International Workshop on Visual Analytics (2011)
S. Miksch and G. Santucci (Editors)

Visual Comparison of Ranked Result Cumulated Gains

N. Ferro[2], A. Sabetta[1], G. Santucci[1], G. Tino[1], and F. Veltri[1]†

[1] Sapienza Universita di Roma, Italy
[2]University of Padua, Italy

Abstract

Ranking is fundamental in Information Retrieval (IR) *and several measures have been developed over the years to assess the quality of a ranked result list, such as those based on the idea of computing the cumulative gain up to a given ranked position and taking into account multiple relevance levels. These measures allow for comparing the performances of different* Information Retrieval System (IRS)*, giving credit to their ability to retrieve highly relevant documents and to rank them topmost in the result list. However, while this approach is able to assess the differences among two or more retrieval systems, it does not allow to easily understand and inspect the reasons of good or bad performances. To this end, this paper presents a* Visual Analytics (VA) *environment that allows for visually exploring the ranked retrieval results, pointing out the search failures and providing useful insights for improving the underlying IRS ranking algorithm.*

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [INFORMATION STORAGE AND RETRIEVAL]: Information Search and Retrieval—Search process; H.3.4 [INFORMATION STORAGE AND RETRIEVAL]: Systems and Software—Performance evaluation (efficiency and effectiveness);

1. Introduction and related work

Ranking is a central and ubiquitous issue in *Information Retrieval (IR)*: *Information Retrieval System (IRS)*, whose *Search Engine (SE)* represent a particularly important example, order the results retrieved in response to a user query according to the estimation of their relevance to the query. When it comes to assessing the performances of an IRS, the IR field has a strong and long-lived tradition that dates back to late 50s/early 60s of the last century. In particular, in the last 20 years, large-scale evaluation campaigns, such as the *Text REtrieval Conference (TREC)*† [HV05] in the United States and the *Cross-Language Evaluation Forum (CLEF)*‡ [AFP*10] in Europe, have conducted cooperative evaluation efforts involving hundreds of research groups and industries, producing a huge amount of valuable data to be analysed, mined, and understood.

Large-scale evaluation campaigns rely mainly on the traditional Cranfield methodology [Cle97] which makes use of shared experimental collections in order to create comparable experiments and evaluate their performance. An experimental collection is a triple $\mathcal{C} = (D, Q, J)$, where: D is a set of documents, called also collection of documents; Q is a set of topics simulating actual user information needs, from which the actual queries are derived; J is a set relevance judgements, i.e., for each topic $q \in Q$ the documents $d \in D$, which are relevant for the topic q, are determined. Note that the relevance judgements J can be not only binary, i.e., relevant or not relevant, but also multi-graded, e.g., highly relevant, partially relevant, not relevant and so on. On the other hand, the relevance judgements J do not define a unique optimal ranking but only which documents are the “correct" answers and so a whole set of optimal rankings is possible by permutating those “correct" answers.

Overall, an experimental collection $\mathcal{C}$ allows the comparison of two retrieval methods, say X and Y, according to some measurements which quantifies the retrieval performances of these methods. The final aim is to compare the

† The reported work has been partially supported by the PROMISE network of excellence (http://www.promise-noe.eu/, contract n. 258191), as part of the 7th Framework Program of the European Commission.

† http://trec.nist.gov/

‡ http://www.clef-campaign.org/

ranked results lists produced by different systems according to various performance measures [RZ07] and statistical analyses [Car09, Hul93].

Many attempts have been made to develop metrics that capture the quality of a ranking and allow to compare it to an optimal/ideal one and model the degree of satisfaction of a user when he scans a result list [RZ07]. We focus our attention on the (normalized) *discounted cumulated gain* (n)DCG family of measures [JK02] because they have shown to be especially well-suited not only to quantify system performances but also to give an idea of the overall user satisfaction with a given ranked list considering the persistence of the user in scanning the list. The overall idea of DCG measures is to assign a gain to each relevance grade (according to the relevance judgements J) 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 raise to a whole family of measure, 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 – e.g., 0 for not relevant documents, 1 for partially relevant ones, and 3 for highly relevant ones – 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 [KJPK08] have tried to assign also negative gains to not relevant documents: this gives raise to performance curves that start falling sooner than the standard ones when not relevant documents are retrieved and let us to better grasp, from the user's point of view, the progression of retrieval towards success or failure.

The contribution of this paper is to improve on the previous work [JK02, KJPK08] by trying 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 an optimal ranking, rank by rank, according to the actual result list produced by an IRS. Our method gives an idea of the distance of an IRS with respect to to its own optimal performances rather than the distance from the best performances possible, setting a goal of improvements that might be more achievable.

The proposed model provides the basis for the development of VA techniques that give us the possibility to get a quick and intuitive idea of what happened in a result list and what determined its perceived performances. A relevant application of the proposed techniques is *failure analysis*, i.e. the detailed and manual analysis to understand the behaviour and variability of retrieval across topics. This is a critical and especially resource demanding task: the most extensive attempt in this respect has been the *Reliable Information Access (RIA)*[§] workshop [HB09] which involved 28 people from 12 organizations for 6 weeks requiring from 11 to 40 person-hours per topic. 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 IRS and SE as components of wider systems. As an example, you can consider the digital library community, where IRS are usually components of wider *Digital Library System (DLS)* used to provide access 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 IRS often have difficulties in understanding and assessing the performances of an actual IRS to be embedded into their systems, since this usually requires too specialistic competencies.

Finally, the idea itself of exploring and applying VA techniques to the experimental evaluation in the IR field is quite innovative since it has never been attempted before and, due to the complexity of the evaluation measures and the amount of data produced by large-scale evaluation campaigns, there is a strong need for better and more effective representation techniques. Moreover, visualizing and assessing ranked list of items, to the best of the authors' knowledge, has not been addressed by the VA community. The few related proposals, see, e.g., [SS04], use rankings for presenting the user with the most relevant visualizations, or for browsing the ranked result, see, e.g., [DCHW03], but do not deal with the problem of observing the ranked item position, comparing it with an ideal solution, to assess and improve the ranking quality.

The paper is organized as follows. Section 2 introduces the metrics and the model underling the system together with their visualization, Section 3 provides an overview of the implemented prototype, and Section 4 concludes the paper, pointing out ongoing research activities.

2. The formal model

According to [JK02] we model the retrieval result as a ranked vector of n documents V, i.e., $V[1]$ contains the document ID of the most relevant document (according to the search engine judgment), $V[n]$ the least relevant one. The ground truth GT function assigns to each document $V[i]$ a value in the relevance interval $(0..k)$, where $k << n$ represents the highest relevance score. Typical values of n and k are 200 and 3, respectively. The basic assumption is that the greater 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 greater 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.

[§] http://ir.nist.gov/ria

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 [JK02]: $DF(V[i]) = \begin{cases} GT(V[i]),\ if\ i \leq x \\ GT(V[i])/\log_x(i),\ if\ i > x \end{cases}$

that reduces, in a logarithmic way, the relevance of a document whose rank is greater than the logarithm base. As an 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 discounted cumulative gain function $DCG(V,i) = \sum_{j=1}^{i} DF(V[j])$ that estimates the information gained by a user that examines the first i documents of V.

The *DCG* function allows for comparing the performances of different search engines, e.g., plotting the $DCG(i)$ values of each engine and comparing the curve behavior.

However, if the user's task is to improve the ranking performance of a single search engine, 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 aim, 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 by $k+1$ 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 $0 \leq r \leq k$, that return the greatest and the lowest indexes of elements in a vector belonging to $OptPerm(V)$ that share the same GT value r. As an 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: $R\_Pos(V[i]) =$

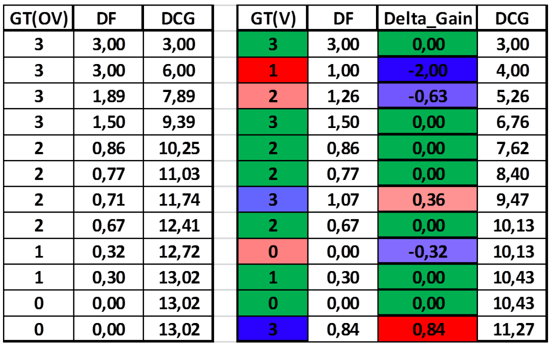
$$\begin{cases} 0,\ \text{if}\ min\_index(V,GT(V[i]) \leq i \leq max\_index(V,GT(V[i]) \\ min\_index(V,GT(V[i]) - i,\ \text{if}\ i < min\_index(V,GT(V[i]) \\ max\_index(V,GT(V[i]) - i,\ \text{if}\ i > max\_index(V,GT(V[i]) \end{cases}$$

$R\_Pos(V[i])$ 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 $\Delta\_Gain(V,i)$ that 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 corresponds 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 of these two metrics allows the user for easily locating misplaced elements and understanding the impact that such errors have on *DCG*.

3. The prototype

The results presented in this paper have been implemented in a web based prototype that visualizes the *R_Pos* and *Delta_Gain* functions, together with the optimal and the actual *DCGs*.



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

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

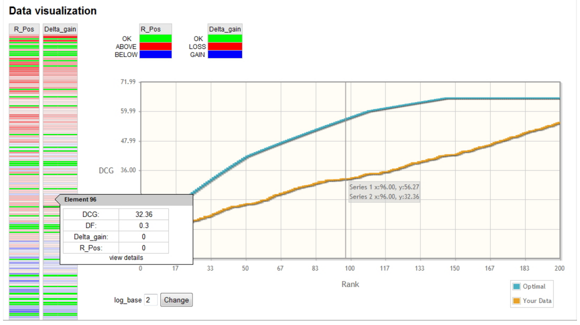
Figure 1: *Visual representation of R_Pos and Δ_Gain.*

Figure 1 shows the visualization choices adopted in the VA prototype. In particular, the leftmost table in the figure represents one of the optimum vector of $OptPerm(V)$. The first column contains the GT values, the second one the DF values (computed using a log_2 function), and the third one the *DCG* function. The rightmost table represents the actual search result V. The first column contains the GT values together with the *R_Pos* function, coded through color shading: 0=green, negative values=red, and positive values= blue. The third 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 fourth column represents the actual *DCG* function.

The prototype allows the end user for comparing the actual result with the optimal one and facilitate the activities of failure analysis, easily locating misplaced elements, blue or red items, that pop up from the visualization together with the extent of their displacement and the impact they have on *DCG*. In this way the analyst can gain insights on the worst errors of the search engine 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 they are far from the optimal position. The vector on the right shows *Delta_Gain* values: light blue/blue codes negative values, light red/red positive 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*, *DF*, *DCG*, together with a link to the document. The rightmost part of the screen shows the *DCG* graphs of *V* and *OV* vectors.

Brushing allows for highlighting relationships between graph and vectors; indeed by placing mouse cursor over colored rows the corresponded point on the graph is highlighted. Finally, through the input panel below the graphs it is possible to change the logarithm base for modeling different discount function according to different class of users.



Element 96

DCG:	32.36
DF:	0.3
Delta_gain:	0
R_Pos:	0

log_base 2

Figure 2: *A screenshot of the prototype.*

At time of writing a user study of the system has not been performed; however, the prototype has been discussed with IR experts that have reported positive feedbacks together with several suggestions for improvement .

4. Conclusions

This paper presents some preliminary results of a VA system for IR evaluation that allows for exploring 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 improve the IRS ranking algorithm effectiveness.

The system builds up 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 result.

We are currently investigating on:

- metrics, algorithms, and visualizations able to locate and visualize the most productive permutations of the result vectors, i.e., heuristic based best flips;
- a way of visually correlate the rank of the documents with the ranking algorithm parameters;
- an extension of the model able to deal with missing information, i.e., with the very common case in which a document has not been assigned a relevance value.

Moreover, we intend to assess our approach with a user study, as the system will incorporate the suggestions and the functionalities raised from the first exploratory interaction with IR experts.

References

[AFP*10] Agosti M., Ferro N., Peters C., de Rijke M., Smeaton A. (Eds.):. *Multilingual and Multimodal Information Access Evaluation. Proceedings of the Int. Conference of the Cross-Language Evaluation Forum (CLEF 2010)* (2010), Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany. 1

[Car09] Carterette B.: On Rank Correlation and the Distance Between Rankings. In *Proc. 32nd Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)* (2009), Allan J., Aslam J. A., Sanderson M., Zhai C., Zobel J., (Eds.), ACM Press, NY, USA, pp. 436–443. 2

[Cle97] Cleverdon C. W.: The Cranfield Tests on Index Languages Devices. In *Readings in Information Retrieval* (1997), Spärck Jones K., Willett P., (Eds.), Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, pp. 47–60. 1

[DCHW03] Derthick M., Christel M. G., Hauptmann A. G., Wactlar H. D.: Constant density displays using diversity sampling. In *Proceedings of the IEEE Information Visualization* (2003), pp. 137–144. 2

[HB09] Harman D., Buckley C.: Overview of the Reliable Information Access Workshop. *Information Retrieval 12*, 6 (2009), 615–641. 2

[Hul93] Hull D. A.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)* (1993), Korfhage R., Rasmussen E., Willett P., (Eds.), ACM Press, NY, USA, pp. 329–338. 2

[HV05] Harman D. K., Voorhees E. M. (Eds.):. *TREC. Experiment and Evaluation in Information Retrieval* (2005), MIT Press, Cambridge (MA), USA. 1

[JK02] Järvelin K., Kekäläinen J.: Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS) 20*, 4 (October 2002), 422–446. 2, 3

[KJPK08] Keskustalo H., Järvelin K., Pirkola A., Kekäläinen J.: Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In *Proc. 31st Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)* (2008), Chua T.-S., Leong M.-K., Oard D. W., Sebastiani F., (Eds.), ACM Press, NY, USA, pp. 675–681. 2

[RZ07] Robertson S. E., Zaragoza H.: On rank-based effectiveness measures and optimization. *Information Retrieval 10*, 3 (July 2007), 321–339. 2

[SS04] Seo J., Shneiderman B.: A rank-by-feature framework for interactive exploration of multidimensional data. In *Proceedings of the IEEE Information Visualization* (2004), pp. 65–72. 2