promise

# PROMISE

Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation

FP7 ICT 2009.4.3, Intelligent Information Management

# Deliverable 5.2
# User interface and Visual analytics environment requirements

Version 1.0, 31st August 2011

SEVENTH FRAMEWORK PROGRAMME

Network of Excellence co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191

# Document Information

| | |
|---|---|
| **Deliverable number:** | D5.2 |
| **Deliverable title:** | User interface and Visual analytics environment requirements |
| **Delivery date:** | 31/08/2011 |
| **Lead contractor for this deliverable** | ROMA1 |
| **Author(s):** | Martina Croce, Emanuele Di Buccio, Emiliano Di Reto, Marco Dussin, Nicola Ferro, Guido Lorenzo Granato, Preben Hansen, Mihai Lupu, Mirko Perlorca, Alessio Pronesti, Alessandro Sabetta, Giuseppe Santucci, Gianmaria Silvello, Giuseppe Tino, Theodora Tsikrika |
| **Participant(s):** | All partners |
| **Workpackage:** | WP5 |
| **Workpackage title:** | Collaboration and Knowledge sharing |
| **Workpackage leader:** | ROMA1 |
| **Dissemination Level:** | PU – Public |
| **Version:** | 1.0 - Final |
| **Keywords:** | Visual Analytics, collaboration. knowledge sharing, user requirements |

# History of Versions

| Version | Date | Status | Author (Partner) | Description/Approval Level |
|---|---|---|---|---|
| 0.1 | 27/07/2011 | Draft | Martina Croce, Emiliano Di Reto, Guido Lorenzo Granato, Mirko Perlorca, Alessio Pronesti, Alessandro Sabetta, Giuseppe Santucci, Giuseppe Tino | Circulated among Roma1 members |
| 0.2 | 13/08/2011 | Draft | Giuseppe Santucci | Circulated to PROMISE reviewers |
| 1.0 | 27/08/2011 | Final | Giuseppe Santucci | Final version, incorporating PROMISE reviewers' comments |

# Abstract

This deliverable reports the requirements for the PROMISE Visual Analytics module, which aims at improving the usage and comprehension of the large amount of data collected during the information retrieval evaluation campaigns.

# Table of Contents

# Executive Summary

Work package 5 (“Collaboration and Knowledge Sharing”) is responsible for designing, developing, and delivering the user interfaces and the annotation service needed to promote the collaboration among the stakeholders of the evaluation infrastructure and foster the knowledge sharing and reuse. Moreover, it is responsible for exploring how to apply information visualization and visual analytics techniques to information retrieval experimental data in order to improve their understanding and allow researchers to effectively cope with huge amount of data.

This deliverable focuses on the PROMISE Visual Analytics component, discussing the kind of data managed by information retrieval evaluation campaigns and the way in which it is possible to analyze such data through suitable visualizations and algorithms.

The deliverable describes a general Visual Analytics environment, specializing it to the actual PROMISE requirements by taking into consideration the work carried out in D2.1 “Initial specification of the evaluation tasks”, D3.1 “Initial prototype of the evaluation infrastructure”, D3.2 “Specification of the evaluation infrastructure based on user requirements”, and D5.1 “Collaborative user interface requirements”.

User requirements are formalized using the IEEE 830-1998 Recommended Practice Software Requirements Specifications standard and the UML Use cases.

These requirements will be the basis for the next deliverables, namely the two prototypes D5.3 Collaborative User Interface Prototype with annotation functionalities (due on M24) and D5.4 Revised Collaborative User Interface Prototype with annotation functionalities (due on M36).

The appendix reports the needed background information about the chosen visualizations.

# 1 Introduction

This deliverable reports the PROMISE user interface and Visual Analytics component requirements, encompassing knowledge sharing and collaboration issues. Requirements are described with UML Use Cases (see PROMISE deliverable 5.1 - Appendix 8.4 for a quick guide about UML Use Cases) and using a textual template derived by the "IEEE standard 830-1998 - Recommended Practice for Software Requirements Specifications" (See PROMISE deliverable 5.1 - Appendix 8.3 for more details about such a standard).

The deliverable is structured as follows. Section 2 provides an introduction about Visual Analytics and Section 3 presents the information retrieval evaluation state of the art with respect to current data analysis, focusing on data structure and common visualizations.

Section 4 describes the overall Promise visualization and analysis requirements and introduce the Visual Analytics component architecture and functionalities, distinguishing, on Sections 4.2.1.2 and 4.2.1.3, between the general Visual Analytics functionalities and the ad-hoc PROMISE predefined tasks. Section 5 formalizes the requirements through UML Use Cases. The appendix provide an overview of the visualizations that will be implemented in the Visual Analytics component.

# 2 Visual Analytics

Today it is possible to store larger and larger amount of data but is not yet possible to analyze them. Around the year 2000, for the purpose of supporting human beings in analyzing data, synergies between Information Visualization (IV) and Data Mining started to be considered, defining Visual Data Mining (VDM) as a new area focused on the explorative analysis of visually represented data. In 2001, the first VDM workshop was held in Freiburg. In 2004, first in the United States, and almost at the same time in Europe, researchers started talking about Visual Analytics [WON04]. Compared to VDM, there is the clear intention to focus on the analysis process that leads to explanation, interpretation and presentation of hidden information in the data, taking advantage of dynamic visualizations. From that moment on, the term VDM is superseded by the term Visual Analytics (VA). Daniel Keim, one of the major European experts in the field, provides the following definition "Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, human factors and data analysis. ..."

On a grand scale, Visual Analytics provides technology that combines the strengths of human and electronic data processing. Visualization becomes the medium of a semi-automated analytical process, where humans and machines cooperate using their respective distinct capabilities for the most effective results. The user has to be the ultimate authority in giving the direction of the analysis along his or her specific task. At the same time, the system has to provide effective means of interaction to concentrate on this specific task since in many applications different people work along the path from data to decision. A visual representation will sketch this path and provide a reference for their collaboration across different tasks and abstraction levels.

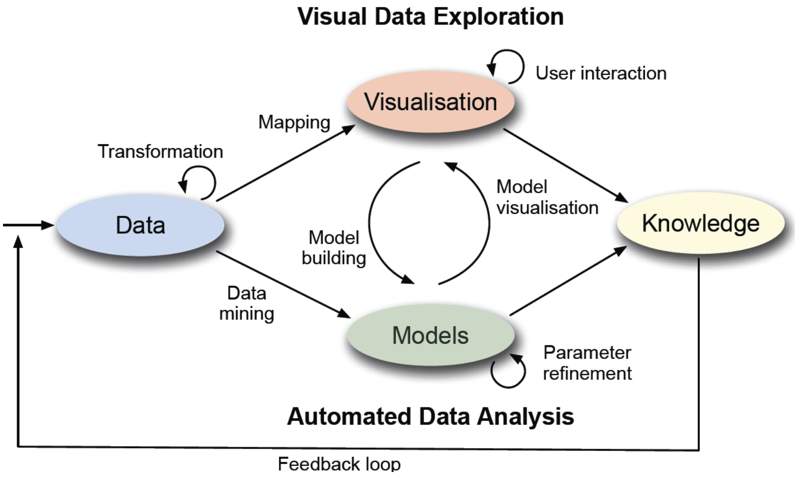


Figure 1: The Visual Analytics process

Figure 1 [AAV2010] schematizes the visual analytics process that combines automatic and visual analysis methods with a tight coupling through human interaction in order to gain knowledge from data. The figure shows an abstract overview of the different stages (represented through ovals) and their transitions (arrows) in the visual analytics process.

The first step is often to preprocess and transform the data to derive different representations for further exploration (as indicated by the Transformation arrow). Other typical preprocessing tasks include data cleaning, normalization, grouping, or integration of heterogeneous data sources.

After the transformation, the analyst may choose between applying visual or automatic analysis methods. Alternating between visual and automatic methods is characteristic for the visual analytics process and leads to a continuous refinement and verification of preliminary results. User interaction with the visualization is needed to reveal insightful information, for instance by zooming in on different data areas or by considering different visual views on the data. In summary, in the visual analytics process, knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts.

With respect to the field of visualization, visual analytics integrates methodology from information analytics, geospatial analytics, and scientific analytics. Especially human factors (e.g., interaction, cognition, perception, collaboration, presentation, and dissemination) play a key role in the communication between human and computer, as well as in the decision-making process." [KMS06]. One of objectives of VA is, thus, to refine the techniques born in the IV context taking care of the cognitive, qualitative and scalability aspects.

Aside from the visualizations previously mentioned, most IV attention has focused on multidimensional data visualization, which is especially relevant in the VA context. A well-known taxonomy of these techniques is in [KK96]. In this context, representation scalability the definition of techniques and metrics that are able to evaluate the quality of a representation appear to be extremely relevant [CH05].

In [ED07] a description of general techniques that permit to partially solve the scalability problem is given, but most of them were not devised explicitly for that problem. In [ED07] sampling techniques such as an explicit solution for this problem are proposed, pointing out that sampling has the advantage of significantly reducing data dimension, without compromising their representation. In [BS06b] Bertini and Santucci introduce non-uniform sampling techniques to reduce, in a selective way, the density of big amounts of data; a sampling led by quality metrics and perceptual considerations. Ware [WA04] has shown how sensorial/pre-attentive processes can influence the type of information that the user understands, thus guiding our comprehension of representations of multivariate discrete data.

There are several studies in cognitive psychology that have tried to analyze which are the fundamental perceptual elements determining a good comprehension of graphs and tables [SP90; SR96; ZT99].

In addition to the study of these variables, different authors [GWC98; EBB05; A06] have pointed out that it is fundamental to understand the role of superior order cognitive factors, like memory or attention (measured also using eye tracking devices), or users' expertise (so has to understand which users' characteristics are fundamental to the understanding of these elements).

# 3 Information retrieval evaluation data analysis

The goal of this section is to describe the information retrieval evaluation state of the art with respect to current data analysis. It is based on the elaboration of the information collected during the WP5 joint meeting hold in Rome on 16-17 May, 2011.

If we want to apply Visual Analytics to PROMISE in order to find solutions that improve visualizations, analysis, and interpretations of experimental data, a preliminary study is needed to understand both the data structures that are actually used within PROMISE community and its usage. According to this aim, this section is divided into two parts:

- data analysis, which has the goal of understanding how which data is useful for the PROMISE activity and how it is organized (Section 3.1);
- visualization analysis, which has the goal of understanding what are the most common visualization patterns and the underlying data feeding them (Section 3.2).

## 3.1 PROMISE data

In order to understand how data are organized and displayed it is important to define a typical scenario in which these data are used. In particular, we focus on information retrieval evaluation campaigns.

An evaluation campaign is an activity intended to support researchers in information retrieval by providing a large test collection and uniform scoring procedures. Within an evaluation campaign there are many tracks like multimedia, multilingual, text, images, and so on. A track includes, in turn, several tasks. A task is used to define the experiment structure specifying a set of documents, a set of topics, and a relevance assessment. For each task the set of document can be structured defining for example a title, keywords, images, and so on.

Some ad-hoc metadata allows for partitioning the set of documents. For example, in the same set we can have European or American documents and a mechanism that allows for choosing only one of these sets. Moreover, it is important to remark that very often in an evaluation campaign the so called closed world assumption holds, which means that the set of documents is finite and known a-priori.

A topic represents an information need. It is structured and its structure can change according to the task at hand. Documents can be assessed as being relevant or not (or more or less relevant) for a given information need (topic). The relevance of a document with respect to a specific topic is independent of the other documents in the collection, based solely on the qualities of that document. In some case we can have different sets of relevance assessment for a set of documents. The relevance assessment can be done manually, automatically, or using online approach like Amazon mechanical Turk.

Basically, an evaluation campaign involves two kinds of actors: organizers and participants. Organizers prepare the campaign establishing, among other things, tracks and tasks. Participants run their searching algorithm(s) according to the actual tasks. Each run produces a (ranked) result set on which different metrics are computed and stored in different kind of tables. In the following we use the terms run and experiment to indicate the systematic application of an algorithm to a task. The computed metrics can be used by organizers or participant. In general, organizers are interested in evaluating the whole campaign, while participants are interested in evaluating their own algorithms, comparing them with algorithms of other participants.

Having introduced these basic notions we can analyze the PROMISE data. We will refer to data which are stored in the DIRECT system developed by Padua University (see PROMISE Deliverable 3.2 Specification of the evaluation infrastructure based on user requirements, Section 6, for details about accessing such data). These data can be represented by the TME (Topics-Metrics-Experiment) cube shown on Figure 2.

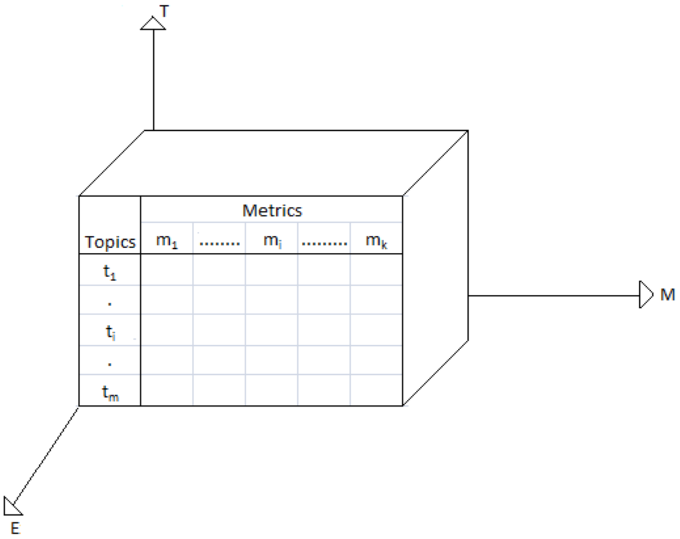


**Figure 2: The PROMISE TME Data cube**

Starting from this cube, we can aggregate or manipulate data in different ways, according to our needs. In particular we are interested in computing four kinds of tables.

The first kind of table describes **a single experiment** *e* in terms of **topics** and **metrics** and is a projection of the TME cube on the Topics-Metrics axes. In particular, this table is represented by a matrix T x M where T is the set of topics and M is the set of metrics. In the following we will refer to this kind of tables as TM(e) tables (topics x metrics table of experiment e). In order to have statistically relevant measures, at least 50 topics are needed (in some case 25 are still enough). However, in the patent domain the number of topics can be up to 4000. This is possible because in this field there is no relevance judgment which is a bottleneck in an evaluation campaign since it is performed by human beings and takes a long time. The number of topics can be increased reducing the set of documents. Metrics are in general 100, but this number can increase or decrease.

Still considering the TME cube we can derive a second kind of tables (see Figure 3), useful to analyze **a single metric** *m* in terms of **topics** and **experiments**. In particular, this table is represented by a matrix T x E where T is the set of topics and E is the set of experiments. In the following we will refer to this kind of tables with the name TE(m) tables (topics x experiments table of metric m). Comparisons are made along rows, to evaluate the behavior of a single topic, or among columns to compare two or more experiments. For the number of topics the same considerations previously discussed hold. The number of experiments depends on how many algorithms are compared.

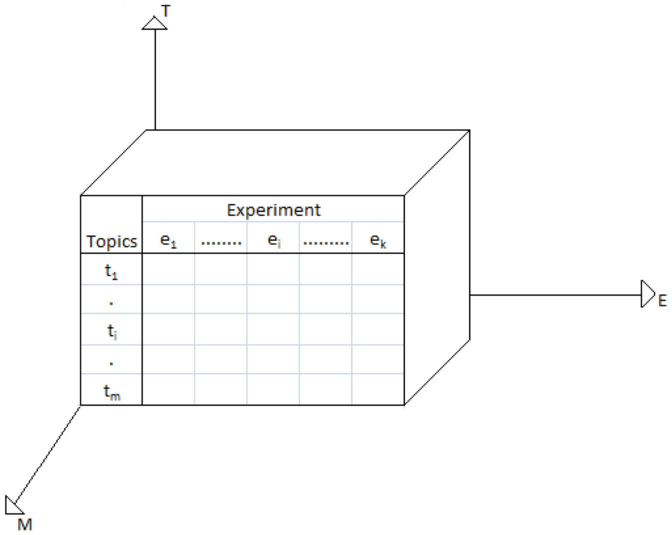


**Figure 3: Projecting the TME data cube on the Topics-Experiments axes**

The third kind of table describes **a single experiment** *e* in terms of **descriptive statistics** and **metrics** In particular, this table is represented by a matrix S x M where S is the set of statistics and M is the set of metrics. In the following we will refer to this kind of table with the name SM(e) table (statistics x metrics table of experiment e). This table is strictly related on the corresponding TM(e) table since values are computed from the TM(e) table's columns. Figure 4 shows an example of how a TM(e) table can be used to calculate values of the SM(m) table.

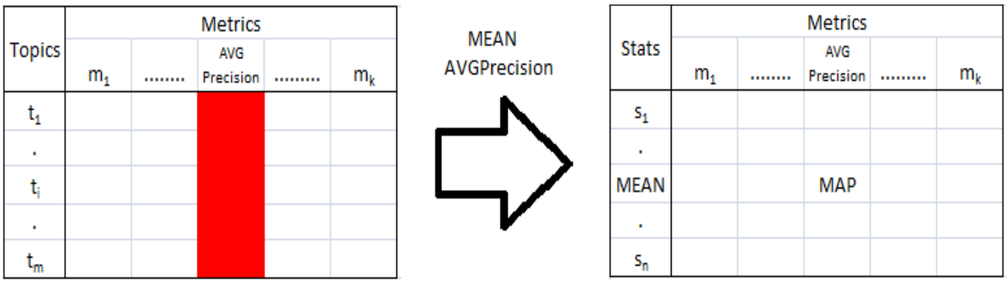


**Figure 4: Relation between TM and SM tables**

As shown on Figure 4 in an SM table there is the same number of metrics as the related TM table. The number of statistics is about 50. While metrics can increase or decrease, the number of statistics is quite stable. If we extend this table with respect to experiments we obtain a new cube, the SME (Statistics-Metrics-Experiment) data cube, shown on Figure 5. With respect to the SME cube an SM table is a projection on the Statistics-Metrics axes.

The last kind of table we consider, allows to inspect **a single metric** *m* in terms of **descriptive statistics** and **experiments**, i.e., it allows for comparing different experiments against a some descriptive statistics computed on a given metric. In particular, this table is represented by a matrix S x E where S is the set of statistics and E is the set of experiments. In the following we will refer to this table as SE(m) table (statistics x experiment table computed on metric m), and it is a projection of the SME cube on the Statistics-Experiments axes (see Figure 5a).

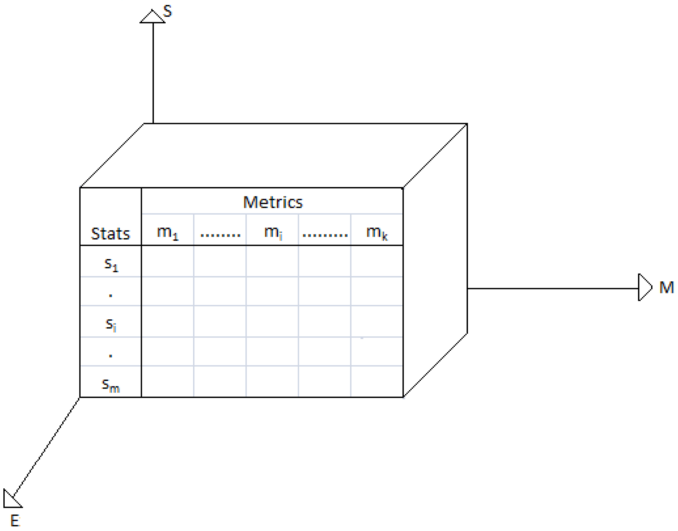


**Figure 5: The SME Data cube**

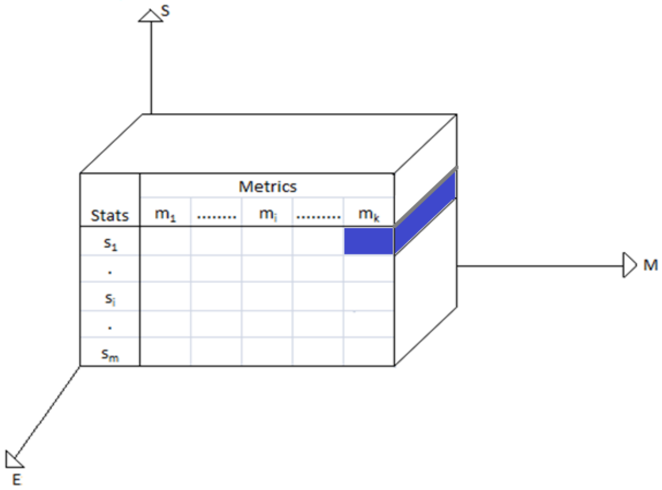


**Figure 5a: The SME Data cube projected on the Statistics-Experiments axes**

To complete our analysis we recall the concept of meta-attribute. A meta-attribute is a categorical attribute that is associated with a cube component (for example the experiments) and it is used to define a further classification of data, with respect to a category. Examples of meta-attributes are: reference track, year, and type of search. Meta-attributes are mainly associated to experiments and documents (see PROMISE Deliverable 3.2 Specification of the evaluation infrastructure based on user requirements), but also topics can have their own meta-attributes (for example the provenance data). Actually there are no meta-attributes defined on metrics. A possible meta-attribute for metrics is the scope of metric, but we will consider this categorization for future developments.

## 3.2 PROMISE visualizations

In this section we present an overview of some significant PROMISE visualizations. The analysis is performed referring the visualizations to the data described in Section 3.1.

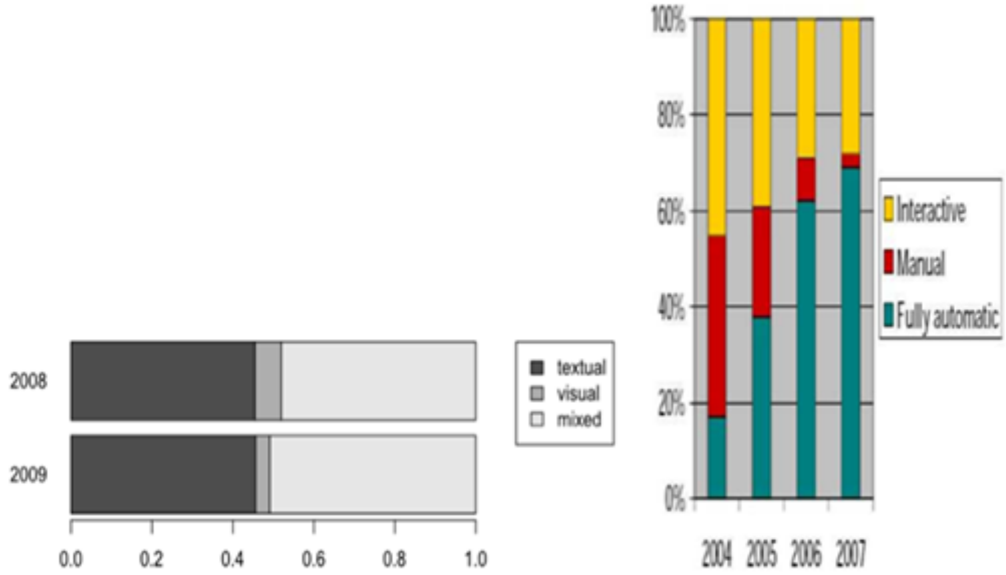


**Figure 6: Experiment analysis across different years**

Figure 6 shows two stacked bar charts (see Appendix, Section 6.2.7) which represent two descriptions of experiments based on different categorical meta-attributes. In particular, in the leftmost chart data are categorized using “type of search” and year; in the rightmost one data are categorized using “type of experiment” and year.

The chart on the left shows the proportion between textual search experiment, visual search experiment and mixed search experiment in 2 different years. It is clear that visual search experiments are quite a minority and that the proportion of the different kinds of searches is quite the same across 2008 and 2009.

The chart on the right shows how the way of managing experiments is changed during years. It is easy to see that the “manual” component decreases from 2004 to 2007, while the "fully automatic" component increases. Both charts describe a proportion, even if the actual visualization does not allows for perceiving precise percentage values.

The information obtained by these charts is useful for organizers.

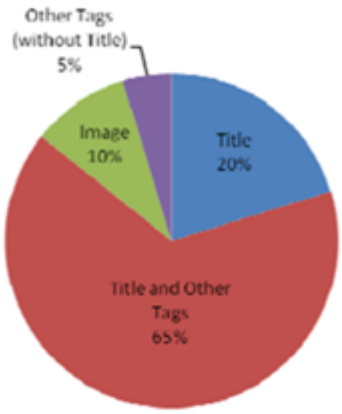


**Figure 7: Analysis of query field usage**

The above image shows a pie-chart that provides information about proportion of query field usage across a set of experiments. The query field is a categorical meta-attribute which is associated to experiments. In 10% of cases images are used as query fields, while title is used in 20% of cases; but the most used way is a composition of "Title and other tags" (65 %).

The information obtained by these charts is useful for organizers.

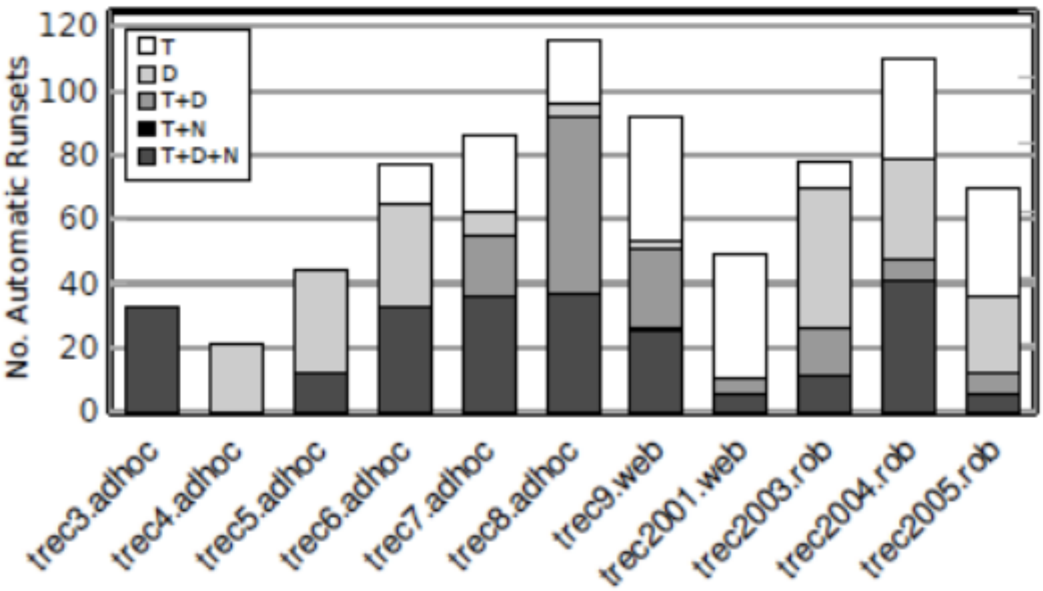


**Figure 8: Analysis of topic field usage**

Figure 8 shows a stacked bar chart that represents, for each “TREC track”, the number of automatic runsets, showing the breakdown (proportion) obtained by considering the combination of the categorical meta-attribute "topic field" used for the query generation. In particular we are considered three topic fields: **T**itle, **D**escription, **N**arrative, and their combinations: Title, Description, Title + Description, Title + Narrative, Title + Description + Narrative, coded with five colors. This representation can be useful for organizers.

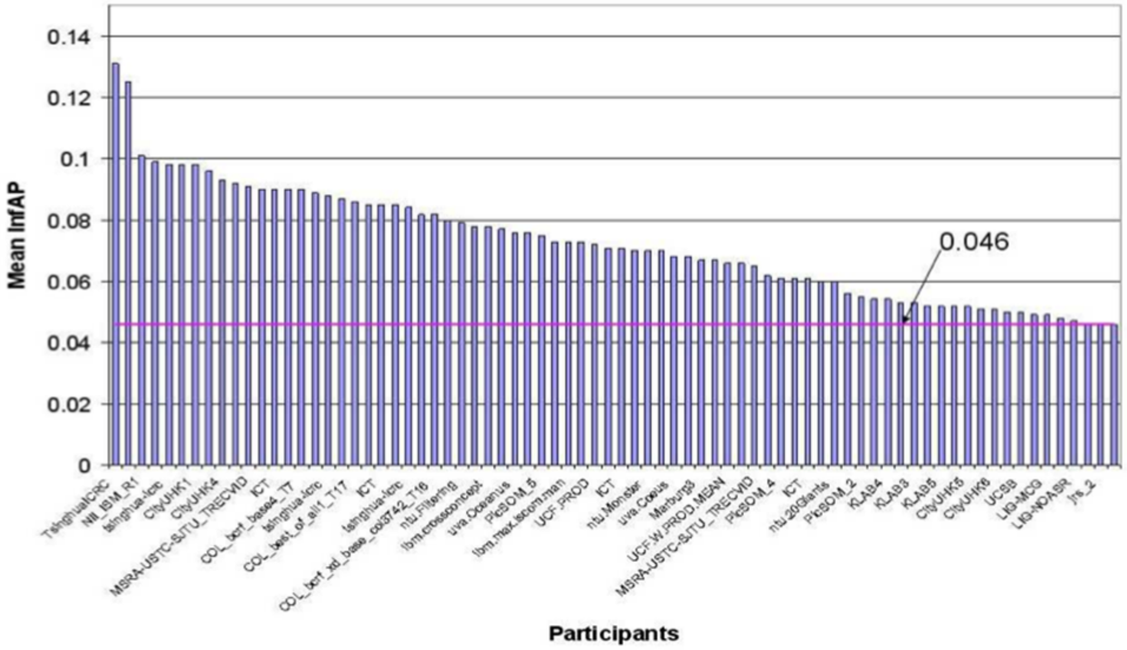


**Figure 9: Comparison of several experiments against the Mean InfAP**

The bar chart on Figure 9 visualizes data coming from a row of a SE(InfAP) table: the x-axis represents a list of experiments and the y-axis a descriptive statistics (in this case, the Mean InfAP). Values are arranged in descending order with respect to the mean inferred average precision. The red line shows the minimum value of such a statistics.

This chart interests participants because it gives to the possibility to compare their performance with other ones and to appreciate the distance with respect to minimum and maximum values.

To generalize this visualization we have to do the following:

**Input:** The SE(InfAP) table

**Mapping:** experiments on x axis, chosen descriptive statistics (Mean) on y axis.

**Step 1:** plot the sorted values of the row corresponding to the chosen statistics (Mean).

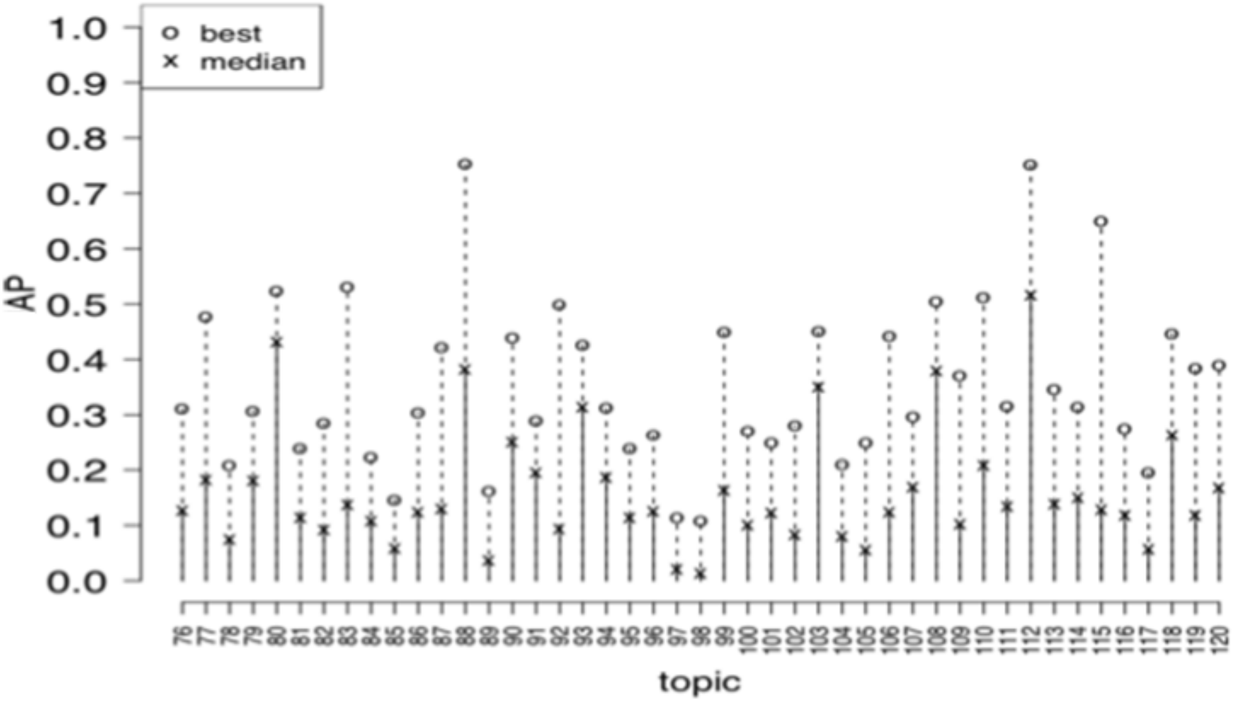


**Figure 10: Analysis of performance of the experiment on the available topics using the AP metric**

The boxplot chart on Figure 10 represents an elaboration of the data which are in a TE(AP) table. In particular, there is a simplified boxplot for every topic displaying only the median and the highest of average precision (AP) across a set of experiments. This visualization interests both participants and organizers, providing an overview of a task performances. To generalize this visualization we have to do the following:

**Input:** The TE(AP) table
**Mapping:** topics on x axis, chosen metric boxplots on y axis.
**Step 1:** for each row (topic) calculate the highest value and the median of all experiments
**Step 2:** display with the chosen style the boxplots (here different markers are used for representing medians and highest values).

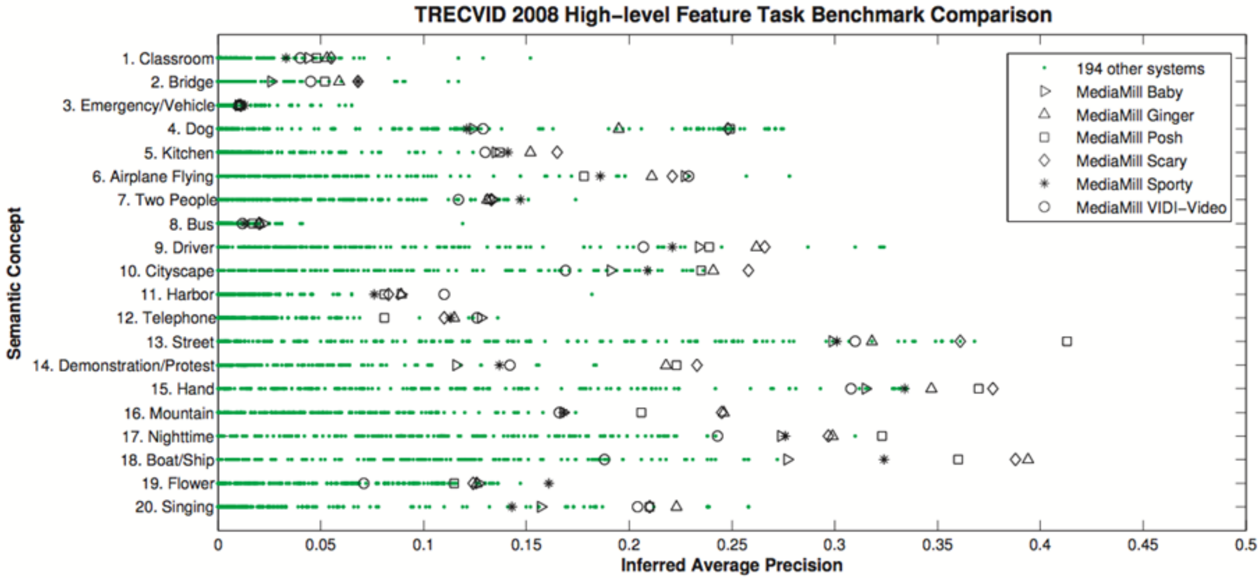


**Figure 11: Comparison between a subset of experiments and all the task experiments using the Inferred AP metric**

The figure on Figure 11 represents an elaboration of a TE(InfAP) tables. In particular, this visualization is a scatter plot which is used to compare, for each topic (Y axis), the inferred average precision of a specific set of experiment (marked with different symbols) with all the other experiments (represented by green dots). This visualization interests participants because allow them to compare the performance of their algorithms with other participants' runs.

To generalize this visualization we have to do the following:
**Input:** the TE(InfAP) table
**Mapping:** topics on y-axis, chosen metric on x-axis (switching the axes will produce a more common visualization).
**Step 1:** choose one or more runs to be highlighted with respect to the other ones.
**Step 2:** plot, for each topic, all the elements in the corresponding row, highlighting the experiments selected at step 1.

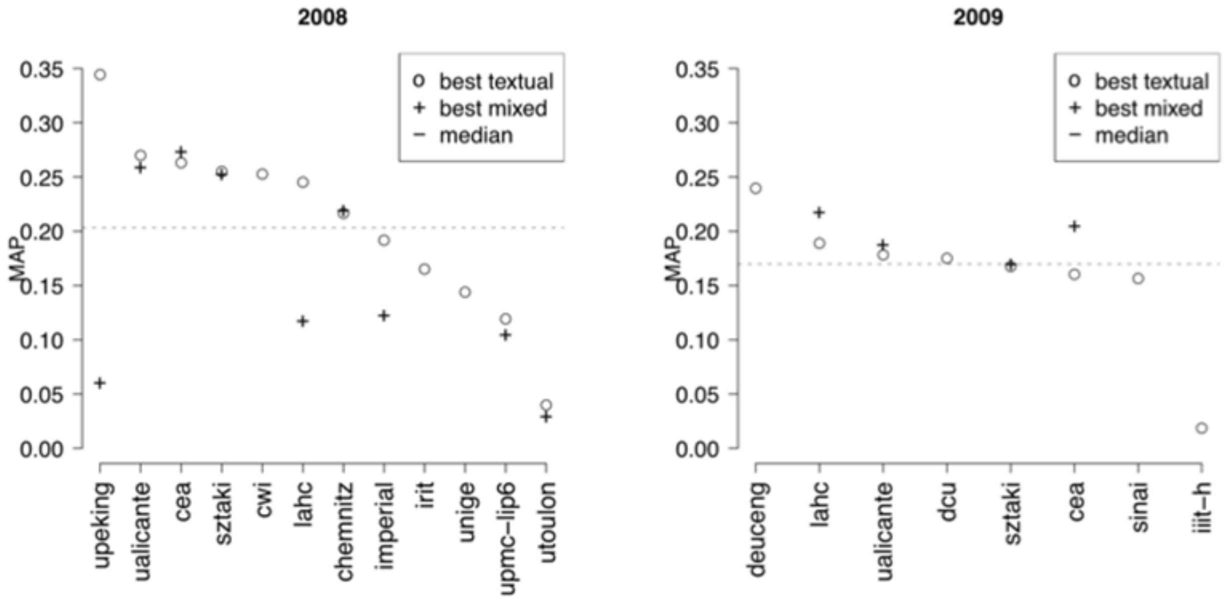


**Figure 12: Analysis of performance of the participants using the MAP statistics**

The simplified boxplot (only max values are displayed) chart on Figure 12 represents an elaboration of a SE(AP) table using meta-attributes: participant and kind of search. A search can be textual or mixed, and runs can be grouped per participants, selecting aggregate values (best values). Participants are mapped on the x axis and for each of them the scatterplot shows the best MAP values for both textual and mixed search experiment (note that in some cases one value is missing). Values are compared with the median and are displayed in decreasing order with respect to the textual search results. Note that the median takes into account also non-displayed results. Moreover values are presented for 2 consecutive years. This visualization interests participants because they can compare their best algorithms with the best algorithms of other participants. It is also possible to evaluate the difference between textual and mixed score and the progress from year to year.
To generalize this visualization we have to do the followings:
**Input:** the SE(AP) table
**Mapping:** participants on x axis ordered by best textual result, the boxplots on the y axis.
**Step 1:** choose the involved categorical attribute (participant, kind of search).
**Step 2:** group each row values using the categorical attribute selected at step, and compute the overall row median.
**Step 3:** display with the chosen style the multiple boxplots associated with each participant (here boxplots are collapsed on the max value, but in general we can display two full boxplots for each participant)

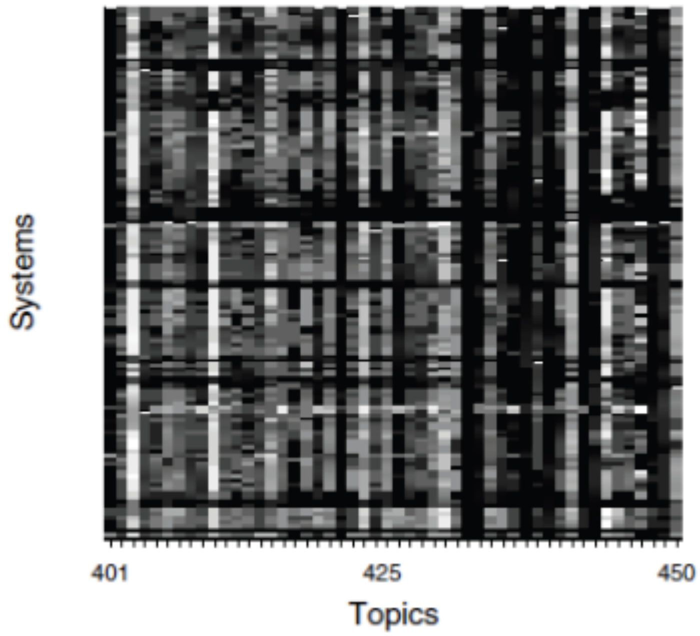


**Figure 13: Analysis of topics difficulties using the AVG precision**

The chart on Figure 13 represents an elaboration of data which are in a TE(AP) table. In particular, this is a bi-dimensional scatterplot which is used to show how much difficult is a topic for a system (algorithm) with respect to the average precision (AP). Topics are mapped on the x axis, ordered by number, and systems on the y axis, ordered by name. The difficult of a topic for a system is encoded with the intensity of color: white for minimum and black for maximum of AP values. This visualization interests participants because they can discover which topics create problems to their systems.

To generalize this visualization we have to do the followings:
**Input:** the TE(AP) table.
**Mapping:** topics on x-axis, systems on y-axis.
**Step 1:** choose a scale of color to encode the chosen metric values.
**Step 2:** Assign a shape to points (squares in the example) and display it.

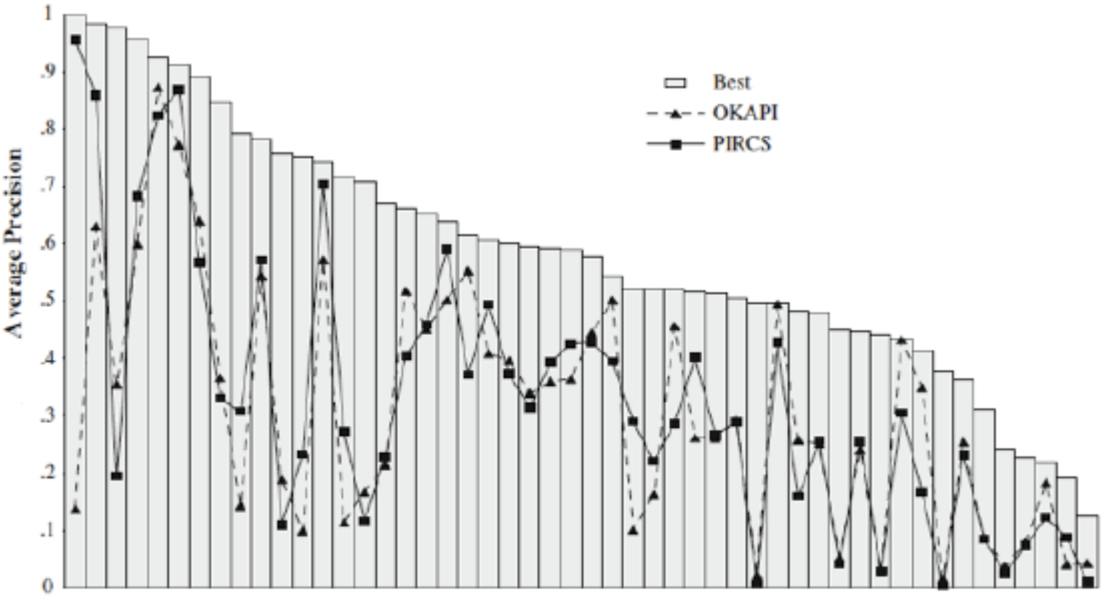


**Figure 14: Analysis of topics difficulties using the AVG precision**

The bar chart on Figure 14 represents data collected by a TE(AP) table. In particular, topics are on the x-axis and bars encode the best AP values for each topic. Moreover, different markers show the best AP values obtained by a subset (two) of participants. This visualization interests participants because they can compare their scores each other and/or with respect to the best result.

To generalize this visualization we have to do the followings:
**Input:** the TE(AP) table.
**Mapping:** topics on x-axis, chosen metrics on y-axis.
**Step 1:** for each row (topic) compute the max value
**Step 2:** select a subset of participants to highlight their best scores.
**Step 3:** group rows by selected participants compute the best value in each partition
**Step 4:** Display the best score for each topic with bars and the best score of the highlighted participants with markers and lines.

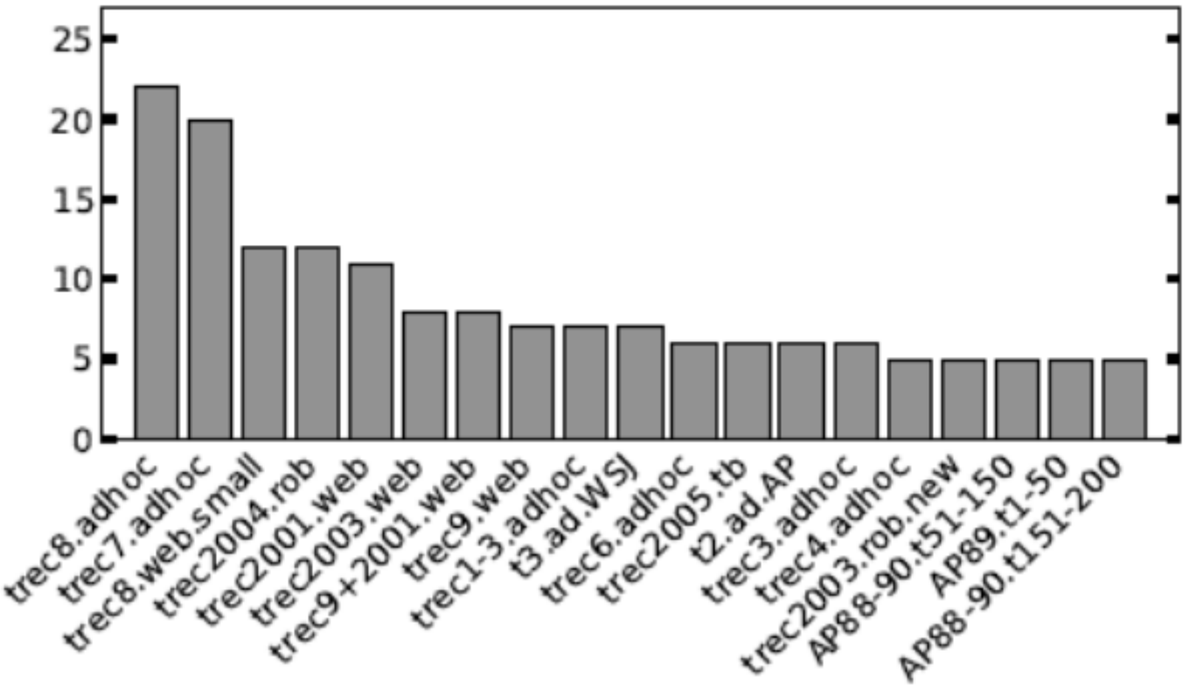


**Figure 15: Analysis of TREC test collection scientific impact**

The bar chart on Figure 15 represents the frequency distribution of usage of TREC test collections in scientific papers. Each bar shows how many papers reports at least one score for that exact collection. The x-axis contains the set of TREC collections, while on the y-axis is mapped the count of times that a collection appears in a paper. In the chart are shown only collections that are used in five or more publications.

This visualization interests organizers to evaluate a scientific impact of a track but also for participants who can see how tracks had a relevant scientific impact.

The following visualizations came from the patent evaluation campaigns that present some differences with respect to other evaluation activities. In particular, they use a controlled vocabulary for searching patents, the International Patent Classification classes, that is administered by the World Intellectual Property Organization (WIPO). The scheme was conceived as an indexing system to organize patent documents from around the world based on the technical field of the invention, thereby providing a retrieval system by subject matter, independent of keyword searching.

The basic structure of an IPC mark is a hierarchy with 8 major sections (A-H). These are:

- Section A – Human Necessities
- Section B – Performing Operations; Transporting
- Section C – Chemistry; Metallurgy

- Section D – Textiles; Paper
- Section E – Fixed Constructions
- Section F – Mechanical Engineering; Lighting; Heating; Weapons; Blasting
- Section G – Physics
- Section H – Electricity

The structure of an IPC classification is made up of a Section, Class, Subclass, Main Group, and Sub-Group: in the following visualization classes are specified only by Section and Class (example A01: section A and class 01).

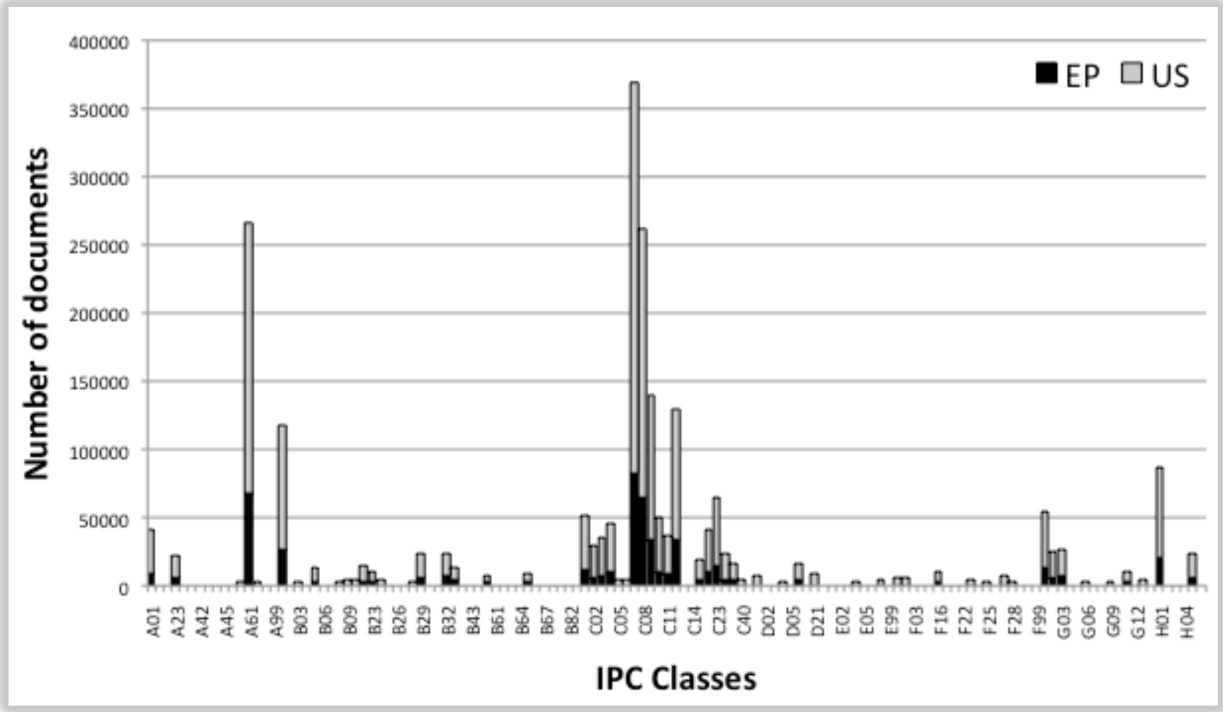


**Figure 16: Analysis document distribution across IPC classes**

Figure 16 is a stacked bar chart where on the X-axis is shown the list of IPC Classes and on the Y-axis there is the number of documents for each class; the color is used to distinguish documents by geographic regions: Europe or United States. It is useful for organizers.

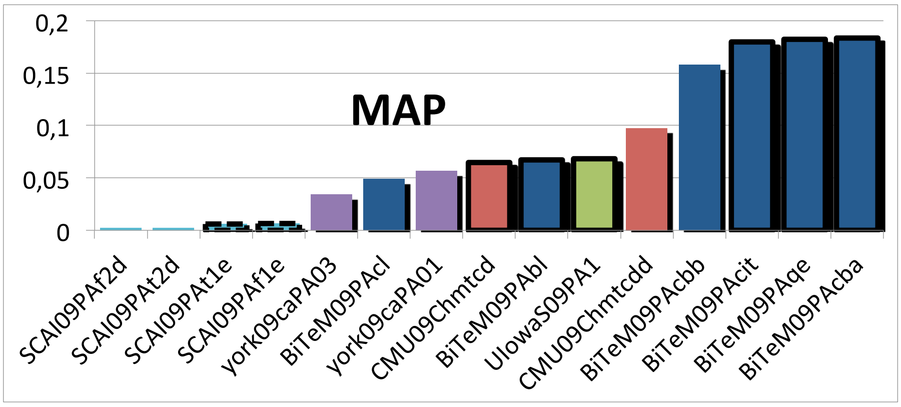
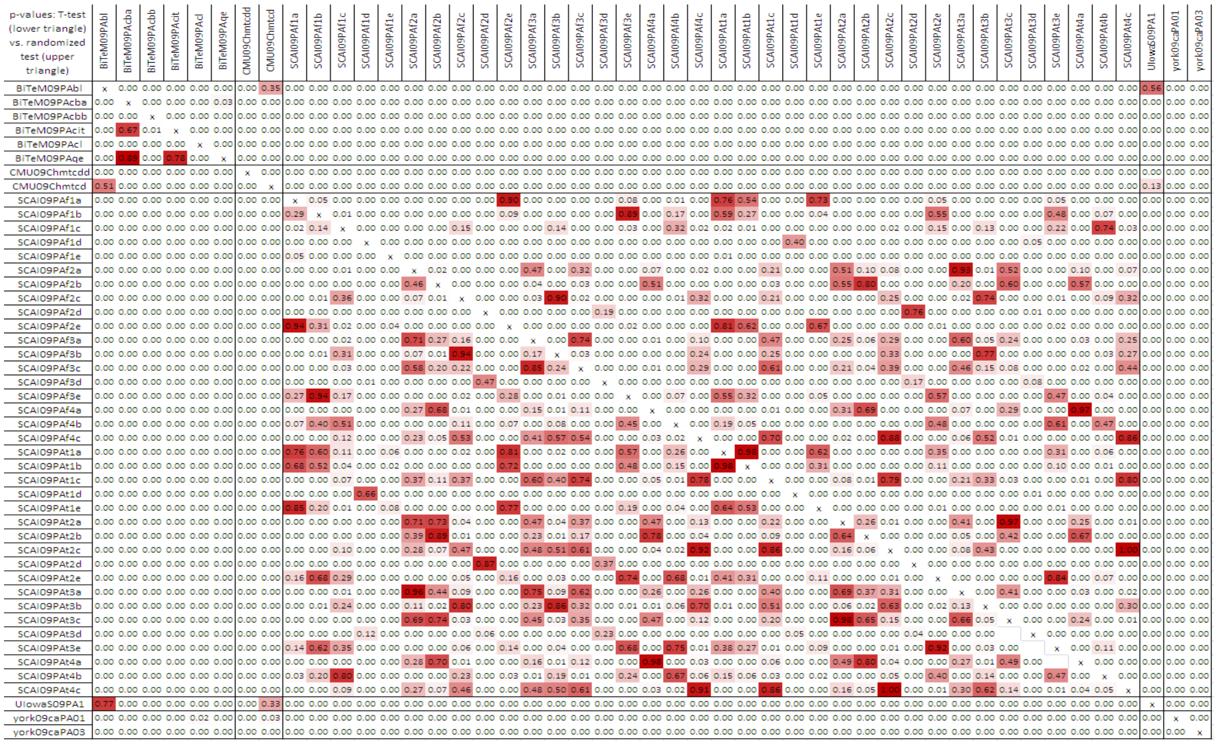


Figure 17: Comparison of 2009 TREC experiments using MAP

The bar chart on Figure 17 visualizes data from SE(AP) tables: the x-axis represents a list of experiments and the y-axis a descriptive statistics (in this case, the Mean AvgPrecision, MAP). Colors are used to represent different experiments of the same participant.

Values are arranged in ascending order to give each participant the possibility to compare his performance with the other ones, for these reasons the information obtained by this visualization is useful for participants.

To generalize this visualization we have to do the following:
**Input:** the SE(AP) table.
**Mapping:** name of experiments on x axis, chosen descriptive statistics on y axis.
**Step 1:** group row (statistics) values for participants assigning to each group a distinguishing color.
**Step 2:** plot the values of the chosen statistics (a row) in ascending order.



| p-values: T-test (lower triangle) vs. randomized test (upper triangle) | BiTeM09PAbl | BiTeM09PAcba | BiTeM09PAcbb | BiTeM09PAcit | BiTeM09PAcl | BiTeM09PAqe | CMU09Chmtcdd | CMU09Chmtcd | SCAI09PAf1a | SCAI09PAf1b | SCAI09PAf1c | SCAI09PAf1d | SCAI09PAf1e | SCAI09PAf2a | SCAI09PAf2b | SCAI09PAf2c | SCAI09PAf2d | SCAI09PAf2e | SCAI09PAf3a | SCAI09PAf3b | SCAI09PAf3c | SCAI09PAf3d | SCAI09PAf3e | SCAI09PAf4a | SCAI09PAf4b | SCAI09PAf4c | SCAI09PAt1a | SCAI09PAt1b | SCAI09PAt1c | SCAI09PAt1d | SCAI09PAt1e | SCAI09PAt2a | SCAI09PAt2b | SCAI09PAt2c | SCAI09PAt2d | SCAI09PAt2e | SCAI09PAt3a | SCAI09PAt3b | SCAI09PAt3c | SCAI09PAt3d | SCAI09PAt3e | SCAI09PAt4a | SCAI09PAt4b | SCAI09PAt4c | UIowaS09PA1 | york09caPA01 | york09caPA03 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BiTeM09PAbl | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.56 | 0.00 | 0.00 |
| BiTeM09PAcba | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| BiTeM09PAcbb | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| BiTeM09PAcit | 0.00 | 0.67 | 0.01 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| BiTeM09PAcl | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| BiTeM09PAqe | 0.00 | 0.89 | 0.00 | 0.78 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CMU09Chmtcdd | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CMU09Chmtcd | 0.51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.00 |
| SCAI09PAf1a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.90 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.01 | 0.00 | 0.76 | 0.54 | 0.00 | 0.00 | 0.73 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf1b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | x | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.89 | 0.00 | 0.17 | 0.00 | 0.59 | 0.27 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.55 | 0.00 | 0.00 | 0.00 | 0.00 | 0.48 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf1c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.14 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.03 | 0.00 | 0.32 | 0.02 | 0.02 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.15 | 0.00 | 0.13 | 0.00 | 0.00 | 0.22 | 0.00 | 0.74 | 0.03 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf1d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.40 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf1e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf2a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | x | 0.02 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 | 0.32 | 0.00 | 0.00 | 0.07 | 0.00 | 0.02 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 0.51 | 0.10 | 0.08 | 0.00 | 0.00 | 0.93 | 0.01 | 0.52 | 0.00 | 0.00 | 0.10 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf2b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.46 | x | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.03 | 0.00 | 0.00 | 0.51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.55 | 0.80 | 0.00 | 0.00 | 0.00 | 0.20 | 0.00 | 0.60 | 0.00 | 0.00 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf2c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.36 | 0.00 | 0.00 | 0.07 | 0.01 | x | 0.00 | 0.00 | 0.03 | 0.90 | 0.02 | 0.00 | 0.00 | 0.00 | 0.01 | 0.32 | 0.00 | 0.00 | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.02 | 0.74 | 0.00 | 0.00 | 0.01 | 0.00 | 0.09 | 0.32 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf2d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.76 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf2e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.94 | 0.31 | 0.02 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.81 | 0.62 | 0.00 | 0.00 | 0.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf3a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.71 | 0.27 | 0.16 | 0.00 | 0.00 | x | 0.00 | 0.74 | 0.00 | 0.00 | 0.01 | 0.00 | 0.10 | 0.00 | 0.00 | 0.47 | 0.00 | 0.00 | 0.25 | 0.06 | 0.29 | 0.00 | 0.00 | 0.60 | 0.05 | 0.24 | 0.00 | 0.00 | 0.03 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf3b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.31 | 0.00 | 0.00 | 0.07 | 0.01 | 0.94 | 0.00 | 0.00 | 0.17 | x | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.00 | 0.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.27 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf3c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.58 | 0.20 | 0.22 | 0.00 | 0.00 | 0.85 | 0.24 | x | 0.00 | 0.00 | 0.01 | 0.00 | 0.29 | 0.00 | 0.00 | 0.61 | 0.00 | 0.00 | 0.21 | 0.04 | 0.39 | 0.00 | 0.00 | 0.46 | 0.15 | 0.08 | 0.00 | 0.00 | 0.02 | 0.00 | 0.44 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf3d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf3e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.94 | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.28 | 0.00 | 0.01 | 0.00 | 0.00 | x | 0.00 | 0.07 | 0.00 | 0.55 | 0.32 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf4a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.68 | 0.01 | 0.00 | 0.00 | 0.15 | 0.01 | 0.11 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.31 | 0.69 | 0.00 | 0.00 | 0.00 | 0.07 | 0.00 | 0.29 | 0.00 | 0.00 | 0.97 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf4b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.40 | 0.51 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.07 | 0.00 | 0.08 | 0.00 | 0.00 | 0.45 | 0.00 | x | 0.00 | 0.19 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.48 | 0.00 | 0.00 | 0.00 | 0.00 | 0.61 | 0.00 | 0.47 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAf4c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.23 | 0.05 | 0.53 | 0.00 | 0.00 | 0.41 | 0.57 | 0.54 | 0.00 | 0.00 | 0.03 | 0.02 | x | 0.00 | 0.00 | 0.70 | 0.00 | 0.00 | 0.02 | 0.00 | 0.88 | 0.00 | 0.00 | 0.06 | 0.52 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt1a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.76 | 0.60 | 0.11 | 0.00 | 0.06 | 0.00 | 0.00 | 0.02 | 0.00 | 0.81 | 0.00 | 0.02 | 0.00 | 0.00 | 0.57 | 0.00 | 0.26 | 0.00 | x | 0.98 | 0.00 | 0.00 | 0.62 | 0.00 | 0.00 | 0.00 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt1b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.68 | 0.52 | 0.04 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.72 | 0.00 | 0.00 | 0.00 | 0.00 | 0.48 | 0.00 | 0.15 | 0.00 | 0.98 | x | 0.00 | 0.00 | 0.31 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt1c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.37 | 0.11 | 0.37 | 0.00 | 0.00 | 0.60 | 0.40 | 0.74 | 0.00 | 0.00 | 0.05 | 0.01 | 0.78 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.08 | 0.01 | 0.79 | 0.00 | 0.00 | 0.21 | 0.33 | 0.03 | 0.00 | 0.00 | 0.01 | 0.00 | 0.80 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt1d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.66 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt1e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.85 | 0.20 | 0.01 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 0.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.19 | 0.00 | 0.04 | 0.00 | 0.64 | 0.53 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt2a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.73 | 0.04 | 0.00 | 0.00 | 0.47 | 0.04 | 0.37 | 0.00 | 0.00 | 0.47 | 0.00 | 0.13 | 0.00 | 0.00 | 0.22 | 0.00 | 0.00 | x | 0.26 | 0.01 | 0.00 | 0.00 | 0.41 | 0.00 | 0.97 | 0.00 | 0.00 | 0.25 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt2b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.39 | 0.89 | 0.01 | 0.00 | 0.00 | 0.23 | 0.01 | 0.17 | 0.00 | 0.00 | 0.78 | 0.00 | 0.04 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 0.64 | x | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.42 | 0.00 | 0.00 | 0.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt2c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 | 0.28 | 0.07 | 0.47 | 0.00 | 0.00 | 0.48 | 0.51 | 0.61 | 0.00 | 0.00 | 0.04 | 0.02 | 0.92 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 | 0.16 | 0.06 | x | 0.00 | 0.00 | 0.08 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt2d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.37 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt2e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.68 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.16 | 0.00 | 0.03 | 0.00 | 0.00 | 0.74 | 0.00 | 0.68 | 0.01 | 0.41 | 0.31 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.01 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.84 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt3a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.96 | 0.44 | 0.09 | 0.00 | 0.00 | 0.75 | 0.09 | 0.62 | 0.00 | 0.00 | 0.26 | 0.00 | 0.26 | 0.00 | 0.00 | 0.40 | 0.00 | 0.00 | 0.69 | 0.37 | 0.31 | 0.00 | 0.00 | x | 0.00 | 0.41 | 0.00 | 0.00 | 0.03 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt3b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.24 | 0.00 | 0.00 | 0.11 | 0.02 | 0.80 | 0.00 | 0.00 | 0.23 | 0.86 | 0.32 | 0.00 | 0.01 | 0.01 | 0.06 | 0.70 | 0.01 | 0.00 | 0.51 | 0.00 | 0.00 | 0.06 | 0.02 | 0.63 | 0.00 | 0.02 | 0.13 | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.30 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt3c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.69 | 0.74 | 0.03 | 0.00 | 0.00 | 0.45 | 0.03 | 0.35 | 0.00 | 0.00 | 0.47 | 0.00 | 0.12 | 0.00 | 0.00 | 0.20 | 0.00 | 0.00 | 0.98 | 0.65 | 0.15 | 0.00 | 0.00 | 0.66 | 0.05 | x | 0.00 | 0.00 | 0.24 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt3d | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 |  | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt3e | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 | 0.62 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.14 | 0.00 | 0.04 | 0.00 | 0.00 | 0.68 | 0.00 | 0.75 | 0.01 | 0.38 | 0.27 | 0.01 | 0.00 | 0.09 | 0.00 | 0.00 | 0.01 | 0.00 | 0.92 | 0.00 | 0.03 | 0.00 |  | x | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt4a | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.28 | 0.70 | 0.01 | 0.00 | 0.00 | 0.16 | 0.01 | 0.12 | 0.00 | 0.00 | 0.98 | 0.00 | 0.03 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.49 | 0.80 | 0.04 | 0.00 | 0.00 | 0.27 | 0.01 | 0.49 | 0.00 |  | x | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt4b | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.20 | 0.80 | 0.00 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 | 0.03 | 0.01 | 0.19 | 0.01 | 0.00 | 0.24 | 0.00 | 0.67 | 0.06 | 0.15 | 0.06 | 0.03 | 0.00 | 0.02 | 0.00 | 0.00 | 0.05 | 0.00 | 0.40 | 0.00 | 0.14 | 0.00 | 0.00 | 0.47 | 0.00 | x | 0.00 | 0.00 | 0.00 | 0.00 |
| SCAI09PAt4c | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.09 | 0.00 | 0.00 | 0.27 | 0.07 | 0.46 | 0.00 | 0.00 | 0.48 | 0.50 | 0.61 | 0.00 | 0.00 | 0.03 | 0.02 | 0.91 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 | 0.16 | 0.05 | 1.00 | 0.00 | 0.01 | 0.30 | 0.62 | 0.14 | 0.00 | 0.01 | 0.04 | 0.05 | x | 0.00 | 0.00 | 0.00 |
| UIowaS09PA1 | 0.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 | 0.00 |
| york09caPA01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x | 0.00 |
| york09caPA03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | x |

**Figure 18: Statistical significance comparison of 2009 TREC experiments**

The table on Figure 18 allows for comparing pairs of experiments, testing their statistical difference according to a chosen metrics (AP), expressed by the *p-value* computed through a T-test (lower triangle) or a randomized test (upper triangle). P-values are coded through a color: strong red emphasizes that the value is close to one.

The information is obtained by this visualization is useful for participants because the table expresses the degree of validation of experimental hypotheses.

To generalize this visualization we have to do the following:
**Input:** the SE table corresponding to the chosen metric (AP).
**Mapping:** name of experiments on x and y axes
**Step 1:** start an automated analysis on all the n(n-1)/2 experiment pairs (T-test and randomized analysis)
**Step 2:** choose a scale of color to encode the p-values returned by step 2.
**Step 3:** plot t-test results in the lower triangle and randomized test in the upper triangle.

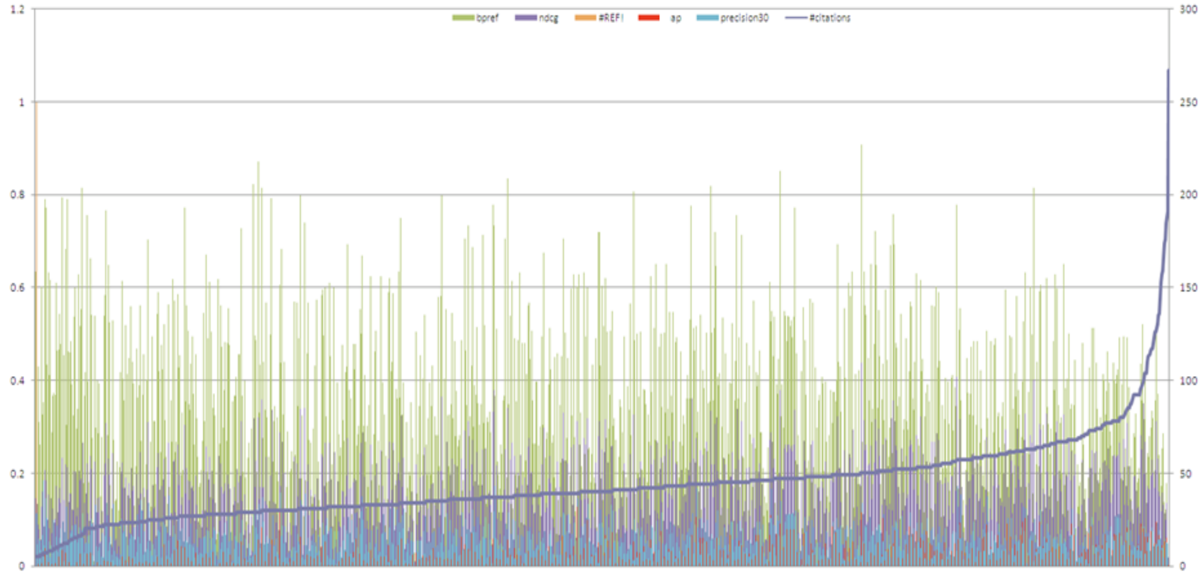


**Figure 19: showing five metrics for thousands of topics**

The crowded bar chart depicted on Figure 19 represents data from a large TM(e) table (thousands of topics) showing, for each topic of five metrics (*bpref, ndcg, #ref, ap, precision_30*) with color and using the left Y-axis together with a frequency distribution, using the line and the Y-axis right scale.

To generalize this visualization we have to do the following:
**Input:** the TM(e) table corresponding to the chosen experiment.
**Mapping:** name of topics on x axis, chosen metrics and count on y axis.
**Step 1:** select the metrics.
**Step 2:** plot the selected metrics for each topic, sorting the topic using the number of associated documents.
The information obtained by this visualization is useful for participants and organizers.

*Source:*

| | |
|---|---|
| EP | 77 |
| US | 923 |

*Year:*

| | |
|---|---|
| 2001 | 151 |
| 2002 | 134 |
| 2003 | 188 |
| 2004 | 162 |
| 2005 | 164 |
| 2006 | 201 |

*Kind:*

| | |
|---|---|
| A1 | 12 |
| A2 | 8 |
| B1 | 488 |
| B2 | 492 |

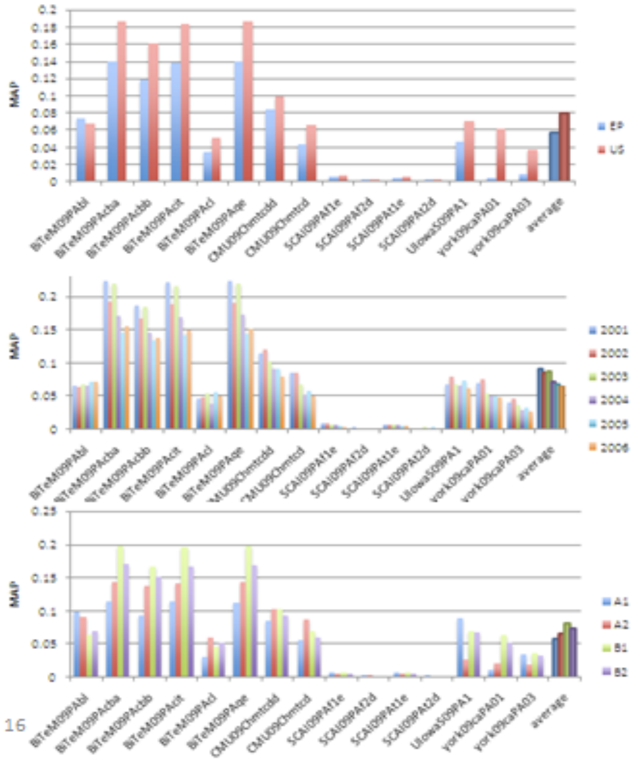


**Figure 20: Analysis of participant performance (MAP) across 6 years**

The visualization on Figure 20 combines tables and bar charts to describe the patent frequency distribution (across years, provenance, and IPC classes) and the experiment performances (evaluated through the MAP statistics). Tables are quite readable, demonstrating that, for few values, numbers are more effective than visualizations. The three multi-value bar charts represent on the X-axis the list of different participants and on the Y-axis the MAP values for the different categorical values, mapped on colors.

As an example in the second bar chart color distinguishes the 6 different values of MAP from 2001 to 2006.

The information obtained by this visualization is useful for organizers.

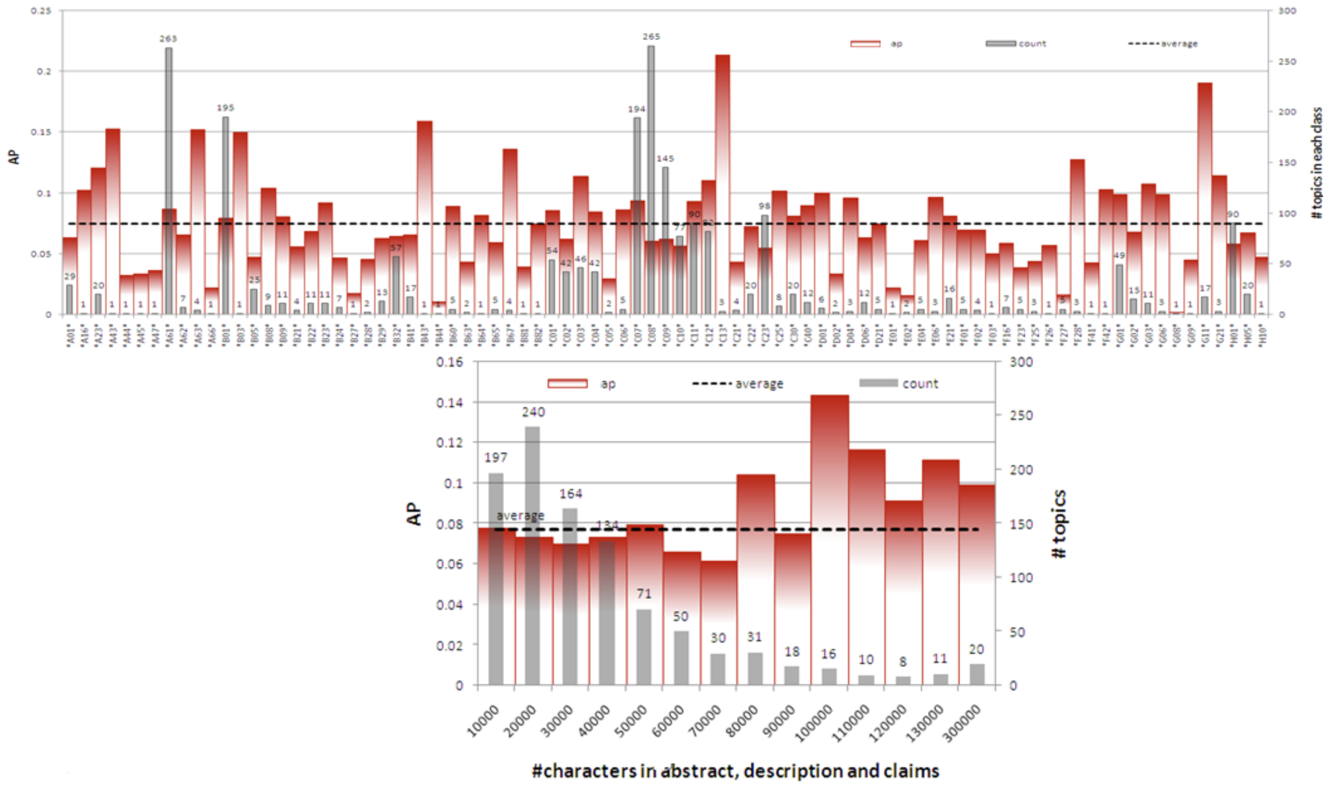


**Figure 21: Analysis of AP values according to IPC classes and number of characters**

The two bar charts depicted on Figure 21 represent information about IPC classes. The uppermost chart plots IPC classes on the X-axis and on the Y-axis there are two different scales: on the left the value of the chosen metric (AP) and on the right the number of topics for each class: red bars for the value of metric and grey bars for the number of topics. A dotted line shows the average.

The second graph is a topic frequency distribution whose bins are based on the number of characters for the entire document (abstract, description and claims) and on the Y-axis there are two different scales as above: on the left the value of the fixed metric (AP) and on the right the number of topics. A dotted line shows the average.

The information obtained by these visualizations is useful for participants and organizers

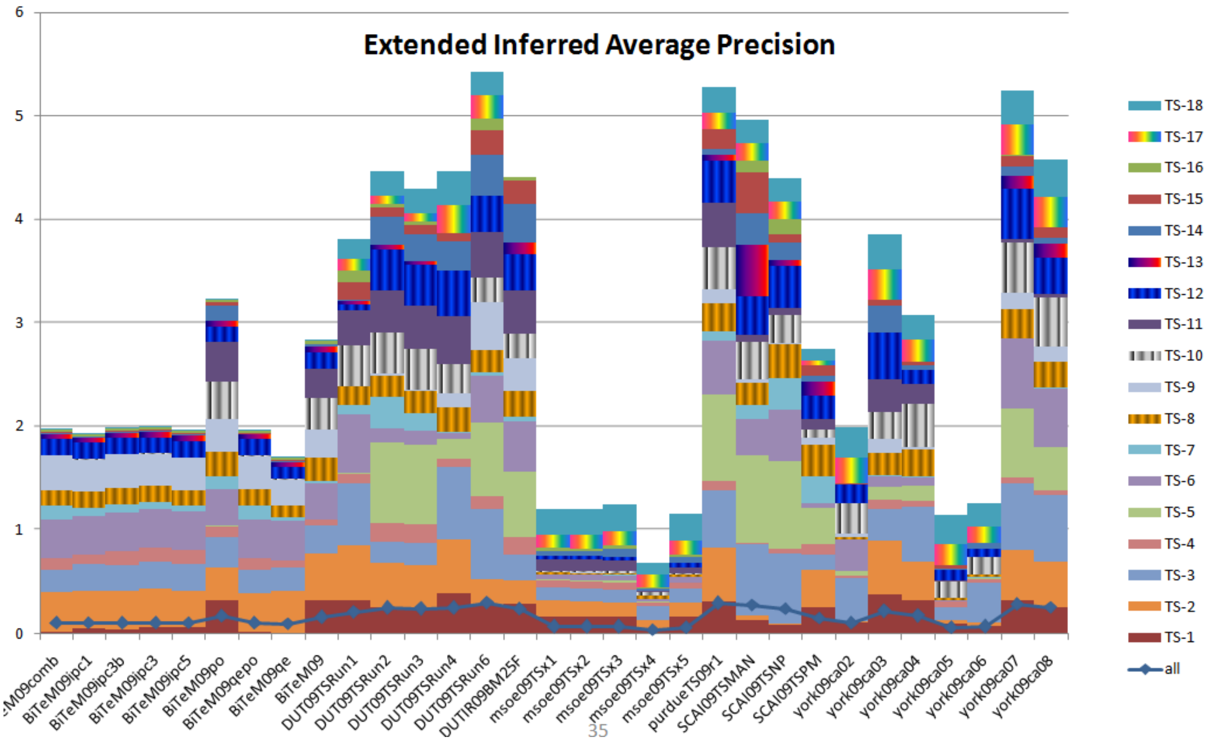


Figure 22: Experiment analysis

This visualization on Figure 22 represents a stacked bar chart where on the X-axis is shown the list of different runs on a fixed set of topics (each topic is coded by color) and on the Y-axis there is the value of the chosen metric (Extended Inferred AP).

This graph gives two different pieces of information: we can evaluate the performance of single algorithm by reading the value on Y-axis or evaluate the performance on a specific topic by looking the associated color on the bars, allowing for understanding the experiment(s) that performed better on a specific topic.

The information obtained by these visualizations is useful for both: participants and organizers.: organizers get an overview on the behavior of various algorithms looking at the metric values on various topics; participants can analyze their own performance and compare it with other competitors.

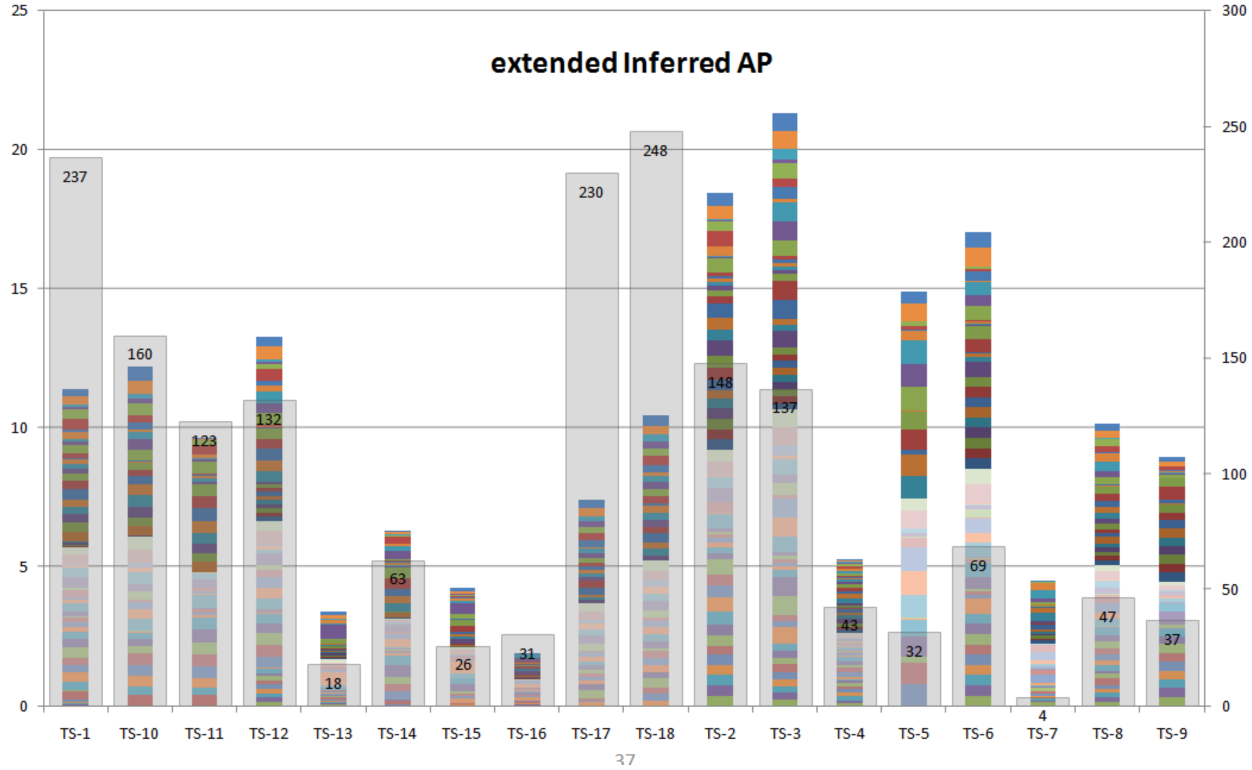


**Figure 23: Topic analysis**

The stacked bar char on Figure 23 is a variant of what depicted on Figure 22. The information represented by X-axis and the color are inverted: on the X-axis there are the topics and the color is coding the different algorithms. On the Y-axis there are two different scales: on the left the value of the fixed metric, on the right the number of relevant documents for each topic, represented by the grey bar.

The information obtained by these visualizations is useful for both: participants and organizers.

**Organizers:** overview on the behavior of various algorithms based on various topics.

**Participants:** analysis of their own performance and the possibility of direct comparison with competitors.

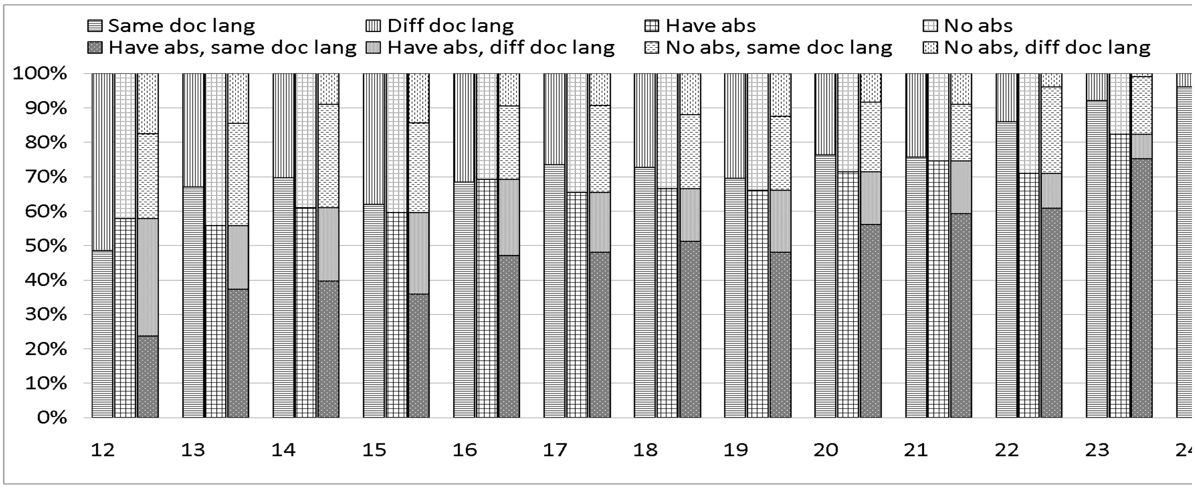


**Figure 24: Source of difficulty in topics**

This graph on Figure 24 deals with sources of difficulty in topics: Y-axis is percentage and the X-axis is split in bins (bin j contains the topics that were answered by j experiments). Each bar visualizes the topic proportions w.r.t three categorical attribute (i.e., [have abstract, no abstract], [same language, different language], [Have abstract-same language, Have abstract-different language, no abstract-same language, no abstract-different language]).

# 4 PROMISE requirements for the Visual Analytics module

The section structure follows the IEEE 830-1998 Recommended Practice Software Requirements Specifications requirement standard described in the Appendix 8.3 of PROMISE Deliverable 5.1.

## 4.1 Introduction

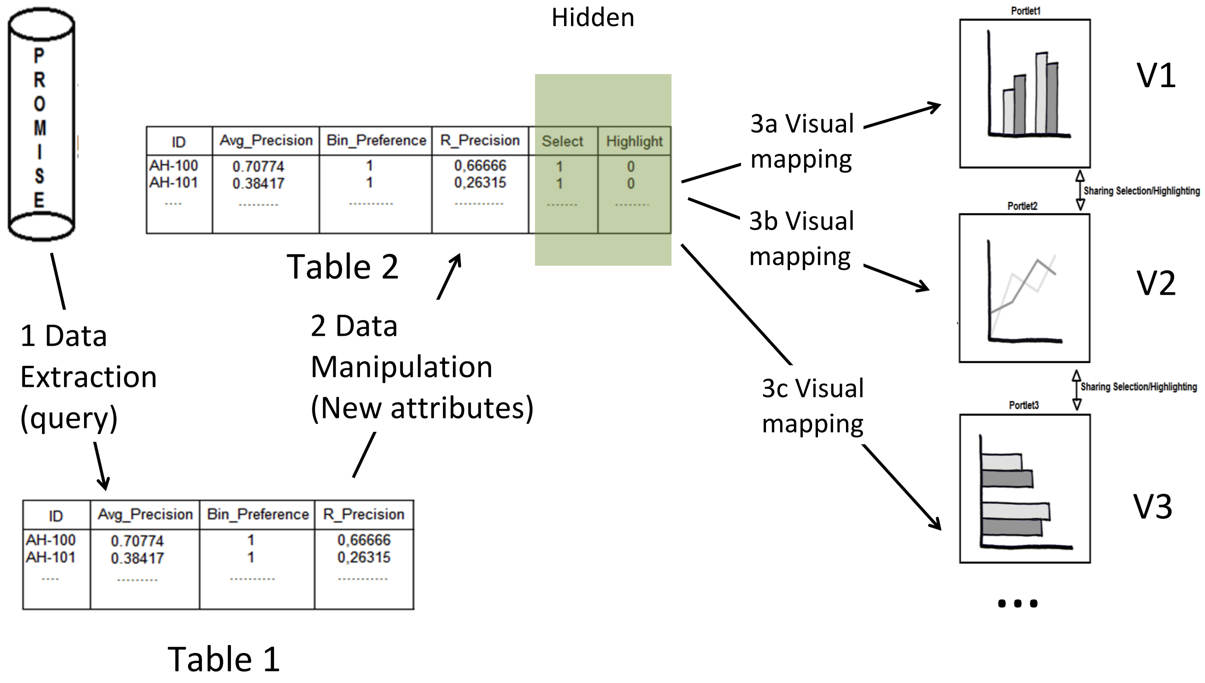
Before discussing the Visual Analytics requirement the overall architecture of the Visual Analytics component is presented, with the main goal of making the following considerations more clear. The overall architecture is depicted on Figure 25 and its structure is totally parametric, without any assumption about the data structure (in the most general case it is contained in non-normalized table). Moreover, there are no assumptions about visualizations (it is possible to obtain any kind of visualization), about the mapping between data and visualizations, and about analytical components.

The most general situation is the one in which the system presents the user with multiple visualizations, each of them working on the same set of data. Visualizations are synchronized using two main interaction mechanisms: selection (it is just a way to focus the attention on a subset of data) and the highlighting (it allows for highlighting a part of the displayed data).

In order to produce a visualization, three main steps are, in principle, needed:

1) data extraction from PROMISE database.
2) data manipulation, i.e., deriving new attributes, applying some aggregation operations , applying some analytical algorithms, etc. During such a process the system adds some hidden attributes to the data, in order to support the selection and the highlighting mechanisms.
3) Mapping the data obtained from step two on one or more visualizations.

The first two steps are optional: in some cases the system will automatically perform them, allowing the user to focus only on the mapping and analysis activities.


PROMISE
Hidden
1 Data Extraction (query)
2 Data Manipulation (New attributes)
Table 2
Table 1
3a Visual mapping
3b Visual mapping
3c Visual mapping
Portlet1
Portlet2
Portlet3
Sharing Selection/Highlighting
Sharing Selection/Highlighting
V1
V2
V3
...

| ID | Avg_Precision | Bin_Preference | R_Precision | Select | Highlight |
|---|---|---|---|---|---|
| AH-100 | 0.70774 | 1 | 0,66666 | 1 | 0 |
| AH-101 | 0.38417 | 1 | 0,26315 | 1 | 0 |
| .... | ......... | .......... | ........... | ....... | ....... |

| ID | Avg_Precision | Bin_Preference | R_Precision |
|---|---|---|---|
| AH-100 | 0.70774 | 1 | 0,66666 |
| AH-101 | 0.38417 | 1 | 0,26315 |
| .... | ......... | .......... | ........... |

**Figure 25: The PROMISE Visual Analytics architecture**

### 4.1.1 Purpose

This section describes the requirements for the Visual Analytics module needed to realize a mapping between data and a suitable visualization chosen by the user, which has also the possibility to rearrange data in order to obtain a significant graph for her visualization purposes. This module will include also the requirements for collaboration and knowledge sharing reported on deliverable 5.1 on Section 6. As an example, annotations are crucial to the extent of reconstructing the operations leading to a visualization. Through annotations one can explain executed operations and can explain, spread, and save particular choices. The same holds for queries executed by the system during the mapping process (the process leading from data to the visualization). To avoid duplication we do not discuss these issues anymore and we refer to PROMISE Deliverable 5.1 Collaborative user interface requirements, Section 6 for further details.

### 4.1.2 Scope

The software that is referred in this section is the PROMISE Visual Analytics Component (PVAC) that allows for establishing a mapping between (a huge amount of not a priori known) PROMISE experimental data and a set of synchronized visualizations. During the analysis phase, according to the general VA process (see Section 1) the module will allow the user for activating some automatic analysis algorithms (e.g., T-test, clustering, etc.) switching back and forth between automatic analysis and visual analysis.

### 4.1.3 Definitions, acronyms, and abbreviations

Mock-up: it is a model of a design used for demonstration, design evaluation and other purpose. A mock-up is called a prototype if it provides at least part of the functionality of a system and enables testing of a design.

VA - Visual Analytics

PVAC - PROMISE Visual Analytics Component

### 4.1.4 References

PROMISE Deliverable 5.1 - Collaborative User Interface Requirements

### 4.1.5 Overview

The rest of the section contains:

- Section 4.2 an overview of some factors affecting the software realizing the visualization and the user features
- Section 4.3 more technical and detailed description of the interfaces and the various required functions.

## 4.2 Overall description

The designer should take into account the data and user characteristics. In particular, the former because the product has to deal with a huge amount of potentially heterogeneous, not known a priori data and not always organized in a suitable way for the user visualization purposes; the latter because of the fact that it will be used from skilled users.

### 4.2.1 Productive perspective

The PVAC is quite an independent module and has only few interfaces with other components.

#### *4.2.1.1 System interfaces*

- The system has an interface with DIRECT, used to retrieve data.

- The system has an interface with the collaboration module, used to store annotations and trigger event notification (see PROMISE deliverable 5.1).

- The system has a **general purpose user interface**, devoted to data manipulation and analysis, that leads to possible visualizations based on data organization and manipulation decided by the user.

- The system has **an ad-hoc user interface**, encompassing a predefined set of common tasks and visualizations

#### *4.2.1.2 General purpose user interface*

In order to obtain an effective visualization of a huge amount of heterogeneous data according to the purposes of the user, there are four main interfaces accessible from the home page: the page devoted to attributes classification, the page devoted to data manipulation (data managing), the page realizing the mapping (from data to visualization), and the page for the ultimate data filtering.

In general we can consider the data as a set of tuples coming from a table and the user is allowed to explore data features, by means of a sampling operation. Data can be rearranged and modified through suitable operations (mathematic or of grouping/reordering) producing a different table, derived from the "original" one.

The above interfaces take different kinds of input and provide different kind of output. In particular:

- The wizard home page (see Figure 25) is devoted to data loading, accessing the PROMISE database or uploaded by the user, and allows for both performing a customized analysis or starting some common and predefined task, as detailed in the next section.

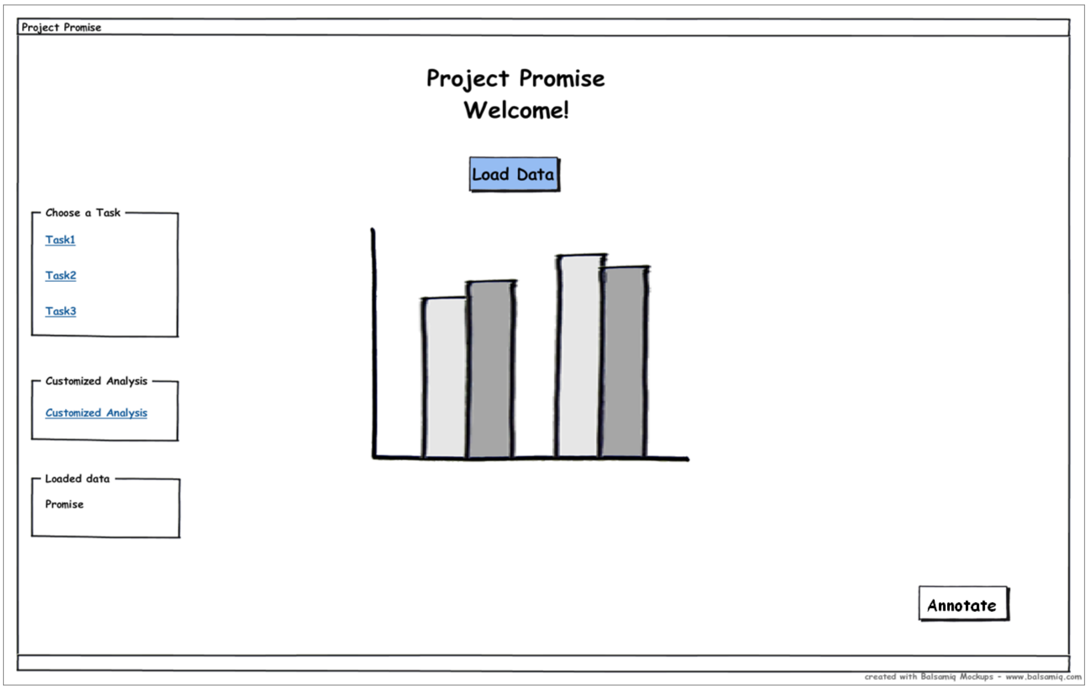
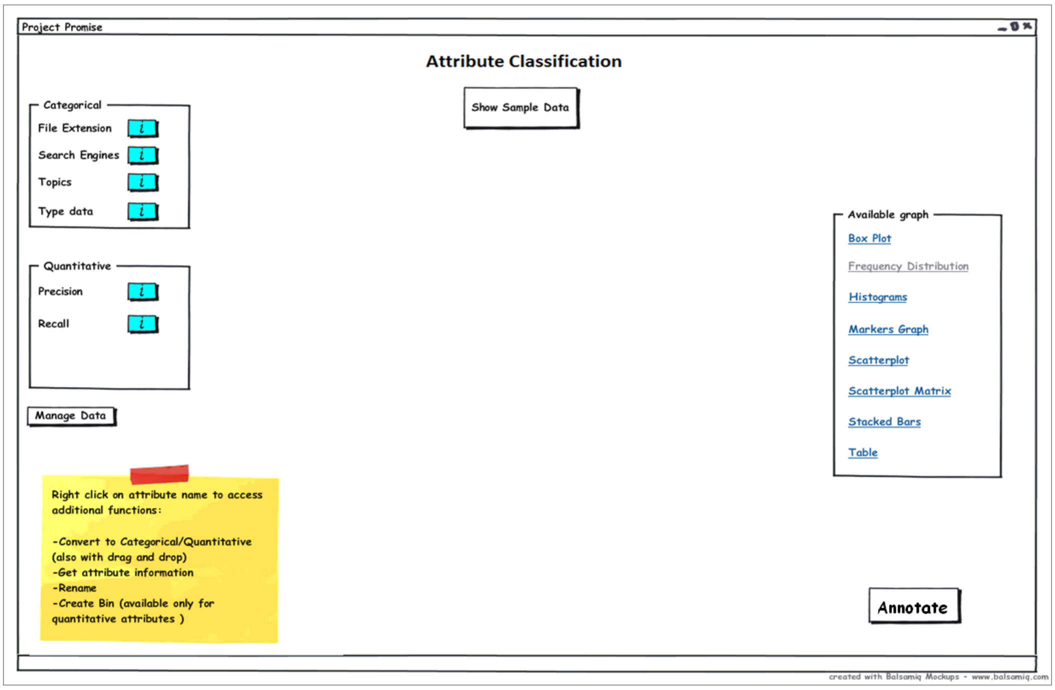


**Figure 25: The wizard home page**

- The page for attributes classification (see Figure 26) takes a list of attributes as input, and provides the same list as output, but with a (possibly) different attribute classification (into quantitative and categorical). It allows also for exploring a data sample, in order to investigate the data structure and meaning.



Right click on attribute name to access additional functions:

-Convert to Categorical/Quantitative (also with drag and drop)
-Get attribute information
-Rename
-Create Bin (available only for quantitative attributes )

**Figure 26: The attribute classification page**

- The interface devoted to data manipulation (see Figure 27) takes as input a table (organized in an arbitrary way) and provides a different table as output, where data are rearranged, and with possibly more columns deriving from the "original" table, through some mathematical or ordering/grouping operation. The system checks whether the manipulated table is suitable for the user visualization purposes (otherwise, there are some warning messages leading the user to a more compliant data organization). The Interface allows also to activate basic data mining algorithms (e.g., T-test, clustering, etc.).

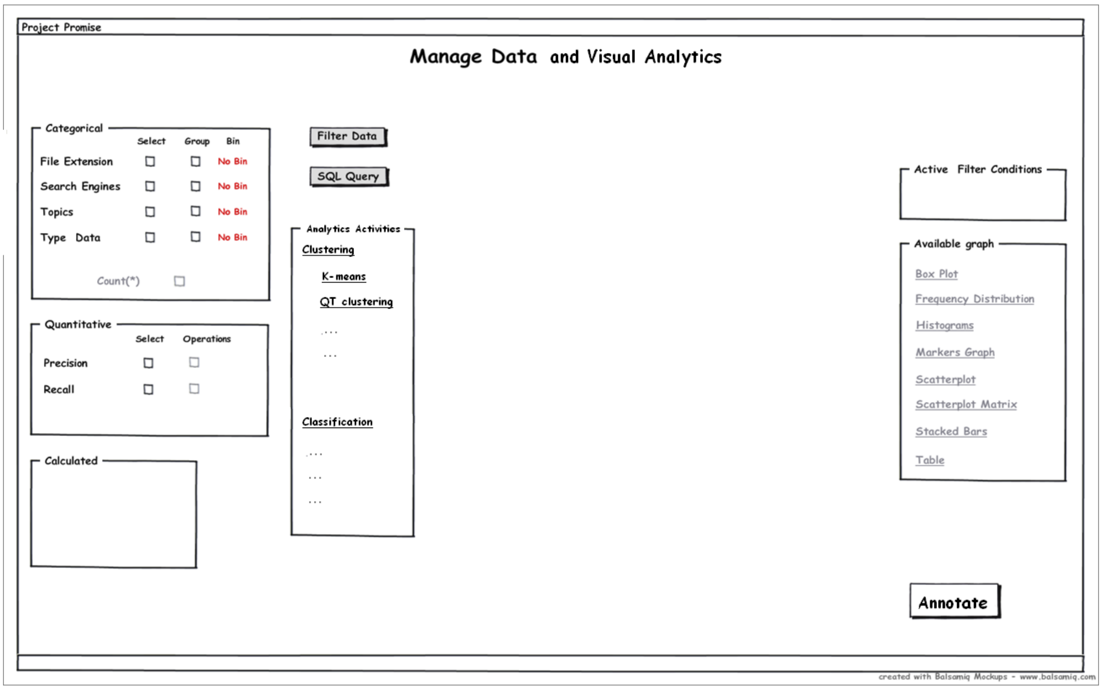


**Figure 27: The data manipulation page**

- The filter interface (see Figure 28) takes as input a set of data coming from a table (manipulated or original), and provide as output a subset of the same data.

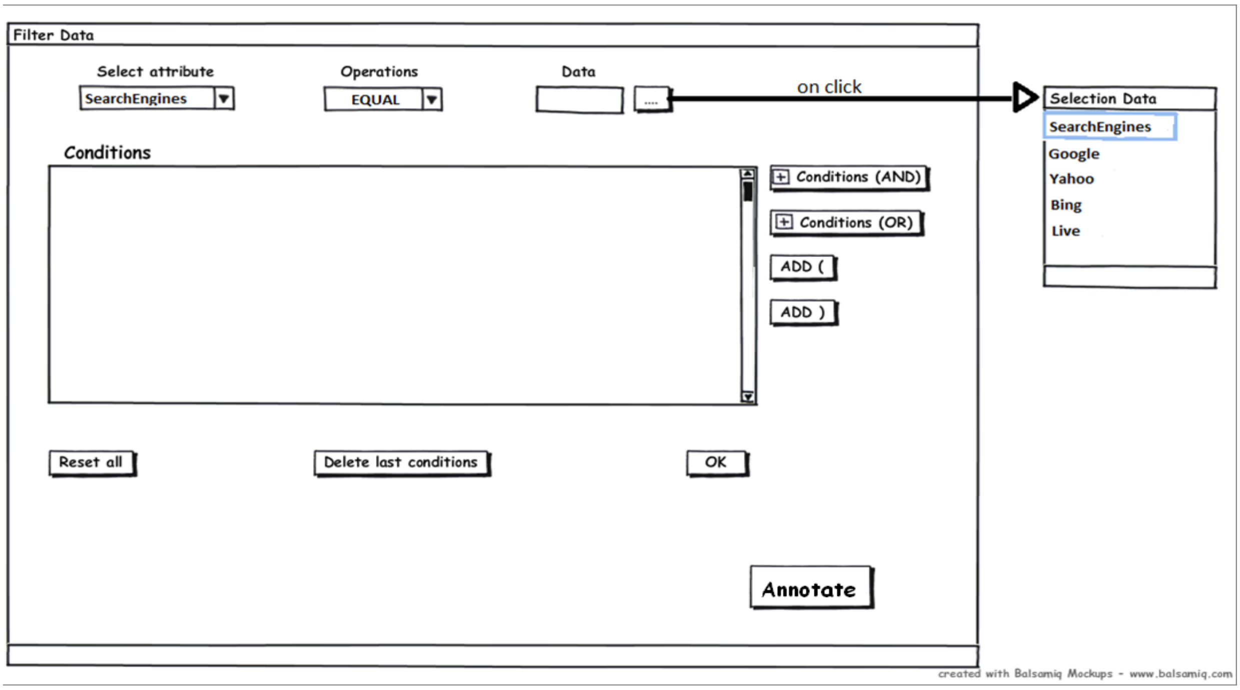
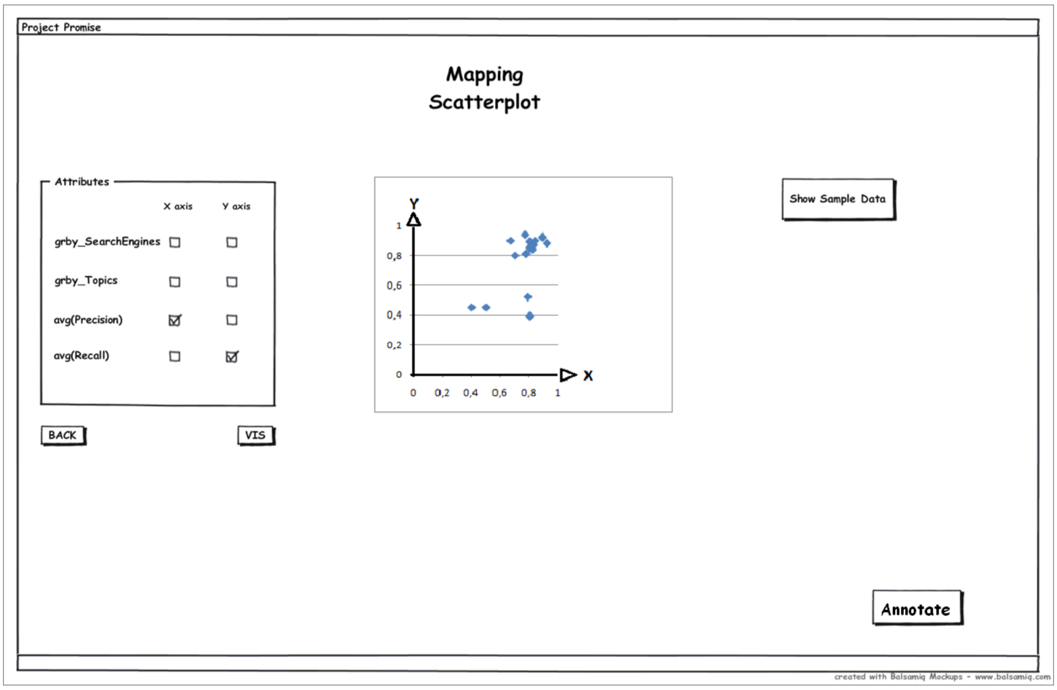


**Figure 28: The data filtering page**

- The interface (see Figure 29) realizing the mapping takes as input the table (manipulated, when needed) and produces as output a graph representing the visualization.

  The mok-up shows an exemplificative mapping, but the approach is the same for all kinds of visualizations.



| | X axis | Y axis |
|---|---|---|
| grby_SearchEngines | □ | □ |
| grby_Topics | □ | □ |
| avg(Precision) | ☑ | □ |
| avg(Recall) | □ | ☑ |

**Figure 29: The mapping page**

Obviously, the above interfaces are part of a sequence of operations, so there is a sort of chain where the output produced by each of them constitutes the input of the following one.

The user expresses the commands by dragging attributes from categorical to quantitative and vice versa, selecting operations by means of checkboxes and slide bars. The greater part of them will be implemented through suitable queries.

All the interfaces have an "Annotate" button, allowing to store an annotation, useful to document and share the analysis choices.

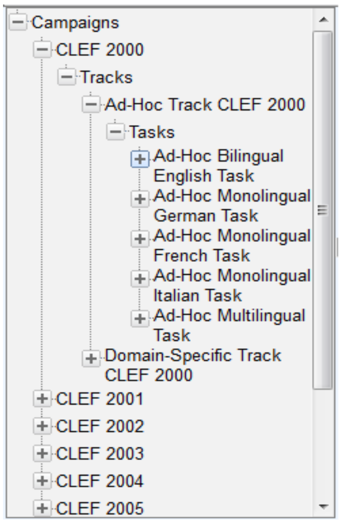
#### 4.2.1.3 Ad-hoc user interface for predefined tasks

The wizard home page allows for accessing a set of predefined ways to perform a data analysis similar to those currently used by users of PROMISE community. In particular we focus on three basic approaches: per topic analysis, per participant analysis, and per experiment analysis. In the following we focus on both the logical steps and the user interfaces, showing, for the purpose of the clarity, some example of visualizations. We do not address the interface issues supporting synchronization and interaction. Moreover, it is important to remark that interaction mechanisms are specific for each visualization. A detailed design of each visualization and its synchronization and interaction mechanisms will be developed in a further design activity.

The initial requirement analysis allows for selecting six visualizations: bi-dimensional scatter-plot, bar charts, stacked bar-charts, box plots, table lens, and frequency distributions. Depending on the chosen approach, the system will present the user with different subset of these visualizations.

In the following we will analyze the three basic approaches to data analysis and present a general description of their own visualizations.

Each of them requires the preliminary selection of a task. This choice corresponds to a selection of an URL connected to the physical address of data. The URL can change depending on whether the user wants to load the entire data set related to a task or just a part of it. The picture on Figure 30 shows a possible way to choose a task.



- Campaigns
  - CLEF 2000
    - Tracks
      - Ad-Hoc Track CLEF 2000
        - Tasks
          - Ad-Hoc Bilingual English Task
          - Ad-Hoc Monolingual German Task
          - Ad-Hoc Monolingual French Task
          - Ad-Hoc Monolingual Italian Task
          - Ad-Hoc Multilingual Task
      - Domain-Specific Track CLEF 2000
  - CLEF 2001
  - CLEF 2002
  - CLEF 2003
  - CLEF 2004
  - CLEF 2005

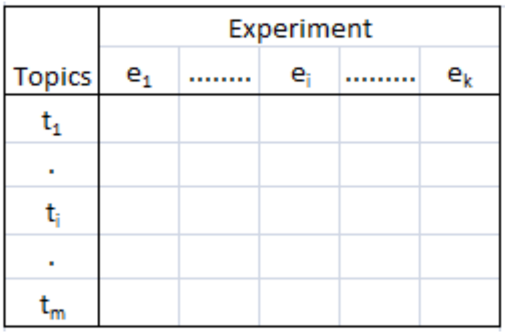
**Figure 30: The page for selecting a task**

According to the typical PROMISE analysis tasks, we foresee a set of ad-hoc visualizations. These visualizations must support synchronization and interaction that are specific for each visualization. Moreover, for each visualization it is needed a mapping mechanism in order to support the user in the creation process. From a technical point of view, designing ad-hoc visualizations implies the design of a module for each visualization. If a user wants to use a visualization he has to select the suitable module and to map on it the desired data. The initial requirement analysis allows for selecting six visualizations: bi-dimensional scatter-plot, bar charts, stacked bar-charts, box plots, table lens, and frequency distributions. Depending on the chosen approach, the system will present the user with different subset of these visualizations.

In the following we will analyze the three basic approaches to data analysis and present a general description of the foreseen visualizations.

### Per topic analysis

In our model, per topic analysis means comparing a set of experiments on each topic with respect to a chosen metric. Therefore the first step for a user is to choose a metric m. Looking at the TME data cube described in the previous section we can note that choosing a metric is equivalent to fix an axis and reduce the set of data to the TE(m) table shown on Figure 31.

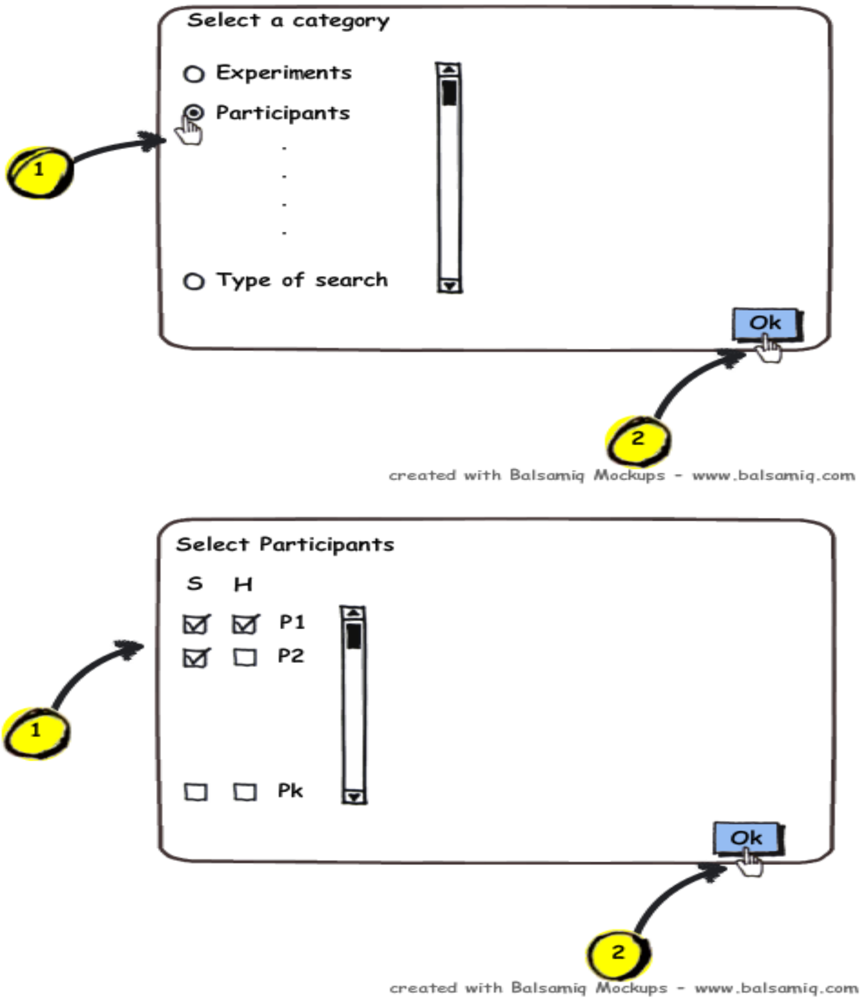


| Topics | Experiment | | | | |
|---|---|---|---|---|---|
| | $e_1$ | ........ | $e_i$ | ......... | $e_k$ |
| $t_1$ | | | | | |
| . | | | | | |
| $t_i$ | | | | | |
| . | | | | | |
| $t_m$ | | | | | |

**Figure 31: The TE(m) table**

Per topic analysis implies a comparison on each topic, so we foresee to represent topics on x-axis in each available visualization.

Having chosen a metric m, the user has to choose the data to display. This choice corresponds to select a subset of the columns of the TE(m) table and that can be performed either by selection or using meta-attribute. For example, the user can select some participants or some experiments. Moreover the user can decide to highlight some elements within the visualizations. Figure 32 show a way in which user can choose a set of column in a table, and highlighting some of them.



Select a category

○ Experiments
● Participants
.
.
.
.
○ Type of search

Select Participants

| S | H | |
|---|---|---|
| ☑ | ☑ | P1 |
| ☑ | □ | P2 |
| □ | □ | Pk |

1
Ok
2
created with Balsamiq Mockups - www.balsamiq.com
1
Ok
2
created with Balsamiq Mockups - www.balsamiq.com

**Figure 32: selecting and highlighting participants and/or experiments**

We foresee four kinds of visualizations for a per topic analysis: bi-dimensional scatter plots, bar charts, box plot charts, and table lens.

In a per topic analysis a bi-dimensional scatter plot presents on x-axis topics and on y-axis the chosen metric. Each metric value is represented by a point. For each topic there are as many points on y-axis as the selected TE(m) columns. To see the trend of a single experiment (a column) you can unify its points with a polyline. To highlight some point it is possible to use color or markers. Figure 33 shows a user provided example of bi-dimensional scatter plot for performing a per topic analysis.

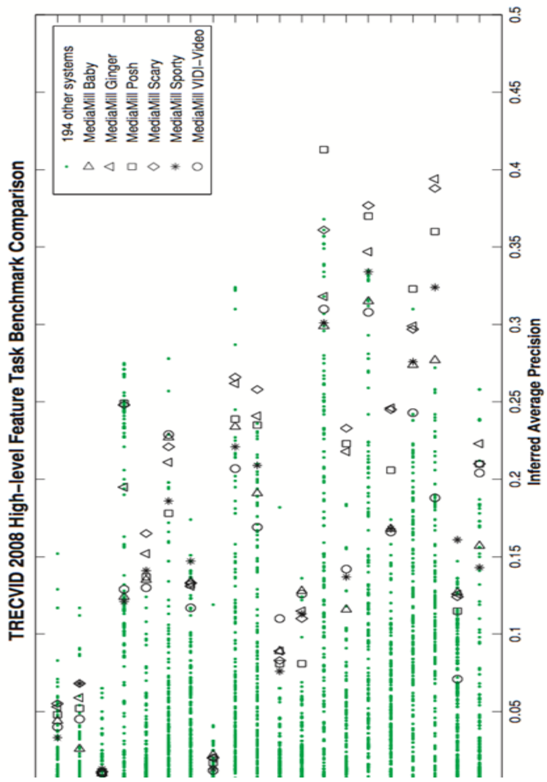


**Figure 33: Per topic analysis bi-dimensional scatterplot**

In a per topic analysis a bar chart can be used to compare two or more experiments on all the topics with respect to a chosen metric. Although it is possible to compare more than two experiments, as the number of experiments (topics) increases the chart representation loses clarity. Possible comparisons are two algorithm of the same participant or the best algorithm of two different participants (e.g., P1 and P2): see Figure 34.

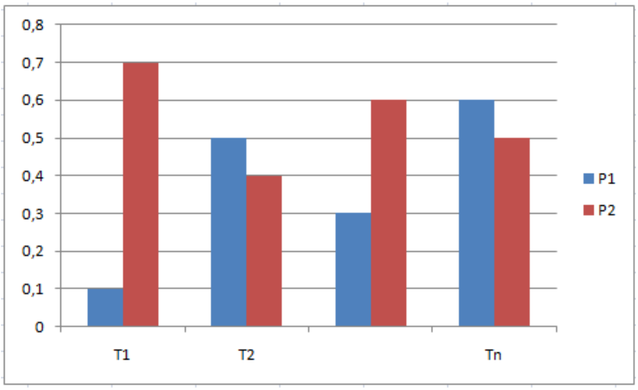


**Figure 34: Per topic analysis bar chart**

In a per topic analysis a box plot chart is used to evaluate the trend of a topic among experiments with respect to a chosen metric. A box plot chart presents a box plot for each topic on x-axis and the chosen metric on y-axis. Looking at table shown on Figure 35, we can say that each box is built calculating statistical indicator on the set of data represented by a single TE(m) row.

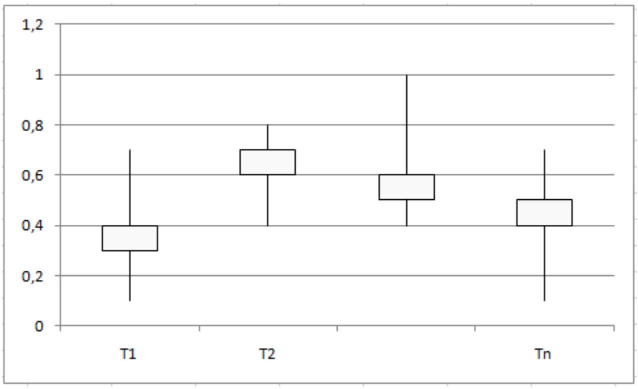
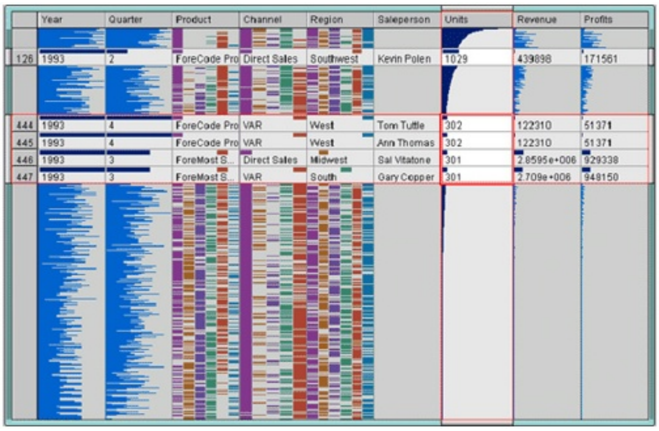


**Figure 35: Per topic analysis box plot chart**

The table lens is a visualization tailored for making sense of large tables. The idea of table lens is to incorporate a graphical representation into a table. The kind of representation depends on the type of attribute represented in a column. For example, if the type is numerical you can use a bar normalized on the maximum value of the column; if the type is categorical you can use a different color to represent each category. User can focus on an area of the table to see real values (focus + context technique, (see, Appendix 6.2.3 for more details).

In a per topic analysis a table lens can be used to represent a TE(m) table. An advantage of table lens with respect to a traditional table is the possibility to see the trend of values in a column and to compare immediately the trend of two different columns. Figure 36 shows a table lens.



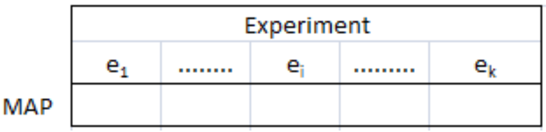
|  | Year | Quarter | Product | Channel | Region | Saleperson | Units | Revenue | Profits |
|---|---|---|---|---|---|---|---|---|---|
| 126 | 1993 | 2 | ForeCode Pro | Direct Sales | Southwest | Kevin Polen | 1029 | 439898 | 171561 |
| 444 | 1993 | 4 | ForeCode Pro | VAR | West | Tom Tuttle | 302 | 122310 | 51371 |
| 445 | 1993 | 4 | ForeCode Pro | VAR | West | Ann Thomas | 302 | 122310 | 51371 |
| 446 | 1993 | 3 | ForeMost S... | Direct Sales | Midwest | Sal Vitatone | 301 | 2.8595e+006 | 929338 |
| 447 | 1993 | 3 | ForeMost S... | VAR | South | Gary Copper | 301 | 2.709e+006 | 948150 |

**Figure 36: Per topic analysis table lens**

### Per participant analysis

In our model per participant analysis means comparing scores of a set of participants with respect to a chosen descriptive statistics. We remember that a descriptive statistics is calculated from a TM(e) table and are aggregated in a SM(e) table (see Section 2).

The first step that a user has to perform is to choose a descriptive statistics. Looking at an SE(m) table we can note that choosing a descriptive statistics (MAP in our example) is equivalent to select a row, obtaining a mono-dimensional array.



|  | Experiment |  |  |  |  |
|---|---|---|---|---|---|
|  | $e_1$ | ........ | $e_i$ | ......... | $e_k$ |
| MAP |  |  |  |  |  |

**Figure 37: Selecting a row in a SE(m) table**

Having chosen a descriptive statistics the user has to select performance measures for the experiments of each participant. When the user chooses a set of participants the experiment data set is partitioned according to her choice. The next Venn diagram shows a partitioning of data set defined by the choice of three participants called P1, P2 and P3.

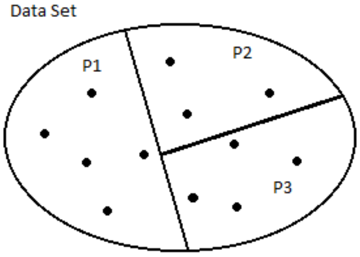


**Figure 38: Partitioning the experiments through participants**

In a per participant analysis on the data set showed in the above figure, it is possible to do other operations. In particular:

- further partitioning using other meta-attribute (textual search experiment, mixed search experiment and so on)
- computing derived data using statistical indicator (best result, median and so on)

For a per participant analysis we foresee four visualizations: bi-dimensional scatter plots, bar charts, box plots, and table lens.

In a per participant analysis a bi-dimensional scatter plot presents on x-axis participants and on y-axis the chosen descriptive statistics. Each value is represented by a point. For each participant there are many points on y-axis as the number of algorithm performed by that participant (in general it is not the same for every participant). Figure 39 shows a scatterplot reporting the MAP values of the experiments of three different participants.

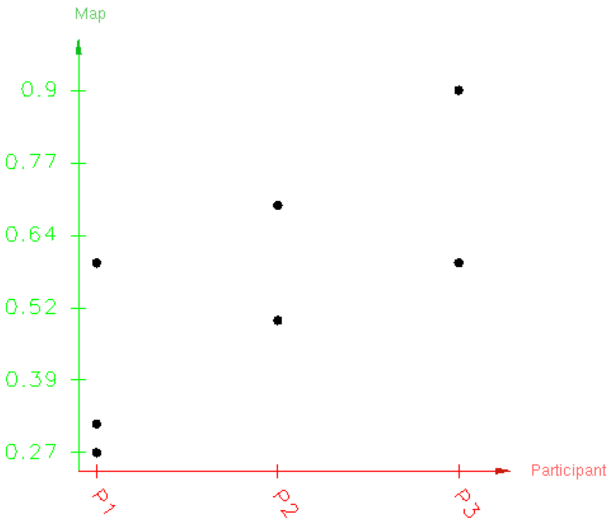


**Figure 39: Per participant analysis scatterplot**

In a per participant analysis a bar chart can be used to compare scores of chosen participants. Colors can be used to group different algorithms from the same participant, producing a mixed per participant/per experiment analysis. If the data set is very large representation problems may rise, so it is possible to reduce this set using statistical indicators, like the best or the worse score. Figure 40 shows an example of bar chart, in which different experiments from the same participant are coded with the same color.

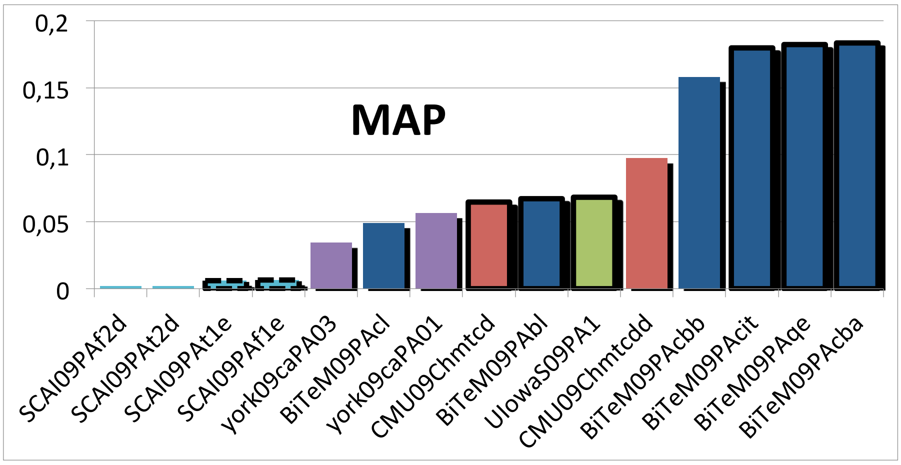


**Figure 40: Per participant analysis bar chart**

In a per participant analysis a box plot can be used to evaluate the trend of a participant among his own algorithm with respect to a chosen descriptive statistics. A boxplot chart presents a boxplot for each participant on x-axis and the chosen statistics on y-axis: each box is built calculating statistical indicators on the set of data represented by a single partition. Figure 41 shows an example of a (simplified) boxplot chart.

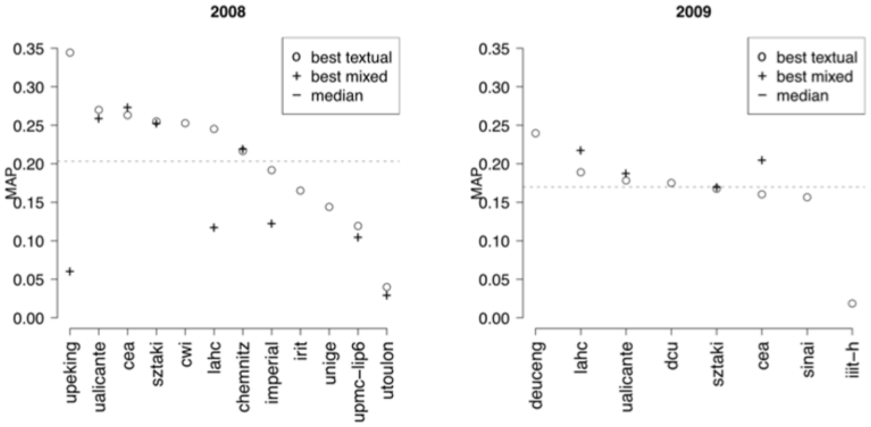
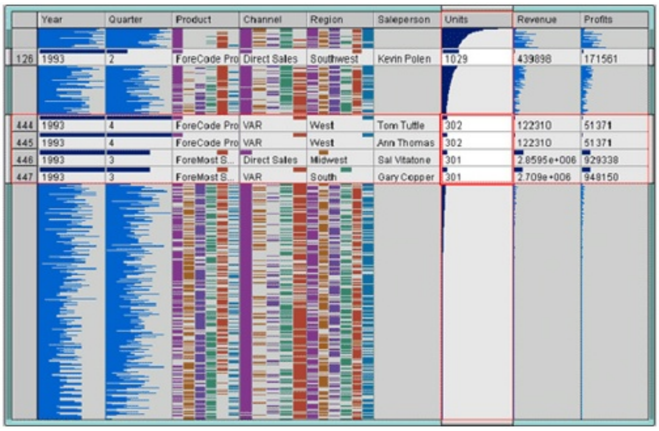


**Figure 41: Per participant analysis box plot chart**

In a per participant analysis a table lens can be used to represent a SE(m) table. An advantage of table lens with respect to a traditional table is the possibility to see the trend of values in a column and to compare immediately the trend of two different columns.



|  | Year | Quarter | Product | Channel | Region | Saleperson | Units | Revenue | Profits |
|---|---|---|---|---|---|---|---|---|---|
| 126 | 1993 | 2 | ForeCode Pro | Direct Sales | Southwest | Kevin Polen | 1029 | 439898 | 171561 |
| 444 | 1993 | 4 | ForeCode Pro | VAR | West | Tom Tuttle | 302 | 122310 | 51371 |
| 445 | 1993 | 4 | ForeCode Pro | VAR | West | Ann Thomas | 302 | 122310 | 51371 |
| 446 | 1993 | 3 | ForeMost S... | Direct Sales | Midwest | Sal Vitatone | 301 | 2.8595e+006 | 929338 |
| 447 | 1993 | 3 | ForeMost S... | VAR | South | Gary Copper | 301 | 2.709e+006 | 948150 |

**Figure 42: Per participant analysis table lens**

## Per experiment analysis

In our model per experiment analysis allows for comparing scores of a set of experiments with respect to a chosen descriptive statistics. Each visualization in a per experiment analysis presents experiments on x-axis and the chosen statistics on y-axis. It can be seen as a particular case of a participant analysis. As said above, in a per participant, analysis we start from the SE(m) table, partitioning it using participant as category. In a per experiments analysis, we partition the set of data using the name of experiment as category. The resulting partitions are constituted by only one element. The Venn diagram on Figure 43 shows a partitioning of data set defined by the name of the experiments.

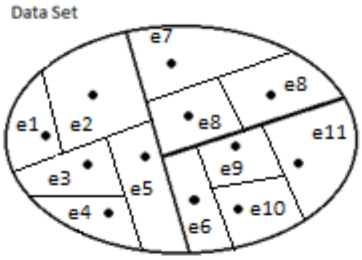


**Figure 43: Partitioning data for per experiment analysis**

As consequence of the above, each visualization in a per experiment analysis has only one value for each point of x-axis.

We foresee four kinds of visualizations for a per topic analysis: bi-dimensional scatter plot, bar chart, frequency distribution, table lens. As regard to bi-dimensional scatter plot, bar chart and table lens, we can say that are similar to those presented for a participant analysis, so we describe only frequency distribution.

In a per experiment analysis a frequency distribution can be used to show the distribution of a set of experiments with respect to a chosen descriptive statistics. In addition to choices of experiments and the descriptive statistics user has to define bins. Defining bins means to divide the set of possible values of the chosen statistics in category transforming the attribute from numerical to categorical. Figure 44 shows an example of frequency distribution.

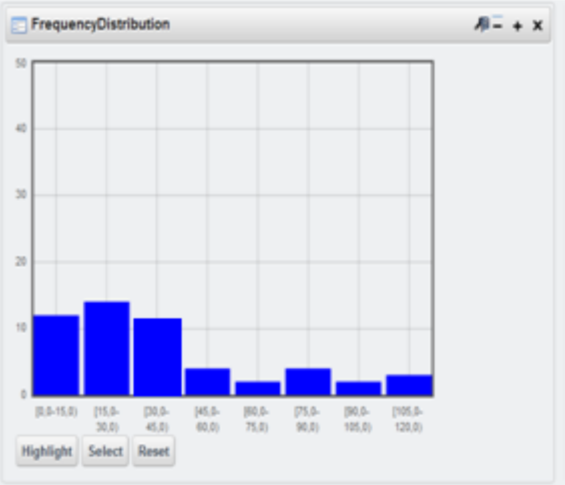


**Figure 44: Per experiment analysis frequency distribution**

#### *4.2.1.4 Software interfaces*

The final implementation will likely make use of portlets, communicating each other and with the rest of the system through REST web services.

#### *4.2.1.5 Communications interfaces*

Not applicable.

#### *4.2.1.6 Operations*

The user has to decide the data organization, according to his/her purposes. He/she is responsible of any data manipulation and reorganization. The system has to manage data based on user's choice and performs a consistency check.

#### *4.2.1.7 Site adaptation requirements*

Not applicable.

### 4.2.2 Product functions

Not applicable.

### 4.2.3 User characteristics

The final user is a technical expert, very skilled in information retrieval evaluation and with a basic level of experience in visualization and visual analytics.

### 4.2.4 Constraints

The developer has to develop the system within the PROMISE target environment, Life Ray, and is limited in her work by the inherent limitations of web applications. Moreover, implementation is based on restful web services using REST (Representational State Transfer), and the developer should take into account all the characteristic of this approach. The access to the PROMISE data will rely on the DIRECT interface (see PROMISE Deliverable 3.2, Section 6.

### 4.2.5 Assumptions and dependencies

Not applicable.

## 4.3 Specific requirements

### 4.3.1 External interfaces

The system extracts data from Direct and, according to user decisions and operations, shows the final visualizations as output.

The data taken as input are organized in a table (possibly in a non-normalized form), and the visualizations returned as output are based on user choices.

### 4.3.2 Functional requirements

The system shall have as “first” input a correct “separation” between categorical and quantitative attributes, because it needs to have a clear distinction for the operations in the rest of the process.

Then, according this distinction, the user can choose the operations (group by, sum, ordering, filtering etc.) to be applied on data, and he/she can manipulate the original table, in order to obtain a suitable table for his/her purposes. Obviously, if data organization satisfies the user, s/he can avoid this step and proceed directly to the visualization.

Then, the user can choose the visualization, with some other adjustment in the data set. If a particular data set leads to a non-significant visualization, same suitable warning messages will notify this situation to the user, with also some indication about what is wrong with it.

### 4.3.3 Performance requirements

The system should support concurrent users, and must be able to handle the data visualization in a way that guarantees a quick interaction with the visualization(s).

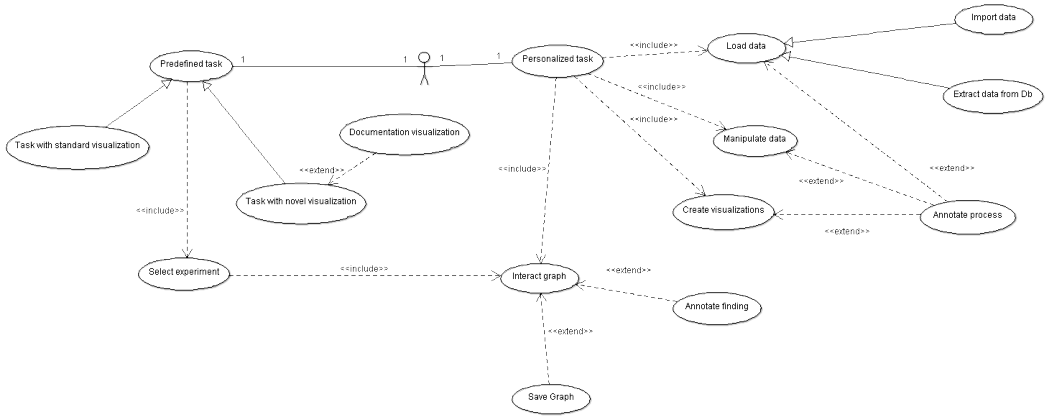
### 4.3.4 Software System attributes

Not applicable.

# 5 UML Use Cases

According to the requirements described in the previous sections, the following UML Use Cases describes the interaction with the Promise Visual analytics component.

The following diagram has been produced with the ArgoUML (ver. 0.30.2) tool that supports UML



<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR1</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Personalized task</td></tr>
<tr><td><b>Primary Actor:</b></td><td>User</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows user for selecting a personalized task</td></tr>
<tr><td><b>Trigger:</b></td><td>A user clicks on a button, labeled ‘Load Data’</td></tr>
<tr><td><b>Preconditions:</b></td><td>An authenticated user</td></tr>
<tr><td><b>Postconditions:</b></td><td>The system will present the user with multiple visualizations</td></tr>
<tr><td><b>Normal Flow:</b></td><td><table>
<tr><td></td><td>Actor Input</td><td>System Response</td></tr>
<tr><td>1</td><td></td><td>System loads data<br>(Use case ‘Load data’)</td></tr>
<tr><td>2</td><td colspan="2">The System and (also) the user manipulate the data<br>(Use case ‘Manipulate data’)</td></tr>
</table></td></tr>
</table>

<table>
<tr><td rowspan="2"></td><td>3</td><td></td><td>The system creates one or more visualizations (use case ‘Create visualizations’)</td></tr>
<tr><td>4</td><td colspan="2">The system and the user interact with graph(s)<br>(Use case ‘interact graph’)</td></tr>
<tr><td><b>Exceptions:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Include:</b></td><td colspan="3">Load data, manipulate data, create visualizations, interact graph</td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td colspan="3">PR2</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td colspan="3">Load data</td></tr>
<tr><td><b>Primary Actor:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td colspan="3">-</td></tr>
<tr><td><b>Description:</b></td><td colspan="3">This UML Use Case allows users for loading data</td></tr>
<tr><td><b>Trigger:</b></td><td colspan="3">When ‘Personalized task’ Use Case (that includes it) starts</td></tr>
<tr><td><b>Preconditions:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Postconditions:</b></td><td colspan="3">A view of data</td></tr>
<tr><td rowspan="3"><b>Normal Flow:</b></td><td></td><td>Actor Input</td><td>System Response</td></tr>
<tr><td>1</td><td></td><td>System extracts data (abstract)</td></tr>
<tr><td>2</td><td></td><td>Extended by ‘Annotate process’ use case<br>Extension point: user wants to write a description of process steps</td></tr>
<tr><td><b>Exceptions:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Include:</b></td><td colspan="3"></td></tr>
</table>

| | |
|---|---|
| **Use Case UML ID:** | PR3 |
| **Use Case UML Name:** | Extract data from Db |
| **Primary Actor:** | - |
| **Secondary Actor(s):** | - |
| **Description:** | This UML Use Case allows users for extracting data from Promise db |
| **Trigger:** | Data in Db |

<table>
<tr><td><b>Preconditions:</b></td><td>-</td></tr>
<tr><td><b>Postconditions:</b></td><td>-</td></tr>
<tr><td><b>Normal Flow:</b></td><td>Point 1 of Use Case UML, labeled Load data
<table>
<tr><th></th><th>Actor Input</th><th>System Response</th></tr>
<tr><td>1</td><td></td><td>System extracts data from dB</td></tr>
</table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR4</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Import data</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows users for importing data from a file</td></tr>
<tr><td><b>Trigger:</b></td><td>-</td></tr>
<tr><td><b>Preconditions:</b></td><td>A file with data in local file system</td></tr>
<tr><td><b>Postconditions:</b></td><td>-</td></tr>
<tr><td><b>Normal Flow:</b></td><td>
<table>
<tr><th></th><th>Actor Input</th><th>System Response</th></tr>
<tr><td>1</td><td></td><td>System read a file</td></tr>
<tr><td>2</td><td></td><td>System populates Db with parsed data</td></tr>
</table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR5</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Manipulate data</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows system and users for manipulating data</td></tr>
<tr><td><b>Trigger:</b></td><td>After ‘Extract data’ Use Case finishes.</td></tr>
<tr><td><b>Preconditions:</b></td><td>A view of data, extracted from Promise Db</td></tr>
<tr><td><b>Postconditions:</b></td><td>A view of data, created by data manipulation</td></tr>
<tr><td></td><td>
<table>
<tr><th></th><th>Actor Input</th><th>System Response</th></tr>
</table></td></tr>
</table>

| | | | |
|---|---|---|---|
| | 1 | | System manipulates data extracted from dB |
| | 2 | The user is directly involved in the manipulation process deriving new attributes, applying some operations, changing the table structure or applying some process of visual analytics on data, according to his/her. | |
| | | | Extended by ‘Annotate process’ use case<br>Extension point: user wants to write a description of process steps |
| **Exceptions:** | - | | |
| **Include:** | | | |

| | | | |
|---|---|---|---|
| **Use Case UML ID:** | PR6 | | |
| **Use Case UML Name:** | Create visualizations | | |
| **Primary Actor:** | - | | |
| **Secondary Actor(s):** | - | | |
| **Description:** | This UML Use Case allows system for creating visualizations | | |
| **Trigger:** | When ‘Manipulate data’ Use Case finishes | | |
| **Preconditions:** | A set of ‘manipulated’ data | | |
| **Postconditions:** | Creation of visualizations | | |
| | | Actor Input | System Response |
| | 1 | User choices a type of graph | |
| | 2 | | The System creates visualizations |
| | 3 | | Extended by ‘Annotate process’ use case<br>Extension point: user wants to write a description of process steps |
| **Exceptions:** | - | | |

| | |
|---|---|
| **Include:** | |

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR7</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Interact graph</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows system for interacting graph</td></tr>
<tr><td><b>Trigger:</b></td><td>When ‘Create visualizations’ Use Case finishes</td></tr>
<tr><td><b>Preconditions:</b></td><td>A visualization on screen</td></tr>
<tr><td><b>Postconditions:</b></td><td>A new visualization on screen</td></tr>
<tr><td></td><td>
<table>
<tr><td></td><td>Actor Input</td><td>System Response</td></tr>
<tr><td>1</td><td>User changes graph parameters</td><td></td></tr>
<tr><td>2</td><td></td><td>System creates visualizations</td></tr>
<tr><td>3</td><td></td><td>Extended by ‘Save graph’ use case<br>Extension point: user wants to save a graph</td></tr>
<tr><td>4</td><td></td><td>Extended by ‘Annotate finding’ use case<br>Extension point: user wants to writes a comment</td></tr>
</table>
</td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

| | |
|---|---|
| **Use Case UML ID:** | PR8 |
| **Use Case UML Name:** | Save graph |
| **Primary Actor:** | - |
| **Secondary Actor(s):** | - |
| **Description:** | This UML Use Case allows user for saving graph in his local directory |
| **Trigger:** | When a user clicks on ‘Save graph’ button in visualizations view |
| **Preconditions:** | A graph on screen |
| **Postconditions:** | A graph saved in local file system |

<table>
<tr><td rowspan="3"></td><td></td><td>Actor Input</td><td>System Response</td></tr>
<tr><td>1</td><td>User choices a directory in local file system</td><td></td></tr>
<tr><td>3</td><td></td><td>System saves a graph in local file system</td></tr>
<tr><td><b>Exceptions:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Include:</b></td><td colspan="3"></td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td colspan="3">PR9</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td colspan="3">Task with predefined visualization</td></tr>
<tr><td><b>Primary Actor:</b></td><td colspan="3">A user</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td colspan="3">-</td></tr>
<tr><td><b>Description:</b></td><td colspan="3">This UML Use Case allows user for selecting a predefined task</td></tr>
<tr><td><b>Trigger:</b></td><td colspan="3">When a user clicks on a predefined task name</td></tr>
<tr><td><b>Preconditions:</b></td><td colspan="3">- It exists at least a saved predefined task<br>- A authenticated user</td></tr>
<tr><td><b>Postconditions:</b></td><td colspan="3">One or more visualizations</td></tr>
<tr><td rowspan="4"></td><td></td><td>Actor Input</td><td>System Response</td></tr>
<tr><td>1</td><td>A user selects an experiment (Use case ‘select experiment’)</td><td></td></tr>
<tr><td>2</td><td colspan="2">It defines a task (abstract)</td></tr>
<tr><td>3</td><td></td><td>System creates visualizations</td></tr>
<tr><td><b>Exceptions:</b></td><td colspan="3">-</td></tr>
<tr><td><b>Include:</b></td><td colspan="3">Select Experiment</td></tr>
</table>

| | |
|---|---|
| **Use Case UML ID:** | PR10 |
| **Use Case UML Name:** | Task with novel visualization |
| **Primary Actor:** | - |
| **Secondary Actor(s):** | - |
| **Description:** | This UML Use Case allows user for selecting a novel task |
| **Trigger:** | When a user click on a novel task name |
| **Preconditions:** | It exists at least a saved novel task |
| **Postconditions:** | - |
| | Point 2 of Use Case UML, labeled Predefined task |

<table>
<tr><td></td><td><table>
<tr><th></th><th>Actor Input</th><th>System Response</th></tr>
<tr><td>1</td><td></td><td>System shows a list of novel task</td></tr>
<tr><td>2</td><td>User choices a novel task</td><td></td></tr>
<tr><td>3</td><td></td><td>Extended by ‘Documentation visualization’ use case<br>Extension point: user wants to read a manual for a novel visualization</td></tr>
</table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR11</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Task with standard visualization</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows user for selecting a standard task</td></tr>
<tr><td><b>Trigger:</b></td><td>When  a user click on a standard task name</td></tr>
<tr><td><b>Preconditions:</b></td><td>It  exists at least a saved standard task</td></tr>
<tr><td><b>Postconditions:</b></td><td>-</td></tr>
<tr><td></td><td>Point 2 of Use Case UML, labeled Predefined task<table>
<tr><th></th><th>Actor Input</th><th>System Response</th></tr>
<tr><td>1</td><td></td><td>System shows a list of standard task</td></tr>
<tr><td>2</td><td>User choices a standard task</td><td></td></tr>
</table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

| | |
|---|---|
| **Use Case UML ID:** | PR12 |
| **Use Case UML Name:** | Documentation visualization |
| **Primary Actor:** | - |
| **Secondary Actor(s):** | - |
| **Description:** | This UML Use Case allows user for reading a manual for a novel visualization |
| **Trigger:** | When  a user click on a button, labeled ‘Documentation’ |

<table>
<tr><td><b>Preconditions:</b></td><td>It  exists at least a document associated to a novel visualization</td></tr>
<tr><td><b>Postconditions:</b></td><td>User read a document</td></tr>
<tr><td></td><td><table><tr><th></th><th>Actor Input</th><th>System Response</th></tr><tr><td>1</td><td></td><td>System shows a document to a user</td></tr></table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td></td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR13</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Select experiment</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows user for selecting a predefined experiment</td></tr>
<tr><td><b>Trigger:</b></td><td>When  a user clicks on a experiment task name</td></tr>
<tr><td><b>Preconditions:</b></td><td>It  exists at least a saved predefined experiment</td></tr>
<tr><td><b>Postconditions:</b></td><td>-</td></tr>
<tr><td></td><td><table><tr><th></th><th>Actor Input</th><th>System Response</th></tr><tr><td>1</td><td></td><td>System loads data of a predefined experiment</td></tr><tr><td>2</td><td></td><td>System shows data</td></tr></table></td></tr>
<tr><td><b>Exceptions:</b></td><td>-</td></tr>
<tr><td><b>Include:</b></td><td>Interact graph</td></tr>
</table>

<table>
<tr><td><b>Use Case UML ID:</b></td><td>PR14</td></tr>
<tr><td><b>Use Case UML Name:</b></td><td>Annotate process</td></tr>
<tr><td><b>Primary Actor:</b></td><td>-</td></tr>
<tr><td><b>Secondary Actor(s):</b></td><td>-</td></tr>
<tr><td><b>Description:</b></td><td>This UML Use Case allows user for annotating steps of process</td></tr>
<tr><td><b>Trigger:</b></td><td>User clicks on a button labeled ‘Annotate’</td></tr>
<tr><td><b>Preconditions:</b></td><td>-</td></tr>
<tr><td><b>Postconditions:</b></td><td>A description of annotate process</td></tr>
<tr><td></td><td><table><tr><th></th><th>Actor Input</th><th>System Response</th></tr><tr><td>1</td><td>User writes a description of process</td><td></td></tr></table></td></tr>
</table>

| | | | |
|---|---|---|---|
| | 2 | | System saves description |
| **Exceptions:** | - | | |
| **Include:** | | | |

| | | | |
|---|---|---|---|
| **Use Case UML ID:** | PR15 | | |
| **Use Case UML Name:** | Annotate finding | | |
| **Primary Actor:** | - | | |
| **Secondary Actor(s):** | - | | |
| **Description:** | This UML Use Case allows user for annotating visualization results | | |
| **Trigger:** | User clicks on a button labeled 'Annotate' | | |
| **Preconditions:** | - | | |
| **Postconditions:** | A description of annotate finding | | |
| | | Actor Input | System Response |
| | 1 | User writes a comment for results, that he has found. | |
| | 2 | | System saves comment of user |
| **Exceptions:** | - | | |
| **Include:** | | | |

# 6 Appendix : Infovis: data and visualizations

The following sections describe the most common Infovis visualizations relevant for the PROMISE environment.

## 6.1 Data

In order to better understand visualizations, in this section some issues about data will be introduced.

There are two types of data: numeric and categorical. The former allows arithmetic operations, the latter allows group, identify, organize but not arithmetic operations. Categorical data may also be expressed in form of numbers, but their meaning is intended not to be numerical, neither they are used for arithmetical purposes.

Another important data related issue, is data dimension. In fact, data can be mono-dimensional, bi-dimensional, three-dimensional, or multi-dimensional, depending on the number of data attributes being represented. In particular, if the data are multi-dimensional, the user need to represent the data with other attributes than X,Y,Z (for example color, dimension, different geometric shapes, etc.).

Moreover, when dealing with numerical data, it is important to avoid distortion effects. Graphs should not provide a distorted picture of the value they portray. There is also a mathematical way to measure the amount of distortion in a graph, this is called “lie factor”, and it is described by the following mathematic relationship:

Lie factor = size of effect in graphic / size of effect in data.

If the above value is greater than 1, the graph is exaggerating the size of the effect.

To avoid this effect, it is recommended to avoid representing single values through 2D or 3D surfaces that amplifies differences and are not well perceived by human beings.

On the other hand, image distortion is useful to solve the focus + context problem (see the table lens below).

## 6.2 Visualizations

### 6.2.1 Scatterplot

A scatterplot uses Cartesian coordinates to display two values for a set of data. It is a useful summary of a set of bivariate data, usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.

The data are displayed as a collection of points (or other geometric shapes), each having the X-Y coordinates set by the two chosen values of data set.

This representation affords an awareness of a general trend, of local trade-offs and of outliers that may be interesting and might not have been anticipated.

One of the most powerful aspects of a scatterplot, however, is its ability to show nonlinear relationships between variables. Furthermore, if the data are represented by a mixture model of simple relationships, these relationships will be visually evident as superimposed patterns.

The figure below has only an exemplificative purpose, like all the others proposed for the various visualization.

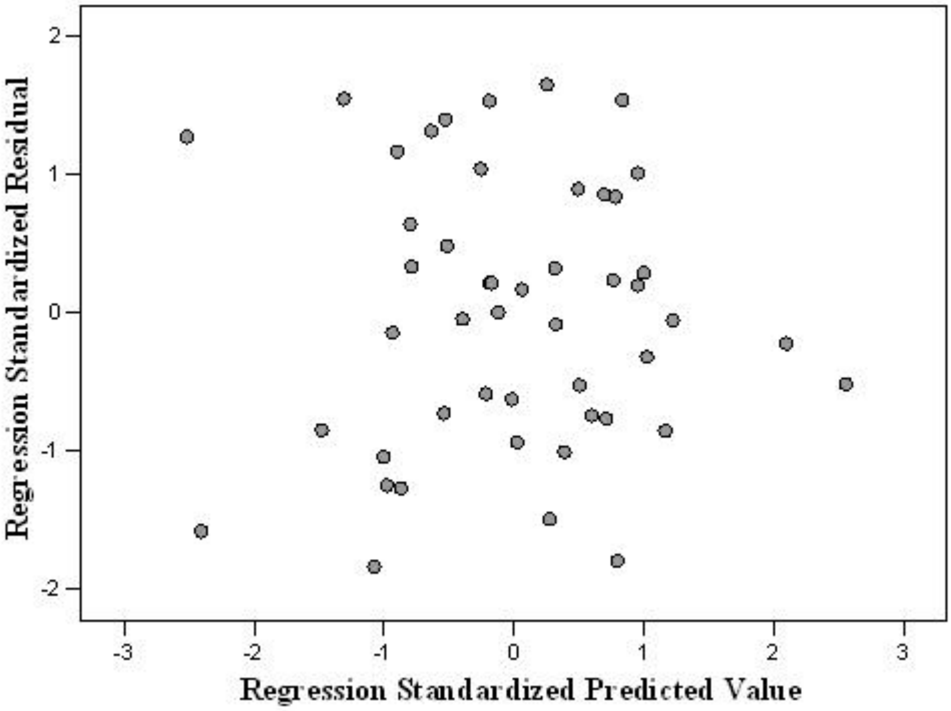


Figure 2: an example of bi-dimensional scatterplot

There exist several variations of the bi-dimensional scatterplots: the extension to the third dimension, the use of size, color, shape, etc., to increase the number of visualized attributes. However, 3D visualizations are affected by occlusion problems, and it is reasonable to assume that the number of attributes a scatterplot can reasonably represent is around five.

### 6.2.2 Scatterplot matrix

A scatterplot matrix is used to represent multidimensional data, i.e. a data set with n>>2 attributes. It is constituted by the combination of all the bi-dimensional scatterplots obtained considering all the n(n-1)/2 pairs of data attributes, and arranged on a single page in a matrix format.

The scatterplot matrix will have n rows and n columns and the ith row and jth column of this matrix is a plot of attributes i and j.

It has the same peculiarities of the scatterplot, but it allows for finding more correlations, having a complete visualizations of the various correlations between several attributes and several scatterplots.

If we are primarily interested in a particular variable, we can scan the row and column for that variable. If we are interested in finding the strongest relationship, we can scan all the plots and then determine which variables are related.

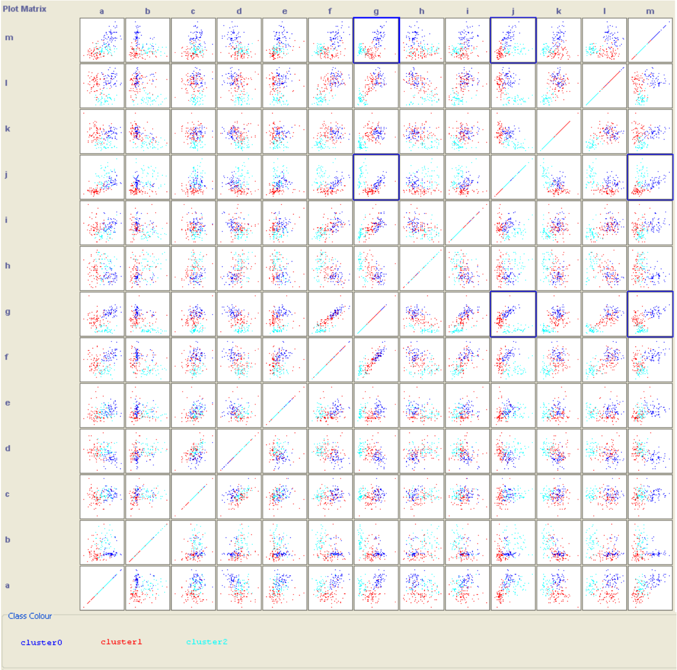


Figure 3 : an example of scatter plot matrix

### 6.2.3 Table lens

The idea of table lens is to incorporate a graphical representation into a table. The kind of representation depends on the type of attribute represented in a column. For example, if the type is numerical it is possible to use bars, normalizing them w.r.t the maximum value of the column; if the type is categorical it is possible to use different colours to represent each category. An advantage of table lens with respect to a traditional table is the possibility to see the trend of values in a column and/or to compare immediately the trend of two different columns.

The user can focus on an area of the table to see real values: the visualization uses a focus+context (fisheye) technique that works effectively on tabular information because it allows display of crucial label information and multiple distal focal areas. With the term focus+context it is intended a technique that supports visualizing an entire information structure at once as well as zooming in on specific items. This interplay between focus and context supports searching for patterns in the big picture and fluidly investigating interesting details without losing framing context.

In addition, a graphical mapping scheme for depicting table contents is provided. The table lens fuses symbolic and graphical representations into a single coherent view that can be fluidly adjusted by the user. In this way, the table lens supports navigating around a large data space easily isolating and investigating interesting features and patterns. This high-bandwidth interactivity enables and extremely powerful style of direct manipulation exploratory data analysis.

Typical interaction strategies allows for altering the table lens layout without bending any rows or columns. Cells in the focal area and the label row and column divide the total focus space of each dimension appropriately. Cells in the context divide the remaining space equally.

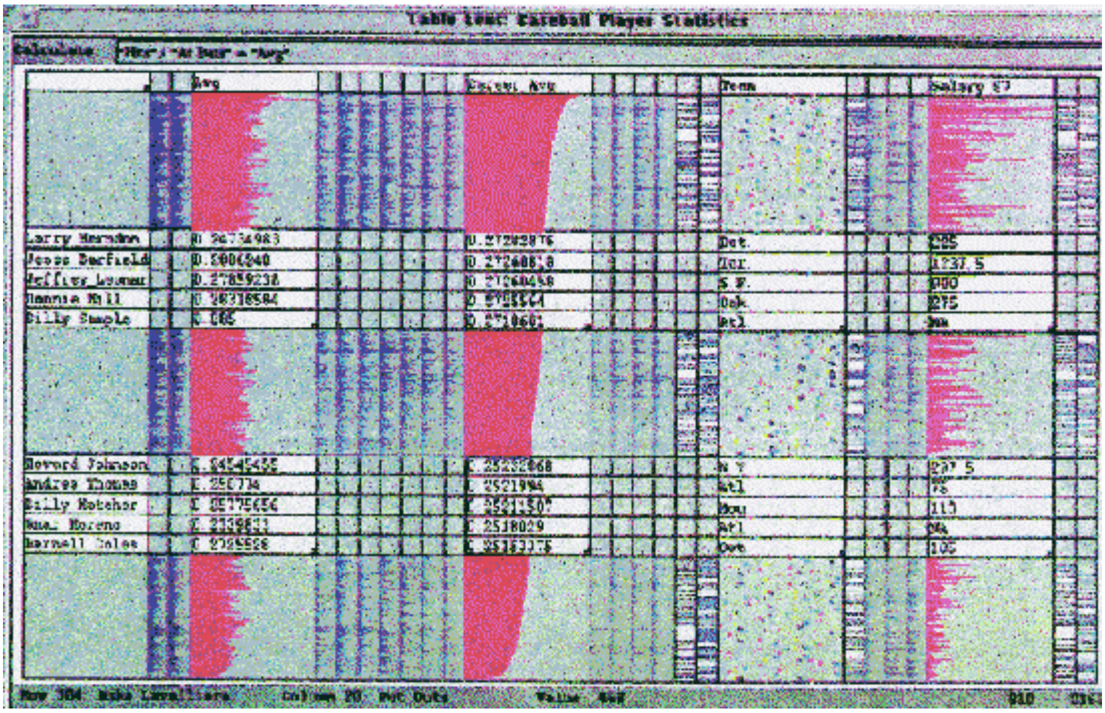


Figure 4: an example of table lens

### 6.2.4 Boxplot

The boxplot is a non-parametric visualization: this means that it displays the data without making any assumption of the underlying statistical distribution. It provides an excellent visual summary of many important aspects of a distribution and it is a tool for conveying location and variation information in data sets, particularly for detecting and illustrating location and variation changes between different groups of data.

More specifically, it shows

1. the median (shown as a line across the box) and the quartiles (the lower quartile is the 25th percentile and the upper quartile is the 75th percentile).
2. plots a symbol at the median (or draw a line) and draw a box (hence the name--box plot) between the lower and upper quartiles; this box represents the middle 50% of the data--the "body" of the data.
3. draws a line from the lower quartile to the minimum point and another line from the upper quartile to the maximum point. Typically a symbol is drawn at these minimum and maximum points, although this is optional.

Thus the box plot identifies the middle 50% of the data, the median, and the extreme points.

A single box plot can be drawn for one batch of data with no distinct groups. Alternatively, multiple box plots can be drawn together to compare multiple data sets or to compare groups in a single data set. For a single box plot, the width of the box is arbitrary. For multiple box plots, the width of the box plot can be set proportional to the number of points in the given group or sample (some software implementations of the box plot simply set all the boxes to the same width).

It takes up less space than a histogram and is therefore particularly useful for comparing distributions between several groups or sets of data. The Infovis research produced several variations of the original idea.

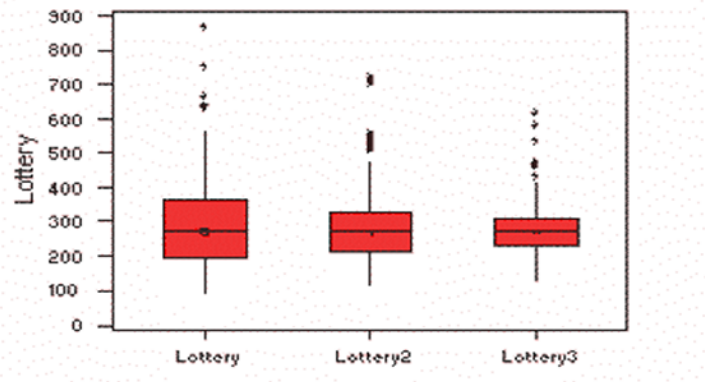


Figure 5: an example of boxplot

### 6.2.5 Histogram

The histogram is a summary graph showing a count of the data points falling in various ranges. If it is computed on a numerical variable the final effect is a rough approximation of the data value distribution, resembling its probability distribution. It represents aggregate properties, or derived values, of the data in a manner that can support both “at a glance” awareness and the need for more precise understanding.

The groups of data are called classes, and in the context of a histogram they are known as bins, because one can think of them as containers that accumulate data and "fill up" at a rate equal to the frequency of that data class.

Histograms are useful data summaries that convey the following information:

- The general shape of the frequency distribution (normal, chi-square, etc.)
- Symmetry of the distribution and whether it is skewed
- Modality - unimodal, bimodal, or multimodal

The histogram of the frequency distribution can be converted to a probability distribution by dividing the tally in each group by the total number of data points to give the relative frequency.

The shape of the distribution conveys important information such as the probability distribution of the data. In cases in which the distribution is known, a histogram that does not fit the distribution may provide clues about a process and measurement problem.

It consists of tabular frequencies, shown as adjacent rectangles, erected over the bins, with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1.

The shape of the histogram sometimes is particularly sensitive to the number of bins. If the bins are too wide, important information might get omitted. On the other hand, if the bins are too narrow, what may appear to be meaningful information really may be due to random variations that show up because of the small number of data points in a bin. To determine whether the bin width is set to an appropriate size, different bin widths should be used and the results compared to determine the sensitivity of the histogram shape with respect to bin size.

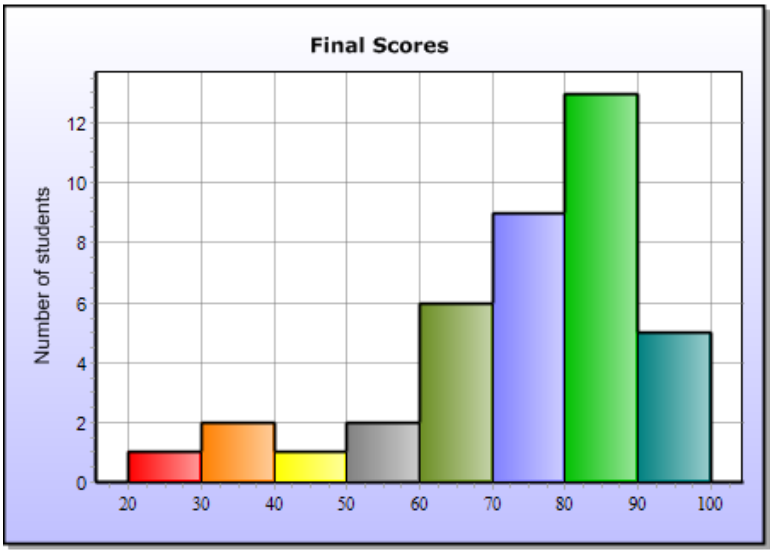


Figure 6: an example of histogram

If we are dealing with a categorical value, this technique leads to a means for representing a proportion and very often histogram values are represented as percentage. In such a case a simple histogram variation, the Pareto diagram, sorted for ascending or descending values is more appropriate.

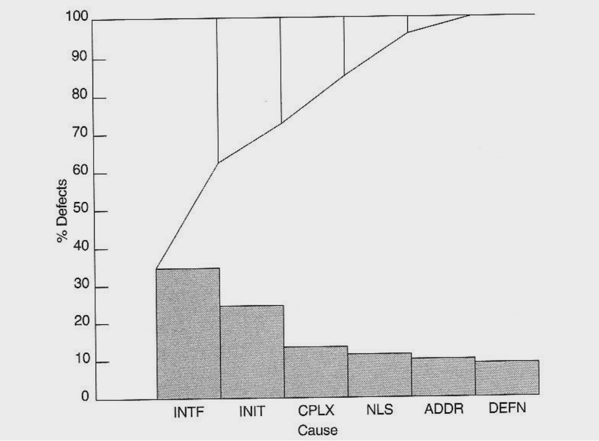


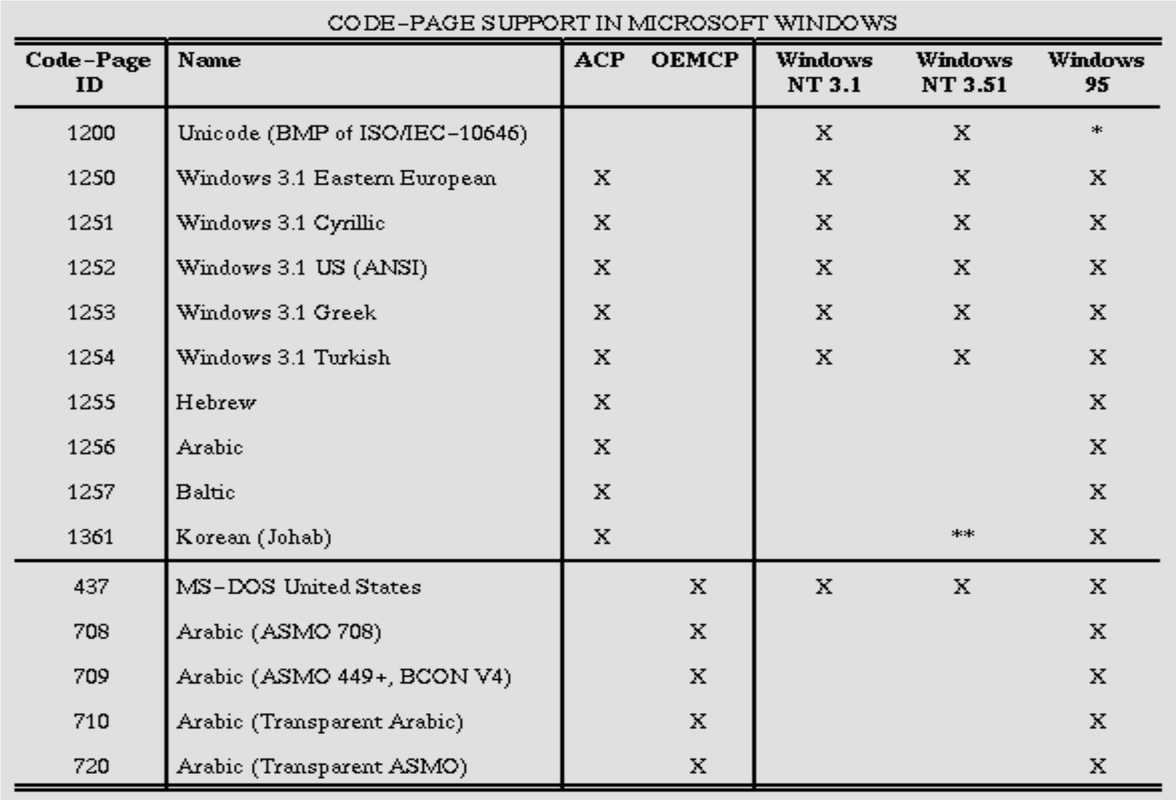
Figura 7: an example of Pareto diagram

### 6.2.6 Table

It is a well known means for arranging data in rows and columns.

The precise conventions and terminology for describing tables varies depending on the context. Further, tables differ significantly in variety, structure, flexibility, notation, representation and use.

Such a presentation is often of limited help, especially if there are many rows and ten or more columns. In the situation where a precise requirement exists, the table can usually been searched, by eye or by some automatic search mechanism, until an entry satisfying those requirements is found (or not). On the other way around, if we have to represent few values a table can be more effective than a visualization



CODE-PAGE SUPPORT IN MICROSOFT WINDOWS

| Code-Page ID | Name | ACP | OEMCP | Windows NT 3.1 | Windows NT 3.51 | Windows 95 |
|---|---|---|---|---|---|---|
| 1200 | Unicode (BMP of ISO/IEC-10646) | | | X | X | * |
| 1250 | Windows 3.1 Eastern European | X | | X | X | X |
| 1251 | Windows 3.1 Cyrillic | X | | X | X | X |
| 1252 | Windows 3.1 US (ANSI) | X | | X | X | X |
| 1253 | Windows 3.1 Greek | X | | X | X | X |
| 1254 | Windows 3.1 Turkish | X | | X | X | X |
| 1255 | Hebrew | X | | | | X |
| 1256 | Arabic | X | | | | X |
| 1257 | Baltic | X | | | | X |
| 1361 | Korean (Johab) | X | | | ** | X |
| 437 | MS-DOS United States | | X | X | X | X |
| 708 | Arabic (ASMO 708) | | X | | | X |
| 709 | Arabic (ASMO 449+, BCON V4) | | X | | | X |
| 710 | Arabic (Transparent Arabic) | | X | | | X |
| 720 | Arabic (Transparent ASMO) | | X | | | X |

**Figure 9: example of table**

### 6.2.7 Stacked bar charts

The stacked bar graph is used to compare the parts to the whole, i.e., to represent proportion with the bars themselves: the bars represent a count across a categorical value and the graph outlines the different components using marks, colors, etc. Often the proportion is expressed as a percentage.

Stacking is another way to represent the third dimension of data, doing this a single bar on the chart can show data for more than one category of data.

By stacking items and assigning a different color to each item, it is effectively possible to display trends among comparable or related items, or visually emphasize a sum of several indicators.

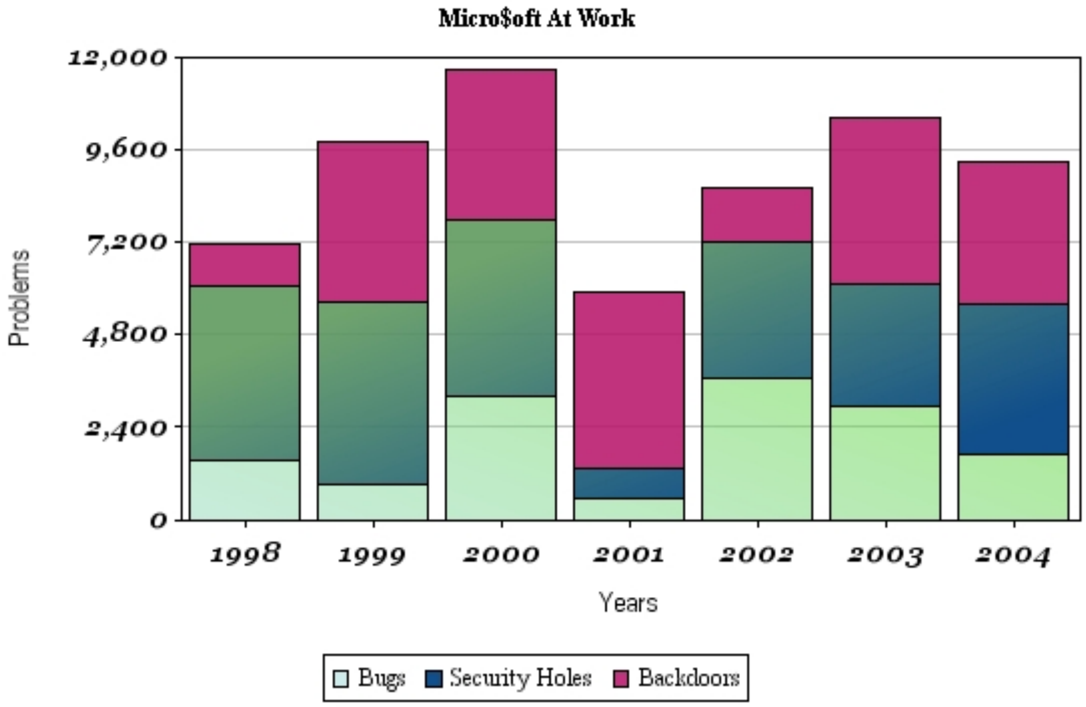
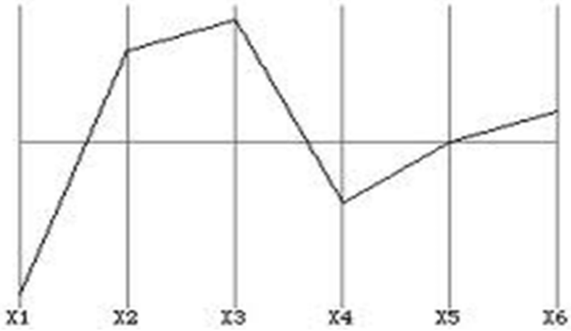


Figure 10: an example of stacked bar

### 6.2.8 Parallel coordinates

It is a common way of visualizing high-dimensional data. To show a set of points in an n-dimensional space, a backdrop is drawn consisting of n parallel lines, typically vertical and equally spaced. A point in n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point.



There are three important considerations:

- **The order of the axes** is critical for finding features, and in typical data analysis many reorderings will need to be tried. Some authors have come up with ordering
- **The rotation** of the axes is a translation in the parallel coordinates and if the lines intersected outside the parallel axes it can be translated between them by rotations. The simplest example of this is rotating the axis by 180 degrees.
- The necessity of **scaling the axes** derives from the fact that the plot is based on interpolation (linear combination) of consecutive pairs of variables. Therefore, the variables must be in common scale, and there are many scaling methods to be considered as part of data preparation process that can reveal more informative views.

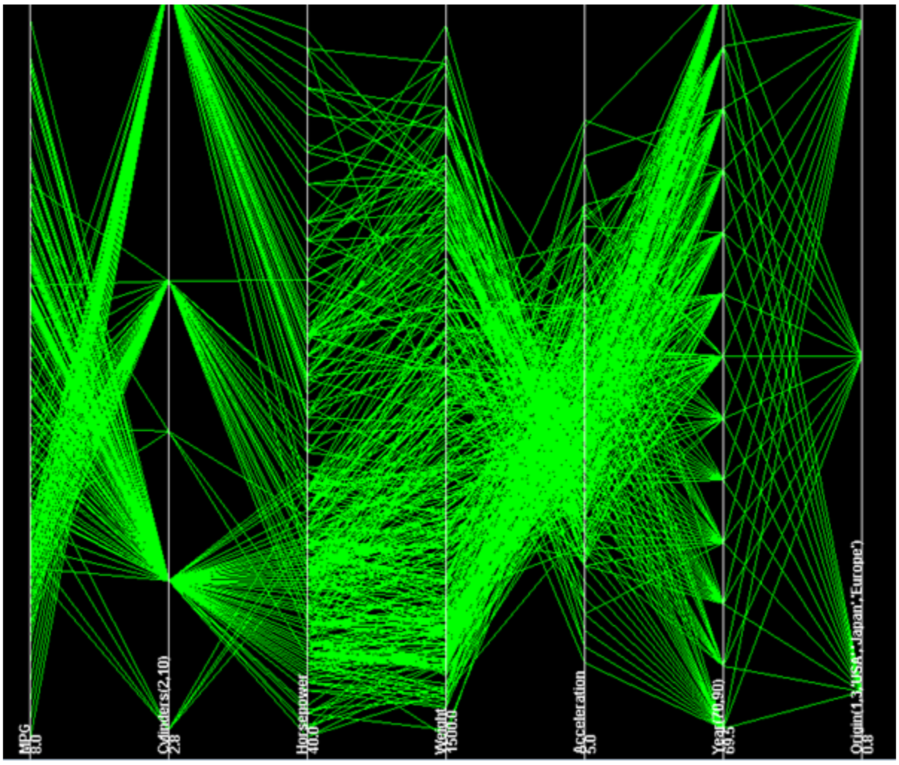
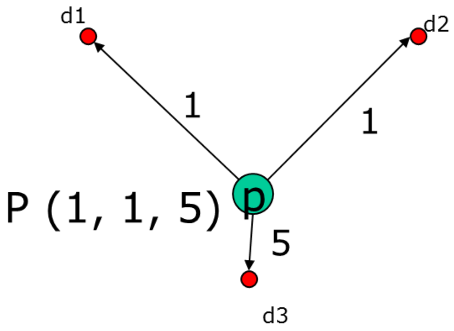


Figure 11: an example of parallel coordinates

### 6.2.9 Radviz

Radviz is a neat non-linear multi-dimensional visualization technique that can display data on three or more attributes in a 2-dimensional projection. The visualized attributes are presented as anchor points equally spaced around the perimeter of a unit circle. Data instances are shown as points inside the circle, with their positions determined by a metaphor from physics: each point is held in place with springs that are attached at the other end to the attribute anchors.



The stiffness of each spring is proportional to the value of the corresponding attribute and the point ends up at the position where the spring forces are in equilibrium. Prior to visualization, attribute values are scaled to lie between 0 and 1. Data instances that are close to a set of feature anchors have higher values for these features than for the others.

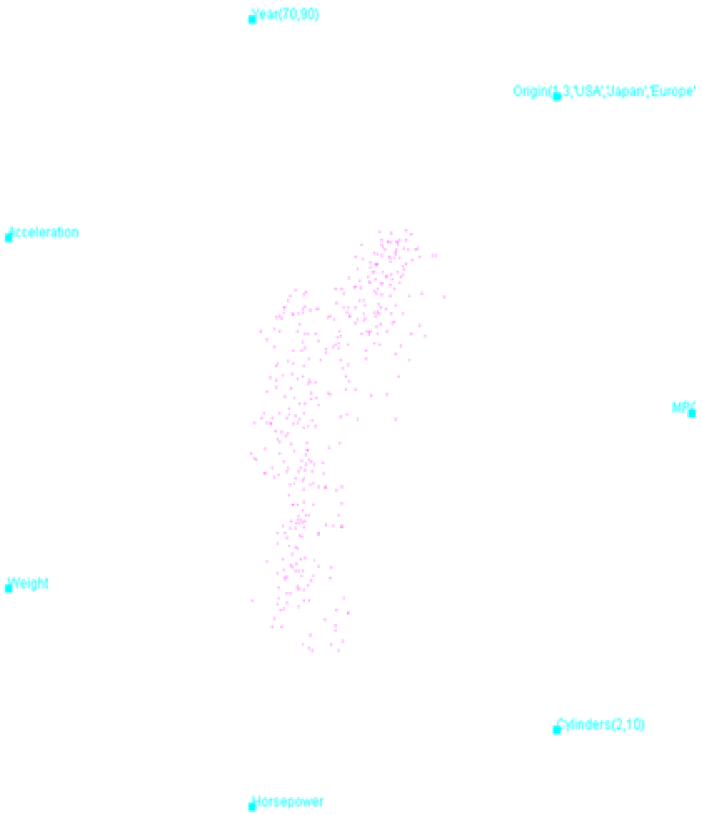


Figure 12: an example of radviz (the same dataset depicted on figure 11)

# References

[A06] Anderson N. H. "A functional theory of cognition". Erlbaum, Mahwah, NJ. Atkins, 2006.

[BE83] Bertin, J. (1983). Semiology of graphics (William J. Berg, Trans.). Madison, Wis.: University of Wisconsin Press.

[BGS07] Bertini E., Di Girolamo A., Santucci G. See what you know: analyzing data distribution to improve density map visualization. In: Proc. of the International Eurovis 2007 conference. Norrkoping, Sweden, May 2007.

[BS03] Benjamin B. Bederson and Ben Shneiderman (2003). The Craft of Information Visualization: Readings and Reflections, Morgan Kaufmann ISBN 1-55860-915-6.

[BS06a] Bertini E., Santucci G. (2006). "Visual Quality Metrics." In: ACM Digital Library - Proc. del International Workshop BELIV'06 BEyond time and errors: novel evaLuation methods for Information Visualization, International workshop of the AVI 2006 International Conference. Venezia, Maggio 2006.

[BS06b] E. Bertini and G. Santucci, "Give chance a chance - modeling density to enhance scatter plot quality through random data sampling", Information Visualisation, 5(2), pp. 95-110, June 2006.

[BRA97] Richard Brath. "Concept demonstration metrics for effective information visualization." In IEEE Symp. on Information Visualization, 1997.

[CH05]Chen, C. Top 10 Unsolved Information Visualization Problems, IEEE Computer Graphics and Applications, July/August 2005,

[CH07]Chang, K. (2007) Introduction to Geographic Information System, 4th Edition. McGraw Hill.

[CJ05] Image Processing and Analysis - Variational, PDE, Wavelet, and Stochastic Methods by Tony F. Chan and Jackie (Jianhong) Shen, (2005).

[CW90] Carswell, C. M., & Wickens, C. D. (1990). The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object integration. Perception & Psychophysics, 47, 157-168.

[DM05] J. Dykes, A. M. MacEachren, "Exploring Geovisualization", Pergamon, 2005.

[EBB05] Engel, D., Bertel, S., & Barkowsky, T. (2005). "Spatial principles in control of focus in reasoning with mental representations, images, and diagrams". In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), Spatial Cognition IV. Reasoning, Action, Interaction. Berlin: Springer.

[ED07] Ellis G P. and Dix A. A "Taxonomy of Clutter Reduction for Information Visualisation." IEEE Trans. on Visualization and Computer Graphics (Proc. Visualization/Information Visualization 2007), Vol.13, No.6, Nov/Dec 2007, IEEE, pp.1216-1223.

[FA85] Falmagne J-C. (1985). "Elements of psychophysical theory". Oxford University Press, NY

[GRI06] H. Goodell, C. Chiang, C. Kelleher, A. Baumann and G. Grinstein. Metrics for analyzing rich session histories. In Proc. of the 2006 AVI Workshop BELIV '06 (Beyond Time and Errors: Novel Evaluation Methods For information Visualization), Venice, Italy.

[GWC98] Gillan, D. J., Wickens, c.D., Hollands, J.G., Carswell, C. M. (1998). Guidelines for Presenting Quantitative Data in HFES Publications. Human Factors, 14, 28-41.

[HMTW83] Hoaglin, D C; Mosteller, F & Tukey, John Wilder (Eds) (1983). Understanding Robust and Exploratory Data Analysis.

[IEEE 1998] Recommended Practice for Software Requirements Specifications, IEEE 830-1998, http://standards.ieee.org/findstds/standard/830-1998.html

[KMS06] Keim, D.A.; Mansmann, F. and Schneidewind, J. and Ziegler, H., "Challenges in Visual Data Analysis." Proc. of Information Visualization (IV06), IEEE, p. 9-16, 2006.

[KO06] Kosslyn, S. M. (2006). Understanding charts and graphs. Applied cognitive psychology, 3, 185-225.

[KR03] Kraak M., "Geovisualization illustrated", ISPRS journal of photogrammetry and remote sensing. 57 (2003) 390- 399 ISSN 0924-2716.

[KR06] Kraak Menno-Jan, "Visualization Viewpoints" Theresa-Marie Rhyne ed.,July/August 2006, IEEE Computer Society.

[KK96] Daniel A. Keim and Hans Peter Kriegel. "Visualization techniques for mining large databases:" A comparison. IEEE Trans. on Knowledge and Data Engineering, 8, 1996.

[LGMR01]P.A. Longley, M.F. Goodchild, D.J. Maguire, D.W. Rhind, Geographical Information System Principles 2 Edition 2001. Wiley.

[PLA04] C. Plaisant, “The Challenge of Information Visualization Evaluation”, In *Proceedings of Working Conf. Advanced Visual Interfaces (AVI ’04)*, pp. 109-116, 2004.

[PROMISE D2.1, 2011] Karlgren, J., Eriksson, G., Frieseke, M., Gäde, M., Hansen, P., Järvelin, A., Lupu, M., Müller, H., Petras, V., and Stiller, J. (2011). *Deliverable D2.1 – Initial specification of the evaluation tasks*. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/5fce1a18-d6c2-4063-958c-bb3e73a27456.

[PROMISE D3.1, 2011] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2011). *Deliverable D3.1 – Initial prototype of the evaluation infrastructure*. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/e0df8a3c-388f-40e8-bfbd-04434a393004.

[PROMISE D3.2, 2011] Agosti, M., Braschler, M., D Buccio, E., Dussin, M., Ferro, N., Granato, G.L., Masiero, I., Pianta, E., Santucci, G., Silvello, G., and Tino, G. (2011). *Deliverable D3.2 – Specification of the evaluation infrastructure based on user requirements*. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/fdf43394-0997-4638-9f99-38b2e9c63802.

[PROMISE D5.1, 2011] Croce, M., Di Reto, E., Granato, G. L., Hansen, P., Sabetta, A., Santucci, G., and Veltri, F. (2011). *Deliverable D5.1 – Collaborative user interface requirements*. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise- noe.eu/documents/10156/50834686-2118-48f8-a57b-8553ec3d7981.

[PU97] Purghé F. (1997), "Metodi di psicofisica e scaling unidimensionale", Boringhieri, Torino

[PW06] Chor Pang Lo, Albert K.W. Yeung (2006), Concepts and Techniques of Geographic Information Systems, Prentice Hall, Inc., Upper Saddle River, NJ, USA.

[SR96] Scaife, M. & Rogers, Y. (1996). External cognition: how do graphical representations work? Internation Jornal of Human-Computer studies, 45, 185-213.

[SP90] Spence, I. (1990) "Visual psychophysics of simple graphical elements". Journal of Experimental Psychology: Human Performance and Perception, 16, 683-692

[ST75] Stevens S. S. (1975). "Psychophysics. Introduction to its perceptual, neural, and social aspects", Wiley, NY.

[TUF83] E. R. Tufte. The Visual Display of Quantitative Information, Graphics Press, 1983.

[VVAA2010] Mastering The Information Age – Solving Problems with Visual Analytics, Eurographics Association, 2010.

[WA04] Ware C. (2004). "Information visualization", Morgan Kaufmann, San Francisco.

[WC95] Wickens, C. D. & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevence to display design. Human Factos, 22, 473-494.

[WM08] Ware & Mitchell, (2008). "Visualizing graphs in three dimensions". ACM Transaction on Applied Perception, 5 (1), article 2.

[WON04] P.C. Wong e J. Thomas, "Visual analytics - guest editors' introduction," IEEE Trans. on Computer Graphics and Applications, vol. 24(5), 20-21, 2004.

[ZT99] Zack & Tversky, (1999). "Bars and Lines: A Study of Graphic Communication". Memory and Cognition, 27, 1073-9.