A Visual Interactive Environment for Making Sense of Experimental Data

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

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

Abstract. We present the *Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE)* which is an interactive and visual system supporting two relevant phases of the experimental evaluation process: performance analysis and failure analysis.

1 Introduction

Developing and testing an *Information Retrieval (IR)* system is a challenging task, in particular when it is necessary to understand the behaviour of the system under different conditions of use in order to tune or improve it to achieve the level of effectiveness needed to meet user expectations. The complex interactions among the components of a system are often hard to trace down and to explain in the light of the obtained results. To this purpose two main activities are carried out in the context of experimental evaluation: performance analysis and failure analysis. The goal of performance analysis is to determine positive and negative aspects of the IR system under evaluation; whereas, the goal of failure analysis [6,8] is to conduct a deeper investigation for understanding the behaviour of a system determining what went well or bad.

Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE) aims at reducing the effort needed to carry out both the performance and failure analyses, both at topic and experiment level, since it allows user to visually interact with and mine the experimental results. An extensive description of the background, analysis of the functionalities and evaluation of VIRTUE can be found in [4]. The running prototype of VIRTUE is available at the following URL: `http://151.100.59.83:11768/Virtue/` while a video showing how the system works is available at the following URL: `http://ims.dei.unipd.it/websites/ecir2014/demo_ecir2014.mp4`.

This paper is organized as follows: Section 2 presents the features of VIRTUE to support performance analysis and Section 3 discusses how VIRTUE enhances failure analysis; finally, Section 4 draws some final remarks.

2 Performance Analysis

Performance analysis is one of the most consolidated activities in IR evaluation and VIRTUE allows for interactive visualization and exploration of the

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 767–770, 2014.

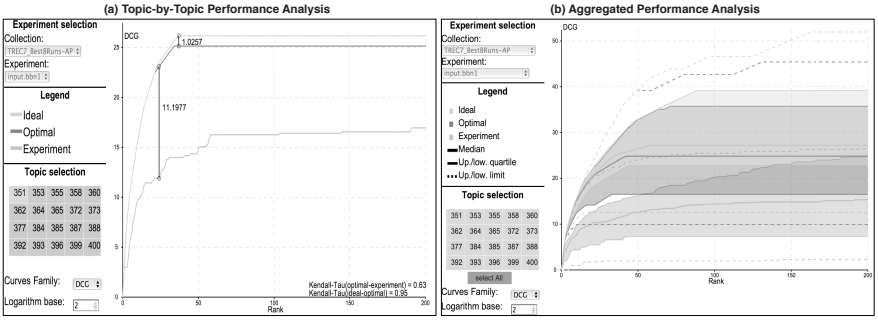

Fig. 1. The performance analysis capabilities provided by VIRTUE

experimental results, according to different metrics and parameters. It provides visual means to grasp whether the system would already have the potential to achieve the best performances or whether a new ranking strategy would be preferred. This analysis can be conducted on a topic-by-topic basis and with aggregate statistics over the whole set of topics.

In order to quantify the performances of an IR system, we adopt the (discounted) cumulated gain family of measures [7] which have proved to be especially well-suited for analyzing ranked results lists because they allow for graded relevance judgments and embed a model of the user behavior while s/he scrolls down the results list which also gives an account of her/his overall satisfaction.

We compare the result list produced by an experiment with respect to an *ideal* ranking created starting from the relevant documents in the ground-truth, which represents the best possible results that an experiment can return – this ideal ranking is what is usually used to normalize the *Discounted Cumulated Gain (DCG)* measures. In addition to what is typically done, we compare the results list with respect to an *optimal* one created with the same documents retrieved by the IR system but with a optimal ranking, i.e. a permutation of the results retrieved by the experiment aimed at maximizing its performances by sorting the retrieved documents in decreasing order of relevance. Therefore, the *ideal ranking* compares the given experiment with respect to the best results possible, i.e. considering also relevant documents not retrieved by the system, while the *optimal ranking* compares an experiment with respect to what could have been done better with the same retrieved documents.

The proposed visualization, shown in Figure 1(a), allows for interaction with these three curves, e.g. by dynamically choosing different measures in the DCG family, adjusting the discounting function, and comparing curves and their values rank by rank. Overall, this method makes it easy to determine the distance of an IR system from both its own optimal performances and the best performances possible and to get an indication about whether the system is going in the right direction or whether a completely different approach is preferable. In order to support this visual intuition, we also provide a Kendall's τ correlation analysis [9] between the three above mentioned curves [5].

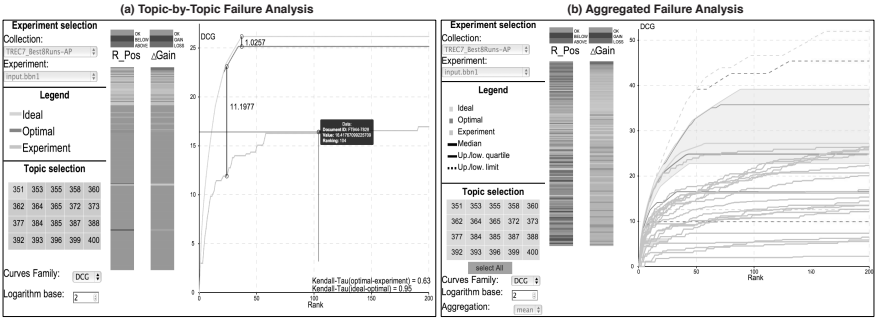

Fig. 2. The failure analysis capabilities provided by VIRTUE

VIRTUE provides also an aggregate representation based on the box-plot statistical tool showing the variability of the three DCG curves calculated either on all the topics considered by an experiment or on those selected by the user. This feature allows a user to interactively choose the topics to whose performances have to be aggregated in order to support the exploration of alternative retrieval scenarios [2]. We can see this feature in Figure 1(b).

3 Failure Analysis

For conducting failure analysis, VIRTUE exploits two indicators, called *Relative Position (RP)* and *Delta Gain (ΔG)* [3], which allow us to visually and numerically figure out the weak and strong parts of a ranking in order to quickly detect failing documents or topics and make hypotheses about how to improve them. RP quantifies the effect of misplacing relevant documents with respect to the ideal case easing the interpretation of the DCG curve, i.e. it accounts for how far a document is from its ideal position. ΔG quantifies the effect of misplacing relevant documents with respect to the ideal case in terms of the impact of the misplacement on the gain at each rank position [1].

These two indicators are paired with a visual counterpart that makes it even easier to quickly spot and inspect critical areas of the ranking. Two bars are added on the left of the visualization, as shown in Figure 2(a): one for the RP indicator and the other for the ΔG indicator. These two bars represent the ranked list of results with a box for each rank position and, by using appropriate color coding to distinguish between zero, positive and negative values and shading to represent the intensity, i.e. the absolute value of each indicator, each box represents the values of either RP or ΔG. For example, in this way, by looking at the bars and their colors a user can immediately identify non-relevant documents which have been ranked in the positions of relevant ones. Then, the visualization allows them to inspect those documents and compare them with the topic at hand in order to make a hypothesis about the causes of a failure.

The techniques described above support and ease failure analysis at the topic level and allow users to identify and guess possible causes for wrongly ranked

documents. The visualization of Figure 2(b) merges the approaches of the visualizations presented in Figure 1(b) and Figure 2(a): it allows users to assess the distribution of the performances of the ideal, optimal, and experiment curves over a set of selected topics or the whole run and it adds the bars reporting the RP and ΔG indicators to ease the interpretation of the performance distribution.

4 Final Remarks

The goal of this work is to provide the researcher and developer with more intuitive and more effective tools to analyse and understand systems behaviour, performances, and failures. VIRTUE eases the interpretation and the interaction with DCG curves, allows for detecting critical areas in ranked lists, and provides an integrated way for combining topic-by-topic and aggregated analyses.

Acknowledgements. The work reported in this paper has been supported by the PROMISE network of excellence (contract n. 258191) project as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

References

1. Angelini, M., Ferro, N., Järvelin, K., Keskustalo, H., Pirkola, A., Santucci, G., Silvello, G.: Cumulated Relative Position: A Metric for Ranking Evaluation. In: Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (eds.) CLEF 2012. LNCS, vol. 7488, pp. 112–123. Springer, Heidelberg (2012)
2. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In: Proc. 4th Symposium on Information Interaction in Context (IIiX 2012), pp. 195–203. ACM Press, New York (2012)
3. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: Improving Ranking Evaluation Employing Visual Analytics. In: Forner, P., Müller, H., Paredes, R., Rosso, P., Stein, B. (eds.) CLEF 2013. LNCS, vol. 8138, pp. 29–40. Springer, Heidelberg (2013)
4. 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 (in print, 2014)
5. Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., Tino, G.: To Re-rank or to Re-query: Can Visual Analytics Solve This Dilemma? In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M. (eds.) CLEF 2011. LNCS, vol. 6941, pp. 119–130. Springer, Heidelberg (2011)
6. Harman, D.K.: Some thoughts on failure analysis for noisy data. In: Proc. 2nd Workshop on Analytics for Noisy unstructured text Data (AND 2008), p. 1. ACM Press, New York (2008)
7. Järvelin, K., Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems (TOIS) 20(4), 422–446 (2002)
8. Savoy, J.: Why do Successful Search Systems Fail for Some Topics. In: Proc. 2007 ACM Symposium on Applied Computing (SAC 2007), pp. 872–877. ACM Press, New York (2007)
9. Voorhees, E.: Evaluation by Highly Relevant Documents. In: Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001), pp. 74–82. ACM Press, New York (2001)