Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 38 (2014) 133 – 137

Procedia Computer Science

10th Italian Research Conference on Digital Libraries, IRCDL 2014

Measuring and Analyzing the Scholarly Impact of Experimental Evaluation Initiatives

Marco Angelini[a], Nicola Ferro[b], Birger Larsen[c], Henning Müller[d], Giuseppe Santucci[a], Gianmaria Silvello[b,*], Theodora Tsikrika[e]

[a] *"La Sapienza" University of Rome, Via Ariosto 25, Rome 00185, Italy*
[b] *University of Padua, Via Gradenigo 6/b, Padua 35131, Italy*
[c] *Aalborg University, A.C. Meyers Vnge 15, Copenhagen 2450, Denmark*
[d] *University of Applied Sciences Western Switzerland (HES-SO), Techno-Ple 3, Sierre 3960, Switzerland*
[e] *Centre for Research and Technology Hellas, 6th km Charilaou-Thermi Road, Thermi-Thessaloniki 57001, Greece*

Abstract

Evaluation initiatives have been widely credited with contributing highly to the development and advancement of information access systems, by providing a sustainable platform for conducting the very demanding activity of comparable experimental evaluation in a large scale. Measuring the impact of such benchmarking activities is crucial for assessing which of their aspects have been successful, which activities should be continued, enforced or suspended and which research paths should be further pursued in the future. This work introduces a framework for modeling the data produced by evaluation campaigns, a methodology for measuring their scholarly impact, and tools exploiting visual analytics to analyze the outcomes.

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
Keywords: Scholarly Impact, Experimental Evaluation, Experimental Data, Visual Analytics

1. Motivations

Experimental evaluation is a fundamental methodology adopted in Information Retrieval (IR) since its inception, which substantially contributed to the scientific advancements of the field. It is based on the Cranfield methodology [1] which makes use of shared experimental collections in order to create comparable experiments and evaluate the performances of different information access systems.

Evaluation activities are very demanding both from the technical and economical point-of-views [2] and to be sustainable and scalable they have been carried out in large-scale evaluation campaigns such as Text REtrieval Conference (TREC) in the United States (`http://trec.nist.gov/`), the Conference and Labs of the Evaluation Forum (CLEF) in Europe (`http://www.clef-initiative.eu/`), and the NII Testbeds and Community for Information access Research (NTCIR) in Asia (`http://research.nii.ac.jp/ntcir/index-en.html`). In order to further facilitate their organization and management, each campaign is usually divided into *tracks* (referred to as *labs* in CLEF) and *tasks*. A lab is an area of focus concentrating on a specific evaluation aspect of a particular domain; for

* Corresponding author. Tel.: +39 049 827 7929; fax: +39 049 827 7799. E-mail address: silvello@dei.unipd.it

Peer-review under responsibility of the Scientific Committee of IRCDL 2014
doi:10.1016/j.procs.2014.10.022

instance, CLEF in 2013 was organized into nine labs comprising, for instance, the "Cross Language Image Annotation and Retrieval" (ImageCLEF) lab (`http://www.imageclef.org/`) concentrating on the experimental evaluation of image classification and retrieval. Each lab may be divided into tasks, each focusing on specific sub-problems concerning the scope of the lab; as an example, ImageCLEF 2013 had four tasks comprising, among the others, the "Photo Annotation and Retrieval" task aimed at studying visual concept detection, annotation and retrieval in the context of diverse collections.

Despite the general agreement about the importance of evaluation campaigns and the experimental data (e.g., collections, measures and statistics) produced by them[3] and the scientific production based on them, no shared methodology for measuring their scientific impact has already been defined. Such a methodology is much needed since measuring the impact of evaluation campaigns is crucial for assessing which aspects have been successful, and thus obtain guidance for the development of improved evaluation methodologies and information access systems.

Given that the contribution of an evaluation campaign to the field is mainly indicated by the research that would otherwise not have been possible, it is reasonable to consider that their success can be measured, to some extent, by the scholarly impact of the research they foster. Previous works aimed at measuring the scholarly impact of evaluation campaigns[3,4] had relied on the scientific production (i.e., the publications relying on datasets created by the evaluation campaigns), but that they did not have a model behind it, they did not make a connection to the experimental data and most of the analyses have been conducted manually. The goal of this work is to introduce the main aspects of a methodology allowing for modeling the experimental data and scientific production related to them. As consequence we propose a new way for measuring the scholarly impact of evaluation campaigns by exploiting (semi) automatic techniques and analyzing the outcomes by means of visual analytics techniques. This paper introduces a methodology that can be adopted for performing joint analysis of experimental data and scientific production and for exploiting advanced data mining algorithms to determine new insights from the data and to characterize the success of evaluation campaigns. Furthermore, we present a first analysis of CLEF evaluations campaigns which shows how the proposed methodology can be employed and how the impact of an evaluation campaign can be measured.

To this purpose, Section 2 presents the bibliographical area of the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) system[5] which is a comprehensive system allowing for managing the experimental data of IR evaluation, providing advanced services over them and defining explicit connections between campaigns and the data produced by them. Section 3 outlines the three main steps to be followed for measuring the scholarly impact of evaluation campaigns. Section 4 shows how the results of the impact analysis can be analyzed through visual analytics tools using the outcomes of the study conducted on CLEF as a use case. Finally, Section 5 draws some final remarks.

2. Modeling Experimental Data of IR evaluation and Scientific Production

The necessity of modeling experimental data and designing a software infrastructure to manage and curate them, led to the development of a rather complex system – i.e., DIRECT – covering all the aspects of experimental evaluation. In this paper we focus on the bibliographical area of DIRECT which is responsible for retaining the relationships between the experimental data and the scientific production based on these data. Furthermore, this area models the bibliometrics (e.g., number of citations, h-index and impact factor) that are used for assessing the impact of evaluation campaigns.

Figure 1 depicts the conceptual schema of the bibliographic area. The central entity is `Contribution` which refers to a published piece of writing; a conference or a workshop paper, a journal article, a book, a technical report, a thesis or a manual are examples of contributions.

Each `Contribution` is associated to a `Concept` that defines its type; e.g., a `Contribution` can be a generic `Publication`, a `Working Note`, or a `Journal`. In general, `Concept` is defined as an idea or notion, a unit of thought; it is used to define the type of relationships in a semantic environment or to create a vocabulary (e.g., contribution types) and it resembles the idea of concept introduced by Simple Knowledge Organization System (SKOS)[6].

Furthermore, each `Contribution` is associated to no, one or more authors (i.e., `User`) via the `Author` relationship and can be described by no, one or more `Metadata` via the `describe` relationship. Similar relationships exist also between `Contribution` and `Task`, `Track` (i.e., a lab in CLEF) and `Campaign` and allow us to explicitly relate contributions with the experimental data.

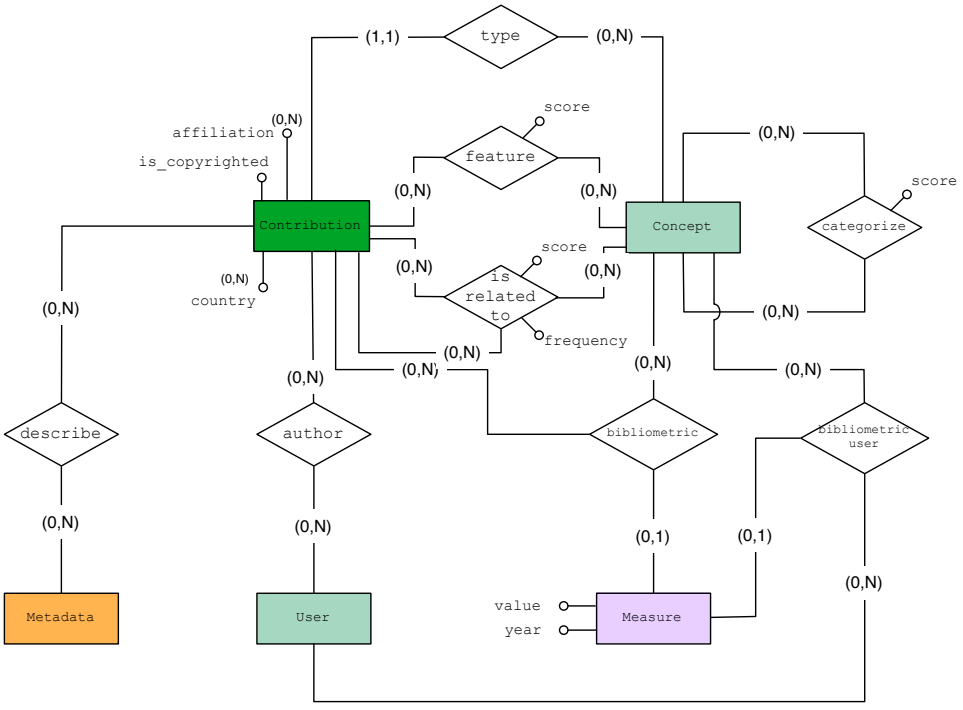

Figure 1. Bibliographical Area relationships

The relationship `feature` relates a `Contribution` to a `Concept` which defines its topic; this allows us to determine the topics of a `Contribution` and its relevance for a given topic. As a consequence, a `Contribution` can feature a `Concept` – e.g., "Digital Library" – and given that contributions are related to experimental data, we can conduct topic-oriented analyses on them; for instance, we can calculate the extent to which a task or a campaign is relevant to the topic "Digital Library". Also the relationship `is related to` is relevant from the scholarly impact point-of-view, because it allows analyses based on the number of citations of a contribution. Indeed, we can say that "Contribution A `cites` Contribution B" where `cites` is a `Concept` relating "Contribution A" with "Contribution B".

Finally, the relationship `bibliometric` relates a `Contribution` to a `Concept` and a `Measure`. This allows us to say that "Contribution A has `impact` of `1.3`"; `impact` is defined as a `Concept` and `1.3` as the value of a `Measure` (e.g., the contribution received 1.3 times as many citations as expected in relation to a given baseline). The relationship `bibliometric user` has the same purpose but oriented to `User` (i.e., author); indeed, through this relationship we can express something like "User A has `h-index` 3", where `h-index` is a `Concept` and 3 is the value of a `Measure`. The entities and relations between them also allow for aggregate indicators to be calculated. For instance, impact factors for a set of contributions in a given time period can be extracted, such as average number of citations per paper up to three years after publication for each of the tracks in an evaluation campaign.

3. Three Steps for Measuring the Scholarly Impact

Starting from the above described model we can conduct bibliometric studies providing a quantitative and qualitative indication of the scholarly impact of a research activity by examining the number of publications derived from it and the number of citations these publications receive. Such studies can be conducted by following these three main steps: (i) Publication data collection; (ii) Citation data collection; (iii) Data analysis.

So, the first step for assessing the scholarly impact of an evaluation campaign is to identify the publications associated with it and collect them in a dataset so that their citation data can then be obtained and analysed.

The second step involves the selection of citation data sources; the most comprehensive are: Thomson Reuters (formerly ISI) Web of Knowledge(`http://wokinfo.com/`), Scopus(`http://www.scopus.com/`) and Google Scholar(`http://scholar.google.com/`). Each of these sources follows a different data collection policy that affects both the publications covered and the number of citations found. Once the citation data sources have been selected, the next step is to query them using the publication data as input so as to obtain the citation data.

The last step regards the analyses that can be performed; they can be along several axes, such as the types of publications and the labs and tasks comprising the evaluation campaign while also drilling down the data into time dimension.

4. An Initial Analysis of the Results via Visual Analytics Tools

The three steps depicted above have been applied to the CLEF (2000-2009) Proceedings publications and to the CLEF (2000-2009) Working Notes publications and detailed results are described in[4,7]. For this study, the relationships between experimental data of IR evaluation and contributions retained by DIRECT allowed us to calculate the measures determining the impact of evaluation initiatives; for instance, it emerged that three labs – i.e., Adhoc, ImageCLEF, and QA@CLEF – clearly dominate in terms of publication and citation numbers and thus have the higher scholarly impact.

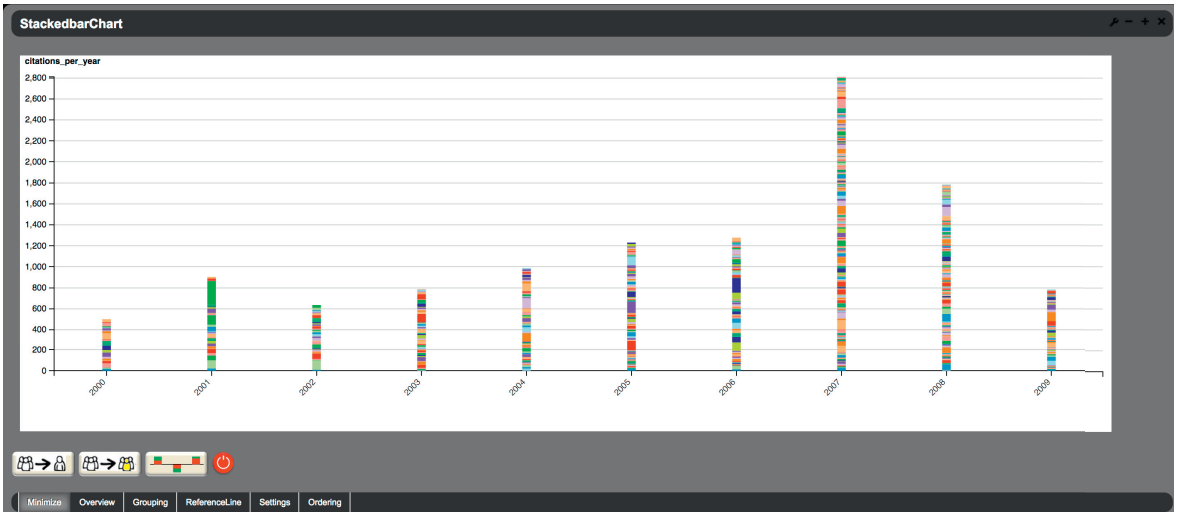

Figure 2. A screen-shot of a part of the interactive visual environment for analysing the results of impact analysis of CLEF.

These conclusions have been drawn thanks to visual analytics tools offered by DIRECT. Figure 2 presents a screen-shot of a part of the visual environment developed for conducting impact analysis. This figure reports the stacked bar chart depicting the number of citations for the CLEF labs and tasks over the years (2000-2009). Each color in the bars represents the number of citations received by the tasks belonging to a specific campaign. This environment allows also for selecting specific tasks and comparing their measures, zooming and highlighting parts of the graphs and to compare citation numbers with other bibliometrics such as the h-index of authors and the impact of publication venues.

By using the analytics possibilities offered by the DIRECT visual environment it is also possible to identify some trends over all labs and tasks; for instance, in many cases there appears to be a peak in their second or third year of operation, followed by a decline[4]. Exceptions include the “Photo Annotation and Retrieval” task of ImageCLEF, which attracted significant interest in its fourth year when it employed a new collection and adopted new evaluation methodologies. Such novel aspects result in renewed interest in labs and tasks, and also appear to strengthen their impact.

5. Final Remarks

In this work we present a general framework for modeling the experimental data and their relationships with scientific contributions. This model sets the ground for calculating bibliometrics to be used for assessing the impact of evaluation activities. We have also introduced the three main steps to be followed for measuring impact starting from scientific publications. Finally, we have shown how visual and interactive tools can be used for conducting impact analysis.

The presented methodology represents the first effort for modelling experimental data and their related scientific contributions. In this paper we show a possible application of the methodology for measuring the impact of some CLEF evaluation labs, but the methodology is general enough to be applied to other evaluation campaigns, tracks and tasks. Furthermore, there is room for applying advanced data mining algorithms and other data analysis tools for measuring the success and the impact of evaluation activities.

Future work will focus on the design and development of more advanced visualizations to interact with and explore the scholarly impact data, as well as improving the automation in gathering and cleaning of further bibliographic data in order to carry out deeper analyses. We also plan to map the presented conceptual model into an RDF schema in order to enable experimental data and their relationships with scientific contributions to be exposed as Linked Data on the Web; this will allow us to reuse existing bibliographic vocabularies and to establish meaningful connections with external datasets and DL as well as to improve interoperability with existing databases.

Acknowledgements

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

References

1. C. W. Cleverdon, The Cranfield Tests on Index Languages Devices, in: K. Spärck Jones, P. Willett (Eds.), Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, CA, USA, 1997, pp. 47–60.
2. B. R. Rowe, D. W. Wood, A. L. Link, D. A. Simoni, 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, 2010.
3. C. V. Thornley, A. C. Johnson, A. F. Smeaton, H. Lee, The Scholarly Impact of TRECVid (2003–2009), Journal of the American Society for Information Science and Technology (JASIST) 62 (4) (2011) 613–627.
4. T. Tsikrika, B. Larsen, H. Müller, S. Endrullis, E. Rahm, The Scholarly Impact of CLEF (2000-2009), in: Proc. of Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Vol. 8138 of LNCS, Springer, 2013, pp. 1–12.
5. M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, G. Silvello, DIRECTions: Design and Specication of an IR Evaluation Infrastructure, in: Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012), LNCS 7488, Springer, Germany, 2012, pp. 88–99.
6. W3C, SKOS Simple Knowledge Organization System Reference – W3C Recommendation 18 August 2009, http://www.w3.org/TR/skos-reference (August 2009).
7. T. Tsikrika, B. Larsen, G. Bordea, P. Buitelaar, Deliverable D6.4 – Report on the impact analysis for the CLEF initiative, PROMISE Network of Excellence, EU 7FP, Contract N. 258191. http://www.promise-noe.eu/documents/10156/9d42701f-7d2f-4450-b6a8-5dec5444a757 (August 2013).