A Visual Analytics Tool for Experimental Evaluation

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. Information Retrieval is a research field strongly rooted in experimentation. Indeed, measuring is a key to scientific progress. Multilingual and multimedia information access systems, such as search engines, are increasingly complex: they need to satisfy diverse user needs and support challenging tasks. It is therefore fundamental to provide automated tools to examine system behaviour, both visually and analytically. This paper 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.

1 Introduction

Information Retrieval (IR) systems, ranging from World Wide Web search engines [4] to enterprise search [5] or expertise retrieval [2] systems and passing through information access components in wider systems such as digital libraries, are key technologies to get access to relevant information items in a context where information overload is day-to-day experience of every user.

To get rid of such huge amount of information, ever increasing, IR systems are getting more and more complex: they rely on very sophisticated ranking models where many different parameters affect the obtained results and are comprised of several components, which interact together in very complex ways to produce a list of relevant documents in response to a user query. Ranking is a central and ubiquitous issue in this context since it is necessary to return the results retrieved in response to a user query according to the estimation of their relevance to that query [14].

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 of use in order to tune or to improve it as to achieve the level of effectiveness needed to meet the user expectations. Moreover, since an IR system does not produce exact answers, but it ranks results by their estimated relevance to a user query, it is necessary to experimentally evaluate its performances in order to assess the actual quality of the produced rankings.

Experimental evaluation [10] is a very strong and long-lived tradition in the IR and the main paradigm is the Cranfield methodology [6] which makes use of shared experimental collections $\mathcal{C} = (D, T, J)$ in order to create comparable experiments and evaluate the performances of different IR systems. D is a collection of documents; T is a set of topics, which simulate actual user information needs; and, J is the set relevance judgements, i.e. a kind of "correct" answers, where for each topic $t \in T$ the documents $d \in D$, which are relevant for the topic t, are determined. Relevance judgements J can be binary, i.e., relevant or not relevant, or multi-graded, e.g., highly relevant, partially relevant, not relevant and so on [12]. Experimental collections constitute the basis which allow for comparing different IR systems and a whole breadth of metrics has been developed over the years to assess the quality of produced rankings [10] as well as statistical approaches are adopted to assess significant differences in IR system performances and the quality of the evaluation metrics themselves.

Experimental evaluation has always been a central topic in IR but it is becoming a more and more important also in the database field. Indeed, [7] proposed an IR-like methodology to empirically evaluate the performances of relational keyword search techniques and in a recent post[3] in the ACM SIGMOD blog, G. Weikum highlighted the importance of shared collections of data for evaluation also in the database field, which has " to make data and experiments an essential part of the academic currency".

Experimental evaluation and large-scale evaluation campaigns provide the means for assessing the performances of IR systems and represent the starting point for investigating and understanding their behaviour. However, the complicated interactions among the components of an IR system are often hard to trace down, to explain in the light of the obtained results, and to interpret in the perspective of possible modifications to be made to improve the ranking of the results, thus making this activity extremely difficult. This activity is usually called, in the IR field, *failure analysis* and it is deemed a fundamental activity in experimental evaluation even if it is too often overlooked due to its difficulty [1].

The contribution of the paper is the design, implementation, and test of a Visual Analytics (VA) system, called Visual Analytics Tool for Experimental Evaluation (VATE^2), which supports all the phases of the evaluation of an IR system, namely performance and failure analysis, greatly reducing the effort needed to carry them out by providing effective interaction with the experimental data. Moreover, VATE^2 introduces a completely new phase in the experimental evaluation process, called *what-if analysis*, which is aimed at getting an estimate of what could be the effects of a modification to the IR system under examination before needing to actually implement it.

The paper is organized as follows: Section 2 discusses related work. Section 3 describes the analytical models for interaction that have been adopted to conduct failure analysis and what-if analysis. Section 4 explains how the visualization

[3] Gerhard Weikum, "Wheres the Data in the Big Data Wave?", 6 March 2013, `http://wp.sigmod.org/?p=786`

and interaction part works and gives and overview of VATE^2. Finally, Section 5 concludes the paper, pointing out ongoing research activities.

2 Related Work

The graded-relevance metrics considered in this paper are based on cumulative gain [11]; the Discounted Cumulated Gain (DCG) measures are based on the idea that documents are divided in multiple ordered categories, e.g. highly relevant, relevant, fairly relevant, not relevant. DCG measures 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.

A work that exploits DCG to support analysis is [16] 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 [11], but using the gap between curves to support analysis as in [16]. Moreover, the models proposed in this paper provide the basis for the development of VA environment that can provide us with: (i) a quick and intuitive idea of what happened in a ranking list; (ii) an understanding of what are the main reasons of its perceived performances; and, (iii) the possibility of exploring the consequences of modifying the system characteristics through an interactive what-if scenario.

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, or for browsing the ranked results [8]; other proposals, see, e.g., [15], use rankings for presenting the user with the most relevant visualizations, or for browsing the ranked result, see, e.g., [8].

However, 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.

3 The Models Behind VATE^2

3.1 Clustering via Supervised Learning

Ranking models highly depends by the tuning of several parameters which in most of the cases is done manually. This is a difficult task especially when the ranking model has many parameters and it is the result of the combination of several other models. To this purpose machine learning based on supervised learning techniques can help because they are effective tools to automatically tune parameters and combine multiple evidences [13]. Supervised learning methods are feature-based and a widely-used list of features usually adopted by these techniques is described in [9]. The discriminative training is an automatic learning process based on the training data with four pillars: the input space (e.g. the object under investigation, usually represented as feature vectors), the output space (e.g. the learning target w.r.t. the input space), the hypothesis space (e.g. the class of functions mapping the input space into the output space), and a loss function (e.g. a function that measures to what degree the prediction is in accordance with the ground truth).

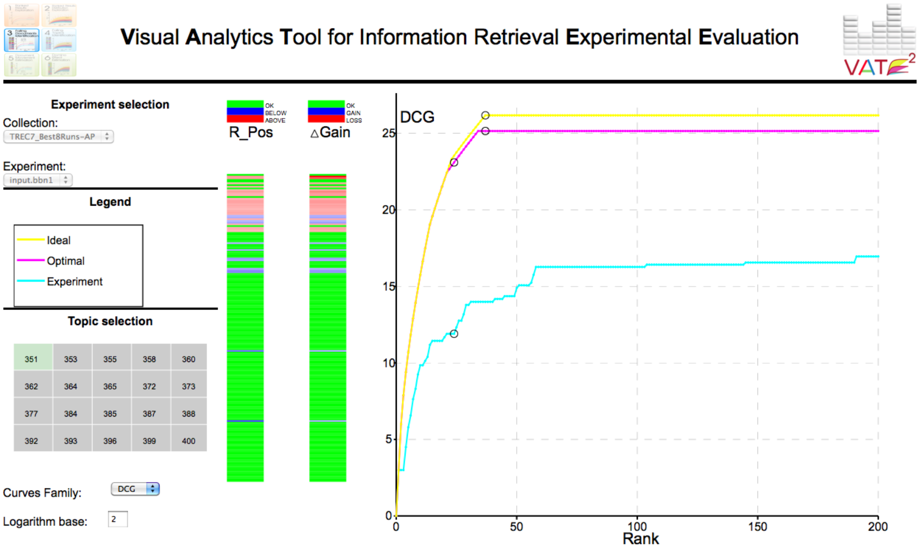
A training set consists of n training queries $q_j (j = 1 \dots n)$, their associated documents represented as feature vectors $\mathbf{x^{(j)}} = \{x_i^{(j)}\}_{i=1}^{m^{(j)}}$ (where $x_i^{(j)}$ is the i^{th} document retrieved for q_j and $m^{(j)}$ is the number of documents retrieved for q_j) and the corresponding relevance values (i.e. $y^{(j)}$). Then a classification algorithm based on regression trees is employed to learn the ranking model corresponding to the way of combining features.

In this work we exploit this framework to learn the ranking model of the IR system under investigation in order to simulate the way in which it ranks the documents. Our aim is to support a "what if" investigation on the ranking list outputted by the system taken into account; the basic idea is to show how the ranking list and the DCG change when we move upward or downward a document in the list. To this purpose, the "cluster hypothesis" saying that "closely associated documents tend to be relevant to the same requests" [17] has to be taken into account; indeed, there can be a correlation in the ranking list between a document and its "closed associated documents". We lever on the hypothesis that if we change the rank of a document also the cluster of documents associated with it will accordingly change their rank.

There are several algorithms for clustering as described in [3]. In this work we focus on the ranking of the considered documents and on how the ranking model can be improved. To this purpose we form the cluster for a target document by grouping together the documents which are similar from the considered ranking model point-of-view. Let us take into account a full result vector FV_j retrieved for a given query q_j, for each document $FV_j[i]$ we create a cluster of documents C_i by:

1. employing a test IR system and submitting $FV_j[i]$ as a query, thus retrieving a result vector FV_i of documents;
2. determining $C_i = FV_j \cap FV_i$;
3. ranking the documents in C_i by employing the learned ranking model.

Therefore, we retrieve a result vector FV_i of relevant documents w.r.t. $FV_j[i]$, then we pick out only those documents which are in the original result vector (say FV_j), and lastly we use the learned ranking model to order these documents accordingly to their "ranking" similarity to $FV_j[i]$. In this way, the higher a document is into the cluster C_i, the more similar it is to the target document $FV_j[i]$. We can see that the similarity measure is based on how the documents



Visual Analytics Tool for Information Retrieval Experimental Evaluation

Experiment selection

Collection: TREC7_Best8Runs-AP

Experiment: input.bbn1

Legend

Topic selection

351	353	355	358	360
362	364	365	372	373
377	384	385	387	388
392	393	396	399	400

Curves Family: DCG

Logarithm base: 2

Fig. 1: A Screen-shot of the failure analysis interface of VATE^2.

are seen by the learned ranking model. It is worthwhile to point out that FV_i usually contains a different set of documents respect to FV_j; we are interested only in the documents belonging to the original rank list (i.e. the documents in FV_j) because we want to specifically evaluate the effect of the tuning of the ranking model and not other aspects related to an IR system as a whole, such as its ability of retrieving relevant documents.

In the end of this process, for each document $FV_j[i]$ obtained by an IR system for a query q_j, we define a cluster of documents C_i ordered by their relevance with respect to $FV_j[i]$.

3.2 Rank Gain/Loss Model.

According to [11] 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. Thus, the higher the index 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 [11]:

$$DF(V[i]) = \begin{cases} GT(V[i]), \;\; if\, i \leq x \\ GT(V[i])/\log_x(i), \;\; if\;\, i > x \end{cases} \tag{3.1}$$

that reduces, in a logarithmic way, the relevance of a document whose index is greater than the logarithm base.

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 $GT(OV[i]) \geq GT(OV[j]), \forall \ \{i,j\} \leq n \bigwedge i < j$, 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.

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

$$R\_Pos(V[i]) = \begin{cases} 0, \ \textit{if} \ min\_index(V, GT(V[i])) \leq i \leq max\_index(V, GT(V[i])) \\ min\_index(V, GT(V[i])) - i, \ \textit{if} \ i < min\_index(V, GT(V[i])) \\ max\_index(V, GT(V[i])) - i, \ \textit{if} \ i > max\_index(V, GT(V[i])) \end{cases} \quad (3.2)$$

$R\_Pos(V[i])$ 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 of these two metrics allows the user to easily locate misplaced elements and understand the impact that such errors have on DCG.

3.3 What-if Analysis Model

The retrieval results are modeled as a ranked vector V containing the first 200 documents of the full result vector FV. The clustering algorithm we described, associates to each document $V[i]$ a cluster C_i of similar documents (we consider only the documents whose relevance with $V[i]$ is greater than a suitable threshold). Moreover, for the sake of notation we define the index cluster set IC_i, i.e., the set of indexes of FV corresponding to elements in C_i: $IC_i = \{j|FV[j] \in C_i\}$. As a consequence, according to the "cluster hypothesis", moving up or down the document $V[i]$ will affect in the same way all the documents in C_i and that might result in rescuing some documents below the 200 threshold pushing down some documents that were above such threshold.

We model the what-if interaction with the system with the operator $Move(i, j)$ whose goal is to move the element in position i in position j. In order to understand the effect on V of such an operation, we have to consider all the C_i elements and the relative position of their indexes, that ranges between $min(IC_i)$ and $max(IC_i)$. Different cases may occur and we analyze them assuming, without loss of generality, that $i < j$, i.e., that the analyst goal is to move up the element $V[i]$ of $j - i$ positions. For the clustering hypothesis that implies that all the C_i elements will move up of $j - i$ positions as well. There are, however, situations in which that is not possible: the maximum upshift is $max(min(IC_i) - 1, j - i)$ and if $j - i > min(IC_i) - 1$ the best we can do is to move up all the C_i elements of just $IC_i - 1$ positions. That corresponds to the situation in which the analyst wants to move up the element in position i of k positions, but there exists a document in C_i whose index is $\leq k$ and, obviously, it is not possible to move it up of k positions. In such a case, the system moves up all the documents in the cluster of $min(IC_i) - 1$ positions, approximating the user intent.

Formally, after applying a $Move(i, j)$ operator, we obtain a permutation FV' of the vector FV. The steps to compute FV' are the following.

```
1. $\triangle = min(min(IC_i) - 1, j - i)$; Initialize $FV'$ to 0;
   $holes = \{k|k \in [min(IC_i), max(IC_i)] \wedge k \notin IC_i\}$
2. $FV'[i] = FV[i],\ if\, i > max(IC_i) \vee i < min(IC_i) - \triangle$;
3. $FV'[i - \triangle] = FV[i],\ if\, i \in IC_i$;
4. Iterate
     – $j = min\{k|FV'[k] = 0\}$ ;
     – $FV'[j] = FV[min(holes)]$;
     – $holes = holes - min(holes)$;

   until $holes = \emptyset$.
```

Step 1 computes the allowed shift, fill FV' of 0s, and computes the set of indexes that corresponds to documents in the range $[min(IC_i), max(IC_i)]$ not belonging to the cluster C_i. Step 2 copies the part of FV that is not affected by the shift and step 3 moves up of $\triangle$ the elements in C_i. Step 4 moves down the documents ousted in step 3.

4 Overview of VATE^2

VATE^2 allows the analyst to perform three main activities: performance analysis, failure analysis and what-if analysis by employing the models described above.

These three main activities can be carried out at the "topic level" or at the "experiment level".

At the topic level VATE^2 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. At the experiment level VATE^2 evaluates the overall quality of the ranking for all the topics of the experiment, focusing on the variability of the results. Basically, at the experiment level VATE^2 shows an aggregate representation based on the boxplot statistical tool showing the variability of the DCG family of metrics calculated on all the topics considered by an experiment. In this way the analyst will have a clearer insight on what to expect from her/his ranking algorithm both in a static way and in a dynamic one (which involves an interactive reordering of the ranked list of documents).

While visually inspecting the ranked list (i.e. failure analysis), 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 (i.e. what-if analysis). This cluster of documents simulates the "domino effect" within the given topic t.

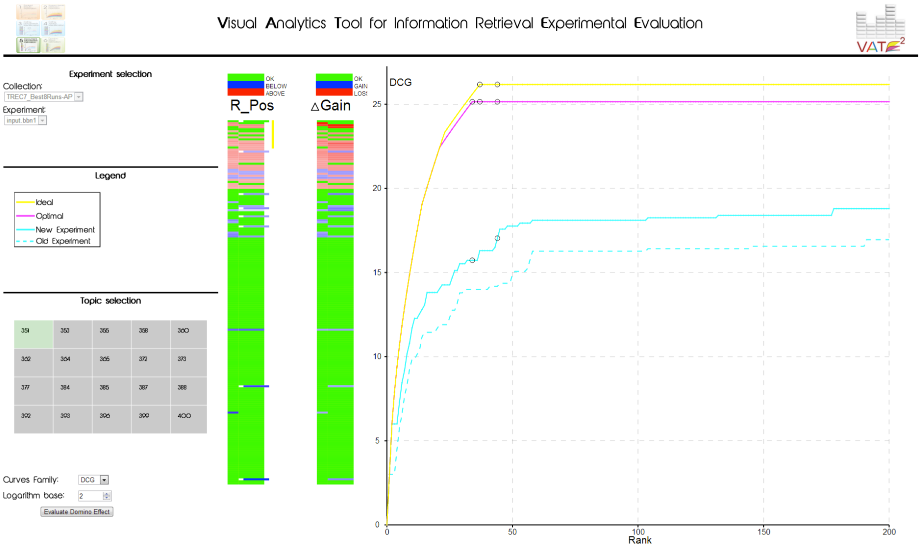
When the analyst is satisfied with the results, i.e. when he has produced a new ranking of the documents that corresponds to the effect that is expected by modifications that are planned for the system, he can feed the Clustering via Supervised Learning model with the newly produced ranked list, obtain 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.

4.1 How to Perform the Failure Analysis

Figure 1 shows the DCG Graph for the topic level analysis. On the left side we can see two vertical bars representing the visualization of the ranking list. The first one represents the *R_Pos* vector. The visualization system computes the optimal ranking list of the documents and assigns to each document a color based on its rank. A green color is assigned to a document at the correct rank w.r.t. the calculated optimal rank; whereas a blue color is assigned to a document ranked below the optimal and a red color is assigned to a document ranked above the optimal. The color intensity gives the user an indication of how far the document is from its optimal rank: a weak intensity means that the document is close to the optimal, a strong intensity means it is far to the optimal. The second vertical bar represents the *Δ_Gain* function values for each document. We adopted the same color code as in the previous vector, but in this case the red color represents a loss and a blue color represents a gain in terms of *Δ_Gain*.

On the right side of Figure 1 we can see a graph showing three curves:

Experiment Ranking refers to the top n ranked results provided by the system under investigation;



Collection: TREC7_Best8Runs-AP

Experiment: input.bbn1

351	353	355	358	360
362	364	365	372	373
377	384	385	387	388
392	393	396	399	400

Curves Family: DCG

Logarithm base: 2

Fig. 2: A Screen-shot of the topic level what-if analysis interface of VATE^2.

Optimal Ranking refers to an optimal re-ranking of the experiment;
Ideal Ranking refers to the ideal ranking of the top n documents in the pool.

The visualization system is built in such a way that if a user selects a document in the *R_Pos* vector, also the DCG loss/gain in the *Δ_Gain* vector and all its contributions to the different curves (i.e. Experiment, Optimal and Ideal) will be highlighted.

The visualization described so far is well-suited to cope with a static analysis of the ranked result: the user can understand if there is the need to re-rank the documents or to perform a re-querying to retrieve a different set of documents with the aim of obtaining a better value of the DCG metric.

4.2 How to Perform the What-if Analysis

The what-if functionality allows the users to interact with the ranked vector of *R_Pos*. The system allows the user to shift a target document t from its actual position to a new one in a "drag&drop" fashion, with the goal of investigating the effect of this movement in the ranking algorithm by inspecting the DCG of the modified ranking list. Clearly, a change in the ranking algorithm will affect not only the target document t, but also all the documents in its cluster.

In Figure 2 it is possible to see the animated phase of interactive re-ranking of the documents at the topic level: after highlighting and moving the target document t from the starting position to a new one, the user will be presented with an animated re-ranking of the documents connected to the target one. Once the new position of the target document has been selected, the system moves it to the new position and the documents in its associated cluster are moved together into their new positions. This leads to the redrawing of the *R_Pos*, *Δ_Gain* and DCG graphs according to the new values assigned to each document involved in the ranking process.

It is possible to see that when a user select a document in the leftest bar, all the documents in its cluster are highlighted in yellow helping the user to understand which documents are involved in a potential movement.

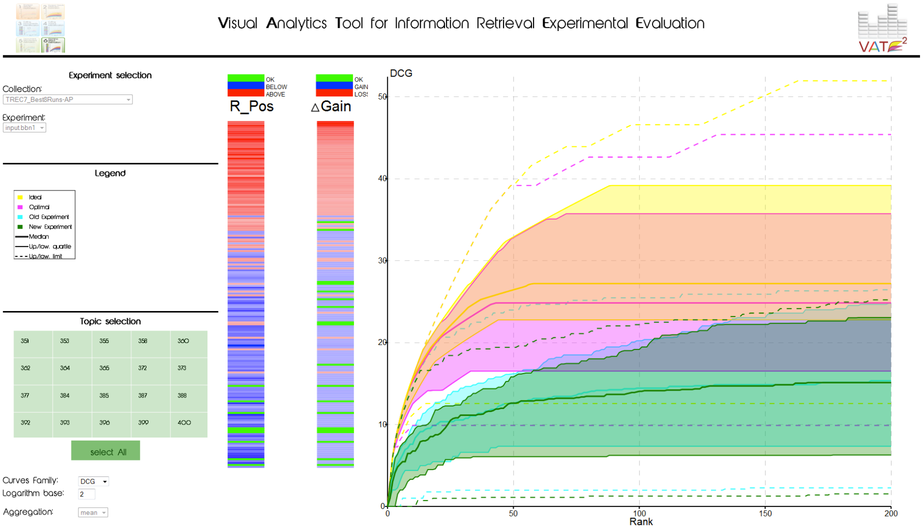
Figure 2 shows also the result of the what-if process: the image presents two new curves, representing the new values assigned for both the experiment curve (purple one) and the optimal curve (orange one). To evaluate the changes in the DCG function, the image shows, in a dash-stroke fashion, the old curve trends. Thanks to this visualization, the user can appreciate the gain or the loss obtained from this particular re-rank. In the case shown in Figure 2 the movements performed by the user improved the performances at the topic level; indeed, the dashed line – i.e. the old experiment curve – is lower than the solid one – i.e. the new experiment curve. This means that we are simulating a change in the system that does improve it.

On top of that, at the experiment level, the change in the ordering of a particular ranking list will result in changing also the other ranking lists within the same experiment: these changes can be intercepted by this graph in terms of variability of the curves and on the raising/declining of the "box" region of the boxplots (showed as filled area in the graph).

To maintain the graph as clear as possible, the choice of not representing the single boxplots, but simply the continuous lines joining the similar points has been taken. So, in the graph area there are five different curves which are: upper limit, upper quartile, median, lower quartile, and lower limit. All these curves are determined for the ideal, the optimal and the experiment cases. For each case, the area between lower and upper quartile is color filled in order to highlight the central area (the box of the boxplot) of the analysis.

In figure 3 we can appreciate that, in this particular case, the optimal and experiment areas do not overlap very much, and the median curve of the experiments is quite far from the one of the optimal. This can be asserted from an aggregate point of view, and not by a specific topic analysis like the one we proposed with the DCG graph. Different considerations can also be made on variability: in this case, while experiment and optimal box areas are quite broad, demonstrating a heterogeneity in values, and also the ideals box area is big meaning a high variability of the data among the different topics.

The domino effect due to the what-if analysis is highlighted by the experiment areas: the old one (before the what-if analysis) is shaded in blue, whereas the new one (after the what-if analysis) is shaded in green. We can see that a change in one topic at the topic level worsens the global performances; indeed, the blu area is better than the green one. This means that the change the user did at the topic level (which improved the local performances) reflects at the experiment (global) level worsening the overall performances of the system.



Experiment selection

Collection: TREC7_Best8Runs-AP

Experiment: input.bbn1

Topic selection

351	353	355	358	360
362	364	365	372	373
377	384	385	387	388
392	393	396	399	400

select All

Curves Family: DCG

Logarithm base: 2

Aggregation: mean

Fig. 3: A Screen-shot of the experiment level what-if analysis interface of VATE^2.

5 Conclusion and Future Work

This paper presented a fully-fledged analytical and visualization model to support interactive exploration of IR experimental results with a two-fold aim: (i) to ease and support deep failure analysis in order to better understand system behaviour; (ii) to conduct a what-if analysis to have an estimate of the impact that possible modifications to the system, identified in the previous step and aimed at improving the performances, can have before needing to actually re-implement the system. Thus, 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.

Future work will concern two main issues: (i) while the informal results about the system usage are quite encouraging we plan to run a more structured user study, involving people that have not participated in the system design; and (ii) we want to improve the way in which the clusters produced by the The Clustering via Supervised Learning methods are used to compute the new ranking and the associated DCG functions.

References

1. M. Angelini, N. Ferro, G. Santucci, and G. Silvello. Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In *Proc. of the*

4th Information Interaction in Context Symposium, IIIX '12, pages 194–203, New York, NY, USA, 2012. ACM.

2. K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si. Expertise Retrieval. *Foundations and Trends in Information Retrieval (FnTIR)*, 6(2-3):127–256, 2012.
3. P. Berkhin. A Survey of Clustering Data Mining Techniques. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data*, pages 25–71. Springer-Verlag, Heidelberg, Germany, 2006.
4. S. Buettcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines.* The MIT Press, Cambridge (MA), USA, 2010.
5. S. Burnett, S. Clarke, M. Davis, R. Edwards, and A. Kellett. *Enterprise Search and Retrieval. Unlocking the Organisation's Potential.* Butler Direct Limited, 2006.
6. C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. In K. Spärck Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 47–60. Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997.
7. J. Coffman and A. Weaver. An Empirical Performance Evaluation of Relational Keyword Search Techniques. *IEEE Transactions on Knowledge and Data Engineering*, 99:1–16, 2012.
8. M. Derthick, Michael G. Christel, Alexander G. Hauptmann, and Howard D. Wactlar. Constant density displays using diversity sampling. In *Proceedings of the IEEE Information Visualization*, pages 137–144, 2003.
9. X. Geng, T.-Y. Liu, T. Qin, and H. Li. Feature Selection for Ranking. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, pages 407–414. ACM Press, New York, USA, 2007.
10. D. Harman. *Information Retrieval Evaluation.* Morgan & Claypool Publishers, USA, 2011.
11. K. Järvelin and J. Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information System (TOIS)*, 20(4):422–446, 2002.
12. J. Kekäläinen and K. Järvelin. Using Graded Relevance Assessments in IR Evaluation. *Journal of the American Society for Information Science and Technology (JASIST)*, 53(13):1120—1129, November 2002.
13. T.-Y. Liu. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
14. S. Mizzaro. Relevance: The Whole History. *Journal of the American Society for Information Science and Technology (JASIST)*, 48(9):810–832, September 1997.
15. J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. In *Proceedings of the IEEE Information Visualization*, pages 65–72, 2004.
16. J. Teevan, S. T. Dumais, and E. Horvitz. Potential for Personalization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(1):1–31, 2010.
17. C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, England, 2nd edition, 1979.