provide users visible options for clarifying and refining queries. This technique has become a popular in commercial search application, particularly for online retailers and libraries.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 and machine learning) 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.

In addition to these techniques more known, there are other search paradigms for data exploration such as Exploratory Search, Exploratory Data Analysis, Exploratory Computing. Exploratory search covers a broader class of activities than typical information retrieval, such as investigating, evaluating, comparing, and synthesizing [1].Exploratory Data Analysis (EDA) is another example of an information exploration activity: is an approach to analyzing data sets to summarize their main characteristics, often with visual methods [24]. Exploratory computing is aninnovative and more recent paradigm that is possible to describe as the step-by-step “conversation” of a user and a system that “help each other” to refine the data exploration process, ultimately gathering new knowledge that concretely fulfills the user needs [23].

Our work aims is, through a critical analysis of the main features of an ideal data exploration system, discover the differences and similarities between the techniques analyzed.

The paper is organized as follow: in section II the main features of an ideal data exploration systemwill be defined. In the section III, IVand V an overview respectively on Faceted Search, Data Mining and other search paradigms. In the section VI a discussion on comparison between the techniques analyzed where we explain what characteristics of the Faceted Search, Data Mining fits the main features defined. Finally in the section VII we conclude the paper with some considerations and future works.

2. MAIN FEATURES OF AN IDEAL DATA EXPLORATION SYSTEM.

In order to understand the similarities and the differences between the several techniques of data exploration it is very useful to define the main featuresfor to obtain an idealdata exploration system. It is clear that the reported 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.There are several approaches of the users during the exploration [6]:

- **Investigation and inspiration seeking**: the user who has an ill-defined idea of what to look forandthrough the exploration of the dataset moves on, she refines, focuses, expands or changes her initial attitude;
- **Researching**: the user who wants to refine or verify some research hypothesis, or she 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.

Regardless if the user is a domain expert or not, when he/she explores a dataset he/she explores the vision (provided by the dataset designer) of the domain and not the domain information itself. Thus, in general the exploration of a dataset is equivalent to search something in a complex unknown environment. Thus, the possibility to organize information in several categorieshelps the user to have a guide to explore the dataset and helps to have a multilevel overview of contained information.

The user needs to be assisted during his/her exploration so the interactivity and the suggestion about the possible next research may be a useful help for the users.

In order to understand the dataset structure and the contained data, it is important to provide to the user analytics information such as count of selected items, percentage, etc., about how many contents fill in the specific category (or categories) selected. This kind of information allows the user to comprehend the dataset and drive his/her next interaction to refine the selection data. In order to provide to the users an intuitive way to explore the data, the explorative portal allows organizing entire information in a subset related to the information categories and their distributions. Thus the user has the perception how the several categories interact on the data and the mutual influences between the categories themselves.As it regards the set definition it is very important to have the possibility to understand what is the current set and to modify the starting set ones user understand its next selection and working with set properties. It is also important to be able to understand characteristics of the set (for example correlation between information also with limits generally not considered by traditional information exploration systems) without change the set itself and to change the set according to the new considerations arise from previous exploration.Starting from the previous considerations, it is possible to define a set of features:

- **Categories search**: 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**: for explore a dataset is necessary to have the possibility of combine several categories to create a complex set, create a new set starting the current one, combine dataset using logical operators
- **Interactivity**: an interactive process that implement mechanisms advanced of Human-Computer Interaction is necessary for 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

3. FACETED SEARCH OVERVIEW

In [7]there is a very interesting definition of Faceted Search "Faceted search is an exploratory approach, which provides an iterative way of refining search results by facets."

The introduction of the faceted concept comes from the Ranganathan that in 1991 describes the multidimensional aspects of a document by defining 5 faceted [8]. Starting from the Ranganathan idea there are several other definition of faceted and a very interesting one is in [9]where faceted are a set o terms related to a specific aspect of a topic. Each term in a facet is an attribute or a category.

Starting form 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 able the user to analyze an iteratively select faceted in orderto obtain the desired result [10]-[11].

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 form the text [12]-

[13].

Another approach to the facet management is the Dynamic Category Sets: using this approachthe results of a search is not only one value but also a set of possible value related to the user search and the user will select the value of interest. The set of values presented to the user are related to a specific taxonomy that link together the terms in the facet taxonomy and a list of possible synonymous useful to solve the vocabulary problem. The faceted search has, as output, not only the categories in the facet but also all the categories in the taxonomy.

In addition to the first idea of Dynamic Category Sets there is the use of Word Net in order to define, using synonymous, hyponyms,etc., network edge useful to the user in order to refine its research.

An important aspect of the facet search is the possibility to put in correlation several faceted. A possibility, expressed in [15] is the use of auser interface where the main element is a bi-dimensional graph that helps user to understand the data distribution in the two dimensions, their correlation and the data intensity. Using mouse over it is possible to obtain the contents.

4. DATA MINING OVERVIEW

The process of analyzing plain dataset including the research and the extraction patterns, the data explorationis known by many names in different environments (knowledge extraction, information discovery, information retrieval, etc.). The Knowledge Discovery in Databases (KDD) indicates the entire process of knowledge discovery from data of databases. Data Mining (DM) is a particular step in the process: the application of specific algorithms for extracting patterns. It, along with the other steps in the KDD process, such as the preparation, selection, data cleaning, storage of previously acquired knowledge, the interpretation of results, ensure and guarantee that extracted knowledge is actually valid.

In the research panorama there are many definitions for DM, in [16]"Data mining involves discovering interesting and potentially useful patterns of different types, such as associations, summaries, rules, changes, outliers, and significant structures". Commonly, data mining and knowledge discovery in database (or KDD) are frequently treated as synonyms, for example in [17] "Data mining popularly known as Knowledge discovery in databases (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for the nontrivial extraction of implicit, previously unknown and potentially useful information from databases". Although some scientists consider data mining to be an integral step in the knowledge discovery process "Data Mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses"[18]and another "Data Mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making"[17].

4.1 Data Mining Techniques and Applications

DM techniques comprise three components: a model, a preference criterion, and a search algorithm. The most common model functions in current data mining techniques include classification, clustering, regression, sequence and link analysis and dependency modeling. Model representation determines both the flexibility of the model for representing the underlying data and the interpretability of the model human terms. This includes decision trees and rules, linear and nonlinear models, example-based techniques such as NN-rule and case-based reasoning, probabilistic graphical dependency models (e.g., Bayesian network) and relational attribute models. The preference criterion is used to determine, depending on the underlying data set, which model to use for mining, by associating some measure of goodness with the model functions. It tries to avoid over fitting of the underlying data or generating a model function with a large number of degrees of freedom. Finally, once the model and the preference criterion are selected, specification of the search algorithm is defined in terms of these along with the given data [19].

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 [16]. The techniques can find novel patterns that may assist an enterprise in understanding the business better and in forecasting [17]. Many data mining techniques are closely related to some of the machine learning techniques that have been developed over the last 50 years. Others are related to techniques that have been developed in statistics, sometimes called exploratory data analysis. These techniques were developed some time ago and were designed to deal with a limited amount of data. The techniques have now been modified to deal with large amounts of data. Yet other techniques are relatively new, for example Web data mining [17].A possible list of Data Mining Techniques is [20]:

- **Classification**
  - Decision Tree based Methods
  - Rule-based Methods
  - Memory based reasoning
  - Neural Networks
  - Genetic Algorithms
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- **Clustering**
- **Association Rules**
- **Sequential Patterns**
- **Regression**

- **Deviation Detection**

Below a brief description of this techniques:

- **Classification**:given a collection of records (training set) where each record contains a set of attributes, one of the attributesis the class; find a model for class attribute as a function of the values of other attributes.Below a list of the methods of classification:
  - Decision Tree based Methods
  - Rule-based Methods
  - Memory based reasoning
  - Neural Networks
  - Genetic Algorithms
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- **Clustering**:is a technique to group the similar data into a cluster and dissimilar data into different clusters[21].
- **Association rules**:a pattern is discovered based on a relationship of a particular item on other items in the same transaction[22].
- **Sequential Patterns**: finds statistically relevant patterns between data examples where the values are delivered in a sequence.
- **Regression**: the regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables.
- **Deviation Detection**: aims to identify any outliers in the data.

The four areas that contributed to the growth of data mining in its current form are Artificial Intelligence, Machine Learning, Statistics Databases [18]. Data Mining is being used for a wide variety of applications. Below a list of Data Mining current trends and applications [17]-[18]: 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. Here we summarize the main functions of data mining.Data Mining allows you to attribute importance to the attributes, that is identifies the attributes that are most important in predicting a target attribute (attribute importance); assigns items to discrete classes and predicts the class to which an item belongs (classification); approximates and forecasts continuous values (regression).

These functions belong to the category of supervised learning of data mining. Supervised learning is also known as directed learning. Directed data mining attempts to explain the behavior of the target as a function of a set of independent attributes or predictors. As part of unsupervised learning, data mining have the following functions: Identifies items (outliers) that do not satisfy the characteristics of "normal" data (anomaly detection); finds items that tend to co-occur in the data and specifies the rules that govern their co-occurrence (association rules); Finds natural groupings in the data (clustering); creates new attributes (features) using linear combinations of the original attribute (feature extraction).

5. SEARCH PARADIGMOVERVIEW

Exploratory Data Analysis, or EDA for short, is a term coined by John W.Tukey in the book “Exploratory Data Analysis” in 1977 [24]. 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” [25].

In 2006, Marchionini [1] postulate 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 state, 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 an 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 [26].

More recently, the research comes back with a new paradigm for access to rich data set, Exploratory Computing. Using this new paradigm, have been developed some Exploratory Portal in several field of interest (archeology, tourism, education, etc. [2]-[3]-[4]). The Exploratory Computing approach as explained in [5] and in its manifesto [6] allows users to investigate of complex dataset composed of rich information. The user can interact with the data and can discover information features that he/she didn’t 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.

6. DISCUSSION

This analysisallows us to take some consideration about these techniques.The identification of the main features of an ideal data exploration system in the techniques of faceted search and data miningcomes from the analysis of the literature and the use of the related systems. A comparison between facet search and data mining is in table1:

Table 1: Comparison between techniques

Main features of an ideal data exploration system	**Faceted Search**	**Data Mining**
Categories search	yes	yes
Set Exploration	no	yes
Interactive process	no	no
Investigation and inspiration seeking	yes	no
Leisure browsing and learning (serendipity)	no	no
Supervision and decision-making	yes	yes
Set Comparison	yes[1]	yes
Correlation between categories	no	yes
Complex answer to simple query	no	yes

From the above comparison, we can see that each technique has several features but not all. To achieve an effective interactive data exploration is needed the use of multiple techniques together through a skillful combination of them. However, the major limitation is in the taxonomies: in the existing systems [5,7] the taxonomies are still created manually by domain experts and this is not good for time consuming and high labor cost.A method to automate the creation of taxonomies is necessary.

7. CONCLUSIONS

In this paper we compare the techniques of data exploration: data mining and faceted search. This analysisallows us to take some important consideration about this research field.We have identified the mainfeatures for an ideal data exploration systemthat allows the user to have a new and more interesting navigational experience and we have highlighted what techniques meet these main featuresand also hypothesized a new requirement, the use of ontologies, to obtain better results from a data exploration.In the next step of our research we plain to make a more deep comparison using existing systems made up by faceted search approach and Data mining approach. Also, a method to automate the creation of taxonomies is necessary. Another aspect to investigate is the use of a knowledge base (ontology) for to support the taxonomies: that would can improve understanding of the domain at inexpert users. The use of ontologies, that provide terms for describing our knowledge about the domain, allows us to obtain better results from a data exploration.

References

[1] G. Marchionini, “Exploratory search: from finding to understanding,” Commun. ACM, vol. 49, no. 4, p. 41, 2006.

[2] N. Di Blas, A. Fiore, L. Mainetti, P. Paolini, R. Vergallo (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)

[3] N. Di Blas, P. Paolini, L. Spagnolo (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

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

[5] 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.

[6] Di Blas, N., Mazuran, M., Paolini, P., Quintarelli, E., &Tanca, L. (2014, October). Exploratorycomputing: a draftManifesto. In Data Science and AdvancedAnalytics (DSAA), 2014 International Conferenceon (pp. 577-580). IEEE.

[7] BifanWei, Jun Liu, QinghuaZheng, Wei Zhang, Xiaoyu Fu, and Boqin Feng. 2013. A survey of facetedsearch. J. Web Eng. 12, 1-2 (February 2013), 41-64

[8] Ranganatan, S.R. Elements of libraryclassification (1st ed). 1991, Bombay, New York: South Asia Books. 168p.

[9] Spiteri, L. A simplifiedModelforFacetAnalysis. CanadianJournal of Information and Library Science, 2008. 23(1-2) p.1-30

[10] Ben-Yitzhak, O., et al., Beyondbasicfacetedsearch, in Proceedings of theinternationalconferenceon Web search and web data mining. 2008, ACM: Palo Alto, California, USA. p. 33- 44.

[11] Dachselt, R., Frisch, M., and Weiland, M., FacetZoom: a continuousmulti-scalewidgetfornavigatinghierarchicalmetadata, in Proceeding of thetwenty-sixthannual SIGCHI conferenceon Human factors in computingsystems. 2008, ACM: Florence, Italy. p. 1353-1356.

[12] Stoica, E., Hearst, M.A., and Richardson, M., AutomatingCreation of HierarchicalFacetedMetadataStructures. Proceedings of the Human LanguageTechnologyConference of the NorthAmerican Chapter of theAssociationforComputationalLinguistics (HLT-NAACL), 2007: p.244-251

[13] Ling, X., et al., Mining multi-facetedoverviews of arbitrarytopics in a textcollection, inProceeding of the 14th ACM SIGKDD internationalconferenceonKnowledgediscovery anddata mining. 2008, ACM: Las Vegas, Nevada, USA. p. 497-505

[14] Tunkelang, D., DynamicCategory Sets: AnApproachforFacetedSearch, in Proceedings of the ACM SIGIR'06 WorkshoponFacetedSearch. 2006: Seattle, WA, USA.

[15] Vladimir Zelevinsky, Breakingdowntheassumption of facetsearch HCIR 2010 Bridging Human-ComputerInteraction and InformationRetrievalWorkshop in conjuctionwithIIiXSun, 22 Aug 2010, New Brunswick, NJ

[16] Mukhopadhyay, Anirban, et al. "A survey of multiobjectiveevolutionaryalgorithmsfor data mining: Part I." EvolutionaryComputation, IEEE Transactionson 18.1 (2014): 4-19, 20-35.

[17] Gupta, G. K. Introductionto data mining with case studies. PHI LearningPvt. Ltd., 2014.

[18] Ramzan, Majid, and Majid Ahmad. "Evolution of data mining: Anoverview." IT in Business, Industry and Government (CSIBIG), 2014 Conferenceon. IEEE, 2014.

[19] Maulik, Ujjwal, Lawrence B. Holder, and Diane J. Cook. Advancedmethodsforknowledgediscoveryfromcomplex data. SpringerScience& Business Media, 2006.

[20] SrivastavaJaidepp, Web Mining: Accomplishements&FutureDirections – http://ieee.org.ar/downloads/Srivastava-tut-pres.Pdf

[21] Verma, Manish, et al. "A comparativestudy of variousclusteringalgorithms in data mining." International Journal of EngineeringResearch and Applications (IJERA) Vol 2 (2012): 1379-1384.

[22] Diwate, Rahul B., and AmitSahu. "Data Mining Techniques in Association Rule: A Review." International Journal of ComputerScience&Information Technologies 5.1 (2014).

[23] Buoncristiano, M., Mecca, G., Quintarelli, E., Roveri, M., Santoro, D., &Tanca, L. Exploratory Computing: What is there for the Database Researcher?

[24] Tukey, John W. "Exploratory data analysis." (1977): 2-3

[25] Behrens, John T., and Chong Ho Yu. "Exploratory data analysis." Handbook of psychology (2003)

[26] White, R. W., Muresan, G., &Marchionini, G. (2006, December). Reporton ACM SIGIR 2006 workshoponevaluatingexploratorysearchsystems. In ACM SIGIR Forum (Vol. 40, No. 2, pp. 52-60). ACM.

[27] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm