Uniform Methodology for Evaluating Information Access Components of Digital Libraries

Giorgio Maria Di Nunzio and Nicola Ferro
Department of Information Engineering
University of Padua – Italy
{dinunzio, ferro}@dei.unipd.it

I. Introduction

The evaluation of *Digital Library Management Systems (DLMSs)* is a non trivial issue that should cover different aspects, such as: the DLMS architecture, the DLMS information access and extraction capabilities, the management of multimedia content, the interaction with users, and so on. In particular, with respect to the classification proposed by [1], we are interested in the evaluation aspects concerned with the technological issues of a DLMS and, more specifically, with the information access and extraction components of a DLMS, which deal with the indexing, search and retrieval of documents in response to a user's query.

Today, the evaluation of the performances of the information access and extraction components of a DLMS is carried out in important international evaluation initiatives, such as the *Text REtrieval Conference (TREC)*[1], the *Cross-Language Evaluation Forum (CLEF)*[2], the *NII-NACSIS Test Collection for IR Systems (NTCIR)*[3], and the *INitiative for the Evaluation of XML Retrieval (INEX)*[4]. All of these initiatives are based on the Cranfield methodology, which makes use of *experimental collections* [2] and measures to quantify the retrieval performances. Besides the evaluation of the performances of the single systems, another type of evaluation is the statistical analysis in order to compare performances among different components. For this reason, a statistical methodology for judging whether measured differences can be considered statistically significant is needed [3].

However, the evaluation forums mentioned above are carried out in a fragmented way and for most of the time individually by each participant: each participant acquires the collection and a set of tasks, performs the tasks locally on his own system, and returns the results to the organizers of the evaluation forum. The organizers make the results of the performances and of the statistical analysis of each participant available, and participants are able to compare the results of their systems with the others. This view shows that different moments of an evaluation forum are carried out and completed separately, and the tools to analyze and compare results are usually different from participant to participant.

[1] http://trec.nist.gov/
[2] http://clef.isti.cnr.it/
[3] http://research.nii.ac.jp/ntcir/index-en.html
[4] http://inex.is.informatik.uni-duisburg.de/

Integrating and uniforming the activities among the different entities involved in the evaluation of the information access components of a DLMS would be of great benefit both for the organizers and for the participants. With "uniform" we intend standard experimental collections that make the experimental results comparable, and standard tools for the analysis of the experimental results that make the analysis and assessment of experimental results comparable, too. The integration is done providing common tools for carrying out each step of the evaluation activities in a networked and distributed manner.

The question of uniforming the methodology for evaluating information access components of digital libraries opens an interesting problem that is usually faced in scientific data curation: the problem of selection of data to be kept. So far, the format in which results are packed in evaluation forums is useful to exchange/transfer them but not to describe/elaborate them. Therefore, the following questions should be asked: what criteria should be applied when selecting data for longer-term retention? How do we know what we should keep? Who sets the selection criteria? How can selection be assessed, when, how often, by whom? Besides these questions, there is also the problem of the right format to use when deciding the format of the record [4], [5].

An innovative system, called *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* [6], [7], has been designed and developed in the context of the CLEF 2005 evaluation campaign. The aim of this system is to address the issues introduced above, by providing:

- the management of an evaluation forum: the track set-up, the harvesting of documents, the management of the subscription of participants to tracks;
- the management of submission of experiments, the collection of metadata about experiments, and their validation;
- the creation of document pools and the management of relevance assessment;
- common statistical analysis tools for both organizers and participants in order to allow the comparison of the experiments;
- common tools for summarizing, producing reports and graphs on the measured performances and conducted analyses;
- an historical vision of the submitted experiments, making them online available to participants for further comparisons and analyses.

Section II describes the adopted evaluation methodologies and statistical analysis techniques in more detail; in Section III, an outlook of curation of scientific data is introduced together with its relationship with the DIRECT system; Section IV provides some insights about the architecture of DIRECT; finally, Section V draws some conclusions.

II. Evaluation Methodologies for the Information Access Components of a DLMS

A. Experimental Collections and Performance Measures

The evaluation of the information access components of a DLMS is, generally, carried out according to the Cranfield methodology, which makes use of *experimental collections* [2]. An experimental collection is a triple $\mathcal{C} = (D, Q, J)$, where:

- D is a set of documents, called also collection of documents;
- Q is a set of queries, called also topics;
- J is a set relevance judgements, i.e. for each topic $q \in Q$ the documents $d \in D$, which are relevant for the query q, are determined.

An experimental collection $\mathcal{C}$ allows the comparison of two retrieval methods, say X and Y, according to some measurements which quantifies the retrieval performances of these methods; in this sense an experimental collection is said to allow the comparison of two information access components X and Y of two different DLMSs (or the same DLMS if more than a type of information access is available). An experimental collection both provides a common test–bed to be indexed and searched by the DLMS X and Y and guarantees the possibility of replicating the experiments. The evaluation happens as follows: the document collection D is indexed by both systems X and Y; each system searches the document collection against topics Q and produces, for each topic q, an ordered list of retrieved documents; finally, the relevance judgements J allow one to check the ordered list of retrieved document and compute the performance of DLMS X and Y in order to compare them.

In order to define a measure to compare performances, we need to introduce the following sets: $A \subseteq D$ is the set of relevant documents according to the relevance judgements J, and $B \subseteq D$ is the set of documents actually retrieved by the *Information Retrieval System (IRS)*. Intuitively, the larger the intersection $A \cap B$, the higher the performance of the IRS. The maximum performance is reached in the ideal case $A = B$.

Given the sets introduced above, a couple of measures for quantifying the performances of an IRS can be defined:

- **recall** $R = \frac{|A \cap B|}{|A|}$: is a measure of the ability of a system to present all relevant items;
- **precision** $P = \frac{|A \cap B|}{|B|}$: is a measure of the ability of a system to present only relevant items;

Precision and recall are set–based measures that is to say they evaluate the quality of an unordered set of retrieved documents. To evaluate ranked lists, precision and recall can be computed for different values of an appropriate parameter; for example, precision can be computed at standard recall levels or after that a given number of documents has been retrieved, which is called precision at different *Document Cut–off Values (DCVs)*.

B. Statistical-analysis

Once the general methodology to follow in order to carry out the evaluation has been defined, the problem of assessing the measured performances still has to be addressed. Hull [3] points out that, in order to evaluate retrieval performances, we do not need only experimental collections and measures for quantifying retrieval performances, but also a statistical methodology for judging whether measured differences between retrieval methods X and Y can be considered statistically significant. In general, the null hypothesis H_0 is that all the retrieval methods being tested are equivalent in terms of performance. The significance test attempts to disprove this hypothesis by determining a p-value, a measurement of the probability that the observed difference could have occurred by chance. Prior to the experiment, a significance level α is chosen, and if the p-value is less than α, one can conclude that the search methods are significantly different. Statistical tests for paired data, such as the sign test or the t-test, can be employed in order to compare two retrieval methods and decide if one is better than the other; on the other hand, tests such as two-way *ANalysis Of VAriance (ANOVA)* can be used in order to simultaneously compare multiple experiments [3], [8].

However, statistical analysis has been usually carried out by the organizers of the evaluation forums mentioned above, who have access to all the submitted results and provide participants with a general analysis of the submitted results. The access to evaluation forum resources is quite limited for participants who have the need to carry out other type of comparison: workshop acts presents only partial and/or summarized performances of all the systems; replicating an experiment of another participant is practically undoable. Moreover, participants may not have the skills, resources and tools needed to perform a statistical analysis even on their own results in order to point out significant differences between the retrieval methods they are proposing. Thus, we could say that in general evaluation forums do not further the systematical employment of statistical analysis from participants.

III. Curation of Scientific Data

In general, "curation" is concerned with long-term care and maintenance of integrity. Data curation is a particular instance where data need maintenance and enhancement together with the promotion of data to potential consumers. Obviously, behind "data curation" there is the assumption that data itself has value. The e–Science Data Curation Report[5] gives a definition of curation which is of particular interest for the problem of evaluation of access components of DLMS:

> Curation: the activity of managing and promoting the use of data from its point of creation, to

[5] http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

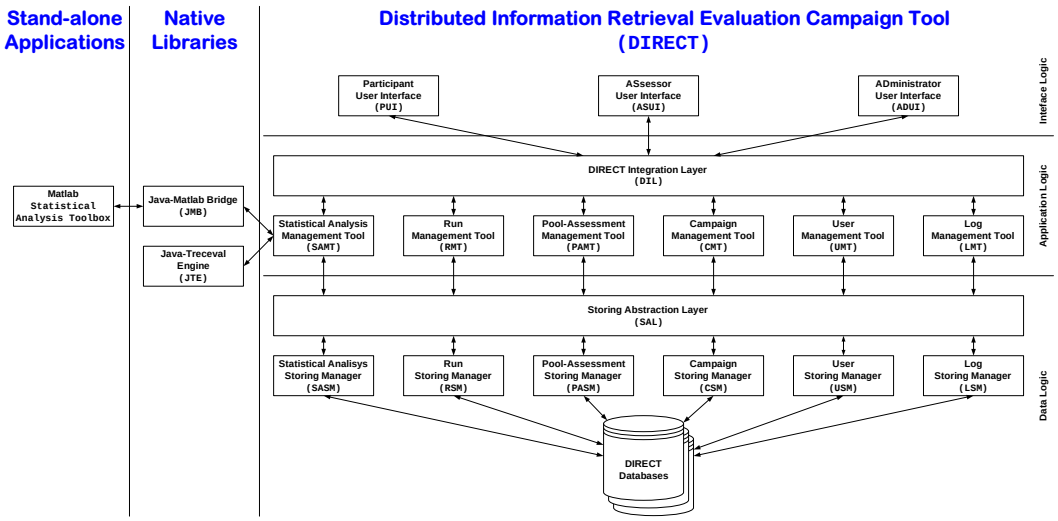

Fig. 1. DIRECT Architecture

ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose.

This definition reflects the importance of some of the many possible reasons for which keeping data is important. For example,

- Re-use of data for new research, including collection based research to generate new science;
- Retention of unique observational data which is impossible to re-create;
- Retention of expensively generated data which is cheaper to maintain than to re-generate;
- Enhancing existing data available for research projects;
- To validate published research results.

If we read these bullet points with the eye of the evaluation forums, we could say that all of these are functional requirements for a system thought to maintain, re–use and promote results of participants. In this sense, it is important to promote the fact that data is not "passive" but can be made "active" by giving the opportunity to maintain links between data, materials and annotations, as well as provenance of information. Moreover, there would be a full trust on data since the continuous double checking done by the participants and organizers who curate the data. Such an organization of data would create a scientific digital library for the evaluation of information access components of DLMS comparable to existing scientific *Digital Library (DL)* like [9].

IV. DIRECT Architecture

Figure 1 shows the architecture of DIRECT. It consists of three layers – data, application and interface logic layers – in order to achieve a better modularity and to properly describe the behavior of DIRECT by isolating specific functionalities at the proper layer. This decomposition makes a clear definition of the functioning of DIRECT by means of communication paths possible. Each communication path is meant to connect the different components when the different layers communicate to each other. In this way, the behavior of the system is designed in a modular and extensible way.

In the following, we briefly describe the architecture shown in figure 1, from bottom to top:

Data Logic

The data logic layer deals with the actual storage of the different information objects coming from the upper layers. There is a set of "storing managers" dedicated to storing the submitted runs, the relevance assessments and so on. These storing managers translate the requests that arrive from the upper layers into *Structured Query Language (SQL)* statements that operate on the underlying *DataBase Management Systems (DBMSs)*. To this end, we adopt the *Data Access Object (DAO)*[6] and the *Transfer Object (TO)*[6] design patterns. The DAO implements the access mechanism required to work with the underlying data source, e.g. it may offer access to a *Relational DBMS (RDBMS)* by using the *Java DataBase Connectivity (JDBC)*[7] technology. The components that rely on the DAO are called *clients* and they use the interface exposed by the DAO, which completely hides the data source implementation details from its clients. Because the interface exposed by the DAO to clients does not change when the underlying data source implementation changes, this pattern allows the DAO to adapt to different storage schemes without affecting its clients. Essentially, the DAO acts as an adapter between the clients and the data source.

Besides the storing manager devoted to guarantee the persistence of the object directly related to the course of the evaluation campaign, there is also the *log storing manager* that allows for the fine tracing of both system and user events.

[6]http://java.sun.com/blueprints/corej2eepatterns/Patterns/

[7]http://java.sun.com/products/jdbc/

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.

Finally, on top of the various "storing managers" there is the *Storing Abstraction Layer (SAL)* which hides the details about the storage management from the upper layers. In this way, the addition of a new "storing manager" is totally transparent for the upper layers.

Application Logic

The application logic layer deals with the flow of operations within DIRECT. It provides a set of tools capable of managing high-level tasks. For example, the *Statistical Analysis Management Tool (SAMT)* offers the functionalities needed to conduct a statistical analysis on a set of runs. Note that the SAMT makes use of the *Java–Matlab Bridge (JMB)* native library[8], a library we developed to allow a Java application to use Matlab[9] as a computational engine. This way, the SAMT can leverage on the Statistics Toolbox provided with Matlab in order to perform the actual statistical analysis. This choice ensures that the statistical tests provided by the SAMT are implemented in a very reliable way, since Matlab is a leader and very consolidated application in the field of numerical analysis which employs state-of-the-art algorithms. As a further example, the *Java-Treceval Engine (JTE)* native library provides an interface for DIRECT towards the `trec_eval` package[10], adopted by TREC for computing the basic performance figures such as precision and recall.

Finally, the *DIRECT Integration Layer (DIL)* provides the interface logic layer with a uniform and integrated access to the various tools. As we noticed in the case of the SAL, thanks to the DIL also the addition of new tools is transparent for the interface logic layer.

Interface Logic

It is the highest level of the architecture, and it is the access point for the user to interact with the system. It provides specialised *User Interfaces (UIs)* for different types of users, that are the participants, the assessors, and the administrators of DIRECT.

V. Conclusions

We outlined some relevant issues that have to be taken into account when evaluating the information access components of a DLMS; furthermore, we have designed a system, called DIRECT, able to address the highlighted issues and we have developed a first prototype of it.

DIRECT has been successfully adopted during the CLEF 2005 campaign, it has been used by nearly 30 participants spread over 15 different nations, who submitted more than 530 experiments, and 15 assessors, who assessed more than 160,000 documents in seven different languages. The system was then used for producing reports and overview graphs about the submitted experiments [10], [11]. On the basis of the experience gained during the CLEF 2005 campaign, DIRECT is going to be further enhanced and tested during the ongoing CLEF 2006 campaign.

[8] `http://java.sun.com/j2se/1.5.0/docs/guide/jni/index.html`

[9] `http://www.mathworks.com/`

[10] `ftp://ftp.cs.cornell.edu/pub/smart/`

References

[1] N. Fuhr, P. Hansen, A. Micsik, and I. Sølvberg, "Digital Libraries: A Generic Classification Scheme," in *Proc. 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, P. Constantopoulos and I. T. Sølvberg, Eds. Lecture Notes in Computer Science (LNCS) 2163, Springer, Heidelberg, Germany, 2001, pp. 187–199.

[2] C. W. Cleverdon, "The Cranfield Tests on Index Languages Devices," in *Readings in Information Retrieval*, K. Spack Jones and P. Willett, Eds. Morgan Kaufmann Publisher, Inc., San Francisco, California, USA, 1997, pp. 47–60.

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

[4] S. Abiteboul, R. Agrawal, P. Bernstein, M. Carey, S. Ceri, B. Croft, D. DeWitt, M. Franklin, H. Garcia-Molina, D. Gawlick, J. Gray, L. Haas, A. Halevy, J. Hellerstein, Y. Ioannidis, M. Kersten, M. Pazzani, M. Lesk, D. Maier, J. Naughton, H.-J. Schek, T. Sellis, A. Silberschatz, M. Stonebraker, R. Snodgrass, J. D. Ullman, G. Weikum, J. Widom, and S. Zdonik, "The Lowell Database Research Self-Assessment," *Communications of the ACM (CACM)*, vol. 48, no. 5, pp. 111–118, 2005.

[5] M. Agosti, G. M. Di Nunzio, and N. Ferro, "A Data Curation Approach to Support In-depth Evaluation Studies," in *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, F. C. Gey, N. Kando, C. Peters, and C.-Y. Lin, Eds. http://ucdata.berkeley.edu/sigir2006-mlia.htm [last visited 2006, August 17], 2006, pp. 65–68.

[6] G. M. Di Nunzio and N. Ferro, "DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns," in *Proc. 8th International Workshop of the DELOS Network of Excellence on Digital Libraries on Future Digital Library Management Systems (System Architecture & Information Access)*, Y. Ioannidis, H.-J. Schek, and G. Weikum, Eds., 2005.

[7] G. M. Di Nunzio and N. Ferro, "DIRECT: a System for Evaluating Information Access Components of Digital Libraries," in *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, A. Rauber, S. Christodoulakis, and A. Min Tjoa, Eds. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005, pp. 483–484.

[8] G. Salton and M. E. Lesk, *Chapter 7: Computer Evaluation of Indexing and Text Processing*. In G. Salton (Ed.), The SMART Retrieval System. Experiments in Automatic Document Processing. Prentice–Hall, Inc., Englewood Cliff, New Jersey, USA, 1971, pp. 143–180.

[9] U. Schindler, J. Brase, and M. Diepenbroek, "Webservices Infrastructure for the Registration of Scientific Primary Data," in *Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, A. Rauber, S. Christodoulakis, and A. Min Tjoa, Eds. Lecture Notes in Computer Science (LNCS) 3652, Springer, Heidelberg, Germany, 2005, pp. 128–138.

[10] G. M. Di Nunzio, N. Ferro, G. J. F. Jones, and C. Peters, "CLEF 2005: Ad Hoc Track Overview," in *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers*, C. Peters, F. C. Gey, J. Gonzalo, G. J. F. Jones, M. Kluck, B. Magnini, H. Müller, and M. de Rijke, Eds. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany, 2006, pp. 11–36.

[11] G. M. Di Nunzio and N. Ferro, "Appendix A. Results of the Core Tracks and Domain-Specific Tracks," in *Working Notes for the CLEF 2005 Workshop*, C. Peters and V. Quochi, Eds. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/appendix_a.pdf [last visited 2006, January 19], 2005.