Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 740–749

Procedia Computer Science

Harnessing the Scientific Data Produced by the Experimental Evaluation of Search Engines and Information Access Systems

Nicola Ferro[a,*], Allan Hanbury[b], Henning Müller[c], Giuseppe Santucci[d]

[a] *University of Padua, Italy*
[b] *Information Retrieval Facility (IRF), Austria*
[c] *University of Applied Sciences Western Switzerland, Switzerland*
[d] *Sapienza University of Rome, Italy*

Abstract

Measuring is a key to scientific progress. This is particularly true for research concerning complex systems, whether natural or human-built. Multilingual and multimedia information systems are increasingly complex: they need to satisfy diverse user needs and support challenging tasks. Their development calls for proper evaluation methodologies to ensure that they meet the expected user requirements and provide the desired effectiveness. In the process, vast amounts of experimental data are generated that beg for analysis tools to enable interpretation and thereby facilitate scientific and technological progress. These scientific data are at the basis of the research and publications in the field and can be better exploited and linked to the scientific literature and production.

We are building a software infrastructure, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*, to manage, curate, enrich, and make these scientific data online accessible and we discuss how this infrastructure can be exploited to directly link the experimental data into papers and publications describing them.

Keywords: information retrieval, multilingual information access, experimental evaluation, large-scale evaluation campaigns, scientific data, DIRECT

1. Motivation and Objectives

Search Engines (SEs) and, more in general, *Information Retrieval Systems (IRSs)* are key systems (or components of wider information systems) in the today's information society to provide access to pertinent and relevant information and to overcome the information overload each of us is exposed to nowadays. Information access systems are becoming increasingly complex: they need to satisfy user needs and carry out tasks that require to cross language and media barriers; moreover, they have to manage increasing amounts of information which is often heterogeneous and demands insightful access to it. If we are to continue advancing the state-of-the-art in information access technologies, we need to understand a new breed of users who need to be able to co-operate and communicate in a way which crosses language and media boundaries and goes beyond separate search in diverse media/languages, but which exploits the interactions between different languages and media [1].

*corresponding author
Email addresses: ferro@dei.unipd.it (Nicola Ferro), a.hanbury@ir-facility.org (Allan Hanbury), henning.mueller@hevs.ch (Henning Müller), santucci@dis.uniroma1.it (Giuseppe Santucci)

1877–0509 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.078

1877–0509 © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Prof. Mitsuhisa Sato and Prof. Satoshi Matsuoka
doi:10.1016/j.procs.2011.04.078

We consider experimental evaluation – both laboratory and interactive – a key means for supporting and fostering the development of multilingual and multimedia information retrieval systems which are more adherent to the new user needs in order to ensure that they meet the expected user requirements, provide the desired effectiveness and efficiency, guarantee the required robustness and reliability, and operate with the necessary scalability.

Moreover, experimental evaluation is an essential part of scientific work and scientific publishing. Relying on the same data sets and same evaluation scenarios, systems can be compared and performances can be better understood. Such evaluation can make publications comparable as well and allows new systems to be compared to the best state of the art and not outdated techniques. This can also be on a component level and not only for entire systems [2, 3].

In this context, large-scale evaluation initiatives provide a significant contribution to the advancement in research and state-of-the-art, industrial innovation in a given domain, and building of strong research communities. They rely mainly on the traditional Cranfield methodology [4] which makes use of shared experimental collections in order to create comparable experiments and evaluate their performance. Experimental collections are made up of documents, topics simulating user information needs, and relevance judgements specifying which documents are relevant to which topics. Relevant and long-lived examples from the information retrieval field are the *Text REtrieval Conference (TREC)*[1] in the United States, the *Cross-Language Evaluation Forum (CLEF)*[2] in Europe, and the *NII-NACSIS Test Collection for IR Systems (NTCIR)*[3] in Japan and Asia. Moreover, new initiatives are growing to support emerging communities and address specific issues, such as the *Forum for Information Retrieval Evaluation (FIRE)*[4] in India. Over the years, they provided qualitative and quantitative evidence as to which methods give the best results in certain key areas, such as indexing techniques, relevance feedback, multilingual querying, and results merging, and so on. Moreover, as reported by [5, p. ES-9], "for every $1 that NIST and its partners invested in TREC, at least $3.35 to $5.07 in benefits accrued to IR researchers. The internal rate of return (IRR) was estimated to be over 250% for extrapolated benefits and over 130% for unextrapolated benefits".

Figure 1 shows the typical cycle of an evaluation campaign: organizers and assessors prepare document collections and topics; then, researchers and developers run their systems on the provided documents and topics and produce their experiments and result lists which are then sampled and pooled in order to produce the relevance judgments; at this point, performance measures, descriptive statistics, and statistical analyses are computed to evaluate the performances of each system and compare the proposed solutions. All of this information is then used for feeding the scientific production and the design and development of next generation systems.

During their life-span, large-scale evaluation campaigns have produced a huge amount of scientific data which are extremely valuable. These experimental and scientific 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 the great amount of effort devoted to their production: [5, p. ES-10] estimates in about 30 million dollars the overall investment in TREC.

Nevertheless, much less attention has been paid over the years to the modelling, management, curation, and access to the produced scientific data, even if the importance of scientific data in general has been highlighted also by many different institutional organizations, such the European Commission [6], the US National Scientific Board [7], and the Australian Working Group on Data for Science [8].

Therefore, the overall goal of our work is to deliver a unified infrastructure and environment collecting data, knowledge, tools, methodologies, and the user community in order to advance the experimental evaluation of complex multimedia and multilingual information systems and support individuals, commercial entities, and communities who design, develop, employ, and improve such complex systems. Part of this wider goal and especially relevant to the executable paper grand challenge, is the possibility of relying on the developed infrastructure to improve the link and the exploitation of the performance measure and analyses in the related scientific literature and production.

1.1. Show Case: Intellectual Property Search

A patent is a complex legal document which is granted by a state to allow an inventor a monopoly in exploiting an invention for a fixed period of time in return for public disclosure of the invention. The number of patent applications

[1] `http://trec.nist.gov/`
[2] `http://www.clef-campaign.org/`
[3] `http://research.nii.ac.jp/ntcir/`
[4] `http://www.isical.ac.in/~clia/`

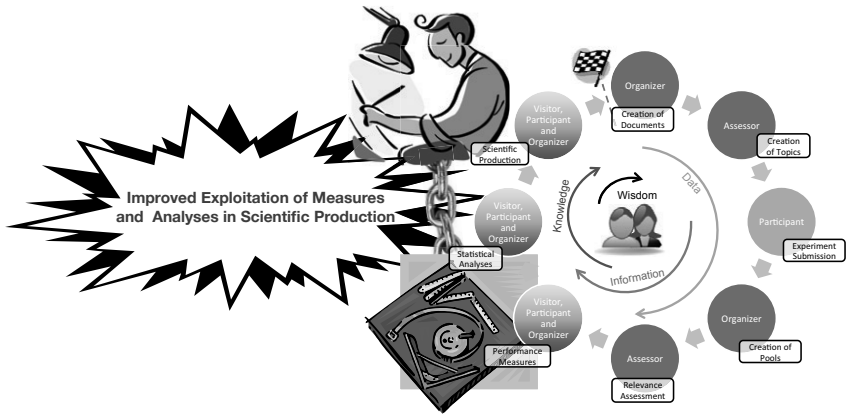


Figure 1: Typical evaluation campaign cycle and objectives of the proposal.

and granted patents is continually increasing. The World Intellectual Property Organisation (WIPO) estimates[5] that the number of patent applications worldwide increased from 926.000 in 1985 to 1.9 million in 2008, and the number of granted patents increased from 395.000 in 1985 to 778.000 in 2008. The increasing number of patent applications creates an increasing workload for patent examiners at patent offices, who have the task of deciding if a patent application meets the requirements for patentability, which are in general that the invention described must be new, non-obvious, and useful or industrially applicable. The WIPO estimates that the total number of pending applications undergoing examination across the world is 3.45 million [9]. Beyond the patent examiners, patent searches are an important tool for companies involved in research and development and for inventors.The following are examples of the types of patent search carried out [10]:

Patentability: The aim is to find prior art to enable a decision on the patentability of an invention. It is similar to the search carried out by a patent examiner at a patent office.

Validity/Invalidity: This search is done to invalidate the claims of an issued patent. It is done if a company is sued for patent infringement or to ensure that royalty payments are justified.

State of the Art: A comprehensive search of all available patent and non-patent literature related to a defined technical field. It may be done to determine research activities or to assist in licensing existing technologies.

A patent document has a well defined structure that can be exploited by search applications. It consists of four parts:

Front page: This contains standard information, including title, patent number, inventor, filing date and patent classification — patents are classified according to areas of technology using hierarchical schemes such as the International Patent Classification (IPC). An example patent front page is shown in Figure 2.

Abstract: Brief summary of the invention.

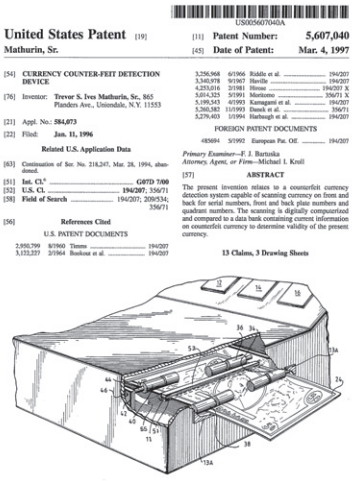
Description: A lengthy description of the proposed invention.

Claims: The claims define the extent of the protection offered by a patent. They are written in a precise legal style.

Patent searches are different in a number of ways to the operation of Internet search engines:

- To validate a patent all information that has been made available to the public in any form before a given date that might be relevant to a patent's claims of originality has to be investigated. This has to cover not only recent publications but more or less the whole documented developments from the last centuries; although practically the search concentrates on issued patents.

[5] http://www.wipo.int/ipstats/en/statistics/patents/



US005607040A

United States Patent [19]
Mathurin, Sr.

[11] **Patent Number: 5,607,040**
[45] **Date of Patent: Mar. 4, 1997**

[54] **CURRENCY COUNTER-FEIT DETECTION DEVICE**

[76] Inventor: **Trevor S. Ives Mathurin, Sr.**, 865 Flanders Ave., Uniondale, N.Y. 11553

[21] Appl. No.: **584,073**

[22] Filed: **Jan. 11, 1996**

Related U.S. Application Data

[63] Continuation of Ser. No. 218,247, Mar. 28, 1994, abandoned.

[51] **Int. Cl.**[6] **G07D 7/00**

[52] **U.S. Cl.** **194/207**; 356/71

[58] **Field of Search** 194/207; 209/534; 356/71

[56] **References Cited**

U.S. PATENT DOCUMENTS

2,950,799	8/1960	Timms	194/207
3,122,237	2/1964	Bouloux et al.	194/207
3,256,968	6/1966	Riddle et al.	194/207
3,340,978	9/1967	Haville	194/207
4,253,016	2/1981	Hirose	194/207 X
5,014,325	5/1991	Morizono	356/71 X
5,199,543	4/1993	Kamagami et al.	194/207
5,260,583	11/1993	Danek et al.	356/71
5,276,403	1/1994	Harbaugh et al.	194/207

FOREIGN PATENT DOCUMENTS

480694	5/1992	European Pat. Off.	194/207

Primary Examiner—F. J. Bartuska
Attorney, Agent, or Firm—Michael I. Kroll

[57] **ABSTRACT**

The present invention relates to a counterfeit currency detection system capable of scanning currency on front and back for serial numbers, front and back plate numbers and quadrant numbers. The scanning is digitally computerized and compared to a data bank containing current information on counterfeit currency to determine validity of the present currency.

13 Claims, 3 Drawing Sheets

Figure 2: Example patent front page.

- Searches tend to be based on lengthy search sessions rather than single queries: result review and query refinement may take several days of work.
- Very high recall is required: not a single relevant document should be missed by the search.

This contrasts with Web search where the aim is high precision: few or no irrelevant documents shall appear among the top ranking Web pages.

The huge and increasing number of documents, the complexity of the searches done and the potential legal and economic repercussions of missing a key document in a search make improving patent search an area of increasing research interest in information retrieval. Patent search also extends beyond patent documents, as a State of the Art search also includes documents from the scientific literature. Boolean search is still the dominant paradigm used by patent professionals, although commercial tools are beginning to offer ranked retrieval. An aspect of patent search that is poorly covered in current systems is searching based on drawings [11].

Information Retrieval evaluation campaigns have recognised the importance of improving techniques for patent search, and have offered a number of tracks in recent years. The first campaign to introduce a patent retrieval task was the NTCIR (NII Test Collection for IR Systems) Project in the NTCIR-3 in 2001–2002. All subsequent NTCIR campaigns have had tasks focused on patents, such as the patent mining and patent translation tasks in the recent NTCIR-7 and NTCIR-8 iterations. The ease of doing research on patent information retrieval was increased by the release of the Matrixware Research Collection (MAREC) in 2009. MAREC consists of 19 million patents from the European Patent Office (EPO), United States Patent and Trademark Office (USPTO), Japan Patent Office (JPO) and the World Intellectual Property Organization (WIPO), stored in a standardised format. Work is underway to include the corresponding patent images in MAREC. The availability of MAREC lead to the first patent search evaluation task in Europe, organised as a task of the Cross language Evaluation Forum (CLEF) 2009, with a prior art search task for patents in English, French and German using a subset of MAREC. This task has continued, with the addition of a patent classification task in 2010, and the addition of a drawing retrieval task in 2011. In 2009, patents and scientific publications in the chemical domain were used in the first TREC Chemistry track. The 2009 and 2010 tracks included Prior Art Search and Technology Survey Search tasks, and a chemical structure recognition task is included in 2011.

These evaluation campaigns have had an impact on the number of groups working on patent retrieval challenges and on the development and evaluation of new solutions, such as solutions for image search in patents (see Figure 3). However, the infrastructure described in this paper will allow even more rapid development in this important area by simplifying access to a realistic dataset of patents and corresponding queries, allowing evaluations to be conducted at any time and allowing results of different search systems to be compared and visualised in an interactive way.

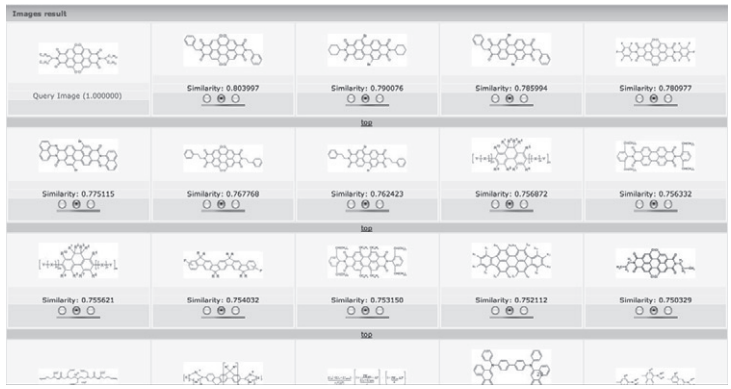


Figure 3: Image similarity search in patents.

2. Approach

As part of recent efforts to shape the future of large-scale evaluation campaigns [12, 13], more attention has been paid to evaluation infrastructures, meant as the information management systems that have to take care of the different steps and outcomes of an evaluation campaign. In this context, we have proposed an extension to the traditional evaluation methodology in order to explicitly take into consideration and model the valuable scientific data produced during an evaluation campaign [14, 15], the creation of which is often expensive and not easily reproducible. Indeed, researchers not only benefit from having comparable experiments and a reliable assessment of their performances, but they also take advantage of the possibility of having an integrated vision of the scientific data produced, together with their analyses and interpretations, as well as benefiting from the possibility of keeping, re-using, preserving, and curating them. Moreover, the way in which experimental results are managed, made accessible, exchanged, visualized, interpreted, enriched and referenced is therefore an integral part of the process of knowledge transfer and sharing towards relevant application communities, such as the *Digital Library (DL)* community, which needs to properly understand these experimental results in order to create and assess their own systems.

Therefore, we have undertaken the design of an evaluation infrastructure for large-scale evaluation campaigns and we have chosen to rely on DL systems in order to develop it, since they offer content management, access, curation, and enrichment functionalities. The outcome is a DL system, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)*[6], which manages the scientific data produced during a large-scale evaluation campaign, as well as supports the archiving, access, citation, dissemination, and sharing of the experimental results [16, 17, 18], as shown in Figure 4. DIRECT has been used, developed and tested in the course of the annual CLEF campaigns since 2005.

Note that this is per se an innovative and valuable effort since, differently from other fields such as bio-engineering, there is no equivalent in the information retrieval field of genome databases or other kind of curated databases and this kind of infrastructures is a pre-requisite for having the possibility of creating "executable papers".

DIRECT now manages the data produced over ten years of CLEF in some of its core tracks, such as the ad-hoc track, which amounts to about 130 million tuples [19]. Table 1 summarizes all the data that can be accessed through the DIRECT system and which are now available to the research and developer communities.

The future challenges for the evaluation campaigns will require an increased attention for the knowledge process entailed by an evaluation campaign. The complexity of the tasks and the interactions to be studied and evaluated will produce, as usual, valuable scientific data, which will provide the basis for the analyses and need to be properly managed, curated, enriched, and accessed. Nevertheless, to effectively investigate these new domains, not only the scientific data but also the information and knowledge derived from them will need to be appropriately treated and managed, as well as the cooperation, communication, discussion, and exchange of ideas among researchers in the field. As a consequence, we have to further advance the evaluation methodologies in order to support the whole

[6] http://direct.dei.unipd.it/

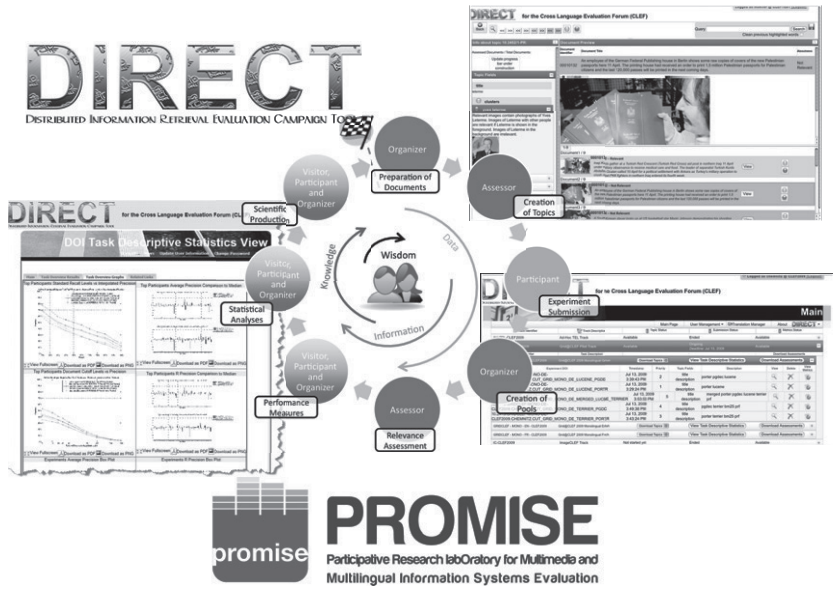


Figure 4: Example of DIRECT functionalities.

Table 1: Summary of the data available in DIRECT.

	Number of items	Number of tuples	Size
Documents	7,205,576	–	36.0 GB
Topics	1,195	20,843	0.0 GB
Experiments	3,574	166,933,261	29.0 GB
Pools	128	3,170,678	0.6 GB
Measures	4,246,372	–	1.5 GB
Statistics	5,407,428	–	1.5 GB
Plots	34,841	–	0.8 GB

knowledge creation process entailed by a large-scale evaluation campaign and to deal with the increasing complexity of the tasks to be evaluated. This requires the design and development of evaluation infrastructures which offer better support for and facilitate the research activities related to an evaluation campaign.

In the perspective of the upcoming challenges, our final goal is to turn the DIRECT system from a DL for scientific data into a virtual research environment, where the whole process which leads to the creation, maintenance, dissemination, and sharing of the knowledge produced during an evaluation campaign is taken into consideration and fostered. The boundaries between *content producers* – evaluation campaign organizers who provide experimental collections, participants who submit experiments and perform analyses, and so on – and *content consumers* – students, researchers, industries and practitioners who use the experimental data to conduct their own research or business, and to develop their own systems – are lowered by the current technologies: considering that we aim at making DIRECT an active communication vehicle for the communities interested in the experimental evaluation. This can be achieved by extending the DL for scientific data with advanced annotation, collaboration, and information visualization functionalities in order to become not only the place where storing and accessing the experimental results take place, but also an active communication tool for studying, discussing, comparing the evaluation results, where people can enrich the information managed through it with their own annotations, tags, etc. and share them in a social evaluation community [20, 21].

This effort is being undertaken by the *Participative Research labOratory for Multimedia and Multilingual Infor-*

mation Systems Evaluation (PROMISE) network of excellence[7] which will provide a virtual and open laboratory for conducting participative research and experimentation in which it will be possible to carry out, advance and bring automation into the evaluation and benchmarking of complex multimedia and multilingual information systems, by facilitating management and offering access, curation, preservation, re-use, analysis, visualisation, and mining of the collected experimental data [22, 23].

3. Relevance to the Executable Paper Grand Challenge

Executability. DIRECT already assigns persistent identifiers to the main entities involved in the experimental evaluation, such as document collections, topics, experiments and their performance measures, statistical analyses and so on. In particular, we have experimented with the use of the *Digital Object Identifier (DOI)* [24]: for example, resolving the DOI `10.2415/AH-BILI-X2BG-CLEF2007.JHU-APL.APLBIENBGTD4` gives online access to the corresponding experiments and all the data and plots related to it.

This provides the opportunity for making a paper executable and two possibilities are foreseen:

- paper authors can directly link the experiments and statistical analyses they are describing by using the provided persistent identifiers; they can even directly reference and cite them, making the experimental data first-class citizens in their scientific production;
- it is possible to develop a light-weight browser plug-in, written in Javascript and actionable, e.g., by means of a bookmarklet, that scans papers opened in a Web browser, recognizes persistent identifiers associated with experiments, statistical analyses, etc., and opens a pop-up window with all the pertinent information from which it will be possible to start the navigation and access further resources and information.

It is worth noting that the novel Visual Analytics techniques that the PROMISE project is introducing in the system allow for a high degree of interactivity, both in data management and visual data exploration. As an example, the Web based prototypical system described in [25], designed for assessing the quality of the ranking of retrieved results according to the estimation of their relevance to the query, allows for loading an experiment result, visualizing it, and interactively locating and inspecting the misplaced elements (see Figure 5).

Moreover, the PROMISE project is going to improve the DIRECT system and make its resources further accessible also as a *REpresentational State Transfer (REST)* Web services [26]. This will open-up the possibility of further integration of the scientific data into the papers concerning them. For example, it will become possible to make focused *Asynchronous JavaScript Technology and XML (AJAX)* calls, to receive small fragments of data represented in *eXtensible Markup Language (XML)* and/or *JavaScript Object Notation (JSON)*, and to contextually render those data in the Web browser while the user is reading a paper.

Overall, the proposed approach allows for a loosely-coupled integration of existing editorial systems with DIRECT which basically happens by means of *HyperText Transfer Protocol (HTTP)* calls to give direct access to the pertinent scientific data in the papers describing them. The proposed approach also has a low impact on existing editorial systems since it does not require to re-design them but it can be achieved by means of lightweight browser plug-ins.

Short and long-term compatibility. The proposed solution relies on standard Web technologies, such as HTTP [27], *Uniform Resource Identifier (URI)* [28], XML [29], AJAX [30] which are inherently cross-platform, scalable, open, and durable over the time. As a consequence, we do not need to develop ad-hoc models for executable files, maybe tied to a specific operating system or platform. Moreover, we can rely on tools already available for the end-users without requesting him to install and use new software, specifically developed for accessing the given executable file format.

Indeed, as discussed above, in the case of papers and documents accessed via a Web browser, it is possible to make the paper executable via a browser plugin or a bookmarklet which opens additional windows with the pertinent and linked information. In the case of other formats, as for example PDF documents, it is possible to insert hypertext links in the PDF source which will turn out to be opened in the Web browser showing the pertinent information.

[7] PROMISE is co-funded by the 7th Framework Program of the European Commission, grant agreement no. 258191; more information are available at: `http://www.promise-noe.eu/`

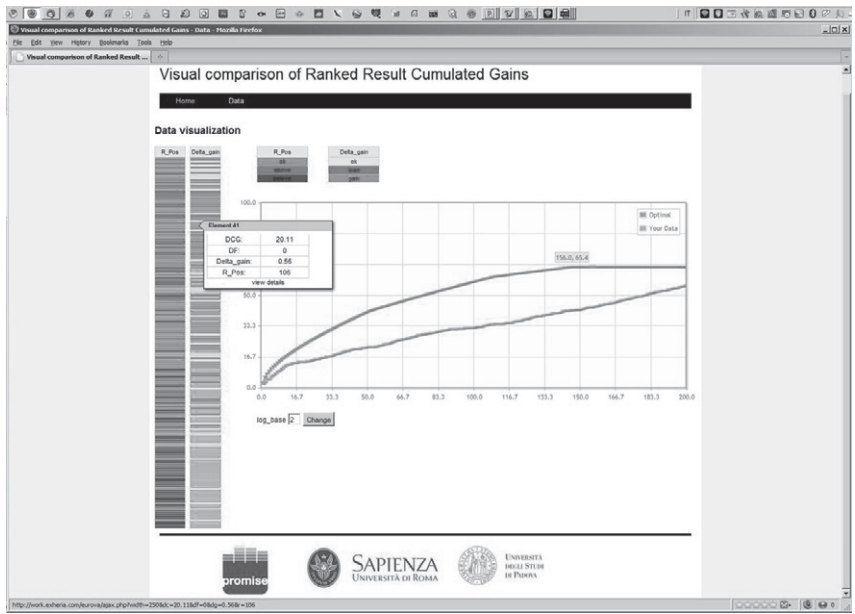


Figure 5: The figure shows a screenshot prototype. Moving the mouse over the vectors' elements triggers a windows with a summary of the relevant metrics, highlighting the document position on the discounted cumulated gain graphs on the right. Moreover, through the input panel below the graphs it is possible to change the logarithm base for modeling different discount function according to different class of users

As a final consideration, all the scientific data will be made available as REST resources; this mean that different representations of the same resource can be transferred upon request by specifying the proper *Multipurpose Internet Mail Extensions (MIME)* [31] in the `Accept` HTTP header. This gives the possibility of transferring not only XML but also JSON, PDF, or other formats that are more suitable for embedding and integration in different platforms of the relevant information to be actioned.

Validation. Data validation is ensured by the rigorous experimental methodology adopted by large-scale evaluation campaigns, by the continuous check that organizers and the community performs on the data, by the automatic checks that DIRECT performs every time a piece of information comes into the system, and by the lively discussions that are carried out when the community meets in the workshops and events organized by large-scale evaluation campaigns where the results are publicly presented.

To increase the collaboration among users, the PROMISE project is going to improve DIRECT by giving users the possibility of annotating the experimental results, as discussed above. This will provide a further and continuous check and turn DIRECT into a curated database. Moreover, the PROMISE project is explicitly addressing the challenging idea of integrating a Visual Analytics component in the system [21], in order to provide reviewers with algorithms and visualizations that allow for managing the overwhelming set of data stored within DIRECT. We foresee to to design both a set of predefined visual exploration patterns, useful for dealing with standard and repetitive evaluation activities, and an interactive visualization environment that provides the means for designing ad-hoc views of the data for exploring novel analysis patterns.

Copyright/licensing. Evaluation resources are usually regulated by specific copyright agreements managed by the organizations which run the large-scale evaluation campaigns that ensure the possibility of using such resources for research purposes.

Specific metadata formats are under development, for example in the META-NET Network of Excellence[8] which cooperates with PROMISE, and the DIRECT system has already the possibility to attach any format of metadata to

[8] `http://www.meta-net.eu/`

the managed resources. Moreover, there is a general agreement on trying to make language and evaluation resources available under Creative Commons[9] alike licenses.

Systems. Experiments submitted at large-scale evaluation campaigns usually represent the outcomes of computations made on large-scale computers. Therefore, they are often not easy to reproduce because of the time and costs needed to produce them. Nevertheless, infrastructures such as the DIRECT system where these experimental results are collected, curated, and made accessible online alleviates the problem. Indeed, researchers and developers have the possibility of both directly comparing their own experiments with these costly-to-reproduce data and downloading these data to make them part of their own experiments, e.g. for data fusion.

While this is a longer term development, initial design of component-based evaluation systems is underway [3]. The aim is to give researchers access to palettes of search system components that can be combined in workflows to build full systems. The development of cloud computing is making large amounts of processing power and storage space available for an ever falling cost. It is foreseeable that the specification of a search system in terms of components can be stored with the results, and can be recreated when requested in a cloud computing environment.

Size. Collections on which information retrieval evaluation experiments are run are often huge, with the largest example being a billion web pages in the ClueWeb collection[10]. In contrast, the results of runs in evaluation campaigns usually only contain lists of references to the documents in the collection, and hence do not present difficulties in terms of file size even if, as you can note from table 1, the amount of managed data is certainly not negligible . Also in this case the Visual Analytics component can provide algorithms and visualizations able to reduce the size of the analyzed data (e.g., through ad-hoc sampling techniques) producing decluttered images that can be used in the visual analysis.

Provenance. The DIRECT system contains a logging infrastructure which fine traces both system and user events. It captures information such as the user name, the *Internet Protocol (IP)* address of the connecting host, the action that has been invoked by the user, the messages exchanged among the components of the system in order to carry out the requested action, any error condition, and so on. This is logging infrastructure is further paired with an infrastructure for keeping the provenance of the managed resources which keeps trace of provenance events, explaining what actions (e.g. create, read, update, delete) have been performed when and by whom on what items and why.

In the scenario envisaged above in the *executability* paragraph, these logging and provenance infrastructures trace everything that happens to the "scientific data side" while what happens to the "paper side" is related to what is already available in the editorial system which manages such papers.

Other issues. The specific goals and data managed by the DIRECT system and foreseen in the PROMISE project are not affected by risks such as viruses and plagiarism. In fact, the data stored in the system does not contain executable files, and plagiarism is not an issue: one of the main objectives of the project is to encourage the reuse and the sharing of both data and evaluation strategies.

Acknowledgements

The authors would like to thank Emanuele Di Buccio for his help in the preparation of the final version of this paper. The work reported in this paper has be partially supported by the PROMISE network of excellence[11] (contract n. 258191) projects, as part of the 7th Framework Program of the European Commission.

References

[1] M. Dussin, N. Ferro, Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns, in: M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (Eds.), Proc. 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009), Lecture Notes in Computer Science (LNCS) 5714, Springer, Heidelberg, Germany, 2009, pp. 63–74.

[9] http://creativecommons.org/

[10] http://boston.lti.cs.cmu.edu/Data/clueweb09/

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

[2] N. Ferro, D. Harman, CLEF 2009: Grid@CLEF Pilot Track Overview, in: C. Peters, G. M. Di Nunzio, M. Kurimo, T. Mandl, D. Mostefa, A. Peñas, G. Roda (Eds.), Multilingual Information Access Evaluation Vol. I Text Retrieval Experiments – Tenth Workshop of the Cross–Language Evaluation Forum (CLEF 2009). Revised Selected Papers, Lecture Notes in Computer Science (LNCS) 6241, Springer, Heidelberg, Germany, 2010, pp. 552–565.
[3] A. Hanbury, H. Müller, Automated Component-Level Evaluation: Present and Future, in: Agosti et al. [32], pp. 124–135.
[4] 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.
[5] 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.
[6] Commission of the European Communities, Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on scientific information in the digital age: access, dissemination and preservation, COMM(2008) 56 Final.
[7] National Science Board, Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (NSB-05-40), National Science Foundation (NSF). http://www.nsf.gov/pubs/2005/nsb0540/, 2005.
[8] Working Group on Data for Science, FROM DATA TO WISDOM: Pathways to Successful Data Management for Australian Science, Report to Ministerís Science, Engineering and Innovation Council (PMSEIC), http://www.innovation.gov.au/Science/PMSEIC/Documents/FromDatatoWisdom.pdf, 2006.
[9] Economics and Statistics Division, WIPO, World intellectual property indicators 2010, Tech. Rep. 941, WIPO (September 2010).
[10] D. Hunt, L. B. Nguyen, M. Rodgers (Eds.), Patent searching: tools & techniques, John Wiley & Sons, 2007.
[11] J. List, How drawings could enhance retrieval in mechanical and device patent searching, World Patent Information 29 (3) (2007) 210–218.
[12] M. Agosti, G. M. Di Nunzio, N. Ferro, D. Harman, C. Peters, The Future of Large-scale Evaluation Campaigns for Information Retrieval in Europe, in: N. Fuhr, L. Kovács, C. Meghini (Eds.), Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007), Lecture Notes in Computer Science (LNCS) 4675, Springer, Heidelberg, Germany, 2007, pp. 509–512.
[13] N. Ferro, CLEF, CLEF 2010, and PROMISEs: Perspectives for the Cross-Language Evaluation Forum, in: N. Kando, K. Kishida (Eds.), Proc. 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, National Institute of Informatics, Tokyo, Japan, 2010, pp. 2–12.
[14] M. Agosti, G. M. Di Nunzio, N. Ferro, A Proposal to Extend and Enrich the Scientific Data Curation of Evaluation Campaigns, in: T. Sakay, M. Sanderson, D. K. Evans (Eds.), Proc. 1st International Workshop on Evaluating Information Access (EVIA 2007), National Institute of Informatics, Tokyo, Japan, 2007, pp. 62–73.
[15] M. Agosti, N. Ferro, Towards an Evaluation Infrastructure for DL Performance Evaluation, in: G. Tsakonas, C. Papatheodorou (Eds.), Evaluation of Digital Libraries: An insight into useful applications and methods, Chandos Publishing, Oxford, UK, 2009, pp. 93–120.
[16] G. M. Di Nunzio, N. Ferro, DIRECT: a System for Evaluating Information Access Components of Digital Libraries, in: A. Rauber, S. Christodoulakis, A. Min Tjoa (Eds.), Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005, pp. 483–484.
[17] M. Dussin, N. Ferro, Design of a Digital Library System for Large-Scale Evaluation Campaigns, in: B. Christensen-Dalsgaard, D. Castelli, J. K. Lippincott, B. Ammitzbøll Jurik (Eds.), Proc. 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008), Lecture Notes in Computer Science (LNCS) 5173, Springer, Heidelberg, Germany, 2008, pp. 400–401.
[18] M. Dussin, N. Ferro, The Role of the DIKW Hierarchy in the Design of a Digital Library System for the Scientific Data of Large-Scale Evaluation Campaigns, in: R. Larsen, A. Paepcke, J. L. Borbinha, M. Naaman (Eds.), Proc. 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008), ACM Press, New York, USA, 2008, p. 450.
[19] M. Agosti, G. M. Di Nunzio, M. Dussin, N. Ferro, 10 Years of CLEF Data in DIRECT: Where We Are and Where We Can Go, in: T. Sakay, M. Sanderson, W. Webber (Eds.), Proc. 3rd International Workshop on Evaluating Information Access (EVIA 2010), National Institute of Informatics, Tokyo, Japan, 2010, pp. 16–24.
[20] M. Agosti, N. Ferro, A Formal Model of Annotations of Digital Content, ACM Transactions on Information Systems (TOIS) 26 (1) (2008) 3:1–3:57.
[21] D. A. Keim, J. Kohlhammer, G. Santucci, F. Mansmann, F. Wanner, M. Schaefer, Visual Analytics Challenges, in: IEEE eChallenges 2009, 2009, pp. 21–23.
[22] M. Braschler, K. Choukri, N. Ferro, A. Hanbury, J. Karlgren, H. Müller, V. Petras, E. Pianta, M. de Rijke, G. Santucci, A PROMISE for Experimental Evaluation, in: Agosti et al. [32], pp. 140–144.
[23] N. Ferro, PROMISE: Advancing the Evaluation of Multilingual and Multimedia Information Systems, ERCIM News 84 (2011) 49.
[24] N. Paskin (Ed.), The DOI Handbook – Edition 4.4.1, International DOI Foundation (IDF). http://dx.doi.org/10.1000/186, 2006.
[25] N. Ferro, G. Sabetta, G. Santucci, G. Tino, F. Veltri, Visual Comparison of Ranked Result Cumulated Gains, in: S. Miksch, G. Santucci (Eds.), EuroVa – International Workshop on Visual Analytics, The Eurographics Association, Switzerland, 2011.
[26] R. T. Fielding, R. N. Taylor, Principled Design of the Modern Web Architecture, ACM Transactions on Internet Technology (TOIT) 2 (2) (2002) 115–150.
[27] R. Fielding, Y. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee, Hypertext Transfer Protocol – HTTP/1.1, RFC 2616 (June 1999).
[28] T. Berners-Lee, R. Fielding, L. Masinter, Uniform Resource Identifier (URI): Generic Syntax, RFC 3986 (January 2005).
[29] W3C, Extensible Markup Language (XML) 1.0 (Fifth Edition) – W3C Recommendation 26 November 2008, http://www.w3.org/TR/xml/ (November 2008).
[30] W3C, XMLHttpRequest – W3C Candidate Recommendation 3 August 2010, http://www.w3.org/TR/XMLHttpRequest/ (August 2010).
[31] N. Freed, N. Borenstein, Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies, RFC 2045 (November 1996).
[32] M. Agosti, N. Ferro, C. Peters, M. de Rijke, A. Smeaton (Eds.), Multilingual and Multimodal Information Access Evaluation. Proceedings of the International Conference of the Cross-Language Evaluation Forum (CLEF 2010), Lecture Notes in Computer Science (LNCS) 6360, Springer, Heidelberg, Germany, 2010.