TREC10 Notebook Paper

Challenges of Multi-Mode IR Software

Gregory B. Newby
UNC Chapel Hill

Abstract

Web track results are presented. A software project, IRTools, is described. IRTools is intended to enable information retrieval (IR) experimentation by incorporating methods for multiple modes of IR operation, such as the vector space model and latent semantic indexing (LSI). Plans for the interactive track are described.

Introduction

For much of the past year, the author and his colleagues have been working towards general-purpose large-scale software for information retrieval experimentation. For TREC 10, the goal was to demonstrate this software's functionality using a "standard" IR approach: vector space retrieval. Functionality demonstrated in prior years' TRECs, notably the information space technique (Newby, 2001) and other approaches related to LSI (described in Rehder et al., 1998) was present but untested for TREC 10.

Submitted runs for the TREC10 Web track were irtLnua and irtLnut:

irtLnua:	Web track, all terms minus stopwords, Lnu.Ltc weighting
irtLnut:	Web track, title only, minus stopwords, Lnu.Ltc weighting

We also did some work on cross language retrieval but did not submit runs. Our work for the interactive track will not be completed in time for presentation at TREC10, but should be ready for the final proceedings.

Software Overview

We believe there is a lack of free, open source, high performance software for information retrieval. We desire to create software with these qualities:

1. Free and open source (e.g., licensed under the General Public License);
2. With implementations for multiple IR methods, including Boolean retrieval, vector space, probabilistic and LSI, as well as variations;
3. Including documentation and examples to enable interested persons to perform experiments, extend the software, or incorporate it with their own tools;
4. Suitable for medium (10GB to 100GB) to large (up to 1000GB) collections of documents; and
5. With a focus on semi-structured documents, including HTML and XML formats, but also compatible with plain text.

Software for IR experimentation that has seen great success includes SMART (Buckley & Walsh, 2001) and Okapi (Robertson & Walker, 2000). However, past versions of such systems have lacked one or more of the qualities above. INQUERY (Allan et al., 2001), like some other successful systems, is not open source. Free search software such as HT://DIG (http://www.htdig.org) offer high performance and open source, but are not readily adaptable for retrieval research.

In contrast to the systems that regularly appear at TREC conferences and other venues, some of the most successful and widely used systems for IR – Web search engines – are prohibitive of most forms of experimental IR research. Despite starting as open or publicly funded projects, popular Web search engines including Lycos, Yahoo and Google do not make their software or algorithms publicly accessible, and there are few opportunities for the utilization of their techniques for TREC-style experimental research.

The above is not intended as criticism of the software or the people behind it. In fact, the list above is indicative of the great success that IR has had in bringing people closer to the information they seek. Nevertheless, there is certainly room for at least one project with the goals above.

The software, which is called the Information Retrieval Toolkit (IRTools), is not intended as a panacea, nor does it propose to supplant existing systems. Instead, as the name implies, it is intended as one possible addition to the modern experimental IR researcher's collection of software and algorithms. The source code for IRTools is available at http://sourceforge.net/projects/irtools .

Web Track Results

Development of IRTools has been steady but slow. File structures, data structures and algorithms have been under constant development and reassessment, and it seems that at any time only part of the software works. To benchmark the software, we wanted to submit runs with fairly standard and well-known approaches. The VSM with Lnu.Ltc weighting was utilized for TREC10. For the pivoted document weights, a constant of 0.25 was chosen based on experiments with TREC9 qrels.

Two runs were submitted, irtLnua and irtLnut. IrtLnua included all non-stopworded terms, and resulted in abysmal results with average precision well under 1%. These results are worse than might be expected if randomly retrieved documents were submitted. There appear to be one or more bugs in the Boolean recombination or term weighting subsystem that resulted in documents with low-value terms being ranked highly. These are disappointing results, but appear to be the outcome of one or more bugs.

IrtLnut is better, though not as good as we expected. As a benchmark run, we anticipated performance similar to our work with post-hoc evaluation of TREC9 runs, in which we typically gained average precision of .25 or so.

For this run, only terms from the <TITLE> section of each topic were used, minus terms on the 622-term stop list (similar to the SMART list). Results are presented in Table 1.

Table 1: irtLnut (judged run) Result Summary

IrtLnut Overall statistics	
Retrieved	46432
Relevant	3363
Rel_ret	838
Exact:	0.0321

Relevant at 1000 docs:	
Runs >= Median	2
Runs < Median	48
Runs with 0 relevant docs	16

Average Precision:	
Runs >= Median	3
Runs < Median	47
Runs with 0 ave_p	22

Generally, topics which other systems found "easy" (in terms of a high median relevant documents at 1000) were found easy in the irtLnut run. Such topics included 509, 513, 527, 530, 544 and 547.

Anomalous topics, in which irtLnut was very low but the median relevant documents found across all participants was high, were 511, 519, 541 and 549. These topics appear to be victims of the same bug that impacted irtLnua – unimportant terms (such as "info" in topic 519) were given higher weights than important terms (such as "frogs").

The best runs for irtLnut included 509 ("steroids what does it do to your body"), 517 ("titanic what went wrong"), 527 ("can you find info on booker t Washington") and 544 ("estrogen why needed"). In all cases, our suspicion is that the initial pre-weighting Boolean set of documents was of sufficiently high quality to offset bugs in term weights and ranking.

Interactive Track Plans

Our work on the interactive track is ongoing. The plan for the study is to test for differences in search results and performance between two versions of the results display interface. The control interface will display results in a traditional list format, whereas the experimental interface will display results in a browseable category hierarchy, based on the Yahoo categories.

Our intent is to produce testable hypotheses about the presentation of a fixed number of results in text and several non-text formats.

There will be 24 participants in the study. Each participant will do two searches on the control system (one fully specified, one partially specified) and two searches on the experimental system (one fully specified, one partially specified). The four topics are distributed evenly across the participants so that each participant is dealing with tasks from only two of the four topics, one partially specified and one fully specified task from each topic.

Searches will be run on Google against the live Web as indexed there. Mapping resulting hits to the Yahoo categories will occur via a proxy server on our local system, using a combination of standard vector space algorithms and some categorization algorithms to address granularity problems (e.g., to make sure we don't present dozens of low-level categories that all share higher-level categories).

As the participants are searching, we will automatically record the URL of each document they view via the proxy server. We will also ask the participants to record the URL(s) of document(s) they believe satisfy the requirements of each task. In addition, we will record the total amount of time each participant spends completing each task.

Each participant will complete a pre-search questionnaire that asks for basic demographic information as well as prior experience with web searching in general and web searching particularly related to the two domains (medical and travel) and two actions (buying online and researching a topic for a project) specified in the tasks. Participants will be given a post-search questionnaire to evaluate each system and express what they like or dislike about each.

Conclusion

The Web track results support our plan to first implement relatively well-known IR techniques in IRTools, in order to gain confidence in our system's performance. As has happened in prior years to other TREC participants, last-minute bugs appear to have thwarted our efforts at reasonable benchmarks.

Integrating well-known techniques into an integrated software package has proven to be challenging. File structures have been particularly problematic, as different data points are required for different IR schemes, yet we desire to minimize disk I/O while having generalized data structures stored to disk. Scaling for sparse-matrix techniques (LSI) as well as dense-matrix techniques (information space) has also been challenging.

Despite these challenges, we anticipate success in achieving our goals for IRTools. We speculate being able to approximate results from best-of-breed IR systems within IRTools, enabling controlled experiments comparing the impact of different manipulations. We hope that these efforts, combined with the work of other research groups, will improve retrieval and scaling for semi-structured textual data.

References

Allan, J.; Connell, M.E.; Croft, W.B.; Feng, F.; Fisher, D. & Li, X. 2001. "INQUERY at TREC9." In Voorhees, Ellen and Harman, Donna (Eds.). The Ninth Text REtrieval Conference (TREC-9). Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-249.

Buckley, Chris & Walsh, Janet. 2001. "SabIR Research at TREC-9." In Voorhees, Ellen and Harman, Donna (Eds.). The Ninth Text REtrieval Conference (TREC-9). Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-249.

Newby, Gregory B. 2001. "Information Space Based on HTML Structure." In Voorhees, Ellen and Harman, Donna (Eds.). The Ninth Text REtrieval Conference (TREC-

9). Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-249.

Rehder, B.; Landauer, T.K; Littman, M.; Dumais, S. 1998. "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing." In Voorhees, Ellen and Harman, Donna (Eds.). The Sixth Text REtrieval Conference (TREC-6). Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-240.

Robertson, S.E. & Walker, S. 2000. "Okapi/Keenbow at TREC-8." In Voorhees, Ellen and Harman, Donna (Eds.). The Eighth Text REtrieval Conference (TREC-8). Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-246.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF award #CCR-0082655).

Many students have worked on the software for this year's results. Their efforts are significant and valued. Some of these students include: Andre Burton, Sheila Denn, Miles Efron, Wei Gao, Xiaosi Li, Monique Lowe, Carolyn Mitchell, Corey Nickens, Zach Sharek, Chang Su, and Yuehong Wang and Jewal Ward. The author, as project director, takes all the blame for bugs and other errors.