Interactive Analysis and Exploration of Experimental Evaluation Results

Emanuele Di Buccio
University of Padua, Italy
dibuccio@dei.unipd.it

Marco Dussin
University of Padua, Italy
dussinma@dei.unipd.it

Nicola Ferro
University of Padua, Italy
ferro@dei.unipd.it

Ivano Masiero
University of Padua, Italy
masieroi@dei.unipd.it

Giuseppe Santucci
Sapienza University of Rome, Italy
santucci@dis.uniroma1.it

Giuseppe Tino
Sapienza University of Rome, Italy
tino@dis.uniroma1.it

ABSTRACT

This paper proposes a methodology based on discounted cumulated gain measures and visual analytics techniques in order to improve the analysis and understanding of IR experimental evaluation results. The proposed methodology is geared to favour a natural and effective interaction of the researchers and developers with the experimental data and it is demonstrated by developing an innovative application based on Apple iPad.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: [Search process]; H.3.4 [**Systems and Software**]: [Performance evaluation (efficiency and effectiveness)]

General Terms

Experimentation, Human Factors, Measurement, Performance

Keywords

Ranking, Visual Analytics, Interaction, Discounted Cumulated Gain, Experimental Evaluation, DIRECT

1. INTRODUCTION

The Information Retrieval (IR) field has a strong and long-lived tradition, that dates back to late 50s/early 60s of the last century, as far as the assessment of the performances of an IR system is concerned. In particular, in the last 20 years, large-scale evaluation campaigns, such as the Text REtrieval Conference (TREC)[1] in the United States and the Cross-Language Evaluation Forum (CLEF)[2] in Europe, have conducted cooperative evaluation efforts involving hundreds of research groups and industries, producing a huge amount of valuable data to be analysed, mined, and understood.

The aim of this work is to explore how we can improve the comprehension of and the interaction with the experimental results by IR researchers and IR system developers. We imagine the following scenarios: (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 understand the experimental results as s/he is listening at 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.

These scenarios call for: (a) proper metrics that allow us to understand the behaviour of a system; (b) effective analysis and visualization techniques that allow us to get an overall idea of the main factors and critical areas which have influenced performances in order to be able to dig into details; (c) for tools and applications that allow us to interact with the experimental result in a both effective and natural way.

To this end, we propose a methodology which allows us to quickly get an idea of the distance of an IR system with respect to both its own optimal performances and the best performances possible. We rely on the (normalized) *discounted cumulated gain* (n)DCG family of measures [7] because they have shown to be especially well-suited not only to quantify system performances but also to give an idea of the overall user satisfaction with a given ranked list considering the persistence of the user in scanning the list.

The contribution of this paper is to improve on the previous work [7,11] by trying to better understand what happens when you flip documents with different relevance grades in a ranked list. This is achieved by providing a formal model that allows us to properly frame the problem and quantify the gain/loss with respect to an optimal ranking, rank by rank, according to the actual result list produced by an IR system.

The proposed model provides the basis for the development of Visual Analytics (VA) techniques that give us the possibility to get a quick and intuitive idea of what happened in a result list and what determined its perceived performances. Visual Analytics [8, 10, 14] is an emerging multi-disciplinary area that takes into account both ad-hoc and classical Data Mining (DM) algorithms and Informa-

[1] `http://trec.nist.gov/`
[2] `http://www.clef-campaign.org/`

Copyright ©2011 for the individual papers by the papers' authors. Copying permitted only for private and academic purposes. This volume is published and copyrighted by the editors of euroHCIR2011.

tion Visualization (IV) techniques, combining the strengths of human and electronic data processing. Visualisation becomes the medium of a semi-automated analytical process, where human beings and machines cooperate using their respective distinct capabilities for the most effective results. Decisions on which direction analysis should take in order to accomplish a certain task are left to final user. While IV techniques have been extensively explored [4, 13], combining them with automated data analysis for specific application domains is still a challenging activity [9]. Moreover, the Visual Analytics community acknowledges the relevance of interaction for visual data analysis, and that the current research activities very often focus only on visual representation, neglecting the interaction design, as clearly stated in [14]. This refers to two different typologies of interaction: 1) interaction within a visualization and, 2), closer to the paper contribution, interaction between the visual and the analytical components.

The idea of exploring and applying VA techniques to the experimental evaluation in the IR field is quite innovative since it has never been attempted before and, due to the complexity of the evaluation measures and the amount of data produced by large-scale evaluation campaigns, there is a strong need for better and more effective representation techniques. Moreover, visualizing and assessing ranked list of items, to the best of the authors' knowledge, has not been addressed by the VA community. The few related proposals, see, e.g., [12], use rankings for presenting the user with the most relevant visualizations, or for browsing the ranked result, see, e.g., [5], but do not deal with the problem of observing the ranked item position, comparing it with an ideal solution, to assess and improve the ranking quality. A first attempt in such a direction is in [1], where the authors explored the basic issues associated with the problem, providing basic metrics and introducing a VA web based system that allows for exploring the quality of a ranking with respect to an optimal solution.

On top of the proposed model, we have built a running prototype where the experimental results and data are stored in a dedicated system accessible via standard Web services. This allows for the design and development of various client applications and tools for exploiting the managed data. In particular, in this paper, we have started to explore the possibility of adopting the Apple iPad[3] as an appropriate device to allow users to easily and naturally interact with the experimental data and we have developed an initial prototype that allows us for interactively inspecting the actual experimental data in order to get insights about the behaviour of a IR system.

Overall, the proposed model, the proposed visualization techniques, and the implemented prototype meet all the (a-c) requirements for the two scenarios introduced above.

The paper is organized as follows. Section 2 introduces the model underlying the system together with its visualization techniques; Section 3 describes the interaction strategies of the system, Section 4 describes the overall architecture of the system, and Section 5 concludes the paper, pointing out ongoing research activities.

2. THE PROTOTYPE

According to [7] we model the retrieval results as a ranked vector of n documents V, i.e., $V[1]$ contains the identifier of the document predicted by the system to be most relevant, $V[n]$ the least relevant one. The ground truth GT function assigns to each document $V[i]$ a value in the relevance interval $\{0..k\}$, where k represents the highest relevance score, e.g. $k = 3$ in [7]. The basic assumption is that the greater the position of a document the less likely it is that the user will examine it, because of the required time and effort and the information coming from the documents already examined. As a consequence, the greater the rank of a relevant document the less useful it is for the user. This is modeled through a discounting function DF that progressively reduces the relevance of a document, $GT(V[i])$ as i increases:

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

The quality of a result can be assessed using the discounted cumulative gain function $DCG(V,i) = \sum_{j=1}^{i} DF(V[j])$ that estimates the information gained by a user that examines the first i documents of V.

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

However, if the user's task is to improve the ranking performance of a single search engine, looking at the misplaced documents (i.e., ranked too high or too low with respect to the other documents) the DCG function does not help: the same value $DCG(i)$ could be generated by different permutations of V and it does not point out the loss in cumulative gain caused by misplaced elements. To this aim, we introduce the following definitions and novel metrics.

We denote with $OptPerm(V)$ the set of optimal permutations of V such as that $\forall OV \in OptPerm(V)$ it holds that $GT(OV[i]) \geq GT(OV[j]) \forall i,j <= n \bigwedge i < j$, that is, OV maximizes the values of $DCG(OV,i) \forall i$. In other words, $OptPerm(V)$ represents the set of the optimal rankings for a given search result. It is worth noting that each vector in $OptPerm(V)$ is composed by $k+1$ intervals of documents sharing the same GT values. As an example, assuming a result vector composed by 12 elements and $k = 3$, a possible sequence of GT values of an optimal vector OV is <3,3,3,3,2,2,2,2,1,1,0,0>; according to this we define the $max\_index(V,r)$ and $min\_index(V,r)$ functions, with $0 \leq r \leq k$, that return the greatest and the lowest indexes of elements in a vector belonging to $OptPerm(V)$ that share the same GT value r. As an example, considering the above 12 GT values, $min\_index(V,2) = 5$ and $max\_index(V,2) = 8$.

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

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

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

According to the actual relevance and rank position, the same value of $R\_Pos(V[i])$ can produce different variations of the DCG function. We measure the contributions of mis-

[3] http://www.apple.com/ipad/

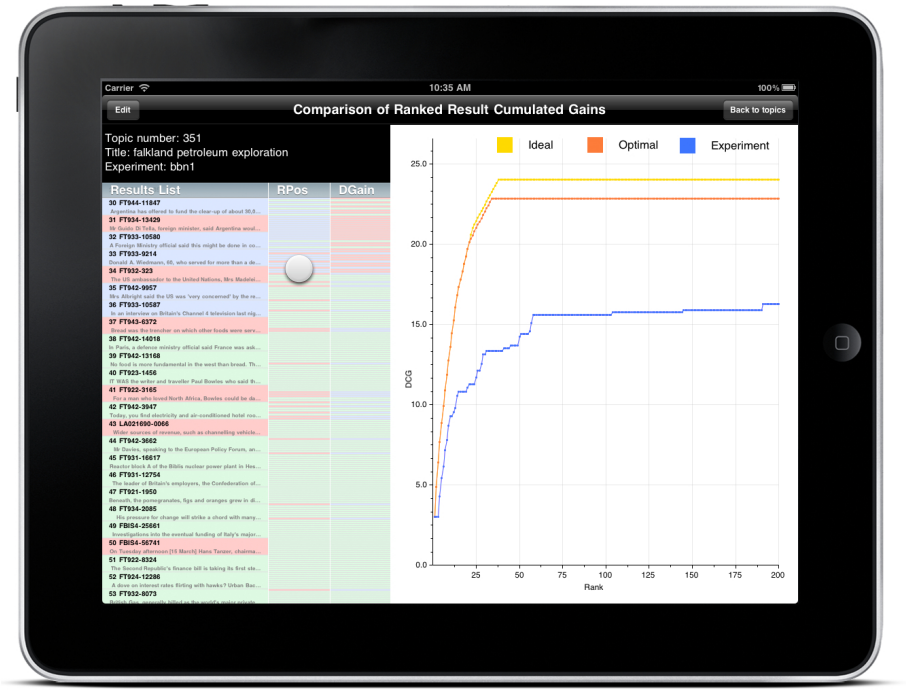

Figure 1: The iPad prototype interface.

placed elements with the function $\Delta\_Gain(V, i)$ that compares $\forall i$ the actual values of $DF(V[i])$ with the corresponding values in OV, $DF(OV[i])$: $\Delta\_Gain(V, i) = DF(V[i]) - DF(OV[i])$.

3. INTERACTION

A multi-touch prototype interface based on the model presented in section 2 has been designed for the iPad device. It has been developed and tested on the iOS 4.2[4] with the integration of the Core Plot[5] plotting framework for the graphical visualization of data. The interface allows the end user for comparing the curve of the ranked results, for a given experiment/topic, with the optimal one and with the ideal one. This facilitates the activities of failure analysis, easily locating misplaced elements, blue or red items, that pop up from the visualization together with the extent of their displacement and the impact they have on *DCG*.

Figure 1 shows a screenshot of the current interface: the main list on the left represents the top $n = 200$ ranked result for a given experiment/topic and it can be easily scrolled by the user. Each row corresponds to a document ID, a short snippet of the content is included in the subtitle of each cell and more information on a specific result (i.e. relevance score, *DCG*, *R_Pos*, Δ*_Gain*) can be viewed by touching the row. On the right side there are two coloured vectors which show the *R_Pos* and Δ*_Gain* functions. The *R_Pos* vector presents the results using different color shadings: light green, light red and light blue respectively for documents that are within, below and above the optimal interval. It allows for locating misplaced documents and, thanks to the shading, understanding how they are far from the optimal position. Similarly, the Δ*_Gain* vector codes the function using colors: light blue refers to positive values, light red codes negative values, and green 0 values. Moreover, if the user touches a specific area of the *R_Pos* vector (that is simulated by the gray round in Figure 1), the main results list automatically scrolls back, providing the end user with a detailed view on the corresponding documents. The rightmost part of the screen shows the *DCG* graphs of the ideal, the optimal and the experiment vector, i.e. the ranking curves. The navigation bar displays a back button on the right which let the user visualize the results for a different topic.

4. ARCHITECTURE

The design of the architecture of the system benefits from what has been learned in ten years of work for the CLEF and in the design and implementation of Distributed Information Retrieval Evaluation Campaign Tool (DIRECT), the system developed in CLEF since 2005 to manage all the aspects of an evaluation campaign [2, 3].

The approach to the architecture is the implementation of a modular design, as sketched in Figure 2, with the aim to clearly separate the logic entailed by the application into three levels of abstraction – data, application, and interface logic – able to reciprocally communicate, easily extensible and implementable using modular and reusable components. The *Data Logic* layer, depicted at the bottom of Figure 2, deals with the persistence of the information coming from the other layers. From the implementation point of view, data stored into databases and indexes are mapped to resources and communicate with the upper levels through the mechanism granted by the Data Access Object (DAO) pattern[6] — see point *(1)* in Figure 2. The *Application Logic*

[4] `http://developer.apple.com/`
[5] `http://code.google.com/p/core-plot/`
[6] `http://java.sun.com/blueprints/corej2eepatterns/`

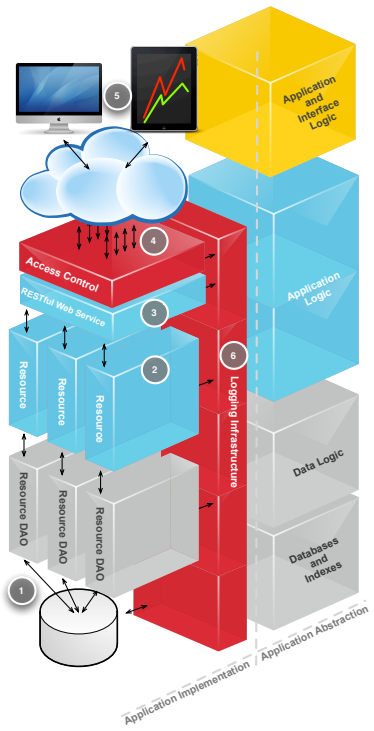

Figure 2: The Architecture of the Application.

layer is in charge of the high-level tasks made by the system, such as the enrichment of raw data, the calculation of metrics and the carrying out of statistical analyses on experiments. These resources *(2)* are therefore accessible via HTTP through a RESTful Web service [6], sketched at point *(3)*. After the validation of credentials and permissions made by the access control mechanism *(4)*, it is possible for remote devices such as web browsers or custom clients *(5)* to create, modify, or delete resources attaching their representation in XML[7] or JSON[8] format to the body of an HTTP request, and to read them as response of specific queries. A logging infrastructure *(6)* grants the tracking of all the activities made at each layer and can be used to obtain information about the provenance of all the managed resources.

5. CONCLUSIONS

We have presented a model and a prototype which allow users to easily interact with the experimental results and to work together in a cooperative way while actually accessing the data. This first step uncovers new and interesting possibilities for the experimental evaluation and for the way in which researchers and developers usually carry out such activities. For example, the proposed techniques may alleviate the burden of certain tasks, such as failure analysis, which are often overlooked due to their demanding nature, thus making easier and more common to perform them and, as a consequence, improving the overall comprehension of system behaviour. This will be explored in the future work.

Patterns/DataAccessObject.html
[7]http://www.w3.org/XML/
[8]http://www.ietf.org/rfc/rfc4627.txt

Acknowledgements

The work reported in this paper has been partially supported by the PROMISE network of excellence (contract n. 258191), as a part of the 7th Framework Program of the European commission (FP7/2007-2013).

6. REFERENCES

[1] N. Ferro, A. Sabetta, G. Santucci, G. Tino, and F. Veltri. Visual comparison of ranked result cumulated gains. In *Proc. of EuroVA 2011*. Eurographics, 2011.

[2] M. Agosti, G. Di Nunzio, M. Dussin, and N. Ferro. 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go. In *Proc. of EVIA 2010*, pages 16–24. Tokyo, Japan, 2010.

[3] M. Agosti and N. Ferro. Towards an Evaluation Infrastructure for DL Performance Evaluation. In *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods*. Chandos Publishing, Oxford, UK, 2009.

[4] S. K. Card and J. Mackinlay. The structure of the information visualization design space. In *Proc. of InfoVis '97*, pages 92–99, Washington, DC, USA, 1997. IEEE Computer Society.

[5] M. Derthick, M. G. Christel, A. G. Hauptmann, and H. D. Wactlar. Constant density displays using diversity sampling. In *Proc. of the IEEE Information Visualization*, pages 137–144, 2003.

[6] R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. *ACM TOIT*, 2:115–150, 2002.

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

[8] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Information visualization. chapter Visual Analytics: Definition, Process, and Challenges, pages 154–175. Springer-Verlag, Berlin, Heidelberg, 2008.

[9] D. Keim, J. Kohlhammer, G. Santucci, F. Mansmann, F. Wanner, and M. Schäfer. Visual Analytics Challenges. In *Proc. of eChallenges 2009*, 2009.

[10] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proc. of IV'06*, pages 9–16, 2006.

[11] H. Keskustalo, K. Järvelin, A. Pirkola, and J. Kekäläinen. Intuition-Supporting Visualization of User's Performance Based on Explicit Negative Higher-Order Relevance. In *Proc. of SIGIR '08*, pages 675–681. ACM Press, NY, USA, 2008.

[12] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. In *Proc. of the IEEE Information Visualization*, pages 65–72, 2004.

[13] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proc. of the 1996 IEEE Symposium on Visual Languages*, pages 336 –343, 1996.

[14] J. J. Thomas and K. A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26:10–13, 2006.