A Visual Analytics Platform to Measure Performance on University Entrance Tests (Discussion Paper)

Daniele Boncoraglio[3], Francesca Deri[3], Francesco Distefano[3], Daniele Fadda[1,2], Giorgio Filippi[3], Giuseppe Forte[3], Federica Licari[3], Michela Natilli[1,2], Dino Pedreschi[1], and Salvatore Rinzivillo[2]

[1] Computer Science Department - University of Pisa, Italy pedre@di.unipi.it
[2] KDDLab, ISTI-CNR Pisa, Italy {daniele.fadda,michela.natilli, rinzivillo}@isti.cnr.it
[3] CISIA - Consorzio Interuniversitario Sistemi Integrati per l'Accesso, Pisa, Italy name.surname@cisiaonline.it

Abstract. Data visualization dashboards provide an efficient approach that helps to improve the ability to understand the information behind complex databases. It is possible with such tools to create new insights, to represent keys indicators of the activity, to communicate (in real-time) snapshots of the state of the work. In this paper, we present a visual analytics platform created for the exploration and analysis of performance data on entrance tests taken by Italian students when entering the university career. The data is provided by CISIA (Consorzio Interuniversitario Sistemi Integrati per l'Accesso), a non-profit consortium formed exclusively by public universities. With this platform, it is possible to explore the performance of the students along different dimensions, such as gender, high school of provenience, type of test and so on.

Keywords: Data visualization · Performance indicators · University test.

1 Introduction

The effective graphical presentation of information is an essential skill in most scientific disciplines. Visualization is intended to clearly convey and communicate information through graphical means, enabling everyone to comprehend data in a much more explicit way[1]. Through visualization, the results of data processing are made more accessible, straightforward, and user-friendly[2]. Communicating information through visual analytic techniques facilitates understanding of information to those who have no specific technical or domain knowledge[3].

A proper way to describe graphically a complex phenomenon is through the use of dashboards. As stated in [4], "compared to visualization modalities for presentation and exploration, dashboards bring together challenges of at-a-glance reading, coordinated views, tracking data and both private and shared awareness". The access to complex phenomena is made possible through the realization of multiple linked displays [5], where sub-dimensions of the whole data are showed in details and the interaction on one display propagate selection and filtering to the other's views. Thus, the complexity of the data to explore is mitigated through the use of interaction with the observer.

In this paper we present a visual analytic platform created for the exploration and analysis of performance data on entry tests taken by Italian students before entering the university career. The data is provided by CISIA[4] (Consorzio Interuniversitario Sistemi Integrati per l'Accesso), a non-profit consortium formed by public universities. Currently, CISIA consortium counts 45 Universities and the Conferences of Engineering, Architecture and Sciences among the consortium members: the CUIA - Italian University Architecture Conference, the CopI - Conference for Engineering and Con.Scienze - National Conference of Presidents and Structure Directors University of Science and Technology.

The Consortium is open to the participation of all Italian universities; among the different statutory purposes, the main is to organize and coordinate the orientation activities for the access to the universities. Entrance test organized by CISIA serve two different purposes:

- for students enrolling the test, it provides a self-assessment of their preparation and aptitude to undertake the chosen field of studies;
- for the faculties and departments, the tests gives a view of the actual knowledge of the students, and let the management to compare it with the minimum skills and capabilities requirements in order to prepare specific orientation and training activities.

For those course with a restricted number of students, the tests are used as screening and ranking tools. CISIA tests are currently available for six areas: Engineering, Economics, Pharmacy, Sciences, Humanities, and Agriculture.

2 The visual analytics platform

2.1 Problem statement

CISIA Online Test (acronym TOLC) is a tool for orientation and assessment of the knowledge required for access to the Study Programs of Italian Universities, which can be used to select students for access, provided on a computerized platform. TOLC is an individual test, different from student to student, automatically composed for each student by a software. The software follows a set of recipes to guarantee that all the tests generated are equivalent in terms of the level of difficulty. The objective of the platform proposed in this paper is

[4] `http://www.cisiaonline.it`

to provide an analytic framework where all the performances on each test may be analyzed. Historical data are maintained since 2012, and they register many dimensions from each test: the number of correct answers, high school education of the student attending the test, the geographical origin, the category of each question in the test. Thus the platform should allow an efficient exploration of the many dimensions and should produce an accessible interface where the observer may interact with the data. During the design phase, we identified two kind of users: public visitor of the website of the consortium, and the site managers of the department that organize the tests. The two profiles are managed through a similar web interface, whereas the users with an authorized account may access more detailed data, for example analyzing the performance of the tests organized in her own site.

2.2 Platform design

Accordingly to Sarikaya et al. [4], we may classify our platform as a Communication Dashboard, whose purpose is to inform the public about Consortium activities. Since it is a public dashboard whose primary purpose is to describe social relevant data, great attention was given to the privacy of the users who performed the tests. All data delivered by the platform are anonymized through aggregation and filtering. The subset of data with few representatives (i.e. few users attending the test) is removed from the analysis. This filtering phase allows us to enforce the statistic robustness of the presented data and also to make not identifiable any student in the dataset. One of the main objectives of the platform was the accessibility even for non-expert users. We select only charts that are widely recognizable, mainly bar chart and line chart. To show in details some distributions we also adopted a box plot chart. To guarantee high readability of the result of the chart, we inserted a very detailed explanation of how to read the graph. The user can explore the data through three dimensions: year, area of interest, geographic origin. It is possible to apply a filter to restrict the analysis only on the top 10% ranking tests (best scores). A subset of charts also allows a specialized dimension on gender. The selection of the dimensions to be analyzed can be managed by the user through an integrated toolbox, where the current selection is always visible. When the user modifies one of the dimensions, the visualization is updated.

A schematic representation of the architecture of the CISIAViz platform is presented in Fig.1. The database layer is organized with two levels. The production database used to manage the tests is kept separated from the CISIAViz platform. An ETL (Extract, Transform, and Load) procedure is scheduled to update a staging database for the visualization. During this phase, sensitive information is removed.

The database is running over an instance of MySQL, the same DBMS adopted for the rest of the TOLC system. The data warehouse schema in Fig.2 shows the relations implemented within the DB. The entity *Test* is the main table containing all the results of each test. Since the sections for all the possible tests are limited, we decided to store them into the field within this table. The test

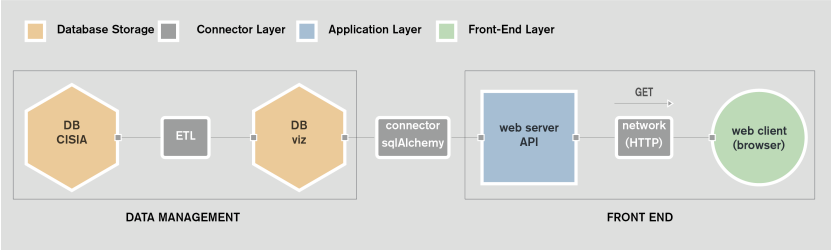

Fig. 1. Web-App schema: Starting from CISIA database, through a process of Extract, Transform, Load (ETL), we build the data warehouse (DB viz). The data warehouse contains only the data that is of interest to the CISIAViz platform. During the ETL phase, new dimensions are calculated (such as the ranking of students by type of test) and some queries are optimized by materializing data (including redundant ones) retrieved from different tables to speed up the subsequent reading phase via web API.

provided in a different area of interests may have different compositions. When a test does not contain a specific section, the corresponding value in the table will be NULL. On one side, this implementation makes the schema of the DB not flexible to accommodate new sections. On the other hand, this allows much faster access during the visualization, as it does not need to create too many joins between the tables.

The relevant dimensions of the star schema are *Geo*, the geographic dimension, and *Time*, the time dimension: the first one is used to explore the origin of the students applying for the test, and the second one allows us to select the time interval to aggregate the tests. Although it is possible to aggregate for an arbitrary temporal interval, to guarantee a sufficient aggregation for sensitive information, the time selection is restricted at the year level.

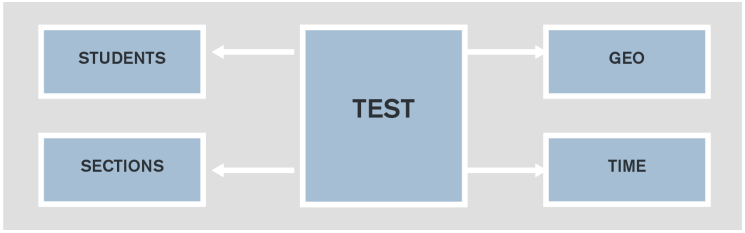

Fig. 2. The schema of the data warehouse. The fact table contains the details of each test performed by each student. This table is linked with the geographical dimension to determine the origin of each student and the temporal dimension to select the time window to analyze.

The access to the database is managed by an application server running a web API implementing the endpoints to access the data. All the interaction of

the user with the visualization dashboard will be translated into the corresponding call from the web application. The access to the web API is managed through HTTP GET calls. The authentication, implemented with an SSO system (Single Sign-On) using an IDP (Identification Provider) of the consortium, is performed with HTTP-POST methods. When the user changes the preferences on the dimensions, the web application generates a set of call to the Web API to retrieve the updated data. The application server is implemented in Python, leveraging the Flask library to handle the implementation of the endpoints. The web application running in the front end is a single page application running in Javascript. The dashboard is implemented using the state of the art libraries for managing of user interfaces: *D3.js* to create custom visualization; *Vue.js* to manage the events and selection of the user and to make the internal data representation reactive to new updates; *Plot.ly* to construct high-level charts using configuration descriptors automatically generated by data.

2.3 Backend

The majority of the endpoints of the Web API share a common parameter structure. This design choice allows us to make the front end flexible by composing the set of parameters automatically to be provided to the server. All the selection a user can make on the web interface are projected into a sequence of parameter values that are sent to the web server. All calls that depend on this filter have the following scheme:

```
@ app.route('/api/function_rest/<year>/<area_interest>/<region>/<quantile>')
```

where `<year>` and `<area_interest>` are the only mandatory parameters. The parameters reflect the possible combination of the dimensions the user may explore:

`<years>` Selection of the year in a range between 2012 and the current year

`<interest>` Select the type of test. Currently, TOLC-I (engineering), TOLC-E (economics), TOLC-F (pharmacy) can be selected.

`<region>` Selection of the region of origin of the candidates to the tests. In case of missed selection (default status) the value `all` is used.

`<quantile>` during the ETL phase, each candidate is associated with a percentile ranking positioning with respect to the score of all the candidates who carried out the test. This filter allows selecting candidates with values higher than a certain percentile. To simplify the exploration for non-expert users, the CISIAViz platform allows selecting only the "best scores", i.e. the score of the top 10th percentile of tests in the DB for the current combination of year and area of interest.

2.4 Front end

The CISIAViz platform is implemented as a single-page web application. The platform is deployed within the website of the consortium[5].

[5] `http://www.cisiaonline.it/area-tematica-cisia/report/`

The *main* script is the entry point to the application and is responsible for aligning the selection on the interface with the data to the DB, by querying the Web API with new data when needed. Here the event handling of the application is orchestrated and it updates all graphs when new data arrives. Every chart is handled by a general class module *chart*, whose input is transformed accordingly to the type of figure to show using a dedicated *data_transformer*. The configuration and the mapping of transformers to charts are described into the *chart_config*. module. The modules *filter*, *toolbar*, *event_handler* take care of sending the requests to the server.

The web application is composed of a central column showing all the statistics, and a sidebar where the user can manage to insert the constraint for the dimensions to explore: year, area of interest, region of origin, best scores.

The central column is organized into sections, each of them focusing on a specific aspect of the data:

- *Volumes*: the first section shows the number of tests executed with reference to the actual selection, presenting the partial sums for sub-periods of the year.
- *Scores*: this section shows the average score per section in the selected tests and the distribution of each section.
- *Repetition of tests*: here we show how the score varies for those students that repeated the test multiple times.
- *Region of origin*: here we cross filter both the geographical origin and the high school diploma of the students attending the test.
- *Historycal data*: here we show the trends of the test along the years. This section is not affected by the year selection in the sidebar since shows always the statistics of the last five years.

3 Use cases

To demonstrate the efficacy of the web application to deliver high-level knowledge extracted from the data, we show the capabilities of the dashboard through a use case of a person interested in the performance of students. For example, let us consider a journalist interested in documenting the skills of students when entering university: a study of how many individuals take the entry test, how they behave in terms of performance and the presence of differences between types of students, could be an engaging topic for an article on young people. Leveraging the dashboard for the different areas of interest, it is possible to get information on the distribution, among gender, of students taking the tests, as shown in figure 3.

From the figure, it can be seen how, also in the testing phase, the women are underrepresented especially among STEM degree (in this case Engineering). The higher representation of female students is found for Economics and Pharmacy.

It is also possible to compare the performance between male and female students in a different section of the test: for this graph (but also for others) there

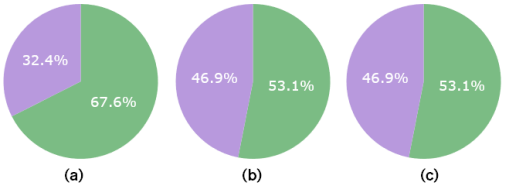

Fig. 3. Distribution of students by gender: percentage of students who took the test in Engineering (a), Economics (b) and Pharmacy (c)

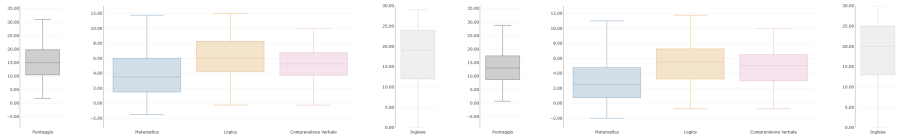

Fig. 4. Distribution of scores on topics: in the graph on left the male distribution of scores, while the graph on the right shows the distribution for female students

are local selectors with which selecting the female, the male or the total population who took a specific test (TOLC). In figure 4 it is shown the distribution of scores among male (top image) and female (bottom image) students. No significant differences can be found between the two genders, except for the results obtained in mathematics, where male students perform better than females. In the figure 4 is reported the distribution of scores for the test for Economics in 2018, but obviously, the same information can be obtained for the other areas (TOLC) and also the trend among years can be seen changing the selector for the years. Another possible comparison can be made among different Italian regions. In figure 5 the average scores obtained to the tests is reported for each region. The results reported in the figure are related to the TOLC in Economics, but the same results can be obtained for Pharmacy or Engineering. In this case average score for each section of the TOLC-Economics of 2018 divided by region and school of origin (all in this specific case). Only regions with a significant number of tests performed are shown. Furthermore, a comparison among different high schools can be made. The detailed average score has been calculated for scientific high school, classical high school, technical institute, technical institute for surveyors, commercial technical institute, professional institute and magistral and pedagogical institute.

The dashboard, therefore, allows you to get a picture of the situation of the students who decide to follow the course of study: a journalist interested in this topic and the differences between students with diverging characteristics (both gender or region of origin), can create his article guided by the data and statistics shown on the portal.

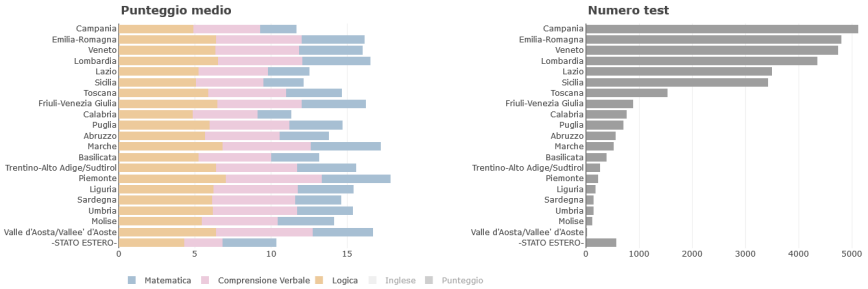

Fig. 5. Average scores among regions: on the left graph the average score made by students in each region, on the right graph the number of administered tests by regions.

4 Conclusions

In this paper we present a visual analytic platform to explore the performances of students when applying for entry tests to the universities. The platform was developed using the state-of-the-art of the technologies for web developing. The visual design of the dashboard was developed to support the user through a set of steps to select sub-dimension of the data. The toolbox which guides the composition of the visual platform is organized in steps and the user may select the relevant parameters to start the exploration. More specific selections, like gender selection or the selection of the high school diploma of origin, are located close to the specific corresponding chart. The system has been deployed and it is used efficiently both by non-expert users and domain experts (with login-based access).

References

1. Lee, M. D., Butavicius, M. A., Reilly, R. E.: Visualizations of binary data: A comparative evaluation. International Journal of Human-Computer Studies, **59**(5), 569-602 (2003).
2. Tao, F., Qi, Q., Liu, A., Kusiak, A.: Data-driven smart manufacturing. Journal of Manufacturing Systems **48**, 157–169 (2018)
3. Gabrielli, L., Rossi, M., Giannotti, F., Fadda, D., Rinzivillo, S.: Mobility Atlas Booklet: an urban dashboard design and implementation. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 4 (2018).
4. Sarikaya, A., Correll, M., Bartram, L., Tory, M., Fisher, D.: What Do We Talk About When We Talk About Dashboards?. IEEE transactions on visualization and computer graphics, 25(1), 682-692 (2019).
5. Andrienko, Gennady and Andrienko, N.: Exploring spatial data with dominant attribute map and parallel coordinates. Computers, Environment and Urban Systems, 25(1), (2019).