DIRECT: a Distributed Tool for Information Retrieval Evaluation Campaigns

Giorgio Maria Di Nunzio and Nicola Ferro

Department of Information Engineering – University of Padua
Via Gradenigo, 6/b – 35131 Padova – Italy
{dinunzio, nicola.ferro}@dei.unipd.it

Abstract. In this paper we describe the architecture of DIRECT, a Distributed Information Retrieval Evaluation Campaign Tool, which is an innovative system for managing evaluation campaigns. In particular, DIRECT deals with the evaluation of the information access and extraction components of a *Digital Library Management System (DLMS)*.

1 Introduction

Digital Library Management Systems (DLMSs) generally manage collections of multi-media digitalized data and include components that perform the storage, access, retrieval, and analysis of the collections of data. The evaluation of 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 [1]. We are interested in the evaluation aspects concerned with the information access and extraction components of a DLMS [2]; this interest ranges from measuring and quantifying the performances of the information access and extraction components of a DLMS to designing and developing an architecture capable of supporting this kind of evaluation in the context of DLMSs.

The paper is organized as follows: Section 2 presents the methodologies and the issues concerned with the evaluation of the information access components of a DLMS and discusses the motivations and the objectives of our system; Section 3 presents the architecture and the functionalities of our system; finally, Section 4 draws some conclusions.

2 Evaluation Issues for the Information Access Components of a DLMS

Today, this type of evaluation is carried out in important international evaluation forums which bring research groups together, provide them with the means for measuring the performances of their systems, discuss and compare their work. The most important forums for *Information Retrieval System (IRS)* are: *Text REtrieval Conference (TREC)*[1], *Cross Language Information Retrieval (CLIR)*[2],

[1] http://trec.nist.gov/
[2] http://clef.isti.cnr.it/

NII-NACSIS Test Collection for IR Systems (NTCIR)[3], and *INitiative for the Evaluation of XML Retrieval (INEX)*[4]. A wide range of questions are covered by these forums like the quality of the information retrieved, the access of multilingual collections of documents, the retrieval of structured documents and the access to asian-language collections. In general, these evaluation campaigns follow the Cranfield paradigm [3] giving the participants one or more test-bed collections, a set of tasks to be performed, and a method by which evaluating performances of their systems with respect to the defined collections and tasks.

A drawback of the approach followed during these evaluation forums is that huge chunks of textual files are shifted from side to side. Document collections usually reside on a single high-loaded server where all participants connect concurrently in a very short limited time in order to download the collections needed to carry out the experiments. The experimental results provided by participants to organizers usually consist of large text files, containing lists of retrieved documents together with their rank and score, that are numerical data. The performance figures computed by organizers and returned to participants consist of text files full of numerical data; in particular, the presentation format adopted by TrecEval[5], the de-facto standard tool for computing the performance figures, is not very suitable for direct processing by a computer program, since it is tailored to be human-readable. These file transfers often requires a mass-mailing between participants and organizers in order to acknowledge the receipt of the files or to correct errors. Moreover, if the performance figures are to be accessed in order to further process them, a lot of textual parsing is needed to transform chunks of text into numerical values, a process which is prone to errors.

Another drawback is that while the performance of IRS are measured by means of traditional IR performance indicators, the analysis of the significance of this results is rarely performed by participants although statistical analysis is a fundamental step in the experimental evaluation, as pointed out by [4]. We identify two main reasons for this: first, the analysis of the whole set of runs submitted is possible only by organizers that collect all the runs, and replicating experiments of other research groups is seldom possible for participants. Second, statistical tools are not easy to handle and the possibility to have non coherent results is high when participants make use of different tools. Moreover, the statistical analysis is burdened by all the textual parsing needed to transform chunks of text into numerical and processable data.

In this paper we want to tackle the problem of a new approach to evaluation campaigns, able to take into account the distributed nature of the entities involved during an evaluation campaign: data collections may reside on different servers, participants are scattered around the world, as well as assessors and organizers. Moreover, DLMSs themselves are distributed systems where the services under evaluation can be developed according to different architectural paradigms, such as *Web Services (WS)*, *Peer-To-Peer (P2P)*, and Grid. Finally,

[3] `http://research.nii.ac.jp/ntcir/index-en.html`
[4] `http://inex.is.informatik.uni-duisburg.de/`
[5] `ftp://ftp.cs.cornell.edu/pub/smart/`

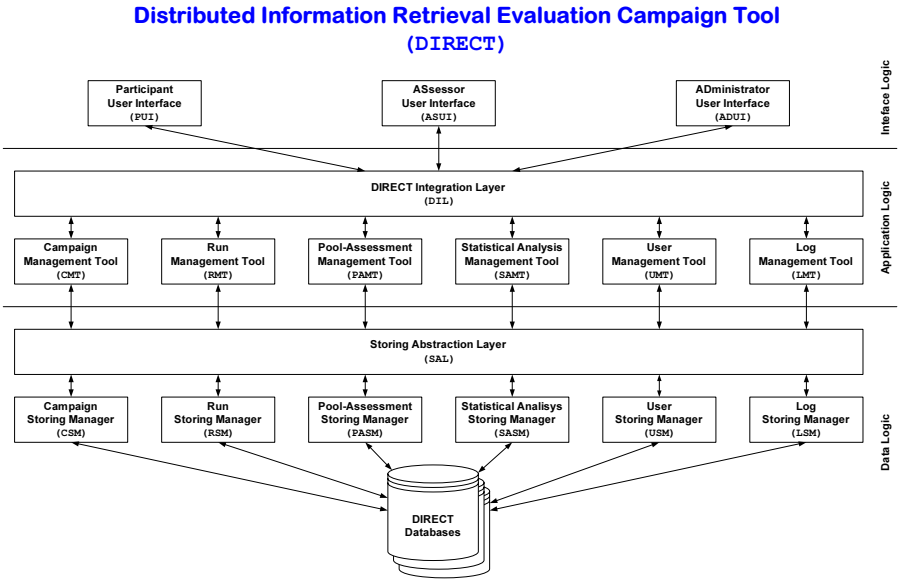

Fig. 1. DIRECT Architecture

another innovative aspect of our approach is to provide participants with a uniform way of performing statistical analysis on their results. In this way, not only participants benefit from standard experimental collections but also they may exploit standard tools for the analysis of the experimental results. This approach, that makes the analysis and assessment of experimental results comparable, is quite innovative since up to now participants employed tools built on their own in order to analyze experimental results, making such analyzes much more diffucult to compare.

An innovative system named *Distributed Information Retrieval Evaluation Campaign Tool (DIRECT)* is being designed and developed to give an alternative to the management of data of these evaluation forums with the aim of integrating the activities among the different entities (both of an evaluation campaign and a DLMS) and giving the tools to make the activities themselves more interactive. The goal will be to create a unified view of this kind of evaluation forums and to propose an innovative architecture able to provide dedicated services and tools to make available data and documents. In particular, the evaluation of information access components of a DLMS will not be calculated by means of standard IR measures only, but also with an integrated tools for statistical analysis available to all participants of evaluation forums.

Since we are are going to provide and manage the technical infrastructure, both hardware and software, for the *Cross-Language Evaluation Forum (CLEF)* 2005 ongoing evaluation campaign, the possibility of testing and evaluating the DIRECT system in real settings will be exploited.

3 DIRECT Architecture and Functionalities

Figure 1 shows the architecture of DIRECT. It consists of three layers – data, application and interface logic layers – which allow us to achieve a better modularity and to properly describe the behavior of DIRECT by isolating specific functionalities at the proper layer. Moreover, this decomposition makes it possible to clearly define the functioning of DIRECT by means of communication paths that connect the different components. 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 Fig. 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" which translate the requests that arrive from the upper layers into *Structured Query Language (SQL)* statements to operate on the underlying *DataBase Management Systems (DBMSs)*. The heart of the data logic is an *Entity–Relationship (ER)* schema that is designed to fulfill the requirements to manage a complex evaluation forums like the ones presented in Section 2. Note that, due to huge quantity of data to be managed, it may be necessary to split the underlying database across different DBMSs, thus dealing with a distributed database. 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; conducting a statistical analysis involves, in the data logic layer, both the *Run Storing Manager (RSM)* to have access to the experimental data, and the *Statistical Analysis Storing Manager (SASM)* to store the results of such analysis. 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.

4 Discussion

We introduced DIRECT, a system for managing information retrieval evaluation campaigns, and described its architecture. DIRECT can be implemented according to different architectural paradigms: for example, the data logic layer can be implemented as a network of P2P "storing managers" in order to distribute the databases. On the other hand, the various tools of the application logic layer could be made available as WSs in order to easily access them and to compose them, if necessary. In conclusion, DIRECT not only allows for managing evaluation campaigns in a distributed manner but it is also a distributed tool itself.

DIRECT can be considered a *scientific databases* manager since it stores scientific data and makes it possible the analysis of the results of computations and data itself. It can become also a *curated database* manager if we partner it with services for annotating its contents, as those described in [5,6], in order to allow users to cooperate and discuss about the performances of different DLMSs in accessing information. In this context, *data provenance* [7] becomes a relevant issues and annotations can be further exploited for data provenance purposes, as described in [8,9].

The main goal of DIRECT is to model the data of the domain of interest (the evaluation forums) and to make available integrated services to operate on this data. The modelling is focussed on the problems explained in Sect.1 in order to efficiently manage the well defined tasks of accessing, (down-)loading, evaluating, submitting data during evaluation forums. In future, the infrastructure of DIRECT can be thought embedded as a part of a more complex system designed by means of scientific workflow management systems like Kepler/PtolemyII[6] [10].

DIRECT is going to be used and tested in the context of the ongoing CLEF evaluation campaign and the outcomes of this test can drive the further development and refinement of it.

Acknowledgements

The work is partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

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

[6] `www.kepler-project.org`

2. Agosti, M., Di Nunzio, G.M., Ferro, N.: Evaluation of a Digital Library System. In Agosti, M., Fuhr, N., eds.: Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries, `http://dlib.ionio.gr/wp7/workshop2004_program.html` (2004) 73–78
3. Cleverdon, C.W.: The Cranfield Tests on Index Languages Devices. In Spack Jones, K., Willett, P., eds.: Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, USA (1997) 47–60
4. Hull, D.: Using Statistical Testing in the Evaluation of Retrieval Experiments. In Korfhage, R., Rasmussen, E., Willett, P., eds.: Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993), ACM Press, New York, USA (1993) 329–338
5. Agosti, M., Ferro, N.: An Information Service Architecture for Annotations. In Agosti, M., Schek, H.J., Türker, C., eds.: Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation, Pre-proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS, Edizioni Libreria Progetto, Padova, Italy (2004) 115–126
6. Agosti, M., Ferro, N., Frommholz, I., Thiel, U.: Annotations in Digital Libraries and Collaboratories – Facets, Models and Usage. In Heery, R., Lyon, L., eds.: Proc. 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004), Lecture Notes in Computer Science (LNCS) 3232, Springer, Heidelberg, Germany (2004) 244–255
7. Buneman, P., Khanna, S., Tan, W.C.: Why and Where: A Characterization of Data Provenance. In Van den Bussche, J., Vianu, V., eds.: Proc. 8th International Conference on Database Theory (ICDT 2001), Lecture Notes in Computer Science (LNCS) 1973, Springer, Heidelberg, Germany (2001) 316–330
8. Buneman, P., Khanna, S., Tan, W.C.: On Propagation of Deletions and Annotations Through Views. In Abiteboul, S., Kolaitis, P.G., Popa, L., eds.: Proc. 21st ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems (PODS 2002), ACM Press, New York, USA (2002) 150–158
9. Bhagwat, D., Chiticariu, L., Tan, W.C., Vijayvargiya, G.: An Annotation Management System for Relational Databases. In Nascimento, M.A., Özsu, M.T., Kossmann, D., Miller, R.J., Blakeley, J.A., Schiefer, K.B., eds.: Proc. 30th International Conference on Very Large Data Bases (VLDB 2004), Morgan Kaufmann (2004) 900–911
10. Ludscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger-Frank, E., Jones, M., Lee, E., Tao, J., Zhao, Y.: Scientific Workflow Management and the Kepler System. Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows (2005) (in print)