Information Retrieval Failure Analysis: Visual Analytics as a Support for Interactive "What-If" Investigation

Marco Angelini
Sapienza University of Roma, Italy

Nicola Ferro
University of Padua, Italy

Guido Granato
Sapienza University of Roma, Italy

Guiseppe Santucci
Sapienza University of Roma, Italy

Gianmaria Silvello
University of Padua, Italy

Abstract

This poster provides an analytical model for examining performances of IR systems, based on the discounted cumulative gain family of metrics, and visualization for interacting and exploring the performances of the system under examination. Moreover, we propose machine learning approach to learn the ranking model of the examined system in order to be able to conduct a "what-if" analysis and visually explore what can happen if you adopt a given solution before having to actually implement it.

1 Introduction

Designing, developing, and testing an Information Retrieval (IR) system is a challenging task, especially when it comes to understanding and analyzing the behavior 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.

Moreover, conducting such analyses 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: this activity is usually called, in the IR field, *failure analysis*.

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

2 Approach and Contributions

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. The ranking model builds on the Discounted Cumulative Gain (DCG) family of measures [1, 2], which are designed to work graded relevance and are well-suited both to quantify system performances and to give an idea of the overall user satisfaction with a given ranked list considering the persistence of the user in scanning the list. We propose a visualization, see for example Figure 2, where the DCG curves for the system ranking, the ideal ranking, and the optimal ranking are displayed in the right area, together with two bars, on the left, which represents the relative position, $R_{pos}(V[i])$, and delta gain, $\Delta\_Gain(V,i)$, with a color coding that allows us to easily spot problematic and misplaced documents.

IEEE Symposium on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

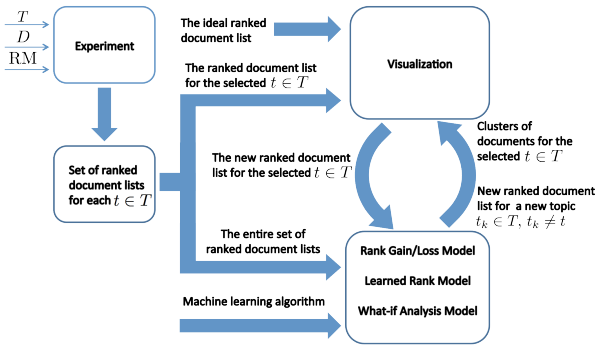

Figure 1: Data pipeline.

The proposed ranking model and the related visualization are quite innovative because applied to the exploration and understanding of the performances and behavior of an IR system. Moreover, our method allows user to compare a system not only with respect to the ideal ranking, but also with respect to the optimal ranking produced with the system results, thus giving the possibility to better interpreting the obtained results.

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.

In order to achieve this goal we have defined two analytical models:the first analytical model is based on learning to rank techniques [3] in order to learn a model of the system under examination from the ranked lists produced for each topic $t \in T$. From the learned model of the system, we then perform clustering in order to understand which documents would be moved together with a selected one, as part of the same cluster according to the system way of working. The second analytical model is devoted to frame what happens when you try to move a document from one position to another one in the ranking, how the other documents in the same cluster move in accordance with the move of the selected document, and how the other documents in the list relocate themselves.

3 The Visual Analytics System Overview

The visual analytics failure analysis system consists of a Web application that retrieves data from a remote server and allows the user to visually analyze it, in a static and/or interactive way.

Figure 1 shows the block diagram describing the pipeline of the data exchanged in the whole process. The ranking model RM generates for each topic $t_j \in T$ a ranked document list RL_j. The whole set of ranked lists constitutes 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 a 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

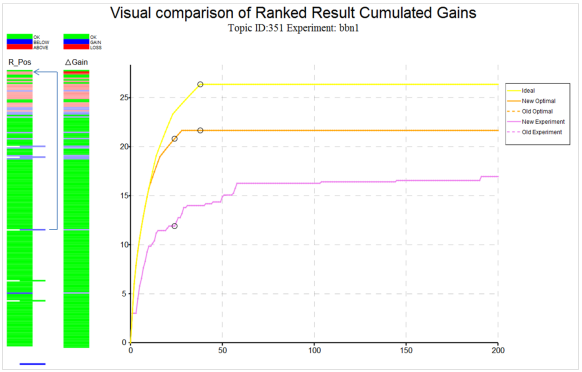

Figure 2: The Visual Analytics prototype.

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, obtaining 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 Visualization

Figure 2 shows a screenshot of the running prototype. Ranked result area, on the left, is compound of two vectors. The first one is the relative position vector. The prototype first compute the optimal ranked list of the documents, and then assign to each document a color based on its position, *R_Pos*, relative to the one that it has in the optimal ranked list. The color intensity gives to the user a visual indication of how far the document is from its optimal position.
The second vector represents the Δ*_Gain* function values for each document, and quantifies the effects of a misplaced/well placed document. The used color codes are the same as the previous ones, but this time a red color represents a loss of Δ*_Gain* based on the actual position of the documents, while a blue color represents instead a gain.

In the right area are displayed three different curves:
Experiment Ranking refers to the top n ranked results provided by the IR approach under consideration;
Optimal Ranking refers to an optimal re-ranking of the experiment ranking where experiment items, namely documents, are ranked in descending order of the degree of relevance according to the judgements (Ground truth) in the pool;
Ideal Ranking refers to the top n ranked documents in the pool, where documents are ranked in descending order of their degree of relevance.
The visualizations, still based on different kinds of data, are interconnected, and is possible to highlight a document in one of them and easily check the values assigned to the other metrics in the other graphs.

This work focuses on a novel what-if functionality, specifically the capability of interacting with the ranked vector of *R_Pos*. The system allows the user to remove a target document t from its actual position and placing it in a new one , in a "drag n drop" fashion, with the goal of investigating the result of a change in the search algorithm, inspecting the DCG of the resulting modified ranked list. This analysis is achieved using an additional set of data retrieved from the Clustering via Learning to Rank Model, that computes for each document the cluster of similar documents.

To evaluate the changes in the DCG function, both the old curves trends will be represented in a dash-stroke fashion.

5 Application Examples

The prototype has been designed in tight collaboration with IR experts that have reported positive feedbacks on the implemented prototype and several suggestions for improvement. The test collection adopted is based on data from the TREC7 Ad-hoc test collection. A subset of all the *topics* 351-400 is considered, specifically those re-assessed in [1]. Moreover, the interaction with domain experts raised different usage situations, discussed in the rest of the section. Now will be presented three different scenario of utlization:

In the first one, namely **Free Evolution** scenario, we can observe what happens when the user selects a document, referred in the following as "Target Document" t, and modifies its position. Without loss of generality we assume that the user's goal is to move t in a lower position. All the similar documents are displaced accordingly (downward) and, after the re-learning phase, the resulting curves show an higher value of DCG.

In second scenario,**Capped Evolution**, we can observe what happens when the user chooses to move the target document t with one or more similar documents positioned above its position. Due to the possibility that the displacement used for the target element cannot be applied to its "higher" similar documents, resulting in a "cap" of the new position desired for the document t, not necessarily the movement will take place as desired by the user. For topic 351, this led to an overall worse value of the DCG for the resulting curves.

In the last scenario,**New Entry Evolution**, we can observe what happens when the user chooses to select a target document t that has one or more similar documents, positioned below the "window" of displayed values of the experiment; such documents might be "called in" by the displacement the user assigns to t. For topic 351, this will not only lead to an improvement of the DCG of the experiment, but also of the one relative to the ideal curve (entry of a more relevant document at the expenses of a less relevant one).

Finally, we inspect **Domino Effect** on other Topics. Based on the previous scenarios of utilization, related to topic 351, here we shows the effects on topic 353 of the modifications on the ranking of topic 351, as estimated after re-learning the ranking model of the system. The modifications on topic 351 have a negative impact on topic 353 since both the experiment and the optimal curves are lower than before these modifications. The lowering of the experiment curve indicates a worsening in the ranking and the lowering of the optimal curve indicates that less relevant documents are retrieved than before the modification on topic 351.

References

[1] K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information System (TOIS)*, 20(4):422–446, October 2002.

[2] H. Keskustalo, K. Järvelin, A. Pirkola, and J. Kekäläinen. Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In T.-S. Chua, M.-K. Leong, D. W. Oard, and F. Sebastiani, editors, *Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 675–682. ACM Press, New York, USA, 2008.

[3] T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.