Visual Analytics and IR Experimental Evaluation

Nicola Ferro and Giuseppe Santucci

Abstract We investigate the application of *Visual Analytics (VA)* techniques to the exploration and interpretation of *Information Retrieval (IR)* experimental data. We first briefly introduce the main concepts about VA and then we present some relevant examples of VA prototypes developed for better investigating IR evaluation data. Finally, we conclude with an discussion of the current trends and future challenges on this topic.

1 Visual Analytics

Around the year 2000, in order to support human beings in analyzing large and complex datasets, synergies between *Information Visualization (IV)* and *Data Mining (DM)* started to be considered. *Visual Data Mining (VDM)* was defined 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 (Wong and Thomas, 2004). Unlike 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 From that moment on, the term VDM was superseded by the term *Visual Analytics (VA)*. Daniel Keim, one of the major European experts in the field, provides the following

Nicola Ferro
Department of Information Engineering, University of Padua, Via G. Gradenigo, 6/B, 35131 Padova, Italy, e-mail: ferro@dei.unipd.it

Giuseppe Santucci
Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy, e-mail: santucci@dis.uniroma1.it

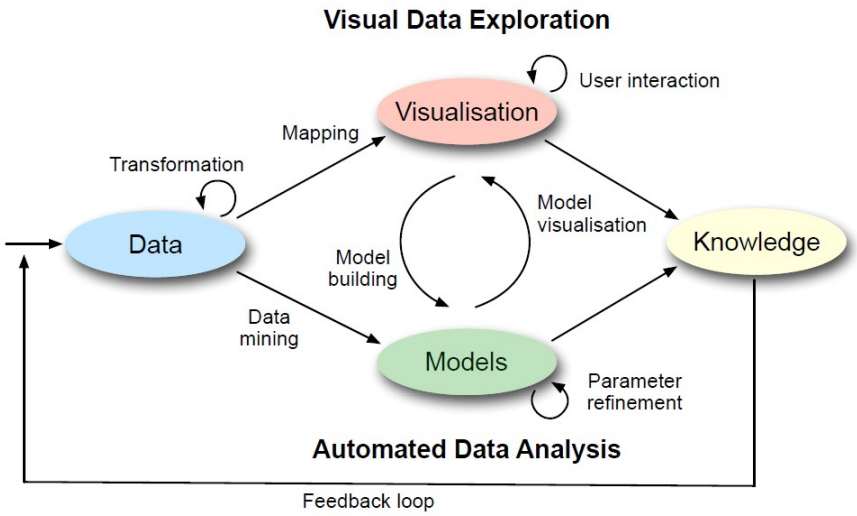

Fig. 1 The Visual Analytics process (Keim et al, 2010).

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, VA 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.

Figure 1 schematizes the VA 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 VA 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 VA 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 VA process, knowledge can be gained from visualization and automatic analysis, as well as the preceding interactions between visualizations,

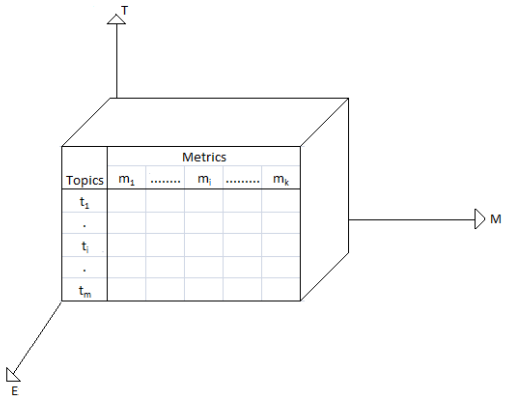

Fig. 2 The overall TME data cube with the TM(e) transformation highlighted.

models, and the human analysts. With respect to the field of visualization, VA integrates methodology from Information Visualization (Card et al, 1999; Chen, 2004; Spence, 2007; Ware, 2012), Visual Data Mining (Keim, 2001), geospatial analytics (Andrienko et al, 2007), and scientific analytics. In particular, 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, see, e.g., Keim et al (2006).

2 The IR Evaluation Data Cube

As shown in Figure 1, the initial step of any analysis is to get a clear understanding of the data involved in the process, in our case the data used within IR evaluation. Despite the strong differences that exist among the different domains targeted by IR applications, IR systems are typically evaluated according to the common Cranfield paradigm (Cleverdon, 1967), which allows us to compare the effectiveness of different IR systems on the same collection. The scientific data produced during evaluation are then arranged across several transformations that are suitable for different analysis patterns. In the European Union project PROMISE[1] these data plus their transformations have been formalized as follows.

The initial view on the data is represented by the *Topics–Metrics-Experiments (TME)* data cube, shown in Figure 2, reporting for each experiment (i.e., an IR system) its performance according to different evaluation measures across a set of topics.

[1] `http://www.promise-noe.eu/`

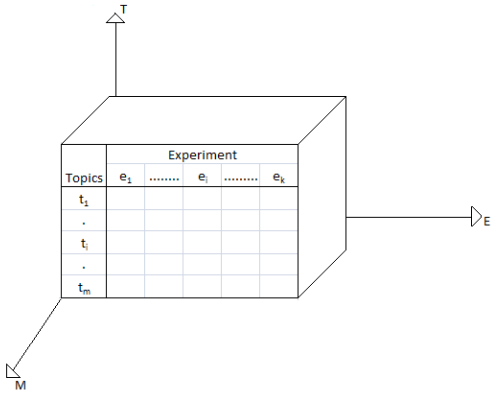

Fig. 3 Projection of the TME data cube on the Topics-Experiments axes with the TE(m) transformation.

Starting from this cube, it is possible to transform data in different ways, according to different analysis objectives. In particular four kinds of transformations have been identified.

The first kind of transformation makes it possible to analyze the performance of a single experiment e, i.e. an IR system, with respect to topics and it is the projection of the TME cube on the Topics–Metrics axes of experiment e. In particular, this table is a matrix $T \times M$, where T is the set of topics and M is the set of metrics. In the following, we refer to this kind of transformation as TM(e) tables (Topics $\times$ Metrics table of experiment e, shown in Figure 2).

A second kind of transformation, shown in Figure 3, is useful to analyze the behavior of a set of experiments, i.e. IR systems, over a set of topics with respect to a single metric m, which is the most common case in IR evaluation. In particular, this table is represented by a $T \times E$ matrix where T is the set of topics and E is the set of experiments. In the following, we refer to this kind of transformation TE(m) tables (Topics $\times$ Experiments table of metric m, shown in Figure 2). Comparisons are made along rows, to evaluate the behavior of a single topic, or among columns to compare two or more experiments.

The third kind of transformation describes a single experiment e in terms of descriptive statistics computed over a set of topics with respect to different metrics. In particular, this table is represented by an $S \times M$ matrix where S is the set of descriptive statistics and M is the set of metrics. In the following, we refer to this kind of transformation as the SM(e) table (Statistics $\times$ Metrics table of experiment e, shown in Figure 4). This table is strictly related to the corresponding TM(e) table since values are computed from the TM(e) table columns. Figure 4 shows an example of how a TM(e) table can be used to calculate values of the SM(e) table.

As shown in Figure 4, in an SM(e) table there is the same number of metrics as in the corresponding TM(e) table. If we extend this table with respect to experiments, we obtain a new cube, the *Statistics–Metrics-Experiments (SME)* data cube, shown

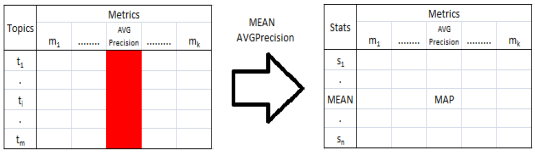

Fig. 4 Relationship between TM and SM tables.

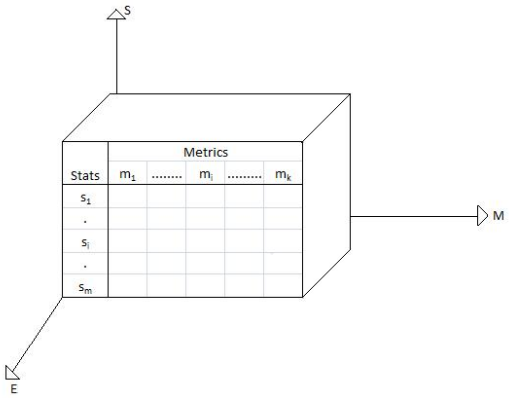

Fig. 5 The SME Data cube.

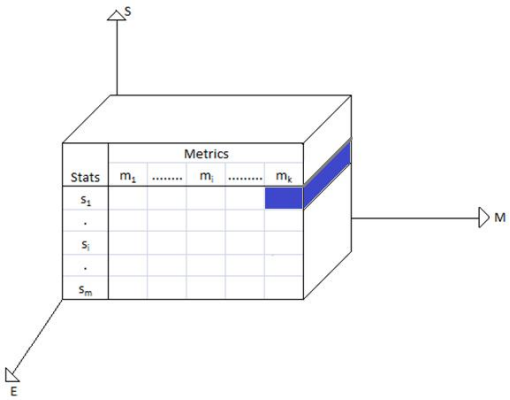

Fig. 6 The SME Data cube projected on the Statistics-Experiments axes.

in Figure 5. With respect to the SME cube, an SM(e) table is a projection on the Statistics-Metrics axes.

The last kind of table we consider, allows us to inspect a single metric m in terms of descriptive statistics and experiments, i.e., it makes it possible to compare different experiments against some descriptive statistics computed on a given metric. In particular, this table is represented by an $S \times E$ matrix where S is the set of statistics and E is the set of experiments. In the following, we refer to this transformation as the SE(m) table (Statistics $\times$ Experiment table computed on metric m, shown in Figure 6) and it is a projection of the SME cube on the Statistics-Experiments axes.

As discussed above, all these data and their transformations constitute the entry step depicted in the leftmost part of Figure 1.

3 Examples of VA Systems on the IR Evaluation Data Cube

In this section, we present some recent examples of systems which exploit VA techniques to improve IR experimental evaluation and to to analyse and interact with IR experimental data. They represent different types of instantiations of the "Models" and "Visualisation" steps depicted in Figure 1.

3.1 VAIRË

Angelini et al (2017) presented a VA environment, called *Visual Analytics for Information Retrieval Evaluation (VAIRË)*, which uses multiple visualizations working on different aspects of the data. Visualizations are synchronized using two main interaction mechanisms: *selection* (a way to focus the attention on a subset of data) and *highlight* (it allows to highlight a part of the displayed data maintaining the context). IR evaluation data cube transformations are then mapped to multiple coordinated visualizations.

Moreover, considering that user activities are quite repetitive and follow several basic analysis patterns, VAIRË provides some ad-hoc, highly automated patterns for analysis: *Per topic analysis* and *Per Experiment* analysis.

The system supports 6 visualizations, listed from the simplest to the most advanced: bi-dimensional scatter-plots, 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.

Per topic analysis it makes it possible to compare a set of experiments on each topic with respect to a chosen evaluation measure. Therefore the first step for a user is to select an evaluation measure m. Looking at the TME data cube described in the

Fig. 7 Per topic analysis: an highlight operation.

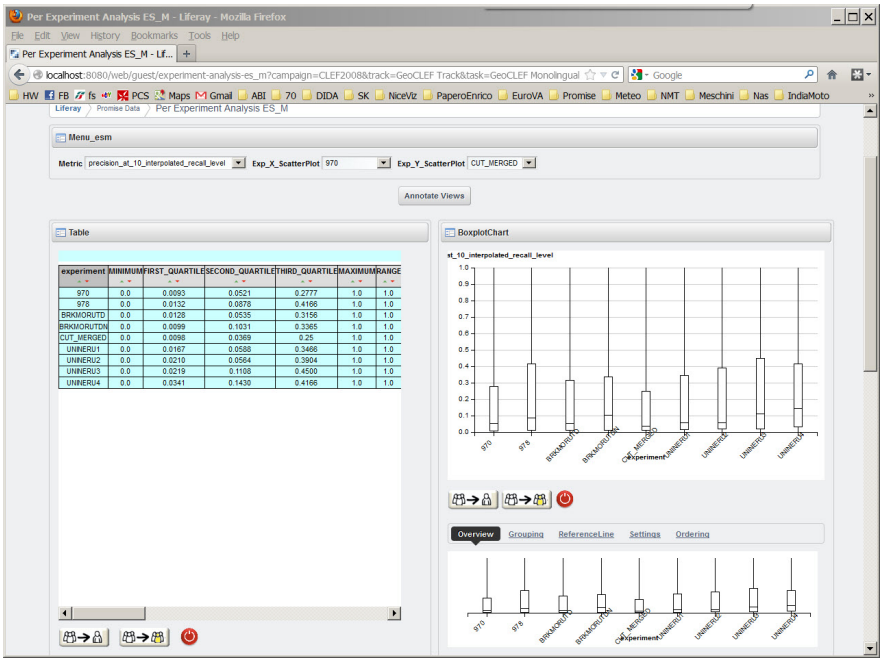

Fig. 8 Per experiment analysis: table and box plot.

previous section, we can note that choosing an evaluation measure is equivalent to fixing an axis and reducing the set of data to the TE(m) transformation. Per topic analysis implies a comparison on each topic, so, by default, we represent topics on the x-axis in each available visualization. We provide four views for a per topic analysis: table lens, a boxplot chart, a scatter plot, and a stacked bar chart.

The user can change the evaluation measure under analysis and restrict her/his focus on data subsets through select and highlight operations. As an example, Figure 7 shows three topics highlighted in all the four visualizations.

Per experiment analysis it makes it possible to analyze an experiment as a whole and/or compare the performance of a set of experiments with respect to a chosen descriptive statistics. As an example, on Figure 8, left side, the table represents an experiment in each row, showing the descriptive statistics of *Average Precision (AP)* (min, max, median, etc.). The box plot chart (McGill et al, 1978) in Figure 8, right side, shows the percentile values of the observed metric for each experiment represented through boxplots.

3.2 VIRTUE

Figure 9 shows the overall framework of *Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE)* to support the evaluation workflow (Angelini et al, 2014): *performance analysis* and *failure analysis* are the traditional phases carried out during experimental evaluation, where VIRTUE contributes to make them more effective and to reduce the needed effort via both tailored visualizations and high interaction with the experimental data.

Topic Level concerns the analysis of the documents retrieved in response to a given topic of a run while *Experiment Level* deals with overall statistics and effects concerning the whole set of topics of a run, i.e., all the different ranked lists of retrieved documents.

In both the topic and experiment level analyses, the user is presented with three curves, reporting the *Discounted Cumulated Gain (DCG)* (Järvelin and Kekäläinen, 2002) in three cases: a) the actual performance (experiment curve), b) the improvement that is possible to achieve reordering the actual result in the optimal way (optimal curve), and c) the best possible score, in which the results contain *all* the relevant documents in the optimal way (ideal curve). On the leftmost part, two bars represent the ranked list of retrieved documents where colors in the leftmost bar indicate how much a document has been misplaced with respect to its ideal position in the ranking and colors in the rightmost bar indicate the gain loss in terms of DCG due to this misplacement.

Therefore, VIRTUE:

- supports performance analysis on a topic-by-topic basis and with aggregate statistics over the whole set of topics;
- facilitates failure analysis to allow researchers and developers to more easily spot and understand failing documents and topics.

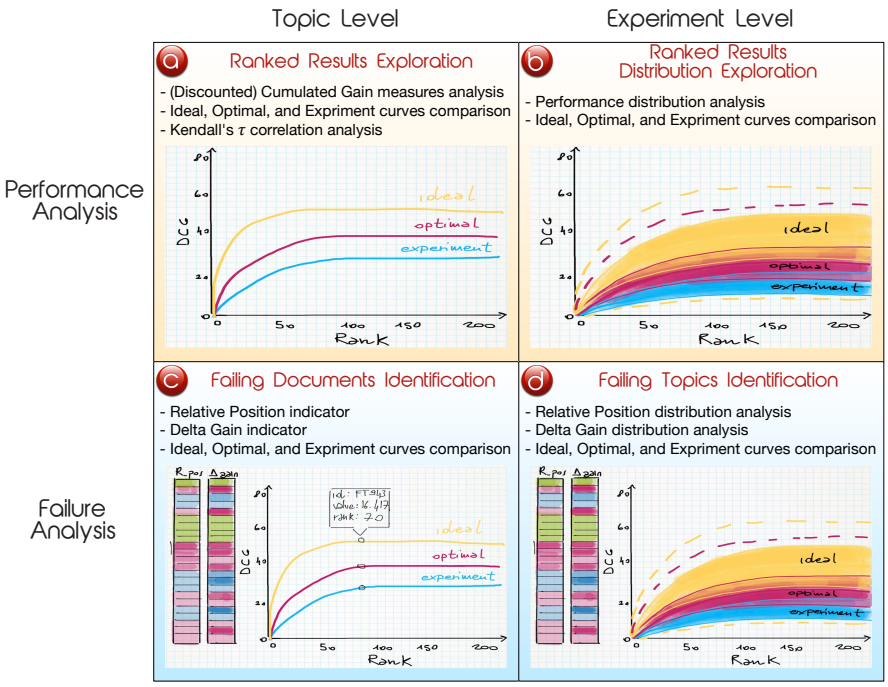

Fig. 9 VIRTUE overall framework.

The main target users of VIRTUE are domain experts, i.e., researchers and developers in the IR and related fields who need to understand and improve their systems. Moreover, VIRTUE can also be useful for educational purposes, e.g. in undergraduate or PhD courses where information retrieval is taught and where explaining how to interpret the performances of an IR system is an important part of the teaching. Finally, it may also find application in production contexts as a tool for monitoring and interpreting the performances of a running system so as to ensure that the desired service levels are met.

3.3 VATE^2

Visual Analytics Tool for Experimental Evaluation (VATE^2*)* (Angelini et al, 2012, 2016b,a) introduced a new phase in the evaluation workflow, called *what-if analysis*. It falls between the experimental evaluation and the design and implementation of the identified modifications. What-if analysis aims at estimating what the effects of a modification to the IR system under examination could be, before actually being implemented. In this way researchers and developers can get a feeling of whether a modification is worth being implemented and, if so, they can go ahead with its

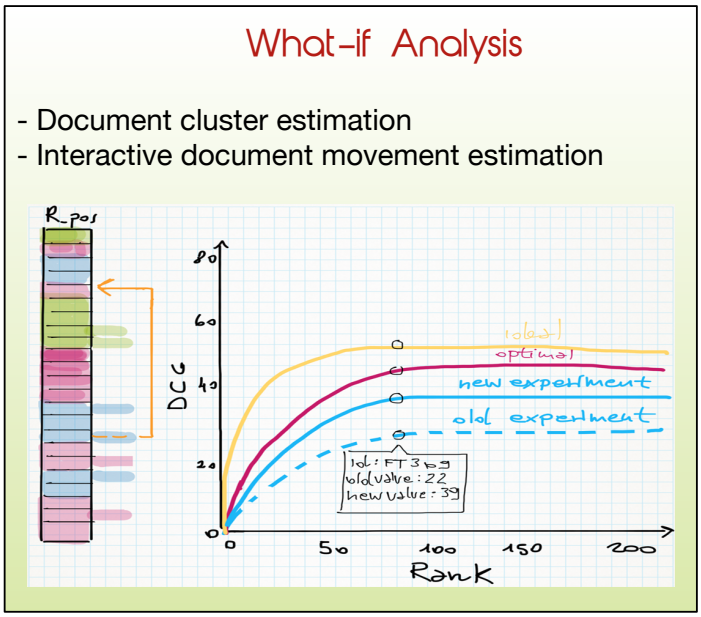

Fig. 10 VATE^2 overview.

implementation followed by a new evaluation and analysis cycle for understanding whether it has produced the expected outcomes.

What-if analysis exploits VA techniques to make researchers and developers: (i) interact with and explore the ranked result list produced by an IR system and the achieved performances; (ii) hypothesize possible causes of failure and their fixes; (iii) estimate the possible impact of such fixes through a powerful analytical model of the system behavior.

Figure 10 shows the mock-up used for designing the VATE^2 user interface whose objective is to provide a rough estimation of what could be the impact of fixing a possible failure on the performances in order to assess if it might be worth implementing it or not. What visualization of Figure 10 offers to the user is: (i) the possibility of dragging and dropping the target document in the desired position of the rank; (ii) the estimation of which other documents would be affected by the movement of the target document and how the overall ranking would be modified; (iii) the computation of the system performances according to the new ranking. Therefore, moving a single target document would actually cause the movement and repositioning of a whole set of documents that share features impacted by the same modification which will affect the target document selected by the user. These complex interactions between documents may generate modifications on the ranking that go well beyond what the user imagined when moving the single target document and which are definitely hard for her/him to guess. Thus, the contribution of the visual-

ization and analytical engine of Figure 10 is to automatically point out to the user all these complex interactions and how they affect the overall ranking.

Once the new ranked list has been produced by using a clustering and movement strategy, the performances of this new ranked list are computed and the corresponding new line is shown to the user so that he can assess whether the hypothesized modification may be beneficial or not. In the former case VATE[2] turns on a green light to indicate to the user that s/he should go on with the fix of the system, otherwise it turns on a red light meaning that the fix may be useless or worsen the system.

3.4 The RETRIEVAL Online Platform

Ioannakis et al (2018) developed RETRIEVAL[2], a Web-based integrated platform for performance evaluation of IR methods, which shares many commonalities with the VAIRË system discussed in Section 3.1.

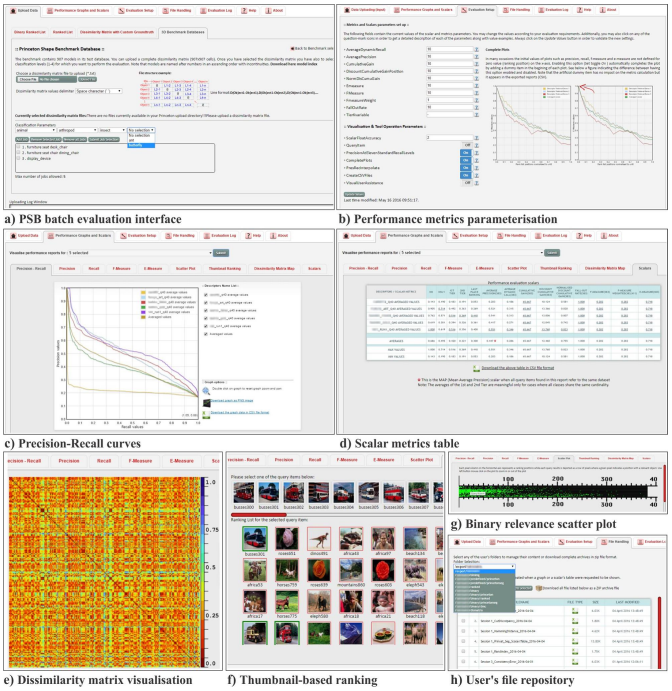

Fig. 11 Example of the RETRIEVAL user interface (Ioannakis et al, 2018). Downloaded from the RETRIEVAL Facebook page (`https://www.facebook.com/RetrievalEvaluationTool/`).

[2] `http://retrieval.ceti.gr/`

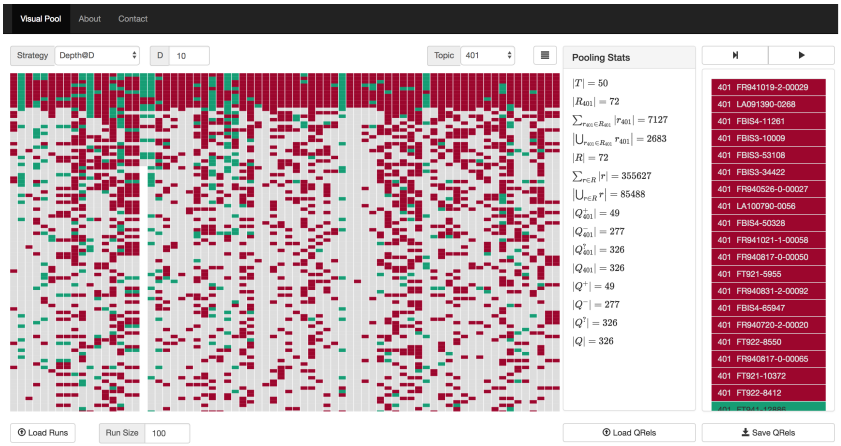

Fig. 12 Example of the Visual Pool user interface (Lipani et al, 2017). Courtesy of Aldo Lipani.

RETRIEVAL allows users to upload their datasets in various formats, converting them into internal data structures which resemble the IR evaluation data cube we described in Section 2. RETRIEVAL supports different evaluation measures, like AP (Buckley and Voorhees, 2005), *Normalized Discounted Cumulated Gain (nDCG)* (Järvelin and Kekäläinen, 2002), *Rank-Biased Precision (RBP)* (Moffat and Zobel, 2008), and many others.

Once the data cube has been created, RETRIEVAL provides several alternative visualisations, shown in Figure 11, such as a precision-recall graph (Figure 11.c), a scatter-plot where each pixel indicates a relevant/not relevant document (Figure 11.g), a dissimilarity matrix map where the user can identify a normalized dis-similarity distance between any two items using an interactive pointer that offers real-time zoom-in functionality (Figure 11.e), a tabular view of the data (Figure 11.d), and more.

3.5 The Visual Pool System

Lipani et al (2017) proposed Visual Pool[3] an IV system to explore alternative pooling strategies to build the ground truth of a test collection.

Figure 12 shows the user interface of Visual Pool. Users can load a set of runs, which are displayed in the left part of the window where each column is a system and each row is a retrieved document. The topmost left button allows users to select among different pooling strategies, whose effects are then interactively displayed. Moreover, users can load an already existing set of relevance judgments whose statistics are reported in the middle of the window. The color coding is as

[3] `http://visualpool.aldolipani.com/`

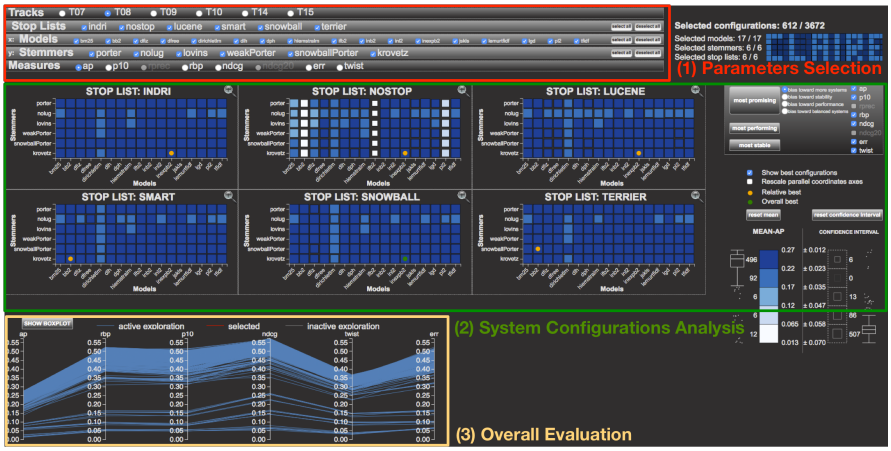

Fig. 13 Example of the CLAIRE user interface (Angelini et al, 2018).

follows with respect to the loaded relevance judgments: red is for not relevant documents; green is for relevant documents; gray is for not pooled documents; and, black is for pooled documents which are not contained in the currently loaded relevance judgments. Finally, the rightmost part of the window shows the details of the currently loaded systems and of the pooling method.

Overall, Visual Pool allows users to interactively experiment alternative pooling strategies over a set of runs, compare their effects with respect to an existing set of relevance assessments, and to assess their intrinsic bias.

3.6 CLAIRE

Angelini et al (2018) developed *Combinatorial visuaL Analytics system for Information Retrieval Evaluation (CLAIRE)*[4], a VA system for exploring and making sense of the performances of a large number of IR systems, in order to quickly and intuitively grasp which system configurations are preferred, what are the contributions of the different components and how these components interact together. In particular, CLAIRE allows users to explore, analize, interact with a *Grid of Points (GoP)* (Ferro and Harman, 2010), i.e. a very large set of IR systems originated from all the possible combinations of targeted components – stop lists, stemmers, and IR models in the case of Figure 13.

The goal of CLAIRE is to avoid the need for complex statistical analyses, such as those based on *ANalysis Of VAriance (ANOVA)* by Ferro and Silvello (2016), while fostering a more natural and intuitive way of making sense of such set of systems.

[4] `http://awareserver.dis.uniroma1.it:11768/claire/`

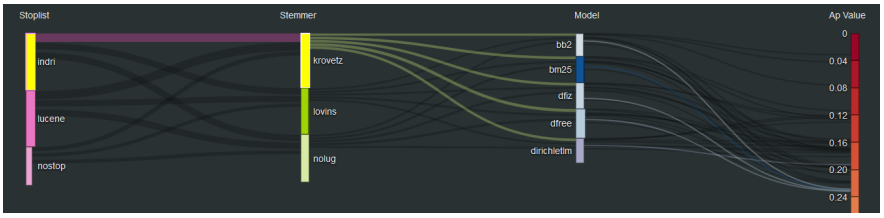

Fig. 14 Example of the Sankey GoP user interface (Rocco and Silvello, 2019). Courtesy of Gianmaria Silvello.

Figure 13 shows the user interface of CLAIRE:

1. The *Parameters Selection* area deals with the exploration coordinates, i.e., collections, stop lists, stemmers, IR models, and evaluation measures;
2. The *System Configurations Analysis* area enables the performance analysis of the system configurations using a specific evaluation measure. The multidimensional performance space is mapped to a bidimensional one by using a set of tiles where the color and the size of the tiles represent, respectively, the average performance and the confidence interval for that performance;
3. The *Overall Evaluation* area, where the system configurations performances are evaluated across the complete set of evaluation measures by using a parallel coordinates plot (Inselberg, 2009).

CLAIRE relies on the multiple coordinated views design, which allows users to propagate the results of the analysis process steps among all these three areas.

3.7 Sankey GoP

Rocco and Silvello (2019) further investigated how to intuitively explore and make sense of a GoP by leveraging a Sankey diagram (Sankey, 1898; Schmidt, 2008).

As shown in Figure 14, Rocco and Silvello replaced the tile-based visualization of CLAIRE with a Sankey diagram which makes it possible to represent the multidimensional performance space as a flow of performance from one component to another in the pipeline constituting an IR system. A single system is represented by a path, i.e. a series of links connecting one component with the next one. The user can select a set of components to highlight the paths of interest. The component columns present a number of rectangles equal to the components selected in the parameter selection area and the size of the rectangle gives a visual idea of the performances of the component it represents.

4 Discussion and Challenges

IV and VA techniques have been traditionally exploited mostly for the presentation and exploration of the results returned by an IR system (Zhang, 2008). The purpose of these components is to increase the ability to fulfill IR tasks where visualization is the natural platform for browsing and query searching. Some examples are: identification of the objects and their attributes to be displayed (Fowler et al, 1991); different ways of presenting the data (Morse et al, 2002); the definition of visual spaces and visual semantic frameworks (Zhang, 2001); using rankings for presenting the user with the most relevant visualizations (Seo and Shneiderman, 2005), for browsing the ranked results (Derthick et al, 2003), or for comparing large sets of rankings (Behrisch et al, 2013). The development of interactive means for IR is an active field which focuses on search user interfaces (Hearst, 2009, 2011), displaying of results (Crestani et al, 2004) and browsing capabilities (Koshman, 2005).

In the context of IR evaluation, IV strategies have been adopted for analyzing experimental runs, e.g. beadplots in (Banks et al, 1999). Each row in a beadplot corresponds to a system and each "bead", which can be gray or colored, corresponds to a document. The position of the bead across the row indicates the rank position in the result list returned by the system. The same color indicates the same document and therefore the plot makes it easy to identify a group of documents that tend to be ranked near to each other and to compare the performance of different systems. As a further example, *Query Performance Analyzer (QPA)* (Sormunen et al, 2002) provides the user with an intuitive idea of the distribution of relevant documents in the top ranked positions through a relevance bar, where rank positions of the relevant documents are highlighted, and it also allows for the comparison between the Recall-Precision graphs of a query and the most effective query formulations issued by users for the same topic.

Nevertheless, much less attention has been generally devoted to applying VA techniques to the analysis and exploration of the performance of IR systems in order to get a better understanding of their behaviour, when and where they fail, and how to improve them.

In Section 3 we have presented some recent examples which start to explore how VA can be applied to improve the IR evaluation workflow and to better interact, analyse, interpret, and understand the performance of IR systems.

We can consider the examples discussed in Section 3 as positive indicators of a rising interest for this topic in the research community, even if the full potential of VA for IR evaluation is still far from being fully unfledged.

Moreover, designing and developing this kind of systems is still extremely challenging because they require not only very specialist competence in both fields – IR and VA – but also a good mutual understanding of what are the main issues, approaches, and techniques in both fields. This sort of cross-disciplinary competencies and reciprocal interest in exploring each other's field is not easy to find. Moreover, joint collaborations must be established between research groups operating in the two fields and willing to invest in something which may be perceived as not mainstream in both fields.

Overall, we think that IR can greatly benefit from using and developing VA techniques to enhance and ease the exploration of the experimental results in order to build better systems. Moreover, the visual interpretation and understanding of IR system performance might even be considered as a community goal in the same way as the explicability and interpretability of IR algorithms is now perceived as a more and more compelling need. On the other hand, IR can be a very relevant domain for VA researchers, especially considering its pervasiveness in daily life. Indeed, IR evaluation poses challenges in terms of the complexity and the huge amount of the data to be analysed as well as the sophistication of the statistical methods used to make sense of the data. Finally, the increasing use of traces for capturing and predicting user behavior is adding a new complexity layer to the whole process, making the call for VA in IR louder.

References

Andrienko G, Andrienko N, Jankowski P, Keim DA, Kraak MJ, MacEachren A, Wrobel S (2007) Geovisual Analytics for Spatial Decision Support: Setting the Research Agenda. International Journal of Geographical Information Science 21(8):839–858

Angelini M, Ferro N, Santucci G, Silvello G (2012) Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In: Kamps J, Kraaij W, Fuhr N (eds) Proc. 4th Symposium on Information Interaction in Context (IIiX 2012), ACM Press, New York, USA, pp 195–203

Angelini M, Ferro N, Santucci G, Silvello G (2014) VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. Journal of Visual Languages & Computing (JVLC) 25(4):394–413

Angelini M, Ferro N, Santucci G, Silvello G (2016a) A Visual Analytics Approach for What-If Analysis of Information Retrieval Systems. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), ACM Press, New York, USA, pp 1081–1084

Angelini M, Ferro N, Santucci G, Silvello G (2016b) What-If Analysis: A Visual Analytics Approach to Information Retrieval Evaluation. In: Di Nunzio GM, Nardini FM, Orlando S (eds) Proc. 7th Italian Information Retrieval Workshop (IIR 2016), CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1653/

Angelini M, Ferro N, Santucci G, Silvello G (2017) Visual Analytics for Information Retrieval Evaluation Campaigns. In: Sedlmair M, Tominski C (eds) Proc. 8th International Workshop on Visual Analytics (EuroVA 2017), Eurographics Association, Goslar, Germany, pp 25–29

Angelini M, Fazzini V, Ferro N, Santucci G, Silvello G (2018) CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. Information Processing & Management 54(6):1077–1100

Banks D, Over P, Zhang NF (1999) Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval 1(1-2):7–34

Behrisch M, Davey J, Simon S, Schreck T, Keim D, Kohlhammer J (2013) Visual Comparison of Orderings and Rankings. In: Pohl M, Schumann H (eds) Proc. 4th International Workshop on Visual Analytics (EuroVA 2013), Eurographics Association, Goslar, Germany

Buckley C, Voorhees EM (2005) Retrieval System Evaluation. In: Harman DK, Voorhees EM (eds) TREC. Experiment and Evaluation in Information Retrieval, MIT Press, Cambridge (MA), USA, pp 53–78

Card SK, Mackinlay JD, Shneiderman B (1999) Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Publishers, San Francisco (CA), USA

Chen C (2004) Information Visualization - Beyond the Horizon. Springer-Verlag, London, UK

Cleverdon CW (1967) The Cranfield Tests on Index Languages Devices. Aslib Proceedings 19(6):173–194

Crestani F, Vegas J, de la Fuente P (2004) A Graphical User Interface for the Retrieval of Hierarchically Structured Documents. Information Processing & Management 40(2):269–289

Derthick M, Christel MG, Hauptmann AG, Wactlar HD (2003) Constant Density Displays Using Diversity Sampling. In: Munzner T, North S (eds) Proc. 9th IEEE Symposium on Information visualization (INFOVIS 2003), IEEE Computer Society, Los Alamitos, CA, USA, pp 137–144

Ferro N, Harman D (2010) CLEF 2009: Grid@CLEF Pilot Track Overview. In: Peters C, Di Nunzio GM, Kurimo M, Mandl T, Mostefa D, Peñas A, Roda G (eds) Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany, pp 552–565

Ferro N, Silvello G (2016) A General Linear Mixed Models Approach to Study System Component Effects. In: Perego R, Sebastiani F, Aslam J, Ruthven I, Zobel J (eds) Proc. 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), ACM Press, New York, USA, pp 25–34

Fowler RH, Lawrence-Fowler WA, Wilson BA (1991) Integrating Query, Thesaurus, and Documents Through a Common Visual Representation. In: Fox EA (ed) Proc. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991), ACM Press, New York, USA, pp 142–151

Hearst MA (2009) Search User Interfaces, 1st edn. Cambridge University Press, New York, NY, USA

Hearst MA (2011) "Natural" Search User Interfaces. Communications of the ACM (CACM) 54(11):60–67

Inselberg A (2009) Parallel Coordinates. Visual Multidimensional Geometry and Its Applications. Springer-Verlag, New York, USA

Ioannakis G, Koutsoudis A, Pratikakis I, Chamzas C (2018) Retrieval–an online performance evaluation tool for information retrieval methods. IEEE Transactions on Multimedia 20(1):119–127

Järvelin K, Kekäläinen J (2002) Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4):422–446

Keim DA (2001) Visual Exploration of Large Data Sets. Communications of the ACM (CACM) 44(8):38–44

Keim DA, Mansmann F, Schneidewind J, Ziegler H (2006) Challenges in Visual Data Analysis. In: Banissi E (ed) Proc. of the 10th International Conference on Information Visualization (IV 2006), IEEE Computer Society, Los Alamitos, CA, USA, pp 9–16

Keim DA, Kohlhammer J, Ellis G, Mansmann F (eds) (2010) Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association, Goslar, Germany

Koshman S (2005) Testing User Interaction with a Prototype Visualization-Based Information Retrieval System. Journal of the American Society for Information Science and Technology (JASIST) 56(8):824–833

Lipani A, Lupu M, Hanbury A (2017) Visual Pool: A Tool to Visualize and Interact with the Pooling Method. In: Kando N, Sakai T, Joho H, Li H, de Vries AP, White RW (eds) Proc. 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017), ACM Press, New York, USA, pp 1321–1324

McGill R, Tukey JW, Larsen WA (1978) Variations of Box Plots. The American Statistician 32(1):12–16

Moffat A, Zobel J (2008) Rank-biased Precision for Measurement of Retrieval Effectiveness. ACM Transactions on Information Systems (TOIS) 27(1):2:1–2:27

Morse EL, Lewis M, Olsen KA (2002) Testing Visual Information Retrieval Methodologies Case Study: Comparative Analysis of Textual, Icon, Graphical, and Spring Displays. Journal of the American Society for Information Science and Technology (JASIST) 53(1):28–40

Rocco G, Silvello G (2019) An InfoVis Tool for Interactive Component-Based Evaluation. arXivorg, Information Retrieval (csIR) arXiv:1901.11372

Sankey HR (1898) Introductory note on the thermal efficiency of steam-engines. Report of the committee appointed on the 31st March, 1896, to consider and report to the council upon the subject of the definition of a standard or standards of thermal efficiency for steam-engines: With an introductory note, Minutes of Proceedings of the Institution of Civil Engineers. 134:278–283 including Plate 5

Schmidt M (2008) The Sankey Diagram in Energy and Material Flow Management. Journal of Industrial Ecology 12(1):82–94

Seo J, Shneiderman B (2005) A Rank-by-Feature Framework for Interactive Exploration of Multidimensional Data. Information Visualization 4(2):96–113

Sormunen E, Hokkanen S, Kangaslampi P, Pyy P, Sepponen B (2002) Query Performance Analyser – a Web-based tool for IR research and instruction. In: Järvelin K, Beaulieu M, Baeza-Yates R, Hyon Myaeng S (eds) Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), ACM Press, New York, USA, p 450

Spence R (2007) Information Visualization: Design for Interaction, 2nd edn. Pearson Education Limited

Ware C (2012) Information Visualization - Perception for Design, 3rd edn. Morgan Kaufmann Publishers, San Francisco (CA), USA

Wong PC, Thomas JJ (2004) Visual analytics - guest editors' introduction. IEEE Computer Graphics and Applications (CG&A) 24(5):20–21

Zhang J (2001) TOFIR: A Tool of Facilitating Information Retrieval - Introduce a Visual Retrieval Model. Information Processing & Management 37(4):639–657

Zhang J (2008) Visualization for Information Retrieval. Springer-Verlag, Heidelberg, Germany