Overview of the INEX 2012 Relevance Feedback Track

Timothy Chappell[1] and Shlomo Geva[2]

[1] Queensland University of Technology
timothy.chappell@qut.edu.au
[2] Queensland University of Technology
s.geva@qut.edu.au

Abstract. The INEX 2012 Relevance Feedback track provided participating organisations with an evaluation system designed to simulate the user of a search engine. Participants provided their own search systems designed to interface with the evaluation platform and receive live feedback from the simulated user showing which parts, if any, of the current document were considered by the user to be relevant.
This version of the track was run in a very different manner compared to the INEX 2011 and 2010 versions of the Relevance Feedback track in an attempt to increase participation and strengthen the quality of the evaluations. While the former goal was not met, the new format of the track allowed the entire Wikipedia collection[5] to be used, as opposed to the small subsets used in 2010 and 2011.
We present the evaluation methodology, its implementation, and experimental results obtained for thirteen submissions from two participating organisations.

1 Introduction

This paper presents an overview of the INEX 2012 Relevance Feedback track. The track was designed to facilitate the development of search engine modules that incorporate focused relevance feedback. The INEX Wikipedia Collection[5], a 50.7GB collection of 2,666,190 Wikipedia articles in XML format was used as the data collection for the track. The search topics and assessments used were collected for the INEX 2009 and 2010 Ad Hoc tracks[8][1].

Organisations participated by supplying executables that would communicate with a supplied evaluation platform through standard operating system I/O pipes. The evaluation platform would provide the search topics and, for each document provided to it by the search module, reply with relevant passages. The search module can then make use of this information to rerank the remaining documents as necessary.

After each topic has been searched, the evaluation platform uploads the document IDs returned by the search module for each topic in the form of a *trec*

eval[2]-compatible submission. The submission is evaluated on a remote server against relevance assessments for the topics and the results are sent back to the evaluation platform.

The evaluation platform had two modes, training and evaluation, with a different set of topics for each. Training mode would run the module over a smaller set of 10 topics and, while the submission would still be uploaded to the remote server, the results would be returned to the user but not recorded as a submission. In contrast, evaluation mode uses a larger set of 50 topics and records all runs submitted. Hence, users can provide submissions to the track simply by executing evaluation runs with the platform.

2 Focused Feedback

This track covers the use of focused feedback, a relevance feedback model wherein users specify segments of the document (usually through some form of selection or highlighting tool) considered relevant to the search topic. This allows users to give more flexible feedback when only portions of the current document are relevant to their search.

More information about focused feedback is available in [6].

3 Evaluation

Submissions to the Relevance Feedback track are evaluated from the perspective of a user searching for information on a number of topics. The user reads each document returned by the search system and highlights sections that are relevant to the current topic. If the document returned is not relevant at all, the user simply skips this document and asks for a new one. Hence, the search system has an opportunity to rerank the unseen documents at every step along the way, taking into account new information about what the user is searching for. However, as the user's search experience is the ultimate indicator of search performance, documents the user has already seen are considered frozen and can not be reranked after they have been presented.

To simulate the user, relevance judgments from the Ad Hoc tracks from INEX 2009 [8] and 2010 [1] are used to provide information about which segments of documents are relevant to which topics. These assessments consist of offset-length pairs, each indicating that the specified segment in the given document is relevant to the given topic. The evaluation platform uses these assessments, returning the segments that match the given document. The relevance feedback modules can then rerank the remaining documents in the collection with information from this and from previous feedback to produce more relevant results for the remaining documents to be presented to the user.

At the end of a run, the evaluation platform compiles the documents, in the order they were presented, into a *trec eval*-compatible submission file, which is

uploaded to a remote server where the evaluation is performed. The results are then returned to the user. This serves to keep the relevance judgments secret; though only to an extent as the relevance judgments for the Ad Hoc track are publicly available and it is trivial to convert them to TREC format.

In the 2010 and 2011 versions of the Relevance Feedback track, the evaluation platform would provide the relevance feedback plugin with the offset and length of each segment of relevant text. This was changed in 2012 to reduce the need for the entire, uncompressed collection to be available to the relevance feedback plugin. Instead, the direct text from the documents, stripped of XML tags, was passed to the relevance feedback plugin. This made it more practical to create Relevance Feedback submissions without a copy of the uncompressed Wikipedia collection or a copy of the archive that makes random access within the collection feasible. As the default form the Wikipedia collection is distributed in (.tar.bz2) is not suitable for random access, it is difficult and time-consuming to extract individual documents as they are required. This step will also make it more feasible to create Relevance Feedback tasks based on other large collections, such as ClueWeb09[3].

The topics used for this collection were the topics for the INEX 2009 [8] and 2010 [1] Ad Hoc tracks. After stripping out the topics that had no relevance judgments attached, the first ten were used as the training set. Out of the remaining topics, every second topic was used to make up the evaluation set until all fifty slots were filled.

4 Task

4.1 Overview

Track participants were tasked with creating relevance feedback modules that would interface with the provided evaluation platform and respond with results in answer to queries. With each result, the evaluation platform would respond with relevant passages from each document and the relevance feedback module would have the opportunity to rerank the remaining results in that topic to deliver better results.

In past iterations of the track, these relevance feedback modules were implemented as dynamic plugins written in Java. These plugins were provided by the track participants as submissions. This approach, while effective at preventing approaches like tuning to specific topics, came with a number of drawbacks. It restricted the implementation environment to Java. In addition, because it would not be feasible for the users to submit their own index of the collection (which can be hundreds of megabytes large) or index the Wikipedia collection in its entirety at the time of evaluation, only subsets of the Wikipedia collection were, making it more difficult to gather realistic performance information.

In the 2011 iteration of the Relevance Feedback track, the same system was used; however, participants were also provided with a Java plugin capable of

interfacing with generic platform-dependent executables over pipes. This made it possible to implement relevance feedback modules in more languages but brought with it issues of compatibility as the resulting module had to be evaluated on a specific operating system and hardware architecture.

The 2012 iteration made an attempt to rectify these issues, keeping the pipe communication aspect from the binary interface plugin but otherwise heavily changing the way the track was run. In the 2012 Relevance Feedback track, participants create a relevance feedback module in whatever language they choose and the only restriction is that the module run on their own hardware. The evaluation platform was rewritten for the new task and would communicate directly with the relevance feedback module over pipes and make submissions to a remote server, set up specifically for the track. Making a submission with the new system was as simple as running the evaluation platform (in the correct mode.)

Separate training and evaluation modes were included to allow participants to test their relevance feedback modules with the code without needing to make a submission. Every evaluation submission was recorded by the server to ensure that, while participants could still tune their code to the evaluation topics, all the results of doing so would be recorded.

4.2 Submission format

When the track was first opened, the evaluation platform was made available from the INEX website.

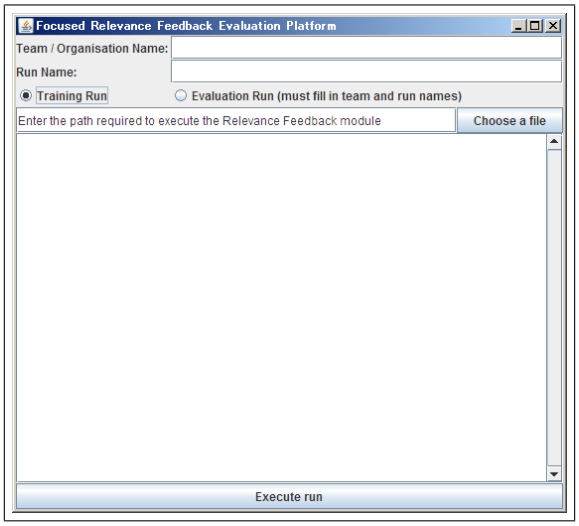

Fig. 1. Evaluation Platform for the INEX 2012 Relevance Feedback track

On supplying a valid path to the relevance feedback module, the evaluation platform would open up an I/O pipe to the module and begin working through the selected topic set.

Choosing the *Evaluation Run* mode would run the module through the 50 topic Evaluation Mode set.

Participating organisations created relevance feedback module executables that adhered to the following specifications, as described in the documentation provided with the evaluation platform:

4.3 Relevance Feedback module interface protocol

The evaluation platform and the module communicate using a pipe, a standard feature of all modern operating systems. Hence, any programming language capable of creating an executable that can read from standard input and write to standard output would be suitable for creating a relevance feedback module for the task.

Each message from the evaluation platform or the relevance feedback module will be in the form of a single line of text ending in a linefeed character. The meaning of the line of text will be derived from the context in which it is submitted.

The evaluation platform communicates first, providing a topic line. This line will either contain the text of the topic or the text EOF, signalling to the module that the evaluation is over and it may exit. The module will respond with a document line. This line will contain either a document ID or the text EOF, signalling to the evaluation platform that the module has finished presenting documents for the current topic and is ready to move on to the next topic. If a document ID is presented, the evaluation platform will respond with feedback.

Feedback will be provided in the form of a line with a number indicating the number of passages of relevant text found in the document. If that number was 0, the document was not relevant and the module should provide the next document ID. Otherwise, the evaluation platform will immediately follow up the number with that many passages of feedback text, each on a single line. After all the lines of feedback have been sent, the module is expected to respond with another document.

4.4 Relevance Feedback module interface format

The topic line supplied by the evaluation platform will be in ASCII text, stripped of characters outside the 32-127 range. The line will be no more than 127 characters long, including the linefeed.

The document ID line returned by the module should contain a number in ASCII text, corresponding to the document ID within the Wikipedia collection of the document to return.

The 'lines of feedback' line returned by the evaluation platform in response to a document ID line will be a number in ASCII text containing the number of segments of relevant text in the document. The feedback will then be followed by lines of text, one for each segment of feedback. The line will be no more than 1048575 characters long, including the linefeed. This, too, will be in ASCII text, stripped of characters outside the 32-127 range.

5 Results

5.1 Submissions

Two groups made a total of 15 submissions to the INEX 2012 Relevance Feedback track, up from four submissions from two groups in 2011. This may be partly due to the new format making it easier to make many submissions as the need for each submission to be packaged into its own Java archive and uploaded was no longer present.

Queensland University of Technology made five submissions using an experimental relevance feedback mode in TopSig[4]. This was originally planned to be the relevance run for the INEX 2012 Relevance Feedback track but due to time constraints this was not possible. The TopSig runs, apart from the baseline run which did not make use of feedback at all, simply used the feedback text as a new query and reranked the remaining documents found by the initial query each time. More information about the signature approach used by TopSig can be found in [7].

The baseline TOPSIG run consisted of an untuned 1024-bit signature search without using collection statistics or relevance feedback returning 100 documents per topic. Subsequent TOPSIG runs incorporated the simple feedback system described earlier. TOPSIG-RF1 reranked the remaining documents not yet presented to the user by using the last line of feedback presented as a new search query. TOPSIG-RF2 kept the same approach but increased the number of documents returned to 1000. TOPSIG-RF3 increased the signature size to 2048 bits and TOPSIG-RF4 changed the feedback approach to use all of the feedback presented instead of the last line. As this is the first experiment performed with using active relevance feedback for signature searching in TopSig, preliminary results are only experimental.

The Universidad Autónoma Metropolitana made 10 submissions using Indri[9] as a base and employing a Markov random field to rerank results with relevance feedback. The BASE-IND run consists of a run with Indri without incorporating relevance feedback while the MF and LF runs consist of the results when adding the 20 most frequent and least frequent terms respectively from the feedback to the query. The RRMRF runs are also based on Indri but employ the Markov random field for reranking. The 100D, 300D and 1000D runs are the results from returning 100, 300 and 1000 documents respectively per topic. The

L values represent the lambda parameter within the reranking approach. More details are available in the Universidad Autónoma Metropolitana's track paper.

5.2 Evaluation

Two sets of topics were made available, not directly to participants but through the evaluation platform. The training set used the first 10 topics from the INEX 2009 Ad Hoc track while the evaluation set used 50 topics chosen from every 2nd topic from the INEX 2009 and 2010 Ad Hoc tracks, excluding the topics used for the training set. Topics without associated relevance judgments were removed from the set beforehand.

All of the submissions were run through *trec eval*[2] using default settings. The results of each run were also presented to the submitter immediately after submission.

Trec eval reports results using a variety of different metrics, including interpolated recall-precision, average precision, exact precision and R-precision. Recall-precision reports the precision (the fraction of relevant documents returned out of the documents returned so far) at varying points of recall (after a given portion of the relevant documents have been returned.) R-precision is calculated as the precision (number of relevant documents) after R documents have been seen, where R is the number of relevant documents in the collection. Average precision is calculated from the sum of the precision at each recall point (a point where a certain fraction of the documents in the collection have been seen) divided by the number of recall points.

Unlike in the previous incarnations of the relevance feedback track, the evaluation platform did not come with the option of producing no-feedback runs. However, both participating organisations created runs that did not utilise feedback, showing where feedback has improved the results of these runs.

5.3 Comparisons

The following tables show the results of each submission in terms of average precision and R-precision.

The charts compare groups of submissions by exact precision. The y axis shows the proportion of relevant documents retrieved and the x axis shows the total number of documents retrieved. Figure 2 shows a comparison of the exact precision of each of the UAM runs submitted. Figure 3 shows a comparison of each QUT run, while figure 4 gives a comparison of the best feedback and non-feedback runs from each group.

Group	Submission	Average Precision	R-Precision
UAM	BASE-IND	0.1015	0.1828
UAM	BASE-INDQE-20tMF	0.0775	0.1396
UAM	BASE-INDQE-20tLF	0.0395	0.0718
UAM	BASE-INDQE-20tMFandLF	0.0728	0.1364
UAM	RRMRF-100D-L03	0.094	0.1612
UAM	RRMRF-100D-L05	0.0946	0.1595
UAM	RRMRF-300D-L03	0.1002	0.1769
UAM	RRMRF-300D-L05	0.1004	0.1805
UAM	RRMRF-1000D-L03	0.1015	0.1824
UAM	RRMRF-1000D-L05	0.1015	0.1824
QUT	TOPSIG	0.1393	0.2059
QUT	TOPSIG-RF1	0.1459	0.2028
QUT	TOPSIG-RF2	0.2015	0.2509
QUT	TOPSIG-RF3	0.2352	0.2747
QUT	TOPSIG-RF4	0.2408	0.2763

Table 1. Average precision and R-precision for submitted runs

Submission	@5	@10	@15	@20	@30	@100	@200	@500	@1000
BASE-IND	0.456	0.41	0.36	0.327	0.3007	0.1844	0.1166	0.0557	0.0292
BASE-INDQE-20tMF	0.396	0.33	0.2907	0.264	0.228	0.1232	0.0789	0.0406	0.0239
BASE-INDQE-20tLF	0.308	0.198	0.148	0.125	0.0967	0.0412	0.0241	0.0114	0.0063
BASE-INDQE-20tMFandLF	0.392	0.304	0.2653	0.237	0.204	0.1136	0.0741	0.0396	0.0225
RRMRF-100D-L03	0.448	0.406	0.3733	0.348	0.3107	0.1846	0.0923	0.0369	0.0185
RRMRF-100D-L05	0.452	0.404	0.372	0.351	0.312	0.1846	0.0923	0.0369	0.0185
RRMRF-300D-L03	0.452	0.398	0.3587	0.337	0.3027	0.1876	0.117	0.0512	0.0256
RRMRF-300D-L05	0.46	0.42	0.368	0.349	0.3	0.1708	0.1157	0.0512	0.0256
RRMRF-1000D-L03	0.456	0.41	0.36	0.328	0.3007	0.1848	0.1166	0.0557	0.0292
RRMRF-1000D-L05	0.456	0.41	0.36	0.328	0.3007	0.1848	0.1166	0.0557	0.0292
TOPSIG	0.448	0.42	0.3827	0.366	0.332	0.232	0.116	0.0464	0.0232
TOPSIG-RF1	0.496	0.44	0.4067	0.384	0.3593	0.232	0.116	0.0464	0.0232
TOPSIG-RF2	0.524	0.46	0.4173	0.398	0.3747	0.242	0.1733	0.0933	0.0569
TOPSIG-RF3	0.56	0.504	0.4813	0.465	0.4187	0.2614	0.1906	0.1049	0.0623
TOPSIG-RF4	0.568	0.52	0.4773	0.459	0.42	0.2656	0.1923	0.1054	0.0623

Table 2. Exact precision of submitted runs

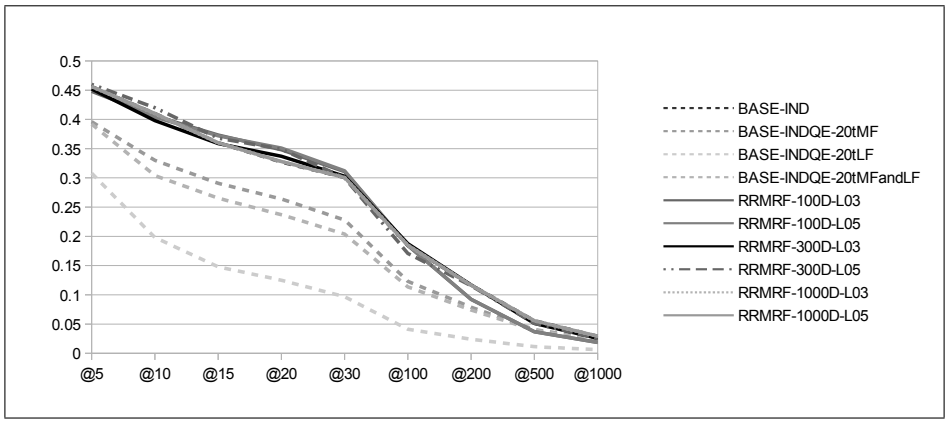

Fig. 2. Exact precision of submissions by the Universidad Autónoma Metropolitana

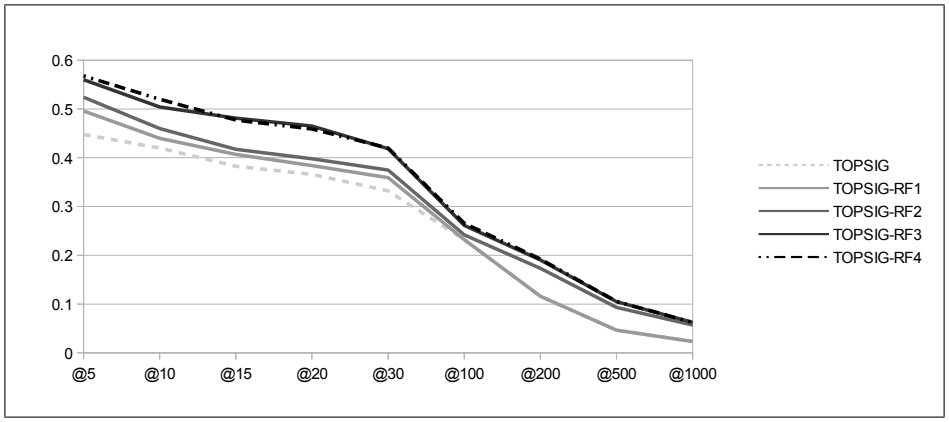

Fig. 3. Exact precision of submissions by the Queensland University of Technology

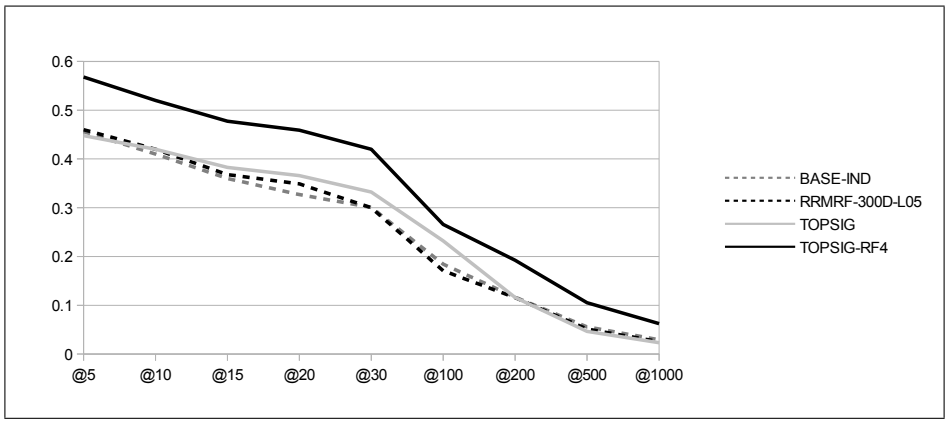

Fig. 4. Exact precision comparison: best non-RF and best RF submissions from each participating organisation

6 Conclusion

We have presented the Relevance Feedback track at INEX 2012.

It is difficult to compare results between different incarnations of the track. While the results are far worse in 2012 from an objective perspective, the large changes in the way the tracks were run between the two years can account for this. In the 2010 and 2011 versions of the Relevance Feedback track, only subsets of the Wikipedia collection were used and these subsets heavily favoured relevant documents. As the burden of finding the results has shifted more to the search systems in the 2012 version of the track the overall results have also declined. The search systems presented at the INEX 2012 Relevance Feedback track are not necessarily worse than those presented in 2011.

While the number of submissions has increased since INEX 2011, the number of participants has not. Lowering the barriers to entry have not resulted in the increased interest in the Relevance Feedback track that was expected. Part of this may be due to the lack of a strong reference submission. In the INEX 2010 and 2011 iterations of the Relevance Feedback track, a relevance feedback module with complete was provided to participants in advance, to be used as a base for other submissions if desired. No equivalent was provided for the INEX 2012 Relevance Feedback track which may have discouraged participation.

7 Acknowledgements

We would like to thank all the participating organisations for their contributions and hard work.

References

1. P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the INEX 2010 Ad Hoc track. *Comparative Evaluation of Focused Retrieval*, pages 1–32, 2011.
2. C. Buckley. trec eval IR evaluation package, 2004.
3. J. Callan, M. Hoy, C. Yoo, and L. Zhao. Clueweb09 data set. *boston. lti. cs. cmu. edu, Jan*, 2009.
4. T. Chappell and S. Geva. TopSig: Topological signature indexing and search engine. *http://www.topsig.org/*, 2012.
5. Ludovic Denoyer and Patrick Gallinari. The Wikipedia XML Corpus. *SIGIR Forum*, 2006.
6. S. Geva and T. Chappell. Focused relevance feedback evaluation. *Simulation of Interaction*, page 9, 2010.
7. S. Geva and C.M. De Vries. Topsig: Topology preserving document signatures. 2011.
8. S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the INEX 2009 Ad Hoc track. *Focused Retrieval and Evaluation*, pages 4–25, 2010.
9. T. Strohman, D. Metzler, H. Turtle, and W.B. Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.