Toward Automated Component-Level Evaluation

Allan Hanbury
Information Retrieval Facility
Palais Eschenbach, Eschenbachgasse 11
1010 Vienna, Austria
a.hanbury@ir-facility.org

Henning Müller
University of Applied Sciences Western
Switzerland (HES SO)
TechnoArk 3
3960 Sierre, Switzerland
henning.mueller@sim.hcuge.ch

ABSTRACT

Automated component-level evaluation of information retrieval is discussed. The advantages of such an approach are considered, as well as the requirements for implementing it. Acceptance of such systems by researchers is discussed.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Performance, Measurement

Keywords

Image retrieval evaluation, future benchmarking

1. INTRODUCTION

The majority of information retrieval evaluation campaigns today are run based on the TREC (Text REtrieval Conference) organisation model. This consists of a yearly cycle in which participating groups are sent data and queries by the organisers, and subsequently submit retrieval results obtained by their system for evaluation. The evaluation produces a set of performance measures, quantifying how each participating group's system performed on the queries.

This approach has a number of disadvantages [2]. One of the main disadvantages is the evaluation at system level only. As each system contains many components (e.g. stemmer, tokeniser, feature extractor, indexer), it is difficult to judge the effect of each component on the final result returned for a query. For this reason, when reviewing a number of years of an evaluation task, it is often difficult to go beyond superficial conclusions based on complete system performance and textual descriptions of the systems. Little information on where to concentrate effort so as to best improve results can be obtained. A further disadvantage of the system-level approach, where the result of an evaluation is a ranked list of participants, is the potential to view the evaluation as a competition. This can lead to a focus on tuning systems to the evaluation tasks, rather than the scientific goal of determining how and why systems perform as they do.

A solution that has been proposed is a component-level evaluation of systems. An example is the MediaMill Challenge [3] in the area of video semantic concept detection. A concept detection system, data and ground truth are provided, where the concept detection system is broken down into feature extraction, fusion and machine learning components. Researchers can replace any of these components with their own components to test the effect on the final results. However, browsing the papers that cite [3] gives the idea that while many researchers make use of the data and ground truth, few use the system framework.

The Grid@CLEF initiative[1] is implementing a component-level evaluation within an evaluation campaign. A basic linear framework consisting of tokeniser, stop list, word decompounder, stemmer and weighting/scoring engine components is specified. Each component should use as input and output XML data in a specified format (CIRCO Schema). This design is an intermediate step between traditional evaluation methodologies and a component-based evaluation — participants run their own experiments, but are required to submit intermediate output from each component.

In this paper, we discuss moving towards a fully automated component-level evaluation. Participation in such an evaluation would consist of registering a number of components at a central server for access over the web. The components would then be called as needed for experiments by the server. Such an idea has already been proposed for CBIR in 2001 [1], in which a communication framework (MRML) was specified, and a web server for running the evaluation by communicating in MRML over a specified port was provided. This system did not receive much use.

In the following sections, we discuss the requirements for an automated evaluation system. As use by researchers of the already proposed systems is often lacking, we pay particular attention to the problem of motivating participants.

2. AUTOMATED EVALUATION

The basic framework for a fully automated component-level evaluation framework follows. An information retrieval system built out of a set of components will be specified (as e.g. for Grid@CLEF and the MediaMill Challenge). Par-

[1] `http://ims.dei.unipd.it/websites/gridclef/`

ticipating groups in the evaluation may choose which components they wish to submit. These components should be written so as to run on the participants' computers, callable through a web interface. Participants register their components on a central server. The central server then runs the experiments using a large number of combinations of components, accessed through their web interfaces. This approach has the following advantages: (1) A large number of experiments can be done. Each participant makes available online components, which are then called from a central server. This reduces the amount of work for each participant in running complete information retrieval experiments. (2) The best performing combination(s) of components can be identified, where components making up this best performing combination could be from different groups. Different search tasks will also possibly be best performed by different constellations of components. (3) Significantly less emphasis will be placed on the final ranking of complete systems. The results will be in the form of which constellations of which components are best suited for which tasks. This will allow participants to concentrate on developing and improving specific components. It also reduces the perceived competitiveness by removing the ranked list of participants.

2.1 System Requirements

To create such a system, the following are needed:

- Software and a central server to run the evaluation.
- Protocols for interfacing with programs over the web, exchanging data and exchanging results.
- As for any IR evaluation: large amounts of data, realistic queries and relevance judgements.

The protocol design is the key challenge. The participants' task will shift from performing the experiments to adapting their code to conform to the protocols. In order to make this attractive to participants, the protocols should be designed to have the following properties:

Stability: The protocols should be comprehensively designed to change little over time — After an initial effort to get their systems compliant, little further "interface work" would have to be done by participants.

Simplicity: The initial effort by participants to get their systems compliant should not be high, as a large initial hurdle could discourage participation. In addition to a specification, code implementing key interface components should be provided.

Wide Applicability: Implementing the protocols should enable groups to achieve more than participation in a single evaluation campaign. Standardising the protocols for different evaluation campaigns and potentially for other uses is therefore important.

These properties can be contradictory. For example, a stable protocol that covers all possible eventualities is less simple. Wide applicability can be obtained through the use of a common web service protocol, however many of these protocols do not meet the requirement for simplicity.

For the control software, as the amount of participation increases and the number of components included in the IR system specification increases, the potential number of component combinations will explode. It will therefore not be feasible to test all possible combinations. Algorithms for selecting potentially good component combinations based on previous experimental results and the processing speeds of components, but with low probability of missing good combinations, will have to be designed. Further difficulties to be considered are the remote processing of large amounts of data, where participants with slower Internet connections may be disadvantaged (an initial solution may be to continue distributing the data to be installed locally). It will also have to be considered how to ensure that participants with less computing capacity are not at a disadvantage.

A current problem in IR evaluation that is not addressed at all in this framework is the provision of sufficient data, queries and relevance judgements. With the potential for more efficient experiments, this problem might become worse.

2.2 Participation

It is important to design the system so that it is accepted and used by the targeted researchers. The system should be designed so that there are clear benefits to be obtained by using it, even though an initial effort is required to adopt it. These benefits should be made clear through a "publicity campaign". Potential benefits include: more extensive experimental results on component performance, the opportunity for each research group to concentrate on research and development of those components matching their expertise, and the reuse of components by other researchers to build a working system. It is expected that web service-based systems will become common and thus many researchers might have an interest in such an interface anyway. With having other research group's components available, the building of systems can become easier.

3. LONG-TERM CONSIDERATIONS

Given the additional experimental data that will become available through such a framework, a long-term aim can be to design a search engine that can be built from components based on the task that a user is carrying out and analysis of his/her behaviour (targeted search, browsing, etc.).

The problem of obtaining a sufficient number of queries and relevance judgements in order to allow large scale experiments should be considered. Innovative approaches to harnessing Internet users for continuously increasing the number of relevance judgements should be examined, such as games with a purpose [5], or remunerated tasks [4].

4. REFERENCES

[1] H. Müller, W. Müller, S. Marchand-Maillet, T. Pun, and D. Squire. A web-based evaluation system for CBIR. In *Proc. ACM Multimedia*, pages 50–54, 2001.

[2] S. Robertson. On the history of evaluation in IR. *Journal of Information Science*, 34(4):439–456, 2008.

[3] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. ACM Multimedia*, pages 421–430, 2006.

[4] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proc. CVPR Workshop on Internet Vision*, 2008.

[5] L. von Ahn. Games with a purpose. *IEEE Computer Magazine*, pages 96–98, June 2006.