VIRTUE: A Visual Tool for Information Retrieval Performance Evaluation and Failure Analysis

Marco Angelini[a], Nicola Ferro[b], Giuseppe Santucci[a], Gianmaria Silvello[c]

[a]*Sapienza University of Rome, Italy*
[b]*University of Padua, Italy*
[c]*Corresponding author.*
Affiliation: University of Padua, Italy.
Address: Via Gradenigo, 6/B, 35131, Padova, Italy.
Tel: +39 049 827 79 29
e-mail: silvello@dei.unipd.it

Abstract

Objective: Information Retrieval (IR) is strongly rooted in experimentation where new and better ways to measure and interpret the behavior of a system are key to scientific advancement. This paper presents an innovative visualization environment: Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE), which eases and makes more effective the experimental evaluation process.

Methods: VIRTUE supports and improves *performance analysis* and *failure analysis*.

Performance analysis: VIRTUE offers interactive visualizations based on well-know IR metrics allowing us to explore system performances and to easily grasp the main problems of the system.

Failure analysis: VIRTUE develops visual features and interaction, allowing researchers and developers to easily spot critical regions of a ranking and grasp possible causes of a failure.

Results: VIRTUE was validated through a user study involving IR experts. The study reports on a) the scientific relevance and innovation and b) the comprehensibility and efficacy of the visualizations.

Conclusion: VIRTUE eases the interaction with experimental results, supports users in the evaluation process and reduces the user effort.

Practice: VIRTUE will be used by IR analysts to analyze and understand experimental results.

Implications: VIRTUE improves the state-of-the-art in the evaluation practice and integrates Visualization and IR research fields in an innovative way.

Keywords: information retrieval, experimental evaluation, visual analytics, performance analysis, failure analysis

1. Introduction

IR systems, which include World Wide Web search engines [19] as well as enterprise search [13], intellectual property and patent search [46] and expertise retrieval systems [7], as well as information access components in wider systems such as digital libraries [14, 28, 71], are key tech-

nologies for gaining access to relevant information items in a context where information overload is a day-to-day experience of every user.

In order to deal such a huge amount of ever increasing information, IR systems are becoming more and more complex: they rely on very sophisticated ranking models where many different parameters affect the results obtained and are comprised of several components, which interact together in 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 and the user information need [49].

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

Experimental evaluation [51, 52] is a strong and long-lived tradition in IR, which highly contributes to the advancements in the field [32]. Nevertheless, it is a very demanding activity, in terms of both time and effort needed to perform it, and it is usually carried out in publicly open and large-scale evaluation campaigns at international level. This allows for sharing the effort, producing large experimental collection, and comparing state-of-the-art systems and algorithms. Relevant and long-lived examples are the Text REtrieval Conference (TREC)[1] in the United States [33], the Conference and Labs of Evaluation Forum (CLEF)[2] initiative in Europe (formerly Cross-Language Evaluation Forum) [16], and the NII Testbeds and Community for Information access Research (NTCIR)[3] in Japan and Asia [39].

During their life-span, large-scale evaluation campaigns have produced huge amounts of scientific data which are extremely valuable. These experimental data provide the foundations for all the subsequent scientific production and system development and constitute an essential reference for all the produced literature in the field. Moreover, these data are valuable also from an economic point of view, due to the great amount of effort devoted to their production: [54] estimates that the overall investment – by NIST and its partners – in TREC in its first 20 years was about 30 million dollars which, as discussed above, produced an estimated return on investment between 90 and 150 million dollars.

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 complex 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. Conducting such analyses is especially resource demanding in terms of time and human effort, since they require, for several queries, the manual inspection of system logs, intermediate outputs of system components, and, mostly, long lists of retrieved documents which need to be read one by one in order 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

[1] http://trec.nist.gov/

[2] http://www.clef-initiative.eu/

[3] http://research.nii.ac.jp/ntcir/

IR field, *failure analysis* [10, 31, 60] and it is deemed a fundamental activity in experimental evaluation and system development even if it is too often overlooked due to its difficulty.

To give the reader an idea of how demanding failure analysis can be, let us consider the case of the the Reliable Information Access (RIA) workshop [30], which was aimed at systematically investigating the behaviour of just one component in a IR system, namely the relevance feedback module [53]. Harman and Buckley in [30] reported that, to analyze 8 systems, 28 people from 12 organizations worked for 6 weeks requiring from 11 to 40 person-hours per topic for 150 overall topics.

This papers aims to reduce the effort needed to carry out both the performance and failure analyses, which are fundamental steps in experimental evaluation, by introducing the possibility of effectively interacting with the experimental results.

The main contribution of the paper is the design, development, and initial validation of an innovative visual analytics environment, called Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE), which integrates and supports the two phases discussed above:

(i) it eases *performance analysis*, which is one of the most consolidated activities in IR evaluation, although it is often the only one performed. This is achieved by interactive visualization and exploration of the experimental results, according to different metrics and parameters, and by providing simple visual means to immediately grasp whether the system would already have the potential to achieve the best performances or whether a complete new ranking strategy would be preferred;

(ii) it explicitly assists *failure analysis*, which is usually overlooked due its laborious nature, and makes failure analysis part of a single and coherent workflow. In particular, it introduces two new indicators, called Relative Position (RP) and Delta Gain (ΔG), which allow us to visually (and also numerically) figure out the weak and strong parts of a ranking in order to quickly detect failing documents or topics and make hypotheses about how to improve them. This greatly reduces the effort needed to carry out this fundamental but extremely demanding activity and promises to make it a much more widespread practice.

The environment has been has been designed using a User-Centered Design (UCD) methodology (see, e.g., [70]) dealing with four IR experts, exploring different visualizations and interaction mechanisms.

Moreover, an original contribution of the paper is to model the phases described above in a single formal analytical framework, not yet present in the literature to the best of our knowledge, where all the different concepts and operations find a methodologically sound formulation and fit all together to contribute to the overall objective of taking a step forward in experimental evaluation. The overall idea of exploiting visual and interactive techniques for exploring the experimental results is quite new to the IR field, since representation and analysis of the experimental results typically happens in static ways or batches. This is confirmed by the lack in the literature of similar proposals and by the experts' opinion that point out the novelty of this approach, as reported in Section 6. Moroever, such an approach is also new to the Visual Analytics (VA) field, since VA techniques are usually applied to the presentation and interaction with the outputs, i.e., the ranked result list and documents [1, 73], produced by an IR system but almost never to the analysis, exploration, and interpretation of the performances and behavior of the IR system itself.

A final contribution of the paper is to have performed an initial validation of the VIRTUE environment with domain experts in order to provide feedback about its innovation potential, its suitability for the purpose and the appropriateness of the proposed solutions.

The paper is organized as follows: Section 2 discusses some related works; Section 3 introduces the conceptual framework proposed in the paper to support and enhance the experimental evaluation methodology and practice by exploiting visual analytics techniques; Section 4 explains the proposed formal analytical framework; Section 5 presents the actual prototype which implements the proposed methodologies; Section 6 discusses the validation of the adopted methodologies and prototype with domain experts; finally, Section 7 draws some conclusions and presents an outlook for future work.

2. Related Works

Visualization in IR is mainly comprised by two components [73]: visual information presentation and visual information retrieval. The purpose of these components is to increase the ability to fulfill IR tasks where visualization is the natural platform for browsing and query searching. There are several works in this area mainly focusing on the identification of the objects and their attributes to be displayed [27], different ways of presenting the data [50], the definition of visual spaces and visual semantic frameworks [72]. The development of interactive means for IR is an active field which focuses on search user interfaces [34], results displaying and browsing capabilities [18, 45]. These approaches do not consider visual tools for dealing with experimental evaluation data and for conducting performance and failure analysis in an interactive way.

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 [61], for browsing the ranked results [21], or for comparing large sets of rankings [9], but they do not deal with the problem of observing the ranked item position, or comparing it with an optimal solution, or assessing and improving the ranking quality. Such a novel idea was initially explored by the authors in [26], where they propose a first formalization of the notion of Relative Position (RP) and Delta Gain (ΔG) and provided a simple visualization for the analysis of a single topic. This approach was extended in [6] to evaluate all the topics of an experiment, allowing the performance of a IR system as a whole to be assessed.

In this work we exploited the (discounted) cumulative gain metrics for performance and failure analysis. In a related work [65], Teevan et al. exploited Discounted Cumulated Gain (DCG) to analyze the curves to derive the potential for personalization. The potential for personalization is the gap between the optimal ranking for an individual and the optimal ranking for a group. The curves plot 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 [37], but using the gap between curves to support analysis as in [65]. Moreover, the framework proposed in this paper provides a VA environment that provides us with a quick and intuitive idea of what happened in a ranked result list, an understanding of what the mains reasons of its performances are by means of novel metrics (RP and ΔG), and comparative analyses between single curves and aggregate curves.

Visualization strategies have been adopted for analyzing experimental runs, e.g. beadplots in [8]. Each row in a beadplot corresponds to a system and each "bead", which can be gray or colored, 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 coloring (from dark red for most relevant to light violet for least relevant) is based on a

reference system, not on graded judgments and the optimal ranking as in our work. In [8] 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. Moreover, several strategies for visualizing runs, metrics and descriptive statistics relative to IR experimental evaluation data have been designed and developed in the context of the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) [3, 4, 25, 29] which is a comprehensive tool for managing all the aspects of the IR evaluation methodology and the experimental data produced. In this context, the focus is on performance analysis, whereas failure analysis is not considered; furthermore, DCG and related metrics for ranking evaluation are not yet considered in the visualization part of DIRECT [2].

Another related work is the Query Performance Analyzer (QPA) [64]. 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.

This paper extends these 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 TREC-7 Ad-hoc track [69] and the pool obtained in [63]. The system works with any available test collection, but it has been evaluated using the data from TREC-7 Ad-hoc track mostly because this collection is well-known and widely adopted by the IR community. Indeed, it has been used in the original formulation of cumulative-gain metrics [37] and in several relevant studies such as for analyzing the effects of ranking of IR systems [41], for evaluating general ranking functions [24], and for predicting query performances [20]. Furthermore, the issues of the systems tested within TREC-7 Ad-hoc track are already known and analyzed in literature [69]. The adoption of this collection for testing VIRTUE had positive effects also for the user validation where the domain experts focused mainly on the visual and interactive functionalities of the system while analyzing system rankings they were confident with.

3. Conceptual Framework

Experimental evaluation in the IR field dates back to late 1950s/early 1960 and it is based on the Cranfield methodology [17] which makes use of shared experimental collections in order to create comparable experiments and evaluate the performances of different IR systems. An experimental collection can be expressed as a triple $\mathscr{C} = (D, T, GT)$, where D is a set of documents, also called collection of documents, which is representative of the domain of interest both in terms of kinds of documents and number of documents; for example, in the case of patent or prior art search you need to use actual patents as provided by the European or US Patent Offices. T is a set of topics, which simulate actual user information needs and are often prepared from real system logs; the topics are then used by IR systems to produce the actual queries to be answered. GT is the ground-truth or the set relevance judgements, i.e. a kind of "correct" answer, where for each topic $t \in T$ the documents $d \in D$, which are relevant for the topic t, are determined. The relevance judgements can be binary, i.e., relevant or not relevant, or multi-graded, e.g., highly relevant, partially relevant, not relevant and so on [42, 62]. 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 [12, 32, 40], according to different user models and tasks. Moreover, statistical approaches are adopted to assess significant differences in IR system performances [35, 59] and the quality of the evaluation metrics and experimental collection themselves [11, 55, 58].

It can be noted that in this paradigm IR systems are dealt with as a kind of "black box", whose internal and intermediate results cannot be examined separately, as also pointed out by Robertson [51, p. 12]: "if we want to decide between alternative indexing strategies for example, we must use these strategies *as part of a complete information retrieval system*, and *examine its overall performance* (with each of the alternatives) directly". As we will discuss in the following, these features of the experimental evaluation process have been explicitly taken into account in modeling, formalizing, designing, and developing VIRTUE.

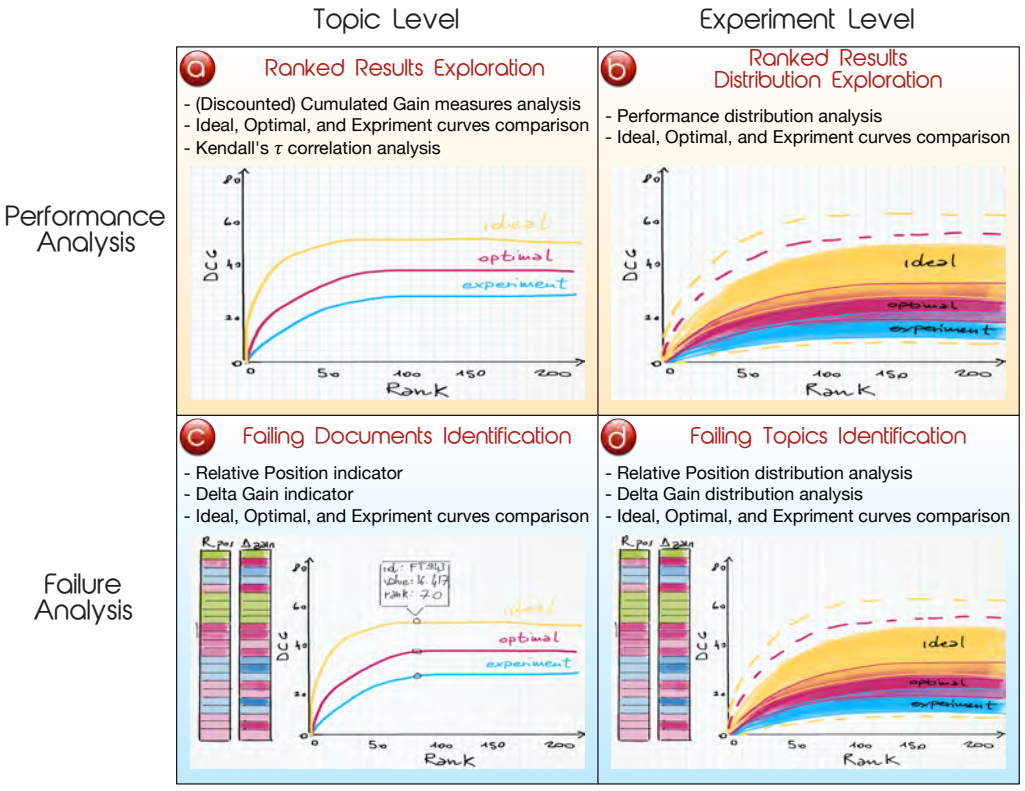

Figure 1: VIRTUE overall framework.

Figure 1 shows the overall framework adopted by VIRTUE to support experimental evaluation. As discussed in Section 1, *performance analysis* and *failure analysis* are the traditional phases carried out during experimental evaluation, where VIRTUE contributes to make them more effective and to reduce the needed effort via both tailored visualizations and measures and high interaction with the experimental data. *Topic Level* concerns the analysis of the documents retrieved in response to a given topic of a run while *Experiment Level* deals with overall statis-

tics and effects concerning the whole set of topics of a run, i.e., all the different ranked lists of retrieved documents. In both cases, the user is presented with three curves, describing a) the actual performance (experiment curve), b) the improvement that is possible to achieve reordering the actual result in the optimal way (optimal curve), and c) the best possible score, in which the results contains *all* the relevant documents in the optimal way (ideal curve). Details about such curves are in Section 4 while the system functionalities and the actual implementation of the four possible analysis are described in Section 5

Therefore, VIRTUE:

- supports performance analysis on a topic-by-topic basis and with aggregate statistics over the whole set of topics;
- facilitates failure analysis to allow researchers and developers to more easily spot and understand failing documents and topics.

The main target users of VIRTUE are domain experts, i.e. researchers and developers in the IR and related fields who need to understand and improve their systems. Moreover, VIRTUE can be also useful for educational purposes, e.g. in undergraduate or PhD courses where information retrieval is taught and where explaining how to interpret the performances of an IR system is an important part of the teaching. Finally, it may find application also in production contexts as a tool for monitoring and interpreting the performances of a running system so as to ensure that the desired service levels are met.

In the following sections, we describe each of these steps of Figure 1 in more detail from top left to bottom right.

3.1. Ranked Results Exploration

In order to quantify the performances of an IR system, we adopt the (discounted) cumulated gain family of measures [38, 44] which have proved to be especially well-suited for analyzing ranked results lists because they allow for graded relevance judgments and embed a model of the user behavior while s/he scrolls down the results list which also gives an account of her/his overall satisfaction. This family of metrics is composed of the Cumulated Gain (CG), the DCG and their normalized versions; as we detail in Section 4, DCG is the standard de-facto for ranking evaluation and, without loss of generality, in the following description we mainly refer to DCG knowing that the same considerations are valid for all the other metrics in the family.

The overall idea of the (discounted) cumulated gain family of measures is to assign a gain to each relevance grade and, for each position in the ranked list, a discount is computed. Then, for each rank, DCG is computed by using the cumulative sum of the discounted gains up to that rank position. 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 – e.g. 0 for non-relevant documents, 1 for partially relevant ones, 2 for fairly 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 patient users in scanning the results list. DCG curves have a typical monotonic non-decreasing behavior: the higher the value of DCG at a given rank position the better the performances and the steeper the slope the better the ranking.

We compare the result list produced by an experiment with respect to an *ideal* ranking created starting from the relevant documents in the ground-truth, which represents the best possible results that an experiment can return – this ideal ranking is what is usually used to normalize the

DCG measures. In addition to what is typically done, we compare the results list with respect to an *optimal* one created with the same documents retrieved by the IR system but with a optimal ranking, i.e. a permutation of the results retrieved by the experiment aimed at maximizing its performances by sorting the retrieved documents in decreasing order of relevance. Therefore, the *ideal ranking* compares the experiment at hand with respect to the best results possible, i.e. considering also relevant documents not retrieved by the system, while the *optimal ranking* compares an experiment with respect to what could have been done better with the same retrieved documents.

The proposed visualization, shown in Figure 1 (a), allows for interaction with these curves, e.g. by dynamically choosing different measures in the DCG family, adjusting the discounting function, and comparing curves and their values rank by rank.

Overall, this method makes it easy to grasp the distance of an IR system from both its own optimal performances and the best performances possible and to get an indication about whether the system is going in the right direction or whether a completely different approach is preferable. Indeed, we support researchers and developers in trying to answer an ambitious question: is it better to invest on improving the ranking of the documents already retrieved by the system or is it better to develop a completely new strategy for searching documents? Or, 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 a new strategy. The former case is when the experiment curve is somewhat removed from the optimal curve but the optimal curve is close to the ideal one; the latter case is when the optimal curve is removed from the ideal one, regardless of how close the experiment curve to the optimal one is.

In order to support the visual intuition, we also provide a Kendall's τ correlation analysis [43, 68] between the three above mentioned curves: each experiment is described by a pair $(\tau_{ideal-opt}, \tau_{opt-exp})$, where $\tau_{ideal-opt}$ denotes the Kendall τ correlation among the ideal and the optimal rankings, while $\tau_{opt-exp}$ denotes the Kendall τ among the optimal and experiment rankings. When the pair is $(1,1)$ the best performance possible is achieved. 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. 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 also an optimal re-ranking of the retrieved document is far from the ideal ranking.

The initial idea of comparing not only the ideal ranking but also the optimal one and of supporting this via Kendall's τ correlation analysis was first proposed in [23]. In this paper, we start from that work and improve it by making it part of an overall workflow and a coherent formal analytical framework.

3.2. Ranked Results Distribution Exploration

The interactive visualization and performance analysis methodology described in the previous section concerns a single topic of an experiment. What is usually needed is to be able to analyze a run as a whole or to analyze a subset of its topics together because, for example, they are considered the hard ones where more problems occurred.

The ranked results distribution exploration, shown in Figure 1 (b), provides an aggregate representation based on the box-plot statistical tool [48, 66, 67] showing the variability of the

three DCG curves calculated either on all the topics considered by an experiment or on those selected by the user. In order to keep the visualization as clear as possible, instead of representing single box-plots concerning the distribution of the performances across different topics for each rank position, a line joining the corresponding points of the various box-plots at different rank positions is used.

Therefore, in the visualization, there are five different curves: 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 of the analysis – what is typically represented with a box in a box-plot. Following this rationale the median lines are thicker in order to be different to the upper/lower quartile ones represented with normal thickness and to upper/lower limit ones represented with dashed lines. Moreover, the visualization allows user to interactively choose the topics to whose performances have to be aggregated in order to support the exploration of alternative retrieval scenarios.

For example, this kind of visualization supports users in understanding whether the optimal and experiment areas overlap to a good extent and the median curve of the experiments tends to the one of the optimal, indicating that the overall performances of a run are close to the best that can be done with that set of retrieved documents. Understanding whether this result is good enough or not is then a matter of understanding how these areas overlap with the area of the ideal curves.

This visualization was first proposed in [6] as a means to offer users an overall view of the systems performances. In this paper, we improve it by framing it in the context of a whole analysis workflow and adding to it further interaction by allowing users to dynamically select different subsets of topics to be explored.

3.3. Failing Documents Identification

As discussed in Section 1, failure analysis is a fundamental but demanding activity. Moreover, when looking at a performance curve, like DCG curve, it is not always easy to spot the critical regions in a ranking. For example, as explained in Section 3.1, DCG is a not-decreasing monotonic function which increases only when you find a relevant document in the ranking. However, when DCG does not increase, this could be due to two different reasons: either you are in an area of the ranking where you are expected to put relevant documents but you are putting a non-relevant one and thus you do not gain anything; or, you are in an area of the ranking where you are not expected to put relevant documents and, correctly, you are putting a non-relevant one, still gaining nothing. So, basically, when DCG stays constant, it is not immediately understandable whether this is due to a failure of the system which is not retrieving relevant documents while it would still be expected to do so, or whether the system is performing properly since there would be nothing to gain at that rank position.

In order to overcome this and similar issues, we introduce two indicators, Relative Position (RP) and Delta Gain (ΔG), which allow us to quantify and explain what happens at each rank position and are paired with a visual counterpart which eases the exploration of the performances across the ranking, so we can immediately grasp the most critical areas.

RP quantifies the effect of misplacing relevant documents with respect to the ideal case, i.e. it accounts for how far a document is from its ideal position. Indeed, the ideal case represents an ordering of the documents in the ground-truth in decreasing degree of relevance whereby, for example, all the highly relevant documents are ranked first, followed by the partially relevant

ones, and then the non-relevant ones, thus creating contiguous intervals of documents with the same degree of relevance. In RP, zero values denote documents which are within their ideal interval; positive values denote documents which are ranked below their ideal interval, i.e. documents of higher relevance degree that are in a position of the ranking where less relevant ones are expected; and, negative values denote documents which are above their ideal interval, i.e. less relevant documents that are in a position of the ranking where documents of higher relevance degree are expected. Overall, the greater the absolute value of RP is, the bigger the distance of the document from its ideal interval.

RP eases the interpretation of the DCG curve. For example, if DCG is constant and RP is negative, this implies that there is a failure of the system which is not retrieving relevant documents while it is still expected to do so. Similarly, if DCG is constant and RP is zero, this implies that the system is performing properly since there is nothing to gain at that rank position.

ΔG quantifies the effect of misplacing relevant documents with respect to the ideal case in terms of the impact of the misplacement on the gain at each rank position. In ΔG zero values indicate document which are within their ideal interval and are gaining what is expected from them; negative values denote documents that are ranked above their ideal interval and are causing a local loss in the gain with respect to what could have been achieved; positive values indicate document that are ranked below their ideal interval and are causing a local profit in the gain. ΔG supports the interpretation of DCG curves in a similar way to RP but provides the additional information about how much gain/loss happened at each rank position with respect to the ideal case.

These two indicators are paired with a visual counterpart that makes it even easier to quickly spot and inspect critical areas of the ranking. Two bars are added on the left of the visualization, as shown in Figure 1 (c): one for the RP indicator and the other for the ΔG indicator. These two bars represent the ranked list of results with a box for each rank position and, by using appropriate color coding to distinguish between zero, positive and negative values and shading to represent the intensity, i.e. the absolute value of each indicator, each box represents the values of either RP or ΔG.

For example, in this way, by looking at the bars and their colors the user can immediately identify non-relevant documents which have been ranked in the positions of relevant ones. Then, the visualization allows them to inspect those documents and compare them with the topic at hand in order to make a hypothesis about the causes of a failure. This greatly reduces the effort needed to carry out failure analysis because: (i) users are not requested to interpret the not always intuitive DCG curve to identify potential problems; (ii) users can grasp the critical areas of the ranking by means of color coding and shading and focus on them, instead of scrolling through almost each rank position to identify potential problems; (iii) once a critical area has been identified, the visualization makes it possible to interactively inspect the failing documents and to readily make guesses about the causes of the failure.

The RP and ΔG indicators first proposed in [26] together with the idea of exploiting them for creating a visual tool to explore the performances of an IR system. Here they are fully formalized in the context of the proposed analytical framework, they are made part of a complete workflow and not used in isolation, and the visualization backing them is improved in terms of interaction with the user and the possibility of exploring the retrieved documents.

This visualization based on RP and ΔG was also exploited in [22] to develop a tablet-based version of it with the purpose of exploring the following scenarios where having interaction and visualization via a tablet can be an added value: (i) a researcher or a developer is attending the workshop of one of the large-scale evaluation campaigns and s/he wants to explore and under-

stand the experimental results as s/he is listening to the presentation discussing them; (ii) a team of researchers or developers is working on tuning and improving an IR system and they need tools and applications that allow them to investigate and discuss the performances of the system under examination in a handy and effective way. This work is not reported here since it is out of the scope of the present paper.

3.4. Failing Topics Identification

The techniques described in the previous section support and ease failure analysis at the topic level and allow users to identify and guess possible causes for wrongly ranked documents. However, an overall picture for a whole run is often needed in order to understand if the critical areas of the ranking identified in the previous step are an isolated case concerning just a given topic or they are common to more topics or even a whole run and thus they have a greater impact.

The visualization of Figure 1 (d) merges the approaches of the visualizations presented in Figure 1 (b) and Figure 1 (c): it allows users to assess the distribution of the performances of the ideal, optimal, and experiment curves over a set of selected topics or the whole run and it adds the bars reporting the RP and ΔG indicators to ease the interpretation of the performance distribution.

In particular, this visualization offers users different strategies according to which RP and ΔG values of the experiment are aggregated for a given rank position over the selected set of topics: for example, the user can choose to compute the average, the median, a quartile, and so on of the RP and ΔG values. In this way, users can not only interactively explore different features of the performance distribution but they can also align the way in which the RP and ΔG values are aggregated to the specific area of the performance distribution they are focusing on. Suppose, for example, that the user is exploring the lower quartile of the performance distribution because his goal is to ensure a minimum level of performances across the topics instead of having some performing very high and some very low. In this case it is preferable to aggregate RP and ΔG values by their lower quartile in order to have a kind of "magnification" of the behavior of the corresponding areas highlighted in the DCG curves.

4. Formal Analytical Framework

4.1. Preliminary Concepts

We formalize the basic notions regarding experimental evaluation in IR by starting from the concepts of relevance and degree (or grade) of relevance of a document with respect to a topic. Then, leveraging on these two concepts we define the basic concepts of ground truth, recall base, and relevance score.

Definition 1. *Let REL be a finite set of* ***relevance degrees*** *and let $\preceq$ be a* total order relation *on REL so that*

$$(REL, \preceq)$$

is a totally ordered set.

We call ***non-relevant*** *the relevance degree* $\mathtt{nr} \in REL$ *such that*

$$\mathtt{nr} = \min(REL)$$

Being a finite totally ordered set, the set of relevance degrees admits the existence of a minimum and a maximum.

Consider the following example reporting a typical IR experimental evaluation setting: the set $REL = \{\texttt{nr}, \texttt{pr}, \texttt{fr}, \texttt{hr}\}$ contains four relevance degrees where `nr` stands for "non relevant", `pr` for "partially relevant", `fr` for "fairly relevant" and `hr` stands for "highly relevant". Then, the total order defined above leads to the following ordering $\texttt{nr} \preceq \texttt{pr} \preceq \texttt{fr} \preceq \texttt{hr}$ as one would expect for the relevance degrees introduced above. In this example `nr` is the minimum and `hr` is the maximum of the *REL* set.

The set of relevance is the starting point for defining the central concept of ground truth; the ground truth associates to a document a relevance degree in the context of a given topic, where a document is the basic information unit considered in experimental evaluation and a topic is a materialization of a user information need. For instance, the ground truth says that document $d_j \in D$ is "highly relevant" for topic $t_i \in T$. We define this concept as a function which associates a relevance degree *rel*, i.e. a relevance judgment, to each document d for each topic t.

Definition 2. *Let D bet a finite set of* documents *and T a finite set of* topics*. The* ***ground truth*** *is a function*

$$\begin{array}{rccc} \mathrm{GT}: & T \times D & \rightarrow & REL \\ & (t,d) & \mapsto & rel \end{array}$$

The definition of ground truth completes the set of concepts needed for defining an experimental collection $\mathscr{C} = \{D, T, \mathrm{GT}\}$. From the ground truth definition we can derive the recall base, which is the total number of relevant documents for a given topic t, where a relevant document is meant by any document with relevance degree above non-relevant (i.e. the minimum of set *REL*).

Definition 3. *The* ***recall base*** *is a function*

$$\begin{array}{rccc} \mathrm{RB}: & T & \rightarrow & \mathbb{N} \\ & t & \mapsto & RB_t = \left| \left\{ d \in D \mid \mathrm{GT}(t,d) \succ \min(REL) \right\} \right| \end{array}$$

The recall base is an important reference for the analysis of experiments in IR; indeed a perfect system should retrieve all the relevant documents and rank them in decreasing order from position one up to the recall base. Several IR metrics are calculated by considering how the system under evaluation behaves at the recall base; relevant examples are *R*-precision [47, p. 161] or *R*-measure [56]. Recall itself is one of the most known IR metrics and it is calculated as number of relevant documents retrieved by a system divided by the recall base for a given topic.

4.2. Runs

Now, we stated all the definitions necessary to define a run as a set of vectors of documents, where each vector $\mathbf{r}_t$ of length N represents the ranked list of documents retrieved for a topic t with the constraint that no document is repeated in the ranked list.

Definition 4. *Given a natural number $N \in \mathbb{N}^+$ called the* length of the run*, a **run** is a function*

$$\begin{aligned} \mathrm{R}: \quad T &\rightarrow D^N \\ t &\mapsto \mathbf{r}_t = (d_1, d_2, \ldots, d_N) \end{aligned}$$

such that $\forall t \in T$, $\forall j,k \in [1,N] \mid j \neq k \Rightarrow \mathbf{r}_t[j] \neq \mathbf{r}_t[k]$ where $\mathbf{r}_t[j]$ denotes the j-th element of the vector $\mathbf{r}_t$, vectors start with index 1, and vectors end with index N.

In the following we introduce two important functions called relevance score and relevance weight which are the basis for calculating metrics and thus evaluating the performances of a system by using VIRTUE.

The relevance score associates the corresponding relevance degree to each element of a run.

Definition 5. *Given a run $\mathrm{R}(t) = \mathbf{r}_t$, the **relevance score** of the run is a function:*

$$\begin{aligned} \widehat{\mathrm{R}}: \quad T \times D^N &\rightarrow REL^N \\ (t, \mathbf{r}_t) &\mapsto \widehat{\mathbf{r}}_t = (rel_1, rel_2, \ldots, rel_N) \end{aligned}$$

where

$$\widehat{\mathbf{r}}_t[j] = \mathrm{GT}\big(t, \mathbf{r}_t[j]\big)$$

From the relevance score it is straightforward to introduce the definition of relevance weight of a run.

Definition 6. *Let $W \subset \mathbb{Z}$ be a totally ordered finite set of integers, REL be a finite set of relevance degrees and let $\mathrm{RW} : REL \rightarrow W$ be a monotonic function which maps each relevance degree ($rel \in REL$) into a relevance weight ($w \in W$).*

*Then, given a run $\mathrm{R}(t) = \mathbf{r}_t$ its **relevance weight** is a function:*

$$\begin{aligned} \widetilde{\mathrm{R}}: \quad T \times D^N &\rightarrow W^N \\ (t, \mathbf{r}_t) &\mapsto \tilde{\mathbf{r}}_t = (w_1, w_2, \ldots, w_N) \end{aligned}$$

where

$$\tilde{\mathbf{r}}_t[j] = \mathrm{RW}\big(\widehat{\mathbf{r}}_t[j]\big)$$

Summing up these two concepts we can say that the relevance score allows us to associate a relevance degree to a document for a given topic – e.g. document d is "highly relevant" (`hr`) for topic t; whereas the relevance weight allows us to associate an integer to a document for a given topic, where this integer reflects the relevance degree given by the relevance score – e.g. `hr` has weight 3, `fr` has weight 2, `pr` has weight 1, and `nr` has weight 0, then the relevance weight function is used to say that document d has weight 3 for topic t.

The relevance score as well as the relevance weight allow us to discern between two main different types of run: the ideal and the optimal run. We define the ideal run for a given topic $t \in T$ as the run where all the relevant documents for t are arranged in the vectors in descending order according to their relevance score. Therefore, the ideal run contains the best ranking of all the relevant documents for each considered topic. In the following definition, condition (1) ensures that all the relevant documents are retrieved in the ideal run while condition (2) guarantees that they are in descending order of relevance, thereby forming intervals of descending quality.

Definition 7. *The **ideal run** $\mathrm{I}(t) = \mathbf{i}_t$ is a run which satisfies the following constraints*

(1) recall base: $\forall t \in T, \left| \left\{ j \in [1,N] \mid \mathrm{GT}(t, \mathbf{i}_t[j]) \succ \min(REL) \right\} \right| = RB_t$

(2) ordering: $\forall t \in T, \forall j,k \in [1,N] \mid j < k \Rightarrow \widehat{\mathbf{i}}_t[j] \succeq \widehat{\mathbf{i}}_t[k]$

From definition 7, it follows that, for each topic $t \in T$ the relevance score of the ideal run $\widehat{\mathbf{i}}_t$ is a monotonic non-increasing function by construction. Therefore, the maximum of the function is at $j = 1$ and it is equal to $\widehat{\mathbf{i}}_t[1] = \max(REL)$ and the minimum is at $j = N$ and it is equal to $\widehat{\mathbf{i}}_t[N] = \min(REL)$.

Following the same line of reasoning, we define the optimal run as a variant of the ideal one. Indeed, the ideal run ranks in descending order all the relevant documents for a given topic t_i and it is the same for every possible run $R(t_i) = \mathbf{r_{t_i}}$; whereas the optimal run directly depends on a given run $R(t_i) = \mathbf{r_{t_i}}$. Indeed, the optimal run orders all documents retrieved by $\mathbf{r_{t_i}}$ in descending order according to their relevance score. This means that the ideal run is the best possible run for a given topic, whereas the optimal run is the best ordering of the documents retrieved by a run. In the following definition, given a run $\mathbf{r_t}$ and its optimal run $\mathbf{o_{r_t}}$, condition (1) guarantees that they contain the same documents and condition (2) guarantees that the documents in $\mathbf{o_t}$ are in descending order of relevance.

Definition 8. *Given a run $R(t) = \mathbf{r_t}$ with length $N \in \mathbb{N}^+$, its **optimal run** $\mathbf{o}_{r_t}$ is a run with length N, which satisfies the following constraints:*

(1) retrieved documents: $\forall t \in T, \forall j,k \in [1,N], \exists! \mathbf{o_{r_t}}[k] \mid \mathbf{r}_t[j] = \mathbf{o_{r_t}}[k]$

(2) ordering: $\forall t \in T, \forall j,k \in [1,N] \mid j < k \Rightarrow \widehat{\mathbf{o_{r_t}}}[j] \succeq \widehat{\mathbf{o_{r_t}}}[k]$

From this definition we can see that the ideal run depends only on the given topic, whereas the optimal run depends on the topic and on a given run. The ideal run tells us the best possible ranking a hypothetic system can return for a given topic, whereas the optimal run tells us the best ordering of the results returned by a real system. When we compare the ideal run with an experimental run, we understand how far the system which produced the experimental run is from the perfect retrieval and how many relevant documents it missed; when we compare the optimal run with an experimental run produced by a tested system, we determine how far the tested system is from a perfect ordering of the retrieved documents.

4.3. (Discounted) Cumulated Gain Metrics

The evaluation metrics considered in this paper exploit the idea that documents are divided into multiple ordered categories [37] and, specifically, they are a family of metrics composed of the CG and its discounted version which is the DCG; both CG and DCG have normalized versions called Normalized Cumulated Gain ((n)CG) and Normalized Discounted Cumulated Gain ((n)DCG) respectively. In VIRTUE we provide the possibility of analyzing the experimental runs on the basis of all the Cumulated Gain metrics. Basically, CG and DCG tell the same story about a run, but DCG is based on a user model which allows us to evaluate a system from the perspective of the patient or impatient user, as we discuss below. Mainly for this reason the adoption of DCG is more diffuse than CG and it is the *de-facto* standard metric for ranking evaluation in IR.

In the following, we exploit the preliminary definitions given above to formally present the cumulative gain metrics.

Definition 9. *Let* $\mathrm{R}(t)$ *be a generic run with length* $N \in \mathbb{N}^+$*, where* $t \in T$ *is a given topic,* RB_t *its recall base, and* $j \leq N$*, then CG[j] is defined as:*

$$CG[j] = cg_{\mathbf{r}_t}[j] = \sum_{k=1}^{j} \tilde{\mathbf{r}}_t[k]$$

We can see that CG (as well as the other cumulated gain metrics) is computed rank-by-rank; this means that it gives a measure of the run at every rank and it does not give a single number summarizing the overall trend of the run like, for instance, precision and recall do. For this reason often two runs on the same topic are compared using the CG at a given rank – i.e. the cut-off value; typical cut-off values are the recall base, 10, 100, and 1000.

The normalized version of the cumulated gain at position j – i.e. nCG[j] – is defined as the ratio between the CG of $\mathrm{R}(t)$ and the CG of the ideal run $\mathrm{I}(t)$:

$$\mathrm{nCG}[j] = \frac{cg_{\mathbf{r}_t}[j]}{cg_{\mathbf{i}_t}[j]}$$

The visualization of (n)CG curves is useful for the analyses conducted via VIRTUE because they are not monotonically non-decreasing curves like the CG ones are. (n)CG curves allow for an "easier analysis of the performances of a run at earlier ranks than CG curves, but the normalized ones lack of the straightforward interpretation of of the gain at each rank given by the CG curves" [37]. For this reason, it is important to be able to pass from one curve to the other dynamically in order to catch the differences between different runs.

To this purpose, the discounted cumulative versions of these metrics are important to give another view of the run, thus providing additional analytic possibilities to the analyst. Indeed, the discounted versions realistically weight down the gain received through documents found later in the ranked results, thus giving more importance to the early positions in ranking. 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.

Definition 10. *Given a run* $\mathrm{R}(t)$ *with length* $N \in \mathbb{N}^+$ *and a log base* $b \in \mathbb{N}^+$*, for all* $k \in [1,N]$ *the* ***discounted gain*** *is defined as:*

$$dg^b_{\mathbf{r}_t}[k] = \begin{cases} \tilde{\mathbf{r}}_t[k] & \text{if } k < b \\ \dfrac{\tilde{\mathbf{r}}_t[k]}{log_b k} & \text{otherwise.} \end{cases}$$

So, the discounted cumulative gain at rank j is defined as:

Definition 11. *Let* $\mathrm{R}(t)$ *be a generic run, then DCG[j] is defined as:*

$$DCG[j] = \sum_{k=1}^{j} dg^b_{\mathbf{r}_t}[k]$$

Typical instantiations of DCG measures make use of positive gains (i.e. relevance scores) 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 results list. DCG is the most used metric of the cumulated-gain family and VIRTUE mainly leverages on it for the study of system performances while supporting all the other metrics in the family.

Lastly, let us see the normalized version of the discounted cumulative gain (nDCG$^b[j]$) that can be defined as:

$$\text{nDCG}^b[j] = \sum_{k=1}^{j} \frac{dg^b_{\mathbf{r}_t}[k]}{dg^b_{\mathbf{i}_t}[k]}$$

4.4. Correlation Analysis

Given a run, for each one of the presented metrics it is possible to draw three curves: the curve of the run, the optimal run curve and the ideal run curve. VIRTUE enables a thorough study of these curves and their inter-relations; to this end, a significant means is Kendall's τ which estimates the distance between two run rankings [41, 68]. Kendall's τ is one of the standard correlation measures adopted in IR; for instance, it is widely used for measuring the correlation between document rankings [15] and between system rankings [56, 68]. In VIRTUE we use Kendall's τ to determine analytically if it is necessary to re-rank the documents in the run or if it is required to re-query to obtain a new set of results. As discussed in Section 3.1, Kendall's τ indicates that it is better to re-rank if there is a high correlation between the ideal and the optimal curve and a low correlation between the optimal and the experimental curve, meaning that the system retrieved many relevant documents, but ranked them poorly; on the other hand, it indicates that it is preferable to re-query if there is a low correlation between the ideal and the optimal curve meaning that the system did not retrieve many relevant documents.

Basically, given two runs with the same length, say $A(t)$ and $B(t)$, we consider the correlation between them on a relevance basis, thus calculating Kendall's τ between the relevance scores of the runs. Given the relevance scores $\widehat{\mathbf{a}}_t$ and $\widehat{\mathbf{b}}_t$ of $A(t)$ and $B(t)$, Kendall's τ is defined by the difference between the number of concordant pairs (relevance degrees in the same order in both rankings) and the number of discordant ones (in reverse order) normalized by the sum of the total number of concordant and discordant pairs [15].

Kendall's τ varies in the $[-1,1]$ range, where $\tau = 1$ means that the two compared rankings are equal, $\tau = -1$ means that one ranking is the reverse of the other (i.e. a perfect disagreement), and $\tau = 0$ means that the two compared rankings are independent of each other. In VIRTUE we have $\tau_{ideal-opt} = 1$ when the system under evaluation retrieves all the relevant documents; indeed, in this case all the relevance degrees in the ideal and optimal rankings are concordant.

4.5. Relative Position and Delta Gain

Relative Position (RP) and ΔG are the two metrics on which VIRTUE bases the "failing documents identification" (Section 3.3) and the "failing topics identification" (Section 3.4). They are complementary to each other; RP quantifies the misplacement of a document in a run ranking with respect to the ideal ranking, and ΔG estimates the effect of this misplacement in the overall calculation of the DCG.

In order to introduce RP we need to define the concepts of minimum rank and maximum rank of a given relevance degree building on the definition of ideal run. Indeed, the minimum

rank is the first position at which we find a document with relevance degree equal to *rel* while the maximum rank is the last position at which we find a document with relevance degree equal to *rel* in the ideal run.

Definition 12. *Given the ideal run* $\mathrm{I}(t)$ *and a relevance degree* $rel \in REL$ *such that* $\exists j \in [1,N] \mid \widehat{\mathbf{i}}_t[j] = rel$*, the* ***minimum rank*** *and the* ***maximum rank*** *are, respectively, a function*

$$\begin{array}{rlcl} \min_{\mathbf{i}_t}(rel): & T \times D^N \times REL & \rightarrow & \mathbb{N}^+ \\ & (t, \mathbf{i}_t, rel) & \mapsto & \min\Big(\big\{ j \in [1,N] \mid \widehat{\mathbf{i}}_t[j] = rel \big\}\Big) \\ \max_{\mathbf{i}_t}(rel): & T \times D^N \times REL & \rightarrow & \mathbb{N}^+ \\ & (t, \mathbf{i}_t, rel) & \mapsto & \max\Big(\big\{ j \in [1,N] \mid \widehat{\mathbf{i}}_t[j] = rel \big\}\Big) \end{array}$$

We can now introduce the RP metric which points out the instantaneous and local effect of misplaced documents and how much they are misplaced with respect to the ideal case $\mathbf{i}_t$. In the following definition, zero values denote documents which are within the ideal interval; positive values denote documents which are ranked below their ideal interval, i.e. documents of higher relevance degree that are in a position of the ranking where less relevant ones are expected; whereas negative values denote documents which are above their ideal interval, i.e. less relevant documents that are in a position of the ranking where documents of higher relevance degree are expected. Note that the greater the absolute value of RP is, the greater the distance of the document from its ideal interval.

Definition 13. *Given a run* $\mathrm{R}(t)$*, the* ***Relative Position (RP)*** *is a function*

$$\begin{array}{rlcl} \mathrm{RP}: & T \times D^N & \rightarrow & \mathbb{Z}^N \\ & (t, \mathbf{r}_t) & \mapsto & \mathbf{rp}_{\mathbf{r}_t} = (rp_1, rp_2, \ldots, rp_N) \end{array}$$

where

$$\mathbf{rp}_{\mathbf{r}_t}[j] = \begin{cases} 0 & \text{if } \min_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) \leq j \leq \max_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) \\ j - \min_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) & \text{if } j < \min_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) \\ j - \max_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) & \text{if } j > \max_{\mathbf{i}_t}\big(\widehat{\mathbf{r}}_t[j]\big) \end{cases}$$

ΔG is a metric which quantifies the effect of misplacing relevant documents with respect to the ideal run. ΔG allows for a deeper comprehension of the behavior of DCG curves indicating, rank-by-rank, how a document contributes to the overall computation of DCG. ΔG has value zero if a document is ranked in the correct position with respect to the ideal case, a positive value if it is ranked above its ideal position and a negative value otherwise. The higher the absolute ΔG value of a document is, the greater its misplacement with respect to the ideal ranking.

ΔG explicitly takes into account the effect of the discounted function and it is calculated by exploiting the discounted gain presented in Definition 10.

Definition 14. *Given the ideal run* $\mathrm{I}(t)$*, a run* $\mathrm{R}(t)$*, and the discounted gains* $dg^b_{\mathbf{i}_t}$ *and* $dg^b_{\mathbf{r}_t}$ *for* $I(t)$ *and* $R(t)$ *respectively. Then,* ***delta gain*** *(*ΔG*) is a function:*

$$\begin{array}{rccl} \Delta\mathrm{G}: & T \times D^N & \rightarrow & \mathbb{Z}^N \\ & (t, \mathbf{r}_t) & \mapsto & \Delta\mathbf{g}_{\mathbf{r}_t} = (\Delta g_1, \Delta g_2, \ldots, \Delta g_N) \end{array}$$

where

$$\Delta\mathbf{g}[j] = dg^b_{\mathbf{r}_t}[j] - dg^b_{\mathbf{i}_t}[j]$$

5. Visual Analytics Environment

In this section, we describe the main characteristics of the system in term of both technological and design choices. VIRTUE has been designed using a UCD methodology (see, e.g., [70]), using requirements and feedback coming from four IR experts, focusing on the usefulness and comprehensibility of the system. The user interface is a direct result of such an activity and represents data according to the expert knowledge (e.g., the vector of retrieved documents is presented in a vertical fashion, using green for correct result, red for a loss a nd blue for a gain). The goal of the system was to assess the scientific validity of the approach and its usefulness; usability issues have been not the main focus of both the design and validation activities. Details about the UCD steps are outside the scope of the paper. VIRTUE has been implemented as a Web application whose home page resembles the structure depicted in Figure 1 and allows for accessing all the system functionalities, namely Performance and Failure analysis, both at Topic and Experiment levels.

5.1. Ranked Results Exploration

Type of analysis: Performance analysis
Granularity level: Topic level

The "Ranked Results Exploration" allows for understanding how the IR system under examination is performing with respect to a specific topic. We can see a screenshot of this functionality in Figure 2. The working area is split into two parts: the *controls area* (left side) and the *graph area* (right side). The control area allows the user to select an experiment and one of the associated topics. Moreover, it is possible to select one of the metrics belonging to the cumulated-gain family that have been implemented in the system: CG, DCG, (n)CG, and (n)DCG, where for the metrics using a discounting function (i.e. DCG and (n)DCG) it is possible to specify the logarithm base.

VIRTUE is designed to deal with few hundreds results because they represent the salient part of a ranking from the user point-of-view; indeed, users are reasonably more interested to understand which system ranks more relevant documents within the higher ranks than to know if there are important documents after rank, say, 100. This assumption is supported by the user model adopted by DCG which assigns low or no gain to the documents placed at low ranks [37]. In particular, in VIRTUE we use the first 200 ranking positions (reported in the x-axis of the graph).

Three different curves are shown in the graph area:

1. *Experiment curve*, displayed in cyan, representing the (discounted) cumulated gain values for the actual list of retrieved documents;

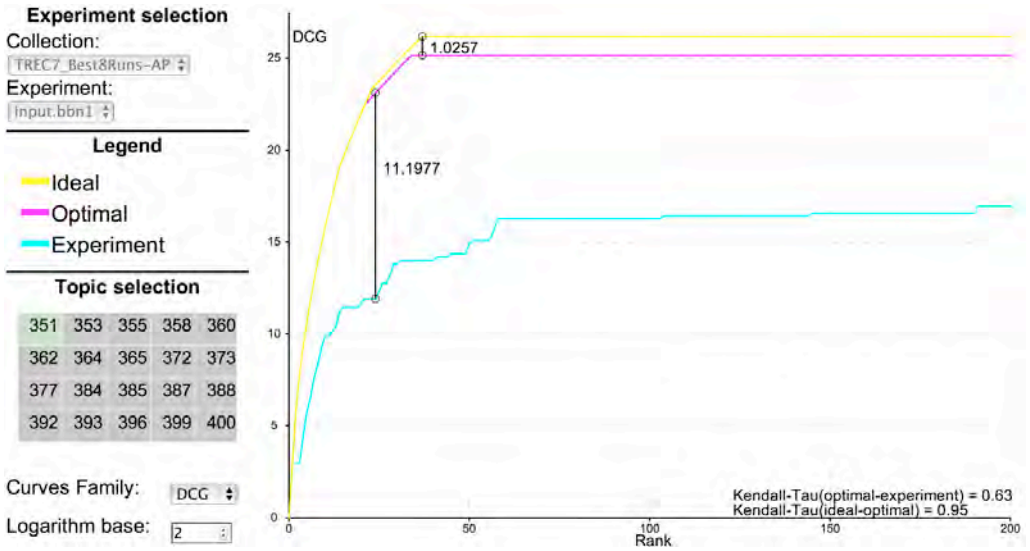

Figure 2: Ranked Results Exploration

2. *Optimal curve*, displayed in magenta, representing the (discounted) cumulated gain values for the optimal run (see Definition 8);
3. *Ideal curve*, displayed in yellow, representing the (discounted) cumulated gain values for the ideal run (see Definition 7).

Some additional graphical indicators are displayed in the graph area. These include two pairs of black circles represent the maximum distance between the experiment and ideal rankings, and between the optimal and ideal rankings. Moreover, a tooltip displaying additional details (e.g. the actual rank, the identifier of the document, and the metric value is activated when the user moves the mouse over one of the curves. Furthermore, the values of Kendall's τ – i.e. $\tau_{ideal-opt}$ and $\tau_{opt-exp}$) – are reported in the lower right part of the graph by providing quantitative clues that help the user to make the decision whether it is more convenient to re-rank or to re-query in order to improve the experiment performances.

5.2. Ranked Results Distribution Exploration

Type of analysis: Performance analysis
Granularity level: Experiment level

The "Ranked Result Distribution Exploration" allows the user to understand the overall system performances. For each rank it shows the distribution of the selected metric on an arbitrary subset of all the topics associated with the experiment.

Figure 3 shows the actual VIRTUE implementation: as an additional feature, the control area allows for selecting a *subset* of the topics (by default all topics are selected) while the metric distribution is rendered by connecting the salient box-plot values computed at each position and for the three reference rankings (experiment, optimal, and ideal). This results in five different curves for each ranking, represented as follows:

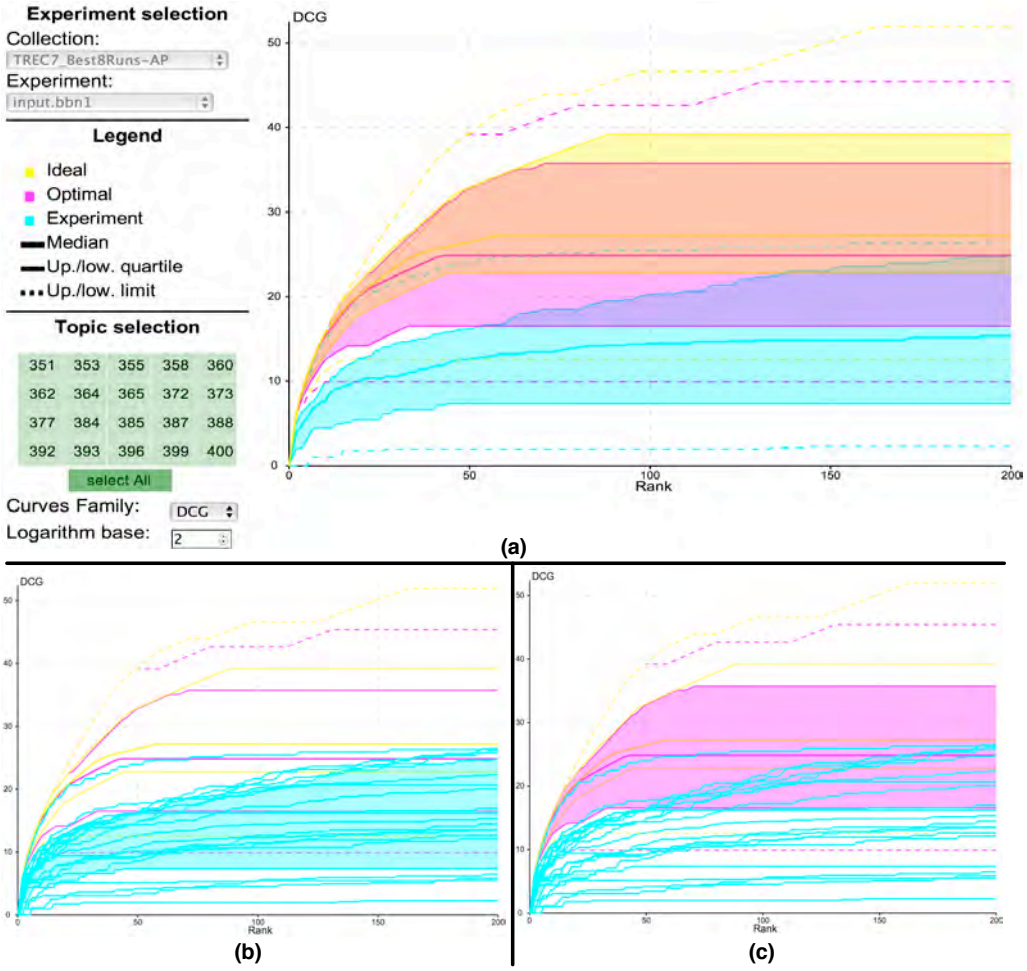

Figure 3: Ranked Results Distribution Exploration

- Upper limit: dash-stroke line;
- Upper quartile: continuous-stroke line;
- Median: thick continuous-stroke line;
- Lower quartile: continuous-stroke line;
- Lower limit: dash-stroke line.

In order to make the general trend of the experiment more evident, the area between the upper and the lower quartile is filled with a solid color, while overlapping areas are rendered with the cyan, magenta and yellow combinations.

To allow the user to further explore the statistics of a single ranking, the system allows for highlighting the upper-lower quartile area regarding the experimental, optimal or ideal curves;

see, for example, Figure 3(b) in which the upper-lower quartile of the experiment area is highlighted, or Figure 3(c) where the optimal area is highlighted. Moreover, to make explicit the contribution of all the topics, the user can select the corresponding label on the legend – for instance, see Figure 3(b) and (c) in which the experimental area is populated by all the curves corresponding to all the analyzed topics. Accordingly to this, the legend on the left presents more entries than before where the additional lines represent box-plot statistics for the selected metric such as the median, the upper/lower quartile and the upper/lower limit.

Figure 3 shows some of the allowed visual analyses; in particular, Figure 3(b) reports the relationships between the experimental curves with respect to the upper-lower quartile of the experimental boxplot and Figure 3(c) the relationships between the experimental curves with respect to the optimal boxplot. From the former we can see for which topics the selected experiment does not perform well, for which ones it is on the average, and for which ones it performs well. The experimental boxplot gives us an immediate overview of the global behavior of the experiment, whereas the single curves indicate where we can improve the experiment and where it is already behaving well. On the other hand, Figure 3(c) shows the same interaction, but it compares the experimental curves with the optimal boxplot. In this case, we can see that many experimental curves are below the optimal boxplot indicating that with a re-rank of the documents the experiment would perform much better because it retrieves many relevant documents, but it ranks them in an ineffective way. The very same analysis can be conducted by comparing the experimental curves with the ideal boxplot in place of the optimal one, which would help us to understand how many topics would be required for our experiment to benefit from a re-query (for instance, if many experimental curves are below the ideal boxplot) instead of a re-rank.

5.3. Failing Documents Identification

Type of analysis: Failure analysis
Granularity level: Topic level

"Failing Document Identification" aims at identifying which documents contribute the most to the performances of the selected experiment. It allows for discovering which documents have been misplaced with respect to the correct ranking given by the ideal run and it makes the consequential loss in the (discounted) cumulated gain function also evident visually.

In order to deal with this failure analysis task, the system presents the user with the usual experiment, ideal, and optimal curves plus a second visualization composed of two color-code bar charts that display the Relative Position (RP) and the Delta Gain (ΔG) values. In particular, the *RP bar* reports, for each document, the relationship that exists between the documents and the ideal ranks as described in Definition 13.

The following color coding has been chosen to encode such relationships:

1. document well placed: green (RP equals to zero);
2. document placed below its ideal position (positive RP values): blue;
3. document placed above its ideal position (negative RP values): red.

The RP bar gives the analyst a hint about the (discounted) cumulative gain behavior; indeed, when the curve goes up we only know that a relevant document has been encountered, but we cannot say if that document should have been placed in another position. The RP bar gives us this information; if the document is associated with a red segment in the RP bar then it should be placed above, if it associated with a blue segment then it should be placed below its actual position.

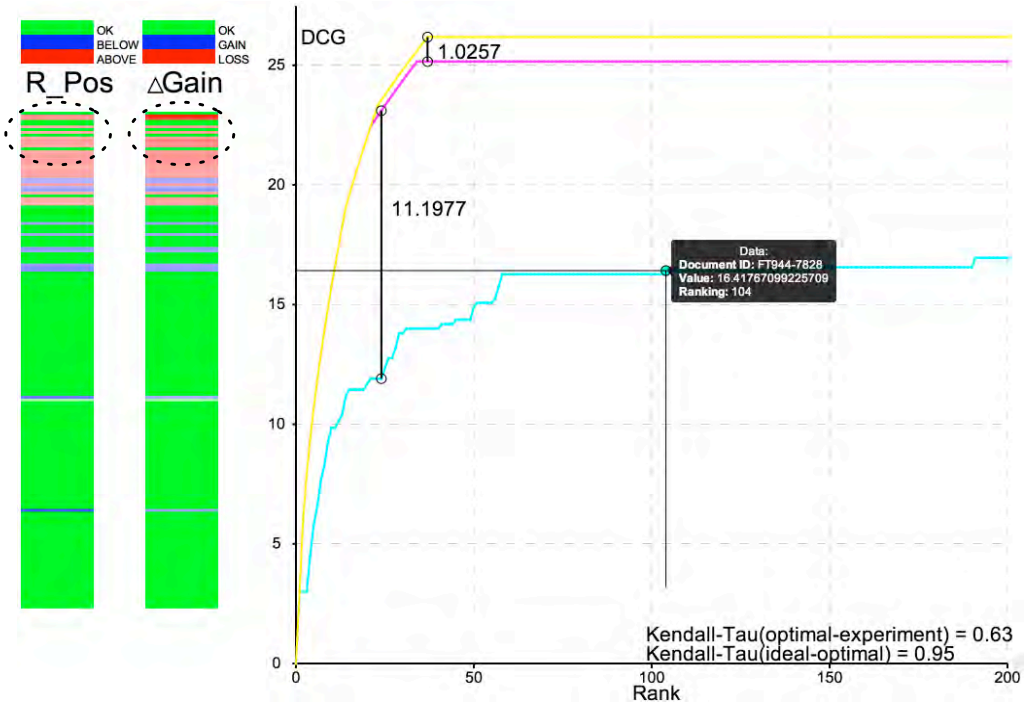

Figure 4: Failing Documents Identification: interaction with the curves.

ΔG complements this information by saying how much a misplaced document contributes to the overall (discounted) cumulated gain metric. Indeed, a document may be misplaced accordingly to RP, but its contribution to the overall performance may be very low; in this case, this misplacement can be ignored, otherwise we should take some action to address the misplacement to reduce its overall impact.

Thus, the *ΔG bar* represents the relationship for each misplaced document existing between its rank and the loss or gain it provides to the evaluation metric. The color coding works as follows: green represents the ideal contribution, red a loss, and blue a gain with respect to the ideal one. Moreover, as with the RP bar, the hue of the color is proportional to the value of loss/gain.

As an example, in Figure 4, looking at the RP bar, the documents in second and third positions, although slightly misplaced, produce a huge loss in score, as documented by a strong hue of red in the ΔG bar. This situation is also visible in the graph, where after a few positions the experiment curve strongly deviates from the optimal and ideal ones with a resulting lower overall score.

Figure 4 shows another interaction with the graph; indeed, by selecting a specific point in the plot a tooltip shows the information about that document and its corresponding segment in the RP and ΔG bars is highlighted. Figure 5 shows another interaction with the system, indeed by selecting a specific point in the RP or in the ΔG bars, its corresponding points in the three curves are spotted by three circles. Furthermore, a tooltip reporting the identifier, content and rank of the selected document is shown on the left; in this way the user can directly analyze the

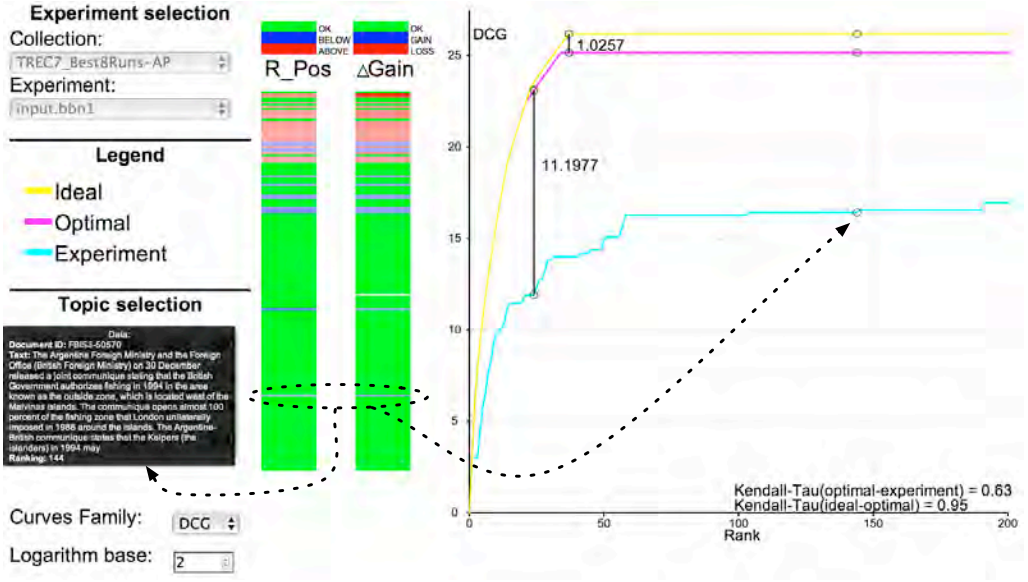

Figure 5: Failing Documents Identification: interaction with the RP and ΔG bars.

misplaced document.

5.4. Failing Topics Identification

Type of analysis: Failure analysis
Granularity level: Experiment level

The "Failing Topics Identification" allows for exploring the contribution of misplaced documents in the context of the whole set of topics (or a chosen subset) thus considering the experiment as a whole. Basically, it adds the analytics functionalities reported in Figure 4 to the visualization shown in Figure 3.

The user can select an aggregation function (e.g. mean, max, min, etc.) summarizing the contribution of all documents to the same rank. In this way, an aggregate vision of the experiment is available to the user where the aggregation is expressed by the plot as happens in Figure 3 and by the RP and ΔG bars.

In this way, a global visualization representing the overall experiment behavior is obtained, both in terms of good/bad ranking of documents (by the graph plot) and in terms of aggregated contributions to the selected metric (by the RP and ΔG bars).

Figure 6 shows this aggregated view where the experimental curves are compared with the ideal boxplot; we can see that only a few curves lie within the shaded ideal area, whereas most of them are below it. This aspect is even more marked if we observe the first rankings where only one experimental curve lies within the highlighted ideal area. This fact is supported by the analysis of the ΔG bar; indeed, the upper segments, indicating high rankings, are colored by a strong red indicating that there is a major loss in terms of (discounted) cumulative gain. The aggregated bars give us a concrete measure of how the experiment behaved when we considered more topics at the same time. Furthermore, we can see that this experiment behaves better when

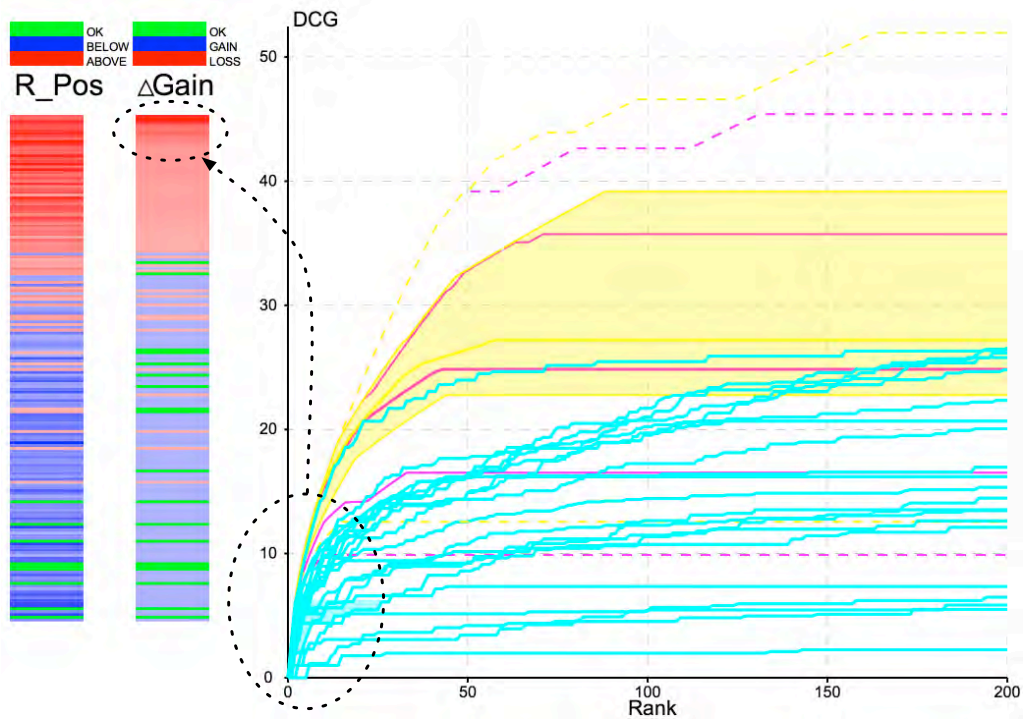

Figure 6: Failing Topics Identification

lower rankings are considered; indeed, the experimental curves intersect the ideal area and there are green areas in the ΔG aggregated bar. We can conclude that in the context of this experiment, the considered system does not behave well from the user point-of-view because it misplaces many documents in higher positions, whereas it behaves better at lower ranks that are generally less useful for the end users.

6. Validation

We conducted a formal user study to evaluate VIRTUE, which involves IR evaluation experts (i.e. academics, post-docs, and PhD students). It is worth noting that such experts are exactly the users the system is intended for: the tool's goal is to assist developers and researchers in understanding and fixing ranking errors produced by a search engine and this activity is not a typical end user task. In particular, 13 experts (7 females and 6 males) were involved in the study, coming from 9 European Countries and working on different aspects of IR experiment evaluation. The goal of the study was to assess a) the VIRTUE scientific relevance and innovation and b) the comprehensibility and efficacy of the proposed visualizations. To this end, we used the well-known and documented [69] TREC-7 dataset for which the failures of the systems are already known in order to understand whether VIRTUE would have been an effective and appropriate tool to ease their detection. However, while usability issues were not the focus of the validation activity, we have collected, through the free part of questionnaires, some issues that can improve the overall system and we are addressing them.

6.1. Methodology

Before starting the study, people were instructed through an oral presentation about the VIRTUE background and a practical use of the system was demonstrated, in order to allow participants to become familiar with the system and to let them understand how to use it. Each visualization was discussed in detail together with the associated automated analysis. Questions about the overall methodology, technical details, and visualizations were answered.

After that, a closed questionnaire was given to the participants; each question of the questionnaire had to be answered by using an interval Likert scale ranging from 1 to 5, in which each numerical score was labeled with a description: {1:not at all, 2:a little, 3:enough, 4:a lot, 5:quite a lot}. The questionnaire was structured in 5 identical sections, one for each visualization described in Section 5 (see Figure 1) plus one for the overall system. An additional open section (optional) was provided for collecting additional comments. Each one of the 5 closed sections was composed of two groups of questions:

Q1 Is the addressed problem relevant for involved stakeholders (researchers and developers)?

Q2 Are the currently available tools and techniques adequate for dealing with the addressed problem?

Q3 Do currently available tools and techniques for dealing with the addressed problem offer interactive visualizations?

Q4 Is the proposed visual tool understandable?

Q5 Is the proposed visual tool suitable and effective for dealing with the addressed problem?

Q6 To what extent is the proposed visual tool innovative with respect to the currently available tools and techniques?

Q7 To what extent will the proposed visual tool enhance the productivity of involved stakeholders (researchers and developers)?

The first three questions, which are visually separated from the others, were aimed at collecting the experts' opinion about the relevance of the addressed problem (Q1), the adequateness (Q2) and the degree of interactiveness (Q3) of other visual tools designed for the same purpose. The last four questions were aimed at assessing the understandability (Q4), suitability (Q5), visual innovativeness (Q6), and efficiency (Q7) of VIRTUE.

The study was conducted by allowing the experts to freely use VIRTUE for an hour, following the path "Performance Analysis and Failure Analysis" (see Section 5) and compiling the questionnaire sections that were arranged in the same order.

6.2. Results

The questionnaire results are depicted in Figure 7, which presents the distribution of the answers assessing the system as a whole, and in Figure 8, which provides details, through averages, on each of the four VIRTUE components.

Considering Figure 7, we can conclude that the addressed problem has been judged as a relevant one from the involved stakeholders (90% of the answers to Q1 are in the range [4, 5]

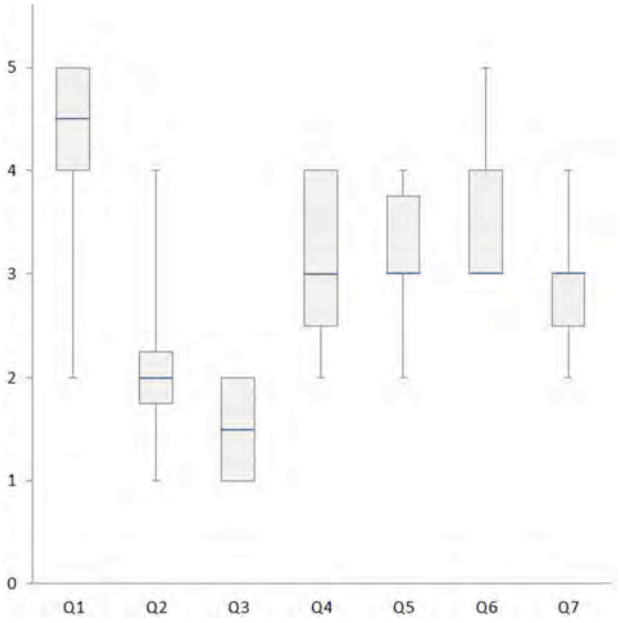

Figure 7: Evaluating VIRTUE as a whole.

with mean=4.3 and STD=0.95) and that there is no other tool doing the work of VIRTUE (Q2 and Q3, in which more than 85% of the answers are in the range [1, 2] with mean=1.9 and STD=0.69 for Q2, and mean=1.50 and STD=0.55 for Q3). It means that, according to the experts' opinion, VIRTUE is proposing something totally new in the field. We can also conclude that the tool is understandable (72% of the answers to Q4 are in the range [3, 4], mean=3.18 and STD=0.87), suitable (90% of the answers to Q5 are in the range [3, 4], mean=3.20 and STD=0.63), innovative (all the answers to Q6 are in the range [3, 4], mean=3.50 and STD=0.76). The last question is about productivity; on average the experts think VIRTUE can improve productivity (71% of the answers to Q7 are in the range [3, 4], mean=2.86 and STD=0.89) but the mean is below 3 and we think that this is due to the time needed to learn how to effectively use the system and by the inherent complexity of the Failure Analysis at the experiment level (Figure 1 (d)).

Such a complexity is confirmed from the detailed results depicted in Figure 8 in which this visualization shows low values for Q4 (mean=2.50) and Q5 (mean=2.25). The other values are closer to the overall means.

6.3. Discussion

While the study results give clear indications of the usefulness and the innovation of the VIRTUE system (we received some enthusiastic comments like P1: "I would love to have this tool, both for research and for teaching purposes" and P8: "If I have had this tool during my PhD thesis writing I would have saved weeks of work"), there are some issues that deserve more attention, requiring a more clear design and a deeper analysis. These considerations rise from some low scores on Q4 and Q5 for the failure analysis at the experiment level, and from the questions the participant raised during the experiment, and from the free comments on the questionnaires. In particular, while the visualization and the analytical models underlying the Failing

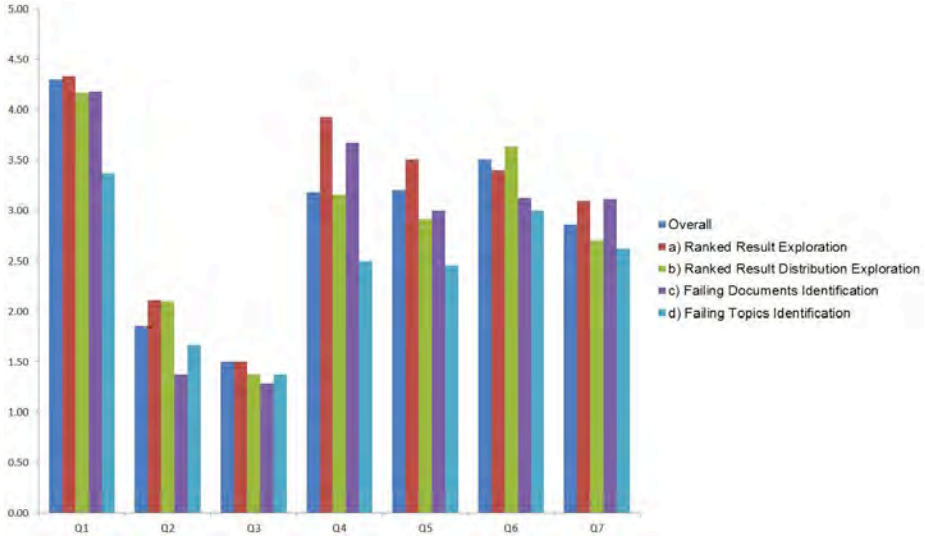

Figure 8: The histogram reporting the averages of the experts' answers to all the sections of the questionnaire.

Documents Identification have been fully understood and positively judged during the evaluation of the system, the same did not happen for the Failing Topics Identification. In particular, we received a few negative comments saying that failure analysis at the experiment level is hard to deal with; for instance, (P7) wrote that "failure analysis is too hard to use [...], experiment level views of performance and failure are difficult to interpret [...]". That gives us the feeling that the visualization we are proposing for Failure Analysis, see Figure 1 (d), contains a lot of visual information, i.e., three levels of analysis: ideal, optimal, and experiment. Interaction, highlighting, alpha blending, and brushing mitigate the problem but require time to be learned and likely some longitudinal studies can provide more insights on how to further improve such visualization. Moreover, we received several comments on basic usability issues like missing on-screen instructions (P1, P3) and on additional required features, like allowing for inspecting details of topics and documents (P1, P5, P7, P9, P11) and having information about the number of relevant documents for a topic (P1); fixing such issues and addressing user suggestions will result in a clear systems improvement. Indeed, we received some useful indications on how to improve the system, pointing out different analysis strategies, e.g. P11: "[...] it would be nice make it the possible to cluster topics by good/bad to look at the chosen group of topics only", giving us some insights on how to refine and improve the actual model. Moreover P9, P10, and P13 suggested to use the system to compare two or more experiments for the same subset of topics.

7. Conclusions and Future Work

IR is a field deeply rooted in evaluation which is carried out to assess the performances of the proposed algorithms and systems and to better understand their behaviour. Nowadays, systems are becoming increasingly complex since the tasks and user requirements addressed are becoming more and more challenging. As a consequence, evaluating and understanding these

systems is an increasingly demanding activity in terms of the time and effort needed to carry it out. The goal of this paper is thus to provide the researcher and developer with better and more effective tools to understand the system behaviour, its performances, and failures.

To this end, we have designed and developed an innovative tool for conducting performance and failure analysis of IR systems. The proposed tool exploits visual analytics techniques in order to foster interaction with and exploration of the experimental data at both topic and experiment level. It improves the state-of-the-art in the evaluation practice by: (i) easing the interaction and interpretation of DCG curves, a very widely adopted way of measuring ranked result lists; (ii) highlighting critical areas of a ranked result list in order to inspect and detect causes of failure; (iii) providing a convenient way to partner the detailed analysis at the topic level with an overall analysis at the global experiment level which support users in spotting critical topics and/or critical rank areas across several topics.

We conducted an evaluation of the proposed tools with IR experts and the outcomes have been encouraging in terms of the usefulness, innovativeness and potential of the proposed approaches.

Concerning future activities, we will improve the current system by incorporating all the suggestions collected during the system evaluation and through a semi-formal longitudinal study: the system will be made available as a support to the regular CLEF evaluation activities, which will allow for having experts using it in an intensive way. Issues and suggestions collected in this way will help us in further shaping and refining the proposed tools.

Moreover, the RP indicator opened the way for designing and developing a brand new metric, called Cumulated Relative Position (CRP), for evaluating the performances of an IR system [5]. The CRP metric is the cumulative sum of the RP indicator and it shares a similar approach to the DCG measures, i.e. cumulating what happened up to a given rank position. The discussion about this metric is out of scope for this paper, but this pointer is given for highlighting how the formal analytical framework proposed here allows new research directions to stem from it, also beyond its original purposes and how the approach taken for the DCG can be straightforwardly applied to other, similar, metrics.

Finally, we plan to extend the system in two main directions. The first is to allow the comparison of two or more experiments at a time in order to assist users in simultaneously assessing the impact of alternative strategies. The second, more ground breaking, direction consists of introducing a completely new phase, called "what-if analysis", in the process aimed at estimating the impact of possible modifications and fixes to a system suggested by the failure analysis, in order to anticipate whether they will have a helpful or harmful effect before actually implementing them in a new system and running another evaluation cycle to assess them.

Acknowledgments

The PROMISE network of excellence[4] (contract n. 258191) project, as part of the 7th Framework Program of the European Commission, has partially supported the reported work. The authors would like to thank the IR evaluation experts involved in the validation study, who provided valuable suggestions about how to improve VIRTUE.

References

[1] Agosti, M., Berendsen, R., Bogers, T., Braschler, M., Buitelaar, P., Choukri, K., Di Nunzio, G. M., Ferro, N., Forner, P., Hanbury, A., Friberg Heppin, K., Hansen, P., Järvelin, A., Larsen, B., Lupu, M., Masiero, I., Müller,

[4] http://www.promise-noe.eu/

H., Peruzzo, S., Petras, V., Piroi, F., de Rijke, M., Santucci, G., Silvello, G., Toms, E., December 2012. PROMISE Retreat Report – Prospects and Opportunities for Information Access Evaluation. SIGIR Forum 46 (2).

[2] Agosti, M., Di Buccio, E., Ferro, N., Masiero, I., Peruzzo, S., Silvello, G., 2012. DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In: [16], pp. 88–99.

[3] Agosti, M., Di Nunzio, G. M., Dussin, M., Ferro, N., 2010. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In: Sakay, T., Sanderson, M., Webber, W. (Eds.), Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010). National Institute of Informatics, Tokyo, Japan, pp. 16–24.

[4] Agosti, M., Ferro, N., 2009. Towards an Evaluation Infrastructure for DL Performance Evaluation. In: Tsakonas, G., Papatheodorou, C. (Eds.), Evaluation of Digital Libraries: An insight into useful applications and methods. Chandos Publishing, Oxford, UK, pp. 93–120.

[5] Angelini, M., Ferro, N., Järvelin, K., Keskustalo, H., Pirkola, A., Santucci, G., Silvello, G., 2012. Cumulated Relative Position: A Metric for Ranking Evaluation. In: [16], pp. 112–123.

[6] Angelini, M., Ferro, N., Santucci, G., Silvello, G., 2012. Visual Interactive Failure Analysis: Supporting Users in Information Retrieval Evaluation. In: Kamps, J., Kraaij, W., Fuhr, N. (Eds.), Proc. 4th Symposium on Information Interaction in Context (IIiX 2012). ACM Press, New York, USA, pp. 195–203.

[7] Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L., 2012. Expertise Retrieval. Foundations and Trends in Information Retrieval (FnTIR) 6 (2-3), 127–256.

[8] Banks, D., Over, P., Zhang, N.-F., May 1999. Blind Men and Elephants: Six Approaches to TREC data. Information Retrieval 1, 7–34.

[9] Behrisch, M., Davey, J., Simon, S., Schreck, T., Keim, D., Kohlhammer, J., 2013. Visual Comparison of Orderings and Rankings. In: Pohl, M., Schumann, H. (Eds.), Proc. 4th International Workshop on Visual Analytics (EuroVA 2013). Eurographics Association, Goslar, Germany.

[10] Buckley, C., 2004. Why Current IR Engines Fail. In: Sanderson, M., Järvelin, K., Allan, J., Bruza, P. (Eds.), Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004). ACM Press, New York, USA, pp. 584–585.

[11] Buckley, C., Voorhees, E. M., 2000. Evaluating Evaluation Measure Stability. In: Yannakoudakis, E., Belkin, N. J., Leong, M.-K., Ingwersen, P. (Eds.), Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000). ACM Press, New York, USA, pp. 33–40.

[12] Buettcher, S., Clarke, C. L. A., Cormack, G. V., 2010. Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press, Cambridge (MA), USA.

[13] Burnett, S., Clarke, S., Davis, M., Edwards, R., Kellett, A., October 2006. Enterprise Search and Retrieval. Unlocking the Organisation's Potential. Butler Direct Limited.

[14] Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., Schuldt, H., December 2007. The DELOS Digital Library Reference Model. Foundations for Digital Libraries. ISTI-CNR at Gruppo ALI, Pisa, Italy, http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf.

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

[16] Catarci, T., Forner, P., Hiemstra, D., Peñas, A., Santucci, G. (Eds.), 2012. Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012). Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany.

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

[18] Crestani, F., Vegas, J., de la Fuente, P., 2004. A Graphical User Interface for the Retrieval of Hierarchically Structured Documents. Inf. Process. Management 40 (2), 269–289.

[19] Croft, W. B., Metzler, D., Strohman, T., 2009. Search Engines: Information Retrieval in Practice. Addison-Wesley, Reading (MA), USA.

[20] Cronen-Townsend, S., Zhou, Y., Croft, W. B., 2002. Predicting Query Performance. In: [36], pp. 299–306.

[21] Derthick, M., Christel, M. G., Hauptmann, A. G., Wactlar, H. D., 2003. Constant density displays using diversity sampling. In: Proceedings of the Ninth annual IEEE conference on Information visualization. INFOVIS'03. IEEE Computer Society, Washington, DC, USA, pp. 137–144.
URL http://dl.acm.org/citation.cfm?id=1947368.1947395

[22] Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., Tino, G., 2011. Interactive Analysis and Exploration of Experimental Evaluation Results. In: Wilson, M. L., Russell-Rose, T., Larsen, B., Kalbach, J. (Eds.), Proc. 1st European Workshop on Human-Computer Interaction and Information Retrieval (EuroHCIR 2011) http://ceur-ws.org/Vol-763/. pp. 11–14.

[23] Di Buccio, E., Dussin, M., Ferro, N., Masiero, I., Santucci, G., Tino, G., 2011. To Re-rank or to Re-query: Can Visual Analytics Solve This Dilemma? In: Forner, P., Gonzalo, J., Kekäläinen, J., Lalmas, M., de Rijke, M.

(Eds.), Multilingual and Multimodal Information Access Evaluation. Proceedings of the Second International Conference of the Cross-Language Evaluation Forum (CLEF 2011). Lecture Notes in Computer Science (LNCS) 6941, Springer, Heidelberg, Germany, pp. 119–130.

[24] Fan, W., Gordon, M. D., Pathak, P., 2004. A Generic Ranking Function Discovery Framework by Genetic Programming for Information Retrieval. Inf. Process. Manage. 40 (4), 587–602.

[25] Ferro, N., Hanbury, A., Müller, H., Santucci, G., 2011. Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems. Procedia Computer Science 4, 740–749.

[26] Ferro, N., Sabetta, A., Santucci, G., Tino, G., 2011. Visual Comparison of Ranked Result Cumulated Gains. In: Miksch, S., Santucci, G. (Eds.), Proc. 2nd International Workshop on Visual Analytics (EuroVA 2011). Eurographics Association, Goslar, Germany, pp. 21–24.

[27] Fowler, R. H., Lawrence-Fowler, W. A., Wilson, B. A., 1991. Integrating Query, Thesaurus, and Documents Through a Common Visual Representation. In: Fox, E. A. (Ed.), Proc. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991). ACM Press, New York, USA, pp. 142–151.

[28] Fox, E. A., Gonçalves, M. A., Shen, R., 2012. Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach. Morgan & Claypool Publishers, USA.

[29] Hansen, P., Bartolini, C., Convertino, G., Santucci, G., Angelini, M., Granato, G., 2012. Collaborative Environment of the PROMISE Infrastructure: an "ELEGantt" Approach. In: Wilson, M. L., Russell-Rose, T., Larsen, B., Kalbach, J. (Eds.), Proc. of the 2nd European Workshop on Human-Computer Interaction and Information Retrieval. Vol. 909 of CEUR Workshop Proceedings. CEUR-WS.org, pp. 55–58.

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

[31] Harman, D. K., 2008. Some thoughts on failure analysis for noisy data. In: Lopresti, D., Roy, S., Schulz, K., Venkata Subramaniam, L. (Eds.), Proc. 2nd Workshop on Analytics for Noisy unstructured text Data (AND 2008). ACM Press, New York, USA, pp. 1–1.

[32] Harman, D. K., 2011. Information Retrieval Evaluation. Morgan & Claypool Publishers, USA.

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

[34] Hearst, M. A., 2011. "Natural" Search User Interfaces. Commun. of the ACM 54 (11), 60–67.

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

[36] Järvelin, K., Beaulieu, M., Baeza-Yates, R., Hyon Myaeng, S. (Eds.), 2002. Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002). ACM Press, New York, USA.

[37] Järvelin, K., Kekäläinen, J., October 2002. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information System 20, 422–446.

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

[39] Kando, N., Ishikawa, D., Sugimoto, M. (Eds.), 2011. Proc. 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. National Institute of Informatics, Tokyo, Japan.

[40] Keen, E. M., 1971. Evaluation parameters. In: [57], pp. 74–111.

[41] Kekäläinen, J., 2005. Binary and Graded Relevance in IR Evaluations–Comparison of the Effects on Ranking of IR Systems. Information Processing & Management 41 (5), 1019–1033.

[42] Kekäläinen, J., Järvelin, K., November 2002. Using Graded Relevance Assessments in IR Evaluation. Journal of the American Society for Information Science and Technology (JASIST) 53 (13), 1120—1129.

[43] Kendall, M., 1948. Rank correlation methods. Griffin, Oxford, England.

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

[45] Koshman, S., 2005. Testing User Interaction with a Prototype Visualization-Based Information Retrieval System. Journal of the American Society for Information Science and Technology 56 (8), 824–833.
URL http://dx.doi.org/10.1002/asi.20175

[46] Lupu, M., Hanbury, A., 2013. Patent Retrieval. Foundations and Trends in Information Retrieval (FnTIR) 7 (1), 1–97.

[47] Manning, C. D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

[48] McGill, R., Tukey, J. W., Larsen, W. A., Februay 1978. Variations of Box Plots. The American Statistician 32 (1), 12–16.
[49] Mizzaro, S., September 1997. Relevance: The Whole History. Journal of the American Society for Information Science and Technology (JASIST) 48 (9), 810–832.
[50] Morse, E. L., Lewis, M., Olsen, K. A., 2002. Testing Visual Information Retrieval Methodologies Case Study: Comparative Analysis of Textual, Icon, Graphical, and Spring Displays. Journal of the American Society for Information Science and Technology (JASIST) 53 (1), 28–40.
[51] Robertson, S. E., 1981. The methodology of information retrieval experiment. In: Spärck Jones, K. (Ed.), Information Retrieval Experiment. Butterworths, London, United Kingdom, pp. 9–31.
[52] Robertson, S. E., 2008. On the history of evaluation in IR. Journal of Information Science 34 (4), 439–456.
[53] Rocchio, J. J., 1971. Relevance Feedback in Information Retrieval. In: [57], pp. 313–323.
[54] Rowe, B. R., Wood, D. W., Link, A. L., Simoni, D. A., July 2010. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. RTI Project Number 0211875, RTI International, USA. http://trec.nist.gov/pubs/2010.economic.impact.pdf.
[55] Sakai, T., 2006. Evaluating Evaluation Metrics based on the Bootstrap. In: Efthimiadis, E. N., Dumais, S., Hawking, D., Järvelin, K. (Eds.), Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006). ACM Press, New York, USA, pp. 525–532.
[56] Sakai, T., 2007. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. Information Processing & Management 43 (2), 531–548.
[57] Salton, G. (Ed.), 1971. The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice–Hall, Inc., Englewood Cliff, New Jersey, USA.
[58] Sanderson, M., Zobel, J., 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In: Baeza-Yates, R., Ziviani, N., Marchionini, G., Moffat, A., Tait, J. (Eds.), Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005). ACM Press, New York, USA, pp. 162–169.
[59] Savoy, J., 1997. Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management 33 (44), 495–512.
[60] Savoy, J., 2007. Why do Successful Search Systems Fail for Some Topics. In: Cho, Y., Wan Koo, Y., Wainwright, R. L., Haddad, H. M., Shin, S. Y. (Eds.), Proc. 2007 ACM Symposium on Applied Computing (SAC 2007). ACM Press, New York, USA, pp. 872–877.
[61] Seo, J., Shneiderman, B., Jul. 2005. A rank-by-feature framework for interactive exploration of multidimensional data. Information Visualization 4 (2), 96–113.
URL http://dx.doi.org/10.1057/palgrave.ivs.9500091
[62] Sormunen, E., 2002. Liberal Relevance Criteria of TREC – Counting on Negligible Documents? In: [36], pp. 324–330.
[63] Sormunen, E., 2002. Liberal Relevance Criteria of TREC: Counting on Negligible Documents? In: Proc. of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, pp. 324–330.
[64] Sormunen, E., Hokkanen, S., Kangaslampi, P., Pyy, P., Sepponen, B., 2002. Query Performance Analyser – a Web-based tool for IR research and instruction. In: [36], p. 450.
[65] Teevan, J., Dumais, S. T., Horvitz, E., 2010. Potential for personalization. ACM Transactions on Computer-Human Interaction (TOCHI) 17 (1), 1–31.
[66] Tukey, J. W., 1970. Exploratory Data Analysis, preliminary edition Edition. Addison-Wesley, USA.
[67] Tukey, J. W., 1977. Exploratory Data Analysis. Addison-Wesley, USA.
[68] Voorhees, E., 2001. Evaluation by Highly Relevant Documents. In: Kraft, D. H., Croft, W. B., Harper, D. J., Zobel, J. (Eds.), Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001). ACM Press, New York, USA, pp. 74–82.
[69] Voorhees, E. M., Harman, D. K. (Eds.), 1998. The Seventh Text REtrieval Conference (TREC-7), Gaithersburg, Maryland, November 09-11, 1998. Vol. Special Publication 500-242. National Institute of Standards and Technology (NIST).
[70] Vredenburg, K., Mao, J.-Y., Smith, P. W., Carey, T., 2002. A survey of user-centered design practice. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '02. ACM, New York, NY, USA, pp. 471–478.
URL http://doi.acm.org/10.1145/503376.503460
[71] Witten, I. H., Bainbridge, D., Nichols, D. M., 2009. How to Build a Digital Library, 2nd Edition. Morgan Kaufmann Publishers, San Francisco (CA), USA.
[72] Zhang, J., 2001. TOFIR: A Tool of Facilitating Information Retrieval - Introduce a Visual Retrieval Model. Inf. Process. Manage. 37 (4), 639–657.
[73] Zhang, J., 2008. Visualization for Information Retrieval. Springer-Verlag, Heidelberg, Germany.