Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation (Extended Abstract)[*]

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. Evaluation has a crucial role in *Information Retrieval (IR)* and developing tools to support researchers and analysts when analyzing results and investigating strategies to improve IR system performance can help make the analysis easier and more effective. To this purpose we present a Visual Analytics-based approach to support the analyst in performing failure and what-if analysis.

1 Introduction

Designing, developing, and testing an IR system is a challenging task, especially when it comes to understanding and analysing the behaviour of the system under different conditions in order to tune or to improve it as to achieve the level of effectiveness needed to meet the user expectations.

Failure analysis is especially resource demanding in terms of time and human effort, since it requires inspecting, for several queries, system logs, intermediate output of system components, and, mostly, long lists of retrieved documents which need to be read one by one in order to try to figure out why they have been ranked in that way with respect to the query at hand.

Considering this, it is important to define new ways to help IR researchers, analysts and developers to understand the limits and strengths of the IR system under investigation. Visual analytics techniques can give assistance to this process by providing graphic tools which interacting with IR techniques may ease the work of the users.

The goal of this paper is to exploit a visual analytics approach to design a methodology and develop an interactive visual system which support IR researchers and developers in conducting experimental evaluation and improving their systems by: (i) reducing the effort needed to conduct failure analysis; (ii) allowing them to anticipate what the impact of a modification to their system could be before needing to actually implement it.

[*] The extended version of this abstract has been published in [1].

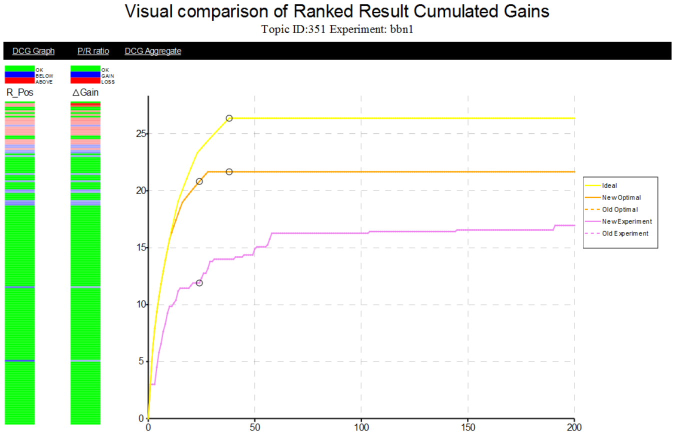

Fig. 1. The Visual Analytics prototype.

2 Failure Analysis

As far as the failure analysis is concerned, we introduce a ranking model that allows us to understand what happens when you misplace documents with different relevance grades in a ranked list. The proposed ranking model is able to quantify, rank by rank, the gain/loss obtained by an IR system with respect to both the ideal ranking, i.e. the best ranked list that can be produced for a given topic, and the optimal ranking, i.e. the best ranked list that can be produced using the documents actually retrieved by the system.

Starting from the *Discounted Cumulative Gain (DCG)* measures, we introduce two functions: the relative position, which quantifies how much a document has been misplaced with respect to its ideal (optimal) position, and the delta gain, which quantifies how much each document has gained/lost with respect to its ideal (optimal) DCG. On top of this ranking model, we propose a visualization, see Figure 1, where the DCG curves for the experiment ranking, the ideal ranking, and the optimal ranking are displayed together with two bars, on the left, representing the relative position and the delta gain. Please note that an equivalent graph can be obtained by using nDCG in the place of DCG.

The proposed ranking model and the related visualization are quite innovative because, usually, information visualization and visual analytics are exploited to improve the presentation of the results of a system to the end user, rather than applying them to the exploration and understanding of the performances and behaviour of an IR system. Secondly, comparisons are usually made with respect to ideal ranking only while our method allows user to compare a system also which respect to the optimal ranking produced with the system results, thus giving the possibility of better interpreting the obtained results [2].

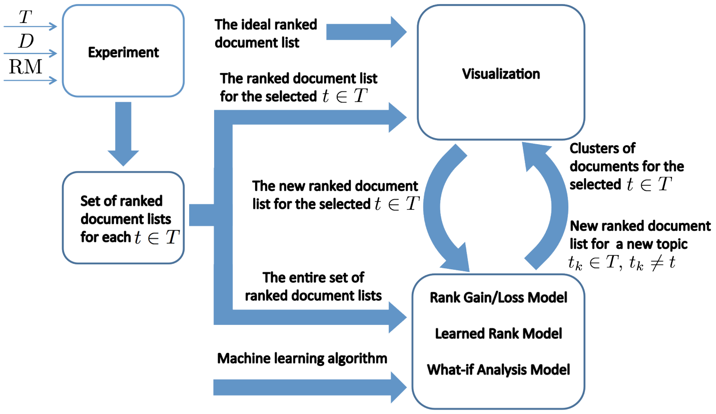
T
D
RM
Experiment
The ideal ranked document list
Visualization
The ranked document list for the selected $t \in T$
Set of ranked document lists for each $t \in T$
The new ranked document list for the selected $t \in T$
Clusters of documents for the selected $t \in T$
New ranked document list for a new topic $t_k \in T$, $t_k \neq t$
The entire set of ranked document lists
Rank Gain/Loss Model
Learned Rank Model
What-if Analysis Model
Machine learning algorithm

Fig. 2. Data pipeline.

3 What-If Analysis

When it comes to the what-if analysis, i.e. allowing users to anticipate the impact of a modification, we allow them to simulate what happens when you change the ranking of a given document for a certain topic not only in terms of which other documents will change their rank for that topic but also in terms of the effect that this change has on the ranking of the other topics. In other terms, we try to give the user an estimate of the "domino effect" that a change in the ranking of a single document can have. Moreover, when you simulate the move of a single document (and all the related documents), you produce a new ranking for a given topic which corresponds to a new version of your system, in our case a bug fixing in a component of the system. However, this new version of the system will now behave differently when ranking documents for the other topics in your experimental collection. Therefore, a change in the system which positively affects the performances on topic t_1 may have the side-effect to be detrimental for the performances on topic t_2 and we would like to give users an estimate also of this kind of "domino effect".

Therefore, the overall goal is to have an initial raw estimate of the effect of a planned modification before actually implementing it in terms of effect both for the topic under examination and for the other topics. This gives researchers and developers the possibility of exploring several alternatives before having to implement them and of determining a reasonable trade-off between the effort and costs for given modifications and the expected improvements.

Figure 2 shows the block diagram describing the pipeline of the data exchanged in the whole process. We consider the general-purpose IR scenario composed by a set of topics T, a collection of documents D, and a ranking model RM; an IR system for a given topic $t_k \in T$ retrieves a set of documents $D_j \subseteq D$.

The ranking model RM generates for each topic $t_k \in T$ a ranked document list RL_j. The whole set of ranked lists constitute the input for building the Clustering via Learning to Rank Model that is in charge of generating, for each document, a similarity cluster. The Visualization deals with one topic t at time: it takes as input the ranked document list for the topic t and the ideal ranked list, obtained choosing the most relevant documents in the collection D for the topic t and ordering them in the best way. While visually inspecting the ranked list, it is possible to simulate the effect of interactively reordering the list, moving a target document d and observing the effect on the ranking while this shift is propagated to all the documents of the cluster containing the documents similar to d. This cluster of documents simulates the “domino effect” within the given topic t.

When the analyst is satisfied with the results, i.e. when he has produced a new ranking of the documents that corresponds to the effect that is expected by modifications that are planned for the system, he can feed the Clustering via Learning to Rank Model with the newly produced ranked list, obtain a new model which takes into account the just introduced modifications, and inspecting the effects of this new model for other topics. This re-learning phase simulates the “domino effect” on the other topics different from t caused by a possible modification in the system.

4 Final Remarks

This paper presented a fully-fledged analytical and visualization model to support interactive exploration of IR experimental results. The overall goal of the paper has been to provide users with tools and methods to investigate the performances of a system and explore different alternatives for improving it avoiding a continuous iteration of trials-and-errors to see if the proposed modifications actually provide the expected improvements.

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. M. Angelini, N. Ferro, G. Santucci, and G. Silvello. Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In J. Kamps, W. Kraaij, and N. Fuhr, editors, *Proc. 4th Symposium on Information Interaction in Context (IIiX 2012)*. ACM Press, New York, USA, 2012.
2. E. Di Buccio, M. Dussin, N. Ferro, I. Masiero, G. Santucci, and G. Tino. To Re-rank or to Re-query: Can Visual Analytics Solve This Dilemma? In *Multilingual and Multimodal Information Access Evaluation. Proc. of the 2nd Int. Conf. of the Cross-Language Evaluation Forum (CLEF 2011)*, pages 119–130. LNCS 6941, Springer, Heidelberg, Germany, 2011.