To Re-Rank or To Re-Query: Can Visual Analytics Solve This Dilemma?

Emanuele Di Buccio[1], Marco Dussin[1], Nicola Ferro[1], Ivano Masiero[1], Giuseppe Santucci[2], and Giuseppe Tino[2]

[1] University of Padua, Italy
{dibuccio,dussinma,ferro,masieroi}@dei.unipd.it
[2] Sapienza University of Rome, Italy
{santucci,tino}@dis.uniroma1.it

Abstract. Evaluation has a crucial role in Information Retrieval (IR) since it allows for identifying possible points of failure of an IR approach, thus addressing them to improve its effectiveness. 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. In this paper we discuss a Visual Analytics-based approach to support the analyst when deciding whether or not to investigate re-ranking to improve the system effectiveness measured after a retrieval run. Our approach is based on effectiveness measures that exploit graded relevance judgements and it provides both a principled and intuitive way to support analysis. A prototype is described and exploited to discuss some case studies based on TREC data.

1 Introduction

Inspecting and understanding the causes for the performances of an IR system is always a difficult and demanding task. For example, *failure analysis*, i.e. the detailed and manual analysis for understanding the behaviour and variability of retrieval across topics is often overlooked due to its complexity. The most extensive attempt in this respect has been the Reliable Information Access (RIA) workshop [1] which involved 28 people from 12 organizations for 6 weeks requiring from 11 to 40 person-hours per topic, which shows just how demanding these tasks are.

In this paper, we investigate a methodology for supporting researchers and developers in getting insights about the performances of their algorithms and systems. The methodology builds on the Discounted Cumulative Gain (DCG) family of measures [2, 3] because they can handle usefulness scores ranging in a non binary scale and have shown they are especially 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 try to better understand what happens when you flip documents with different relevance grades in a ranked list. This is achieved by providing a formal

model that allows us to properly frame the problem and quantify the gain/loss with respect to both an optimal and an ideal ranking, rank by rank, according to the actual result list produced by an IR system. This means that we compare the actual result list with respect to an optimal one created with the same documents retrieved by the IR system, but with an optimal ranking; we also compare the actual result list with respect to an ideal ranking created starting from the relevant documents in the pool (this ideal ranking is what is usually used to normalize the DCG measures). This differs in two ways from what is usually done: firstly, the analysis is conducted rank by rank and not by the overall performances or the area of the difference under two performance curves; secondly, the comparison is done with respect to an optimal ranking created with the same results of the IR system under examination and not only with respect to to an ideal ranking, created with the best results possible, i.e. also considering relevant documents not retrieved by the system.

Our method gives an idea of the distance of an IR system with respect to both its own optimal performances and the best performances possible. The method is adopted as basis for Visual Analytics (VA) techniques that allow analysts to get an intuitive idea through diverse visualizations on possible strategies that could be adopted to improve the IR system performance measured after a retrieval run. In this way, we support researchers and developers in trying to answer an ambitious question: is it better to invest on re-ranking the documents already retrieved by the system or is it better to issue a modified query in the entire collection? In other terms, the proposed techniques allow us to understand whether the system under examination is satisfactory from the recall point of view but unsatisfactory from the precision one, thus possibly benefiting from re-ranking, or if the system also has a too low recall, and thus it would benefit more from re-querying.

Moreover, these visualizations are suitable not only for specialists in the IR field, such as researchers and system developers, but also for users and stakeholders belonging to other communities which employ IR system as components of wider systems. As an example, you can consider the Digital Library (DL) community, where IR systems are usually components of wider DL Systems used to provide access to and retrieval of the multilingual and multimedia cultural heritage assets managed by the system. This is especially important if you consider that such communities which adopt IR system often have difficulties in understanding and assessing the performances of an actual IR system to be embedded into their systems, since this usually requires too specialistic competencies.

This paper describes a prototype that exploits the model and diverse visualizations of it; the prototype is then adopted to analyze several experiments carried out on the TREC7 Ad-hoc track. The paper is organized as follows. Section 2 discusses related work. Section 3 introduces the metrics and the model underling the system together with their visualization and a description of the implemented prototype. Section 4 describes an experiment of the system usage; Section 5 concludes the paper, pointing out ongoing research activities.

2 Related Work

The overall idea of DCG measures is to assign a gain to each relevance grade and for each position in the rank a discount is computed. Then, for each rank, DCG is computed by using the cumulative sum of the discounted gains up to that rank. This gives rise to a whole family of measures, depending on the choice of the gain assigned to each relevance grade and the used discounting function. Typical instantiations of DCG measures make use of positive gains and logarithmic functions to smooth the discount for higher ranks – e.g. a $\log_2$ function is used to model impatient users while a $\log_{10}$ function is used to model very patient users in scanning the result list. More recent works [3] have also tried to assign also negative gains to not relevant documents: this gives rise to performance curves that start falling sooner than the standard ones when non relevant documents are retrieved and let us better grasp, from the user's point of view, the progression of retrieval towards success or failure.

A work that exploits DCG to support analysis is [4] where the authors propose the potential for personalization curve. The potential for personalization is the gap between the optimal ranking for an individual and the optimal ranking for a group. The curves plots the average nDCG's (normalized DCG) for the best individual, group and web ranking against different group size. These curves were adopted to investigate the potential of personalization of implicit content-based and behavior features. Our work shares the idea of using a curve that plots DCG against rank position, as in [2], but using the gap between curves to support analysis as in [4].

The model proposed in this paper provides the basis for the development of VA techniques that can provide us with a quick and intuitive idea of what happened in a result list and what determined its perceived performances. Visual Analytics [5] is an emerging multi-disciplinary area that takes into account both ad-hoc and classical Data Mining (DM) algorithms and Information Visualization (IV) techniques, combining the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where human beings and machines cooperate using their respective distinct capabilities for the most effective results. Decisions on which direction analysis should take in order to accomplish a certain task are left to final user. Although IV techniques have been extensively explored [6, 7], combining them with automated data analysis for specific application domains is still a challenging activity [8]. In the VA community previous approaches have been proposed for visualizing and assessing a ranked list of items, e.g. using rankings for presenting the user with the most relevant visualizations [9], or for browsing the ranked results [10].

Visualization strategies have been adopted for analyzing experimental runs, e.g. beadplots in [11]. Each row in a beadplot corresponds to a system and each "bead", which can be gray or coloured, 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. The colouring scheme uses spectral (ROYGBIV) coding; the ordering adopted for colouring (from dark red for most relevant to light violet for least relevant) is based on a reference system, not on graded judgements and the optimal ranking as in our work. Moreover, in [11] the strategies are adopted for a comparison between the performance of different systems, i.e. the diverse runs; our approach aims at supporting the analysis of a single system, even though it can be generalized for systems comparison.

Another related work is the Query Performance Analyzer (QPA) [12]. This tool 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; our VA approach extends the QPA relevance bar by providing an intuitive visualization for quantifying the gain/loss with respect to both an optimal ranking. QPA 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; in contrast, the curves considered in this work allow the comparison between the system performance with the optimal and ideal ranking that can be obtained from a result list.

None of these works deal with the problem of observing the ranked item position, comparing it with an ideal solution, to assess and improve the ranking quality. In [13] the authors explore the basic issues associated with the problem, providing basic metrics and introducing a VA web-based system for exploring the quality of a ranking with respect to an optimal solution. This paper extends such results, allowing for assessing the ranking quality with both the optimal and the ideal solutions and presenting an experiment based on data from runs of the TREC7 Ad-hoc track and the pool obtained in [2].

3 The formal model

According to [2] we model the retrieval results as a ranked vector of n documents V, i.e. $V[1]$ contains the identifier of the document predicted by the system to be most relevant, $V[n]$ the least relevant one. The ground truth GT function assigns to each document $V[i]$ a value in the relevance interval $\{0..k\}$, where k represents the highest relevance score. The basic assumption is that the higher the position of a document the less likely it is that the user will examine it, because of the required time and effort and the information coming from the documents already examined. As a consequence, the higher the rank of a relevant document the less useful it is for the user. This is modeled through a discounting function DF that progressively reduces the relevance of a document, $GT(V[i])$ as i increases. We do not stick with a particular proposal of DF and we develop a model that is parametric with respect to this choice. However, to fix the ideas, we recall the original DF proposed in [2]:

$$DF(V[i]) = \begin{cases} GT(V[i]), \text{ if } i \leq x \\ GT(V[i])/\log_x(i), \ if \ i > x \end{cases} \quad (1)$$

that reduces, in a logarithmic way, the relevance of a document whose rank is greater than the logarithm base. For example, if $x = 2$ a document at position

16 is valuable as one fourth of the original value. The quality of a result can be assessed using the function $DCG(V,i) = \sum_{j=1}^{i} DF(V[j])$ that estimates the information gained by a user who examines the first i documents of V. This paper exploits the variant adopted in `trec_eval` where GT is divided by $\log_x(i+1)$.

The DCG function allows for comparing the performances of different IR systems, e.g. plotting the DCG(i) values of each IR system and comparing the curve behavior. However, if the user's task is to improve the ranking performance of a single IR system, looking at the misplaced documents (i.e. ranked too high or too low with respect to the other documents) the DCG function does not help, because the same value DCG(i) could be generated by different permutations of V and because it does not point out the loss in cumulative gain caused by misplaced elements. To this end, we introduce the following definitions and novel metrics. We denote with $OptPerm(V)$ the set of optimal permutations of V such that $\forall OV \in OptPerm(V)$ it holds that $GT(OV[i]) \geq GT(OV[j]) \forall i,j <= n \bigwedge i < j$, that is, OV maximizes the values of DCG(OV,i)$\forall i$. In other words, $OptPerm(V)$ represents the set of the optimal rankings for a given search result.

It is worth noting that each vector in $OptPerm(V)$ is composed of $k+1$ intervals of documents sharing the same GT values. As an example, assuming a result vector composed by 12 elements and $k=3$, a possible sequence of GT values of an optimal vector OV is $< 3,3,3,3,2,2,2,2,1,1,0,0 >$; according to this we define the $max\_index(V,r)$ and $min\_index(V,r)$ functions, with $0 \leq r \leq k$, which return the greatest and the lowest indexes of elements in a vector belonging to $OptPerm(V)$ that share the same GT value r. For example, considering the above 12 GT values, $min\_index(V,2) = 5$ and $max\_index(V,2) = 8$.

Using the above definitions we can define the relative position $R\_Pos(V[i])$ function for each document in V as follows:

$$R\_Pos(V[i]) = \begin{cases} 0, \ \textit{if } min\_index(V, GT(V[i]) \leq i \leq max\_index(V, GT(V[i]) \\ min\_index(V, GT(V[i]) - i, \ \textit{if } i < min\_index(V, GT(V[i]) \\ max\_index(V, GT(V[i]) - i, \ \textit{if } i > max\_index(V, GT(V[i]) \end{cases} \quad (2)$$

$R\_Pos(V[i])$ allows for pointing out misplaced elements and understanding how much they are misplaced: 0 values denote documents that are within the optimal interval, negative values denote elements that are below the optimal interval (pessimistic ranking), and positive values denote elements that are above the optimal (optimistic ranking). The absolute value of $R\_Pos(V[i])$ gives the minimum distance of a misplaced element from its optimal interval.

According to the actual relevance and rank position, the same value of $R\_Pos(V[i])$ can produce different variations of the DCG function. We measure the contributions of misplaced elements with the function $\Delta\_Gain(V,i)$ which compares $\forall i$ the actual values of $DF(V[i])$ with the corresponding values in OV, $DF(OV[i])$: $\Delta\_Gain(V,i) = DF(V[i]) - DF(OV[i])$. Note that while DCG($V[i]$) $\leq$ DCG($OV[i]$) the $\Delta\_Gain(V,i)$ function assumes both positive and negative values. In particular, negative values correspond to elements that are presented too early (with respect to, their relevance) to the user and positive values to elements that are presented too late. Visually inspecting the values

Optimal vector

i	GT(OV)	DF	DCG[i]
1	3	3,00	3,00
2	3	3,00	6,00
3	3	1,89	7,89
4	3	1,50	9,39
5	2	0,86	10,25
6	2	0,77	11,03
7	2	0,71	11,74
8	2	0,67	12,41
9	1	0,32	12,72
10	1	0,30	13,02
11	0	0,00	13,02
12	0	0,00	13,02

Experiment vector

i	GT(V)	DF	DeltaGain	DCG[i]
1	3	3,00	0,00	3,00
2	1	1,00	-2,00	4,00
3	2	1,26	-0,63	5,26
4	3	1,50	0,00	6,76
5	2	0,86	0,00	7,62
6	2	0,77	0,00	8,40
7	3	1,07	0,36	9,47
8	2	0,67	0,00	10,13
9	0	0,00	-0,32	10,13
10	1	0,30	0,00	10,43
11	0	0,00	0,00	10,43
12	3	0,84	0,84	11,27

Fig. 1. Visual representation of R_Pos and Δ_Gain.

of these two metrics allows the user to easily locate misplaced elements and understand the impact that such errors have on DCG.

3.1 The prototype

The results presented in this paper have been implemented in a web based prototype that for a given topic q visualizes the R_Pos and $Delta\_Gain$ functions, together with the DCGs plotted against the rank position for the experiment, the optimal ranking and the ideal ranking where:

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 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.

Basically, *optimal* refers to the best ranking the system could have provided on the basis of the retrieved documents, while *ideal* refers to the best ranking the system could have provided on the basis of the knowledge of all the relevant documents in the pool. From now on the curves obtained by interpolating the DCGs at the diverse rank positions for the experiment, the optimal, and the ideal ranking will be named respectively *experiment*, *optimal*, and *ideal curve*.

Figure 1 shows the visualization choices adopted in the VA prototype. The leftmost table in the figure represents one of the optimal vectors of $OptPerm(V)$. The second column of the table contains the GT values, the third one the DF values (computed using a log_2 function), and the fourth one the DCG function.

The rightmost table represents the experiment result V. The second column contains the GT values together with the R_Pos function, coded through color shading: values on correct position=green, values on above positions=blue, and values on below positions=red. The third column contains the DF values. The fourth column contains the Δ_Gain function, where negative values are coded in red, positive values are coded in blue, and 0 values are coded in green. The fifth column represents the experiment DCG function.

The prototype allows researchers and analysts to compare the experiment result with both the optimal and the ideal result. This facilitates the activities of failure analysis, making it easy to locate misplaced elements, blue or red items, that pop up from the visualization as well as the extent of their displacement and the impact they have on DCG. In this way the analyst can gain insights into the worst errors of the IR system and devise suitable recovering actions.

Figure 2 shows a screenshot of the prototype: the vector on the left represents the R_Pos function through color shadings: green, light red/red, and light blue/blue. It allows for locating misplaced documents and, thanks to the shading, understanding how far they are from the optimal position. The vector on the right shows $Delta\_Gain$ values: light blue/blue codes positive values, light red/red negative values, and green 0 values. A mouse-over triggered interactive pop-up window allows for inspecting the numerical values of single documents: R_Pos, $Delta\_Gain$ and DCG, together with a link to the document. The rightmost part of the screen shows the DCG graphs of the ideal vector, of the optimal vector and of the experiment vector, namely the ranking curves. The points of maximum distance between the experiment and the optimal curves and between the optimal and the ideal curves (highlighted by red circles) can also be seen. A useful popup appears when the mouse is over the graph and displays information about the DCGs of the curves and the distance between them at the rank identified by the mouse position. Brushing allows for highlighting relationships between graph and vectors; indeed, by placing the mouse cursor over colored rows the corresponding point on the graph is highlighted. Finally, through the input panel below the graphs the logarithm base can be changed to model different discount functions according to different classes of search users.

4 Experimentation

The objective of the VA approach introduced in this paper is to support a researcher or analyst investigating how to improve the effectiveness of an IR approach, when the results for one or more queries on the same topic are available. Let us consider, for instance, the case of a retrieval run on a test collection for which the IR approach under evaluation is not effective for one or more topics when considering the top n — in this work $n = 1000$; let us focus on one of these topics. Possible causes of poor performance can be a lack of capability of the system in either: (i) retrieving relevant documents, e.g. a low recall is observed; or (ii) ranking highly relevant document at high rank positions, when the measure of effectiveness adopted is based on graded relevance judgements, as

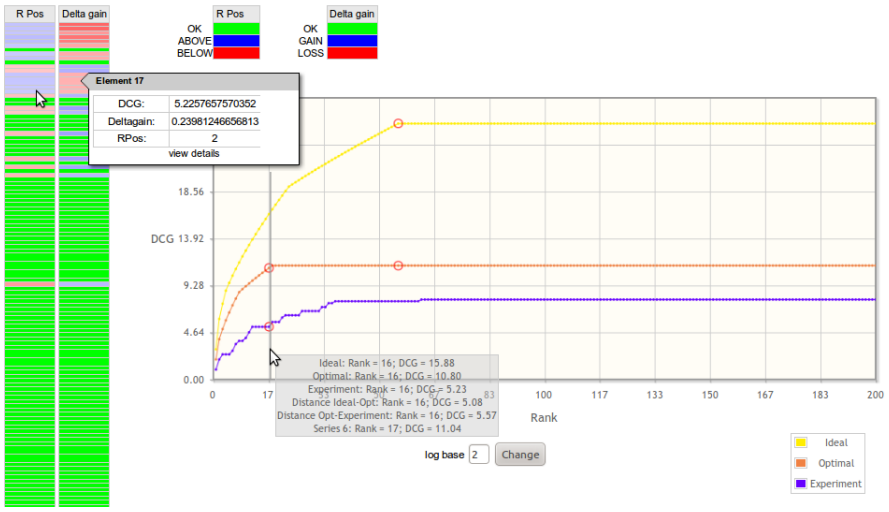

Fig. 2. A screenshot of the prototype.

for the family of measures considered in this work; or both of these. A possible approach to address the former issue is to perform a new modified query on the entire collection in order to gather additional relevant documents among the top n. In contrast, if a high recall is observed but the system was unable to rank the documents in descending degree of relevance, a more effective choice to improve the effectiveness of that run could be performing top n document re-ranking in order to achieve the optimal ranking. In this paper we will focus on visualization to support the selection of the strategies to improve system performance, not on the actual implementation of these strategies. In this section we will show how the proposed VA approach can help address the following question: *given a ranked result list obtained in response to a query submitted to the system, should we re-rank the top n documents in the retrieved result list or issue a new modified query on the entire collection?*

The remainder of this section will discuss how the prototype can be adopted to address this question, specifically considering some case studies based on data of the Ad-hoc Track of the TREC7 evaluation campaign.

Dataset 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 [2]. Indeed, the *relevance judgements* adopted are those obtained by the evaluation activity carried out in that paper. All the relevant documents of 20 TREC7 topics and 18 TREC8 topics were re-assessed together with 5% of documents judged as not relevant, where assessment was performed using a four graded relevance scale; details on the re-assessment procedure can be found in [2]. The TREC7 Ad-hoc test collection together with this set of judgements

were used because of the family of measures adopted in our VA approach, namely DCG. The way the VA approach can be adopted to support researchers and analysts during the evaluation is based on runs submitted to the TREC7 Ad-hoc Track. In order to be consistent with the choice adopted in [2] we will visualize the curves for top $k = 200$ rank positions.

DCG Curves to Support Per-Topic Analysis A possible approach for addressing the above research question is to examine the DCG curves, specifically looking at the distance between them. Let us consider, for instance, the experiment `KD71010q` whose data is visualized in Figure 2 for topic `365`. The existence of a gap between the experiment and the optimal curve suggests that an improvement in terms of DCG can be obtained by investigating an optimal re-ranking for the set of retrieved documents for the IR approach under evaluation and the considered topic. Indeed, the distance between experiment and optimal curve at a given rank position indicates the maximum increment in terms of gain that can be achieved by an optimal re-ranking; for instance, at rank 16 the maximum increment that can be achieved in terms of gain is $\Delta = 10.80 - 5.23 = 5.57$ — this Δ differs from Δ_Gain.

In general, if a gap exists between experiment and optimal curve, an improvement in terms of effectiveness can be accomplished by investigating a strategy for optimal re-ranking of the retrieved document set. However re-ranking is not necessarily the best strategy to adopt. Indeed, an analysis of the optimal and the ideal curves reported in Figure 2 shows that a large gap exists between them, which indicates that the system retrieved a low number of relevant documents among those present in the pool, namely a low recall. Therefore, the researcher can opt for investigating strategies based on automatically modified queries, for instance exploiting feedback strategies, and issued on the entire collection, in order to increase the number of relevant documents retrieved instead of trying to optimally re-rank those currently retrieved.

In contrast the curves concerning experiment `mds98td` and topic `387`, whose data are visualized in Figure 3, suggest that investigating the re-ranking strategy can be beneficial if we are interested in improving effectiveness at high rank positions. A gap exists between the optimal and ideal curves, thus indicating that we can further improve recall, but the curves basically overlap in the top 10 rank positions and $\Delta < 1$ is observed for rank positions from 10 to 40. These values suggest that the IR system was able to retrieve highly relevant documents among the top 1000, but it was not able to rank them at high rank positions. The IR system was therefore effective in supporting a first stage prediction, i.e. the first of a series of search episodes needed by the search user in order to achieve his goal in multiple steps. Re-ranking is the best choice in this case. Another example is that depicted in Figure 4. The visualized data concern run `Brkly25` and topic `358`. Also in this case, the gap between the curves suggests re-ranking could be the best choice: the ideal and optimal curves are overlapped up to rank 40, namely $\Delta_{rank \leq 40} = 0$, and $0.19 < \Delta_{rank > 40} < 0.54$. Both runs depicted in Figure 3 and Figure 4 can benefit from re-ranking, as visually suggested by

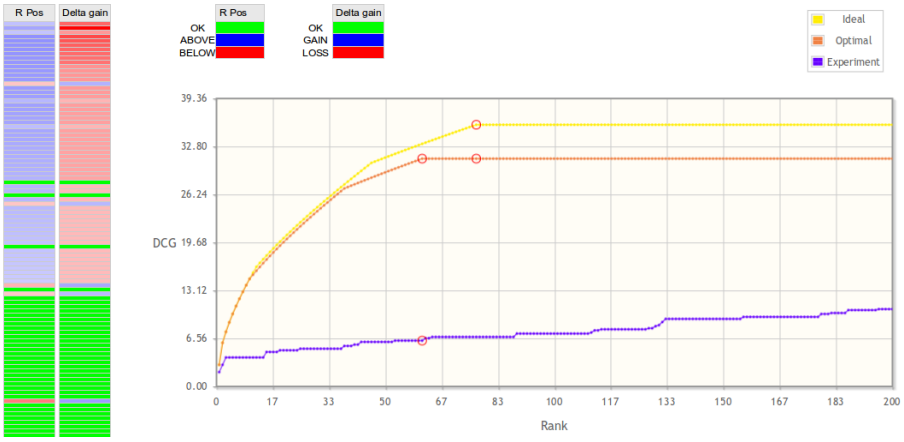

Fig. 3. Curves for experiment `mds98td` when considering topic `387`.

the gap among curves. But the comparison of Δ_Gain vectors, specifically the difference in terms of shading and in number of green entries, shows that there are more documents in the top-most positions that are far from their optimal position in the former case than in the latter. The analyst can interact with the worst ranked documents by a click, thus inspecting the document in order to understand which of its properties were possible causes of failures.

Towards a VA-based Methodology to Support IR Experiment Analysis The previous section discussed how the described prototype can support an analysis on a per-topic basis. An additional issue is the automatic categorization of topics according to the possible causes of failures of the system when searching from them. The approach we adopted to identify possible topics that can benefit for "re-ranking" or "re-query" is based on the correlation between vectors describing experiment, optimal and ideal ranking. Each topic is described by a pair $(\tau_{ideal-opt}, \tau_{opt-exp})$, where $\tau_{ideal-opt}$ denotes the Kendall τ correlation between the ideal and the optimal vectors of gains, while $\tau_{opt-exp}$ denotes the Kendall τ between the optimal and experiment vectors of gains. When the pair is $(1, 1)$ the best performance is achieved: this is the case of run `KD71010q` for topic `385`. A pair where $\tau_{ideal-opt}$ is high and $\tau_{opt-exp}$ is low suggests that re-ranking could probably improve effectiveness, since there is a strong correlation between ideal and optimal ranking, thus suggesting that the IR approach was quite effective in retrieving relevant documents, but not in the document ranking. This is for instance the case of run `mds98td` and topic `387` depicted in Figure 3 where the τ pair is $(0.88, 0.07)$. A pair where $\tau_{ideal-opt}$ is low or negative suggests "re-query" on the entire collection as a possible strategy to improve retrieval effectiveness, since an optimal re-ranking of the retrieved document is far from the ideal ranking. This is for instance the case of run `KD71010q` and topic `365` depicted in Figure 2 where the τ pair is $(0.59, 0.45)$. τ pairs can be adopted in a three step

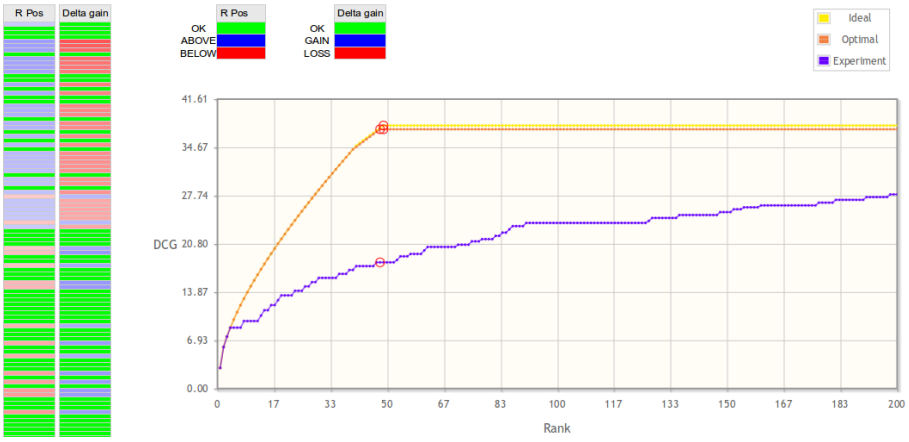

Fig. 4. Curves for experiment `Brkly25` when considering topic `358`.

methodology: (i) the pair values allow a first approximation to be obtained when identifying possible causes of failures and topics for which the approach failed; (ii) ranking curves analysis allows for a more in-depth investigation on a per-topic basis, and (iii) Δ_Gain and R_Pos vectors allows for an analysis on a per document basis.

5 Conclusion and Future Work

This paper presents some preliminary results of a VA system for IR evaluation able to explore the quality of a ranked list of documents. The challenging goal of the system is to point out the location and the magnitude of ranking errors in a way that provides insights that contribute to improving the ranking algorithm effectiveness. The system builds on existing and novel metrics that capture the quality of a ranking and allow us to compare it to the optimal one constructed starting from the actual results produced by the system, modeling the degree of satisfaction of a user when s/he inspects those search results. The comparison of the ranking curves, as well as the Δ_Gain and R_Pos vectors, provides an intuitive tool to support IR researchers when conducting retrospective analysis.

Future versions of the prototype could exploit Δ_Gain and R_Pos vector visualization as entry points for more complex user interaction, e.g. manual ranking modification. In [14] we reported on the design and the implementation of a prototype that accesses experimental data via standard Web services from a dedicated system. Access via a web service is adopted in order to allow for the design and development of various client applications and tools for exploiting those data; the prototype described in this paper is an instance of such applications. The prototype in [14] has been developed for a touch device and will be adopted to support the user study we intend to carry out to assess the methodology proposed in this paper. We are currently investigating the metrics,

algorithms and visualizations able to locate and visualize the most productive permutations of the result vectors, i.e. heuristic based best flips, and ways of visually correlating the rank of the documents with the ranking algorithm parameters. Lastly, the limitations observed for the Kendall τ in IR evaluation suggest using more suitable variants, e.g. those able to exploit graded relevance scale as proposed in [15].

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. Harman, D., Buckley, C.: Overview of the reliable information access workshop. Information Retrieval **12** (December 2009) 615–641
2. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information System **20** (October 2002) 422–446
3. Keskustalo, H., Järvelin, K., Pirkola, A., Kekäläinen, J.: Intuition-supporting visualization of user's performance based on explicit negative higher-order relevance. In: Proceedings of SIGIR '08, New York, NY, USA, ACM (2008) 675–682
4. Teevan, J., Dumais, S.T., Horvitz, E.: Potential for personalization. ACM Transactions on Computer-Human Interaction (TOCHI) **17**(1) (2010) 1–31
5. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melançon, G.: Information visualization. Springer-Verlag, Berlin, Heidelberg (2008) 154–175
6. Card, S.K., Mackinlay, J.: The structure of the information visualization design space. In: Proceedings of InfoVis '97, Washington, DC, USA, IEEE Computer Society (1997) 92–99
7. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, Washington, DC, USA, IEEE Computer Society (1996) 336–343
8. Keim, D., Kohlhammer, J., Santucci, G., Mansmann, F., Wanner, F., Schäfer, M.: Visual Analytics Challenges. In: Proceedings of the eChallenges 2009. (2009)
9. Seo, J., Shneiderman, B.: A rank-by-feature framework for interactive exploration of multidimensional data. Information Visualization **4** (July 2005) 96–113
10. Derthick, M., Christel, M.G., Hauptmann, A.G., Wactlar, H.D.: Constant density displays using diversity sampling. In: Proceedings of InfoVis'03, Washington, DC, USA, IEEE Computer Society (2003) 137–144
11. Banks, D., Over, P., Zhang, N.F.: Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval **1** (May 1999) 7–34
12. Sormunen, E., Hokkanen, S., Kangaslampi, P., Pyy, P., Sepponen, B.: Query performance analyser -: a web-based tool for ir research and instruction. In: Proceedings of SIGIR '02, New York, NY, USA, ACM (2002) 450–450
13. Ferro, N., Sabetta, A., Santucci, G., Tino, G., Veltri, F.: Visual comparison of Ranked Result Cumulated Gains. In: Proceedings of EuroVA 2011. (2011)
14. Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., Tino, G.: Interactive analysis and exploration of experimental evaluation results. In: Proceedings of EuroHCIR 2011, To Appear. (2011)
15. Melucci, M.: Weighted rank correlation in information retrieval evaluation. In: Proceedings of AIRS '09, Berlin, Heidelberg, Springer-Verlag (2009) 75–86