# Visual Analytics for Information Retrieval Evaluation (VAIRË 2015)

Marco Angelini[1], Nicola Ferro[2], Giuseppe Santucci[1], and Gianmaria Silvello[2]

[1] "La Sapienza" University of Rome, Italy
{angelini,santucci}@dis.uniroma1.it
[2] University of Padua, Italy
{ferro,silvello}@dei.unipd.it

**Abstract.** Measuring is a key to scientific progress. This is particularly true for research concerning complex systems, whether natural or human-built. The tutorial introduced basic and intermediate concepts about lab-based evaluation of information retrieval systems, its pitfalls, and shortcomings and it complemented them with a recent and innovative angle to evaluation: the application of methodologies and tools coming from the *Visual Analytics (VA)* domain for better interacting, understanding, and exploring the experimental results and *Information Retrieval (IR)* system behaviour.

## 1 Scope and Learning Objectives

The tutorial addressed the topic of experimental evaluation, which has been a core topic in *Information Retrieval (IR)* since its inception. However, the tutorial faced this topic mixing basic and indispensable concepts on IR evaluation with a new angle that comes from applying information visualization and *Visual Analytics (VA)* methods and techniques to improve the interpretation and interaction with the experimental data, with the final goal of better understanding the system behaviour.

The overall aim of the tutorial was to improve the skills and practices of junior researchers (but also senior ones were welcome) in carrying out a thorough evaluation of IR system, providing them with both solid knowledge of IR evaluation and its pitfalls and with an innovative angle, coming from the application of visual analytics techniques to the understanding of and interaction with experimental data.

The specific learning objectives were: (i) to learn basic and intermediate competencies on IR evaluation and its pitfalls; (ii) to learn basic competencies on VA; (iii) to learn how VA techniques can be fruitfully applied to IR evaluation; (iv) to learn to implement basic VA components for IR evaluation.

## 2 Contents

The lecture in the first half-day was constituted by three modules. The objective of this first half-day was to provide attendees with needed methodological notions to achieve the learning objectives described above.

A. Hanbury et al. (Eds.): ECIR 2015, LNCS 9022, pp. 809–812, 2015.

The first module started introducing the main motivations and goals for experimental evaluation [1] and explained the basic concepts of the experimental evaluation according to the Cranfield paradigm, namely experimental collections, ground-truth creation and pool, evaluation campaigns and their typical life-cycle [2].

Evaluation measures have been introduced and discussed, also from in relation to what is usually done in the (representational) theory of measurement [3], their main constituents (user models, ...) were presented, and some caveats about scale types and the allowed operations with them have been raised. Some examples of well-known measures, such as *Average Precision (AP)*, *Normalized Discounted Cumulated Gain (nDCG)*, *Rank-Biased Precision (RBP)* and so on, have been discussed [4].

Failure analysis was then introduced and explained as a fundamental but extremely demanding activity by providing examples from well-known exercises, such as the *Reliable Information Access (RIA)* workshop [5].

The second module introduced the goals and the motivations underlying the emerging VA discipline, detailing the concepts and the basic techniques that are currently adopted in such a research field. In particular, the canonical steps of internal data representation and data presentation have been described, together with an overview of the most used visualization techniques [6,7].

Issues associated with the correct evaluation of VA systems were introduced and discussed. In particular, the tutorial analysed the user centered design methodology and the evaluation through questionnaires [8]. Examples have been given by the application of such techniques to the VA prototype developed within the IR evaluation PROMISE[1] Infrastructure [9,10].

The third module dealt with advanced applications of VA to experimental evaluation, where theoretical notions were complemented with examples from actually implemented prototypes. In particular, we: (i) described how to provide better support for carrying out an effective failure analysis [11,12]; (ii) introduced a new phase in the evaluation, we called it "what-if analysis", aimed at getting an estimate of the possible impact of modifications to an IR system on its performances [13,14].

The hands-on session in the second half-day were constituted by three modules. The objective of this second half-day was to provide attendees with a concrete feeling about how to develop and implement the methodological notions introduced in the first half-day.

The first module let attendees play with a running prototype of a VA system for IR evaluation, the VATE system, in order to let them experience what you can aim at for such kind of systems, how they can work, and how you can benefit from them for better understanding the experimental results. They then went trough a questionnaire for evaluating the used system. This had a two-fold goal: first, to stimulate critical thinking about what the attendees have just experienced; second, to provide them with a concrete example of what evaluating

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

VA systems means and a starting point whether they will have to evaluate their own VA systems.

The second module explained how to evaluate the output of an IR system using standard experimental collections. In particular, it provided a step-by-step example using the open source freely available MATTERS[2] library, a MATLAB toolkit for computing standard evaluation measures and carrying out analyses (previous knowledge about MATLAB is not required), and ad-hoc *Conference and Labs of the Evaluation Forum (CLEF)* collections [15,16].

The third module introduced the basics of the Web based visualization library D3[3], providing a step-by-step comprehensive example for representing and presenting a dataset containing IR evaluation data, focusing on user interaction in order to quickly get insights from coordinated visualizations.

## 3 Schedule

The schedule of the lecture part (half-day) was organized as follows:

- *Information Retrieval and its Evaluation*: basics on laboratory-based IR evaluation [1,2]; basics on IR evaluation measures [4,3]; failure analysis [5].
- *Visual Analytics*: basics on Visual Analytics [6,7]; basics on evaluation of Visual Analytics systems [8]; application of Visual Analytic to IR evaluation and running examples with the PROMISE Infrastructure prototype [9,10].
- *Advanced Applications of Visual Analytics for IR Evaluation*: Visual Analytics for Failure Analysis and running examples with the VIRTUE prototype [11,12]; Visual Analytics for What-if Analysis and running examples with the VATE prototype [13,14].

The schedule of the hands-on part (half-day) was organized as follows:

- *Experiencing with VA for IR Evaluation*: use and trial of the VATE prototype; evaluation questionnaire on the VATE prototype.
- *Example of Building Blocks for VA applied to IR evaluation (part 1 of 2)*: use of the MATTERS evaluation library to assess the performances of an IR system and produce experimental data to analyse;
- *Example of Building Blocks for VA applied to IR evaluation (part 2 of 2)*: use of the D3 library to develop interactive plots and process the experimental data produced in part 1.

## References

1. Harman, D.K.: Information Retrieval Evaluation. Morgan & Claypool Publishers, USA (2011)
2. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) 4, 247–375 (2010)

[2] `http://matters.dei.unipd.it/`
[3] `http://d3js.org/`

3. Fenton, N.E., Bieman, J.: Software Metrics: A Rigorous & Practical Approach, 3rd edn. Chapman and Hall/CRC, USA (2014)
4. Büttcher, S., Clarke, C.L.A., Cormack, G.V.: Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press, Cambridge (2010)
5. Harman, D., Buckley, C.: Overview of the Reliable Information Access Workshop. Information Retrieval 12, 615–641 (2009)
6. Thomas, J.J., Cook, K.A. (eds.): Illuminating the Path: The Research and Development Agenda for Visual Analytics, National Visualization and Analytics Center, USA (2005)
7. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F. (eds.): Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association, Goslar (2010)
8. Kang, Y.: a., Görg, C., Stasko, J.: Evaluating Visual Analytics Systems for Investigative Analysis: Deriving Design Principles from a Case Study. In: May, R., Kohlhammer, J., Stasko, J., van Wijk, J. (eds.) Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST 2009), pp. 139–146. IEEE Computer Society, Los Alamitos (2009)
9. Angelini, M., Ferro, N., Santucci, G., Garcia Seco de Herrera, A.: Deliverable D5.4 – Revised Collaborative User Interface Prototype with Annotation Functionalities. PROMISE Network of Excellence, EU 7FP, Contract N. 258191. `http://www.promise-noe.eu/documents/10156/a61967be-2b23-461b-b72a-302766c942e3` (2013)
10. Ferro, N., Berendsen, R., Hanbury, A., Lupu, M., Petras, V., de Rijke, M., Silvello, G.: PROMISE Retreat Report – Prospects and Opportunities for Information Access Evaluation. SIGIR Forum 46, 60–84 (2012)
11. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: VIRTUE: A visual tool for information retrieval performance evaluation and failure analysis. Journal of Visual Languages & Computing (JVLC) 25, 394–413 (2014)
12. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: A Visual Interactive Environment for Making Sense of Experimental Data. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) ECIR 2014. LNCS, vol. 8416, pp. 767–770. Springer, Heidelberg (2014)
13. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: 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), pp. 195–203. ACM Press, New York (2012)
14. Angelini, M., Ferro, N., Granato, G.L., Santucci, G., Silvello, G.: Information Retrieval Failure Analysis: Visual analytics as a Support for Interactive "What-If" Investigation. In: Santucci, G., Ward, M. (eds.) Proc. IEEE Conference on Visual Analytics Science and Technology (VAST 2012), pp. 204–206. IEEE Computer Society, Los Alamitos (2012)
15. Ferro, N.: CLEF 15th Birthday: Past, Present, and Future. SIGIR Forum 48, 31–55 (2014)
16. Ferro, N., Silvello, G.: CLEF 15th Birthday: What Can We Learn From Ad Hoc Retrieval? In Kanoulas, E. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) CLEF 2014. LNCS, vol. 8685, pp. 31–43. Springer, Heidelberg (2014)