EuroVis Workshop on Visual Analytics (2017)
M. Sedlmair and C. Tominski (Editors)

Visual Analytics for Information Retrieval Evaluation Campaigns

M. Angelini[1] , N. Ferro[2], G. Santucci[1], and G. Silvello[2]

[1]University of Rome "La Sapienza", Italy
[2]University of Padua, Italy

Abstract

Information Retrieval (IR) has been deeply rooted in experimentation since its inception, allowing researchers and developers to understand the behavior and interactions within increasingly complex IR systems, such as web search engines, which have to address ever increasing user needs and support challenging tasks. This paper focuses on the innovative Visual Analytics (VA) approach realized by the Participative Research labOratory for Multimedia and Multilingual Information Systems Evaluation (PROMISE) environment, which simplifies and makes more effective the experimental evaluation process by allowing a formal and structured way to explore the complex data set of measures produced along an evaluation campaign. The system uses the result produced within the Conference and Labs of the Evaluation Forum (CLEF) [Cle].

1. Introduction

International large-scale evaluation campaigns for information access and retrieval systems provide fundamental contributions to the advancement of state-of-the-art techniques through common evaluation procedures, regular and systematic evaluation cycles, comparison and benchmarking of the adopted approaches, and spreading of knowledge. In the process, vast amounts of experimental data are generated that are in need of analysis tools to enable interpretation and thereby facilitate scientific and technological progress. The main goal is to evaluate submitted experiments (in the form of information retrieval algorithm results) in order to assess improvements and or losses on different retrieval tasks and tracks.

The PROMISE[†] project aims at providing a virtual laboratory for conducting participative research and experimentation to carry out, bringing automation into the evaluation and benchmarking of such complex information systems, by facilitating management and offering access, curation, preservation, re-use, analysis, visualization, and mining of the collected experimental data. The contribution of this paper is to bring a radical innovation into the experimental evaluation practices by developing an appropriate VA environment allowing stakeholders for effectively analyzing and understanding the huge amount of experimental data gathered in the evaluation infrastructure; the VA environment relies on multiple synchronized views that are selected by navigating the structure of a data metrics cube, built on the experiment, topic, metric axes, described in the next section. The VA environment uses mainly traditional visualization paradigms (e.g., scatterplot, bar chart, box plots, etc.), according to the IR community culture; however some of them introduce novelties to cope with specific IR evaluation objectives.

2. Information Retrieval Evaluation

Large-scale evaluation campaigns rely mainly on the traditional Cranfield methodology [Cle97] which makes use of shared experimental collections in order to create comparable experiments and evaluate their performance. An experimental collection is a triple $\mathcal{C} = (D, Q, J)$, where: D is a set of documents, called also collection of documents; Q is a set of topics (ranging from 25 to 4000) simulating actual user information needs, from which the actual queries are derived; J is a set relevance judgements, i.e., for each topic $q \in Q$ the documents $d \in D$, which are relevant for the topic q, are determined.

Overall, an experimental collection $\mathcal{C}$ allows the comparison of two retrieval methods, say X and Y, according to some measurements which quantifies the retrieval performances of these methods. Therefore, an experimental collection both provides a common test-bed to be indexed and searched by the Information Retrieval System (IRS) X and Y and guarantees the possibility of replicating the experiments. The most common figures adopted for quantifying the performances are the *recall*, which is a measure of the ability of a system to present all relevant items, and the *precision*, which is a measure of the ability of a system to present only relevant items; the aggregation of such metrics (e.g., average) is used as well.

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

[†] http://www.promise-noe.eu/, contract n. 258191

assessment. For each task the set of document can be structured defining for example a title, keywords, images, and so on.

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 result set on which different metrics (around 100) are computed. 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: organizers are interested in evaluating the whole campaign, while participants are interested in evaluating their own algorithms, comparing them with algorithms of other participants. This data can be represented by the Topic Metric Experiment (TME) cube.

From this cube, the users of the system (organizers and participants) can aggregate or manipulate data in different ways, according to their needs. As an example, if the user is interested in a single experiment e in terms of $topics$ and $metrics$ this correspond to a projection TM(e) (topics and metrics data of experiment e) of the cube. A different scenario is the analysis of a single experiment e in terms of descriptive statistics and metrics: this data is represented by a matrix S x M where S is the set of statistics and M is the set of metrics and we refer to it as SM(e) (statistics and metrics of experiment e). This data is strictly related on the corresponding TM(e) since values are computed from the TM(e) table's columns.

These examples provide clues on the typical analytical transformations that the visual analytics system, described in the next section, provides to end users, in order to explore and compare metrics produced within complex evaluation campaigns.

3. Related Work

Information Retrieval evaluation is a well established field, which main contributions regarding methodologies and experiments are described in [Har11] [CKC*08] [VH*05]. To the best of the authors' knowledge, most proposals regarding application of visualization to Information Retrieval target the visualization and exploration of a query result [WC02], or the query formulation, neglecting the problem of comparing the measures of multiple experiments in order to support an experimental evaluation. Requirements about how Visual Analytics can support this task can be found in [Zha07] [San13] [FHMS11]. A prominent task for Information Retrieval data visualization is to represent large amount of data (e.g. the scores of all the experiments for multiple topics on multiple evaluation metrics) on two dimensional charts; we use in our solution techniques mutuated by the broader field of multivariate data visualization, as suggested in [War94]. Additionally, as in [CVW11] our solution proposes an environment that can produce well-known visualization paradigms as novel and visually more complex ones. Various Visual Analytics systems for analyzing document data have been proposed; in [WTP*95] an information visualization approach that involves spatializing text content for enhanced visual browsing and analysis is presented. [CLRP13] presents a visual analytics solution for supporting the topic modeling activity, using the topic concept as in our solution. However in that work the focus was to derive topics from the corpora, addressing the problem that in an evaluation campaign the topic modeling activity can take several months and the topics are continuously refined by domain experts analysis. The work in [RY12] is close to our proposal and presents a system aiming at to visually supporting information retrieval evaluation; however, differently to our approach, the authors focus only on citations and bibliometrics analysis. [MHH96] proposes a solution for visualizing the results coming from a search engine: in that approach only the ranking of documents is taken into consideration, while our solution proposes an environment for evaluating an Information Retrieval system with multiple metrics and statistics.

4. Prototype

This section presents the prototype of the VA Environment that presents the user with multiple visualizations, each of them working on different aspects of the data. Visualizations are synchronized using two main interaction mechanisms: *selection* (it is a way to focus the attention on a subset of data) and *highlight* (it allows for highlighting a part of the displayed data maintaining the context). In order to produce a set of view, three main steps are needed: (1) data extraction from a database (e.g., the historical CLEF (Conference and Labs of the Evaluation forum) data); (2) data manipulation, i.e., deriving new attributes, applying some aggregation operators or 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; and (3) mapping the data obtained from step two on one or multiple visualizations.

On the other hand, most of the IR evaluation activities are quite repetitive and follow several basic analysis patterns. To this aim, it is useful to provide some ad-hoc, highly automated pattern analysis, in which most of the previously described steps are automatically performed according to the analysis pattern at hand. In this scenario the set of available visualizations and their mappings with the data are wired in the interface, according to the most used analysis patterns: *Per topic analysis* and *Per Experiment* analysis.

Per topic analysis Per topic analysis allows for comparing a set of experiments on each topic with respect to a chosen metric. Therefore the first step for a user is to select 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). Per topic analysis implies a comparison on each topic, so, by default, we represent topics on x-axis in each available visualization. We provide four views for a per topic analysis: a table, a boxplot chart, a scatter plot, and a stacked bar chart.

The user can change the metric under analysis and restrict her focus on data subsets through select and highlight operations. As an example, Figure 1 shows three topics highlighted in all the four visualizations.

Per experiment analysis Per experiment analysis allows for analyzing an experiment as a whole and/or comparing the performances of a set of experiments with respect to a chosen descriptive statistics.

According to the typical PROMISE analysis tasks, we designed a set of ad-hoc visualizations. These visualizations must support synchronization and interaction that are specific for each visualization.

Figure 1: *Per topic analysis: an highlight operation.*

Moreover, for each visualization it is needed a mapping mechanism in order to support the user in the creation process.

The actual VA Environment supports 6 visualizations, listed from the simplest to the most advanced: bi-dimensional scatterplots, stacked bar-charts, box plots, table lens, enhanced frequency distribution, and the Precision-Recall-chart, all of them particularly suited for evaluation tasks in IR. Depending on the chosen type of analysis, the system will present the user with different subsets of these visualizations. Nonetheless, the user can customize the environment by simply removing a visualization and dragging a new one from a menu.

The characteristics of the 6 visualizations the following:

1. Table lens chart: it is particularly useful to cope with the high cardinality of the experiments. Moreover, the presence of a *classical* data table is a useful way to understand the data the user is coping with.
2. Box Plot chart: it displays an aggregated visualization of values that each topic/experiment (depending on type of analysis) reached with respect to the chosen metric.
3. Scatter Plot chart: it is used to compare two elements of the same family (experiments/topics/metrics based on the type of analysis chosen) in order to find correlations.
4. Stacked Bar chart: it allows to conduct a quantitative analysis highlighting topics with the highest scores and showing visually how the same experiment behaves on all the topic set (stable experiment results vs. variable experiment results).
5. Precision-Recall chart: this novel chart can be applied to both Per topic analysis with fixed topic and Per topic analysis with fixed experiment: the plotted curves represent, for the former case, the trends of the interpolated precision-recall curve for all the different experiments with respect to a fixed topic; for the latter, instead, they represent the trends of the interpolated precision-recall curve for a single experiment with respect to all the topics. Selecting the quartiles view, the visualization switches from the set of different represented curves, to an aggregated visualization of the data set, with the trends aggregated in 3 areas of quality (*bad*, in red color, *medium*, in blue color, and *good*, in green color) with respect to a quality metric based on the distribution of points in the area. This particular visualization will help the user in situation in which a large number of curves is drawn with the ability, by simply clicking on one of the areas, to display its associated curves.
6. Frequency Distribution chart: this novel chart presents the user with a frequency distribution of experiments with respect to a fixed topic and metric in the EM(t) analysis and aggregated by topics with respect to a fixed metric in the TE(m) analysis. The initial choice on the number and extension of the bins is set by default as the result of Sturges Law [Stu26]; however, by simply dragging the red circles that delimit each bin the user will be able to change the extension of the bin itself, even making it disappear and so reducing the number of bins of the distribution. The color coding assigned to the distribution (ranging from deep red to light green) allows for immediate judging on the results obtained by the investigated experiment/topic with respect to the chosen metric. A novel enhanced aggregated view is present, in which each of the topics is decomposed and superimposed with all the experiment results tested against it, in order to show which experiments contribute the most to the final evaluation of the topic and what instead diverge from that. An example is shown in Figure 3. Hovering the mouse on the distribution of experiments allow to display in the top-right corner the label of

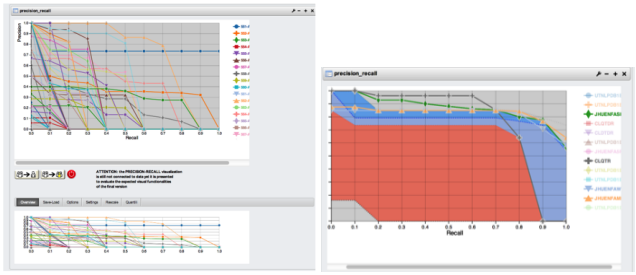

Figure 2: *Precision-Recall chart: the high number of curves represented (left) are not suited for immediate analysis. The system provides the quartiles representation, where the curves are aggregated computing the boxplots for each of the recall values; the lines that connects the same point for each boxplot (e.g., the median) form different areas that represent an aggregated evaluation of the experiments. Clicking on an area will draw the single experiments associated to that area again.*

the topic, the label of the experiment selected and the numerical scoring value, each of them in the related color coding.

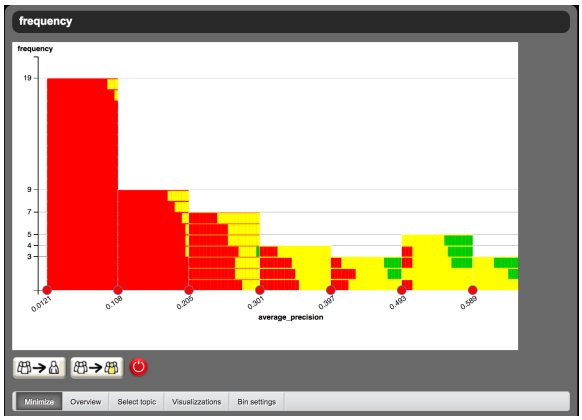

Figure 3: *Frequency distribution chart with additional ranked Experiments evaluation; each bin is decomposed in its associated topics, represented as rectangles. The bin are sorted with respect to the topic performances (the best on top, the worst on bottom). On each topic the scores for each experiment are reported. It is possible to see that the first bin on the left contains only bad performing topics (all the experiments are predominantly red), while on the right there are topics with good performing experiments (predominant green and yellow elements)*

All the visualizations supports selection or highlight of data subset, focus+context, layout editing and plotting of reference lines, and annotations, allowing organizers to assess and rank the performances of the experiments submitted by participants with different kind of analyses:

1. Per topic analysis with fixed experiment, comparing a chosen experiment on each topic and on each metric, in order to evaluate a single experiment scores on all the topics of the task.
2. Per topic analysis with fixed topic, comparing the whole set of experiments on each metric for a selected topic (comparative analysis of all the experiments on a selected highly meaningful topic). This task is good also for evaluating the topic itself based on the experiments scores.
3. Per topic analysis with fixed metric, comparing a set of experiments on each topic with respect to a chosen metric (exploratory analysis of the experiments that performs better on all the topics for a chosen metric).
4. Per experiment analysis with fixed metric, comparing experiments among them, computing all statistical indicators on a fixed metric, in order to obtain a precise and comprehensive analysis of the experiments under examination.
5. Per experiment analysis with fixed experiment, exploring the results that a single experiment obtains in term of all evaluation metrics and statistics (From exploration to detailed analysis of the results for a single experiment).
6. Per experiment analysis with fixed statistic, exploring the results that the experiments obtain in term of a single statistic computed for all the evaluation metrics (useful for understanding the distribution and score stability of submitted experiments for a single statistic function).

5. Conclusions & Future Work

In this paper we have explored the application of VA techniques to the problem of IR experimental evaluation proposing a solution that join together visual representations of evaluation metrics in a comprehensible form for the organizers and capability for the user, through interactions, to steer data aggregations in order to meta-evaluate the results, in the form of extended aggregated precision-recall chart (evaluating variability of precision for different recall levels for the selected experiments) and visual topic evaluation in the frequency distribution chart. We have seen how joining a powerful analytical framework with a proper visual environment can foster the automation into the evaluation and benchmarking of the complex data that is generated by experimental evaluations.

While a systematic user study has not been carried out, the system has been used during different CLEF campaigns, getting positive feedback from both researchers and stakeholders. As an example, we report some sentences we have in our minutes about our informal tests: "I would love to have this tool, both for research and for teaching purposes", " If I have had this tool during my PHD thesis writing I would have saved weeks of work"; moreover we have got several comments on basic usability issues like missing on screen instructions and on additional required features, like allowing for inspecting details of topics and documents and having information about the number of relevant documents for a topic and suggesting alternatives for performing more accurate topic analysis: "... it would be nice to give the possibility to cluster topics by good/bad to look at the chosen group of topics only".

As future work we foresee to extend the VA environment in order to support data coming from IR engine applications, looking at correlations between the search engine performances and adopted engine internal components (e.g., stemmer, stop list, etc.).

References

[CKC*08] CLARKE C. L., KOLLA M., CORMACK G. V., VECHTOMOVA O., ASHKAN A., BÜTTCHER S., MACKINNON I.: Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (2008), ACM, pp. 659–666. 2

[Cle] Conference and labs of the evaluation forum. URL: http://www.clef-initiative.eu/. 1

[Cle97] CLEVERDON C. W.: The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval* (1997), Spärck Jones K., Willett P., (Eds.), Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, pp. 47–60. 1

[CLRP13] CHOO J., LEE C., REDDY C., PARK H.: Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 1992–2001. 2

[CVW11] CLAESSEN J. H., VAN WIJK J. J.: Flexible linked axes for multivariate data visualization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2310–2316. 2

[FHMS11] FERRO N., HANBURY A., MÃIJLLER H., SANTUCCI G.: Harnessing the scientific data produced by the experimental evaluation of search engines and information access systems. In *Procedia Computer Science* (2011), vol. 4, pp. 740–749. 2

[Har11] HARMAN D.: Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services 3*, 2 (2011), 1–119. 2

[MHH96] MUKHERJEA S., HIRATA K., HARA Y.: Visualizing the results of multimedia web search engines. In *Information Visualization'96, Proceedings IEEE Symposium on* (1996), IEEE, pp. 64–65. 2

[RY12] RORISSA A., YUAN X.: Visualizing and mapping the intellectual structure of information retrieval. *Information processing & management 48*, 1 (2012), 120–135. 2

[San13] SANTUCCI G.: Visual analytics and information retrieval. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7757 LNCS* (2013), 116–131. 2

[Stu26] STURGES H. A.: The choice of a class interval. *Journal of the american statistical association 21*, 153 (1926), 65–66. 3

[VH*05] VOORHEES E. M., HARMAN D. K., ET AL.: *TREC: Experiment and evaluation in information retrieval*, vol. 1. MIT press Cambridge, 2005. 2

[War94] WARD M. O.: Xmdvtool: Integrating multiple methods for visualizing multivariate data. In *Proceedings of the Conference on Visualization '94* (Los Alamitos, CA, USA, 1994), VIS '94, IEEE Computer Society Press, pp. 326–333. URL: http://dl.acm.org/citation.cfm?id=951087.951146. 2

[WC02] WHITING M. A., CRAMER N.: Webtheme: Understanding web information through visual analytics. In *International Semantic Web Conference* (2002), Springer, pp. 460–468. 2

[WTP*95] WISE J. A., THOMAS J. J., PENNOCK K., LANTRIP D., POTTIER M., SCHUR A., CROW V.: Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the 1995 IEEE Symposium on Information Visualization* (Washington, DC, USA, 1995), INFOVIS '95, IEEE Computer Society, pp. 51–. URL: http://dl.acm.org/citation.cfm?id=857186.857579. 2

[Zha07] ZHANG J.: *Visualization for information retrieval*, vol. 23. Springer Science & Business Media, 2007. 2