The TREC Files: The (Ground) Truth Is Out There

Savvas A. Chatzichristofis Konstantinos Zagoris Avi Arampatzis
Department of Electrical and Computer Engineering
Democritus University of Thrace, Xanthi 67100, Greece
{schatzic,kzagoris,avi}@ee.duth.gr

ABSTRACT

Traditional tools for information retrieval (IR) evaluation, such as TREC's `trec_eval`, have outdated command-line interfaces with many unused features, or 'switches', accumulated over the years. They are usually seen as cumbersome applications by new IR researchers, steepening the learning curve. We introduce a platform-independent application for IR evaluation with a graphical easy-to-use interface: the TREC_Files Evaluator. The application supports most of the standard measures used for evaluation in TREC, CLEF, and elsewhere, such as MAP, P10, P20, and bpref, as well as the Averaged Normalized Modified Retrieval Rank (ANMRR) proposed by MPEG for image retrieval evaluation. Additional features include a batch mode and statistical significance testing of the results against a pre-selected baseline.

Categories and Subject Descriptors: D.0 [**Software**]: General
General Terms: Experimentation, Measurement
Keywords: TREC files, Evaluation Measurements

1. THE TREC_FILES EVALUATOR

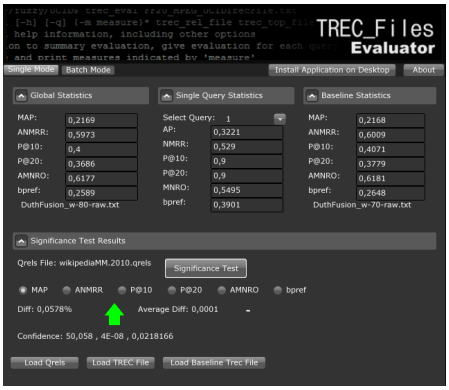


Figure 1: The TREC_Files Evaluator.

The program is developed in Silverlight technology which is an application framework for writing rich Internet applications. Its run-time environment is available for most browsers and it is compatible with Microsoft Windows, Mac OS X, Linux, FreeBSD, and other *nix open source platforms through the Moonlight plugin. The TREC_Files Evaluator (Fig. 1) is freely available at:
`http://thetrecfiles.nonrelevant.net`

Copyright is held by the author/owner(s).
SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

A user must first load the relevance assessments file (qrels), which represents the ground truth of the experiment. Next, the user loads the experimental results file. Both files must be in the standard TREC format. Then, all the available standard evaluation measures can be calculated with a press of button. All calculations are taking place locally, i.e. at the client side.

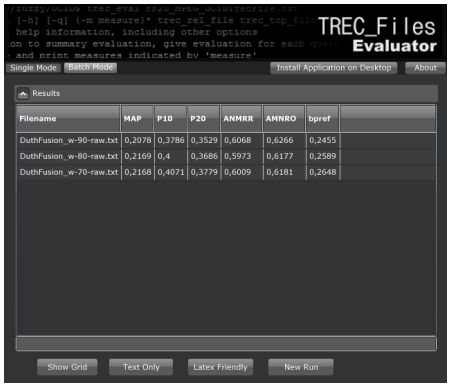


Figure 2: The output of the program in Batch Mode.

The *batch mode* (Fig. 2) allows a user to evaluate several TREC files together and calculate all the measures without manual intervention. At the same time, the system can check statistical significance against a selected baseline run, using a bootstrap, one-tailed test, at significance levels 0.05, 0.01, and 0.001. The results are presented in an interactive table and can be sorted on any measure. Additionally, the TREC_Files Evaluator can export the results in a LaTeX friendly format.

2. CONCLUSIONS

Given that MAP, P10, P20 and bpref constitute main evaluation criteria for the retrieval results in TREC, CLEF, and elsewhere, the TREC_Files Evaluator can be used in order to help either new researchers or organizers of multi-group controlled experiments to evaluate the submitted results more easily. Moreover, the inclusion of ANMRR evaluation metric [1] will encourage researchers from the image retrieval field to employ the standard TREC format for evaluating their experiments.

Future plans include the implementation of more evaluation metrics, such as R-prec and nDCG, as well as enabling users to plug in their custom implementation of metrics.

3. REFERENCES

[1] MPEG-7. *Subjective Evaluation of the MPEG-7 Retrieval Accuracy Measure (ANMRR)*. ISO/WG11, Doc. M6029, 2000.