TREC2002 Notebook Paper

Progress in General-Purpose IR Software

Gregory B. Newby
UNC Chapel Hill

Abstract

TREC 2002 experiments were run, but not submitted in time for inclusion in the official conference results. Post-hoc analysis will be presented in the final conference proceedings paper. Progress on general-purpose IR software for experimental use has been very good, and some features of the software are described. A new focus on IR for grid computing, GridIR, is described.

Introduction

The IRTools software developed by the author and his colleagues was used again this year. Both Lnu.Ltc and LSI were used. Options for automatic query expansion and pseudo relevance feedback were available, as well as a variety of widgets for stoplist processing etc.

Unfortunately, runs were completed just a few hours too late and so were not included in the TREC conference results.

Completed runs for TREC 2002 included:

CLIR: Monolingual Arabic
Web: Topic Distillation

Interactive track data collection is not yet completed, but see the poster and demo session for more information about the systems being evaluated.

Software Overview

IRTools is intended to be a general-purpose toolkit for information retrieval research. It was funded in part by the NSF through an Information Technology Research grant. The source code for IRTools is available at **`http://sourceforge.net/projects/irtools`**.

In 2002, IRTools is nearing maturity. It offers high-performance indexing and retrieval, and many of the features found in other experimental IR systems – but with more of an emphasis on allowing the programmer to change parameters, extend functionality, etc. Completion of IRTools for public release is scheduled for May 2003. At that time, modules to be included are:

- Indexing for multiple document type: XML, text and HTML
- Processing for English, Arabic, Chinese and other languages
- Local file indexing as well as remote harvesting
- Several fundamental IR techniques:
    - Enhanced Boolean
    - VSM

    - LSI
    - Information space
- Several fundamental IR enhancements:
    - Query expansion
    - Document summarization

The toolkit uses the BerkeleyDB for the back end database and Michael Berry's SVDPACKC for eigensystems. Other components are home-grown. The system runs on Unix and Linux systems with the GCC compiler and has been tested extensively on Linux and Solaris systems.

CLIR Arabic Monolingual Results

Two monolingual Arabic runs were completed. One utilized the entire document; the other utilized the title only (intended for high early precision). Basic tools provided by the track coordinators were applied to modify the topic character set to match the document set, but no other processing was done (i.e., no stemming, stopwords, or analysis of document structure). This "bag of words" approach was envisioned as a starting point for further experimentation.

For this run, the VSM was used with Lnu.Ltc weighting (pivoted document length normalization with the cosine measure of association).

Note that in the title only run, all other terms were ignored. Indexing for both title only and whole document ran as part of one IRTools indexing program and took about an hour for the 895MB of text (383K documents with 660K unique terms). Summary results are presented in Tables 1 and 2.

Table 1: Monolingual Arabic for irtta (title only)

```
      Total number of documents
over all queries
          Retrieved:      335
          Relevant:      3055
          Rel_ret:        121
      Interpolated Recall -
Precision Averages:
          at 0.00       0.5461
          at 0.10       0.1937
          at 0.20       0.0122
          at 0.30       0.0115
          at 0.40       0.0000
          at 0.50       0.0000
          at 0.60       0.0000
          at 0.70       0.0000
          at 0.80       0.0000
          at 0.90       0.0000
          at 1.00       0.0000
```

```
      Average precision (non-
interpolated) for all rel
docs(averaged over queries)
                        0.0352
      Precision:
        At    5 docs:   0.2889
        At   10 docs:   0.2111
        At   15 docs:   0.1741
        At   20 docs:   0.1611
        At   30 docs:   0.1333
        At  100 docs:   0.0667
        At  200 docs:   0.0336
        At  500 docs:   0.0134
        At 1000 docs:   0.0067
      R-Precision (precision after
R (= num_rel for a query) docs
retrieved):
          Exact:        0.0537
```

Table 2: Monolingual Arabic for irtba (whole document)

```
      Total number of documents
over all queries
          Retrieved:     2411
          Relevant:      4370
          Rel_ret:        730
      Interpolated Recall -
Precision Averages:
          at 0.00       0.4789
          at 0.10       0.1498
          at 0.20       0.1216
          at 0.30       0.0679
          at 0.40       0.0671
          at 0.50       0.0633
          at 0.60       0.0547
          at 0.70       0.0513
          at 0.80       0.0000
          at 0.90       0.0000
          at 1.00       0.0000
```

```
      Average precision (non-
interpolated) for all rel
docs(averaged over queries)
                        0.0709
      Precision:
        At    5 docs:   0.2059
        At   10 docs:   0.1882
        At   15 docs:   0.1804
        At   20 docs:   0.1691
        At   30 docs:   0.1529
        At  100 docs:   0.0894
        At  200 docs:   0.0671
        At  500 docs:   0.0411
        At 1000 docs:   0.0215
      R-Precision (precision after
R (= num_rel for a query) docs
retrieved):
          Exact:        0.0970
```

As hoped, the title-only run yielded higher early precision, but (also as expected) failed entirely for a number of queries. Of the 50 TREC topics, only 19 yielded any results for this run (indicating that there were no Arabic collection documents with all query terms in the title). Topics with some relevant document retrieved included AR37, AR44, AR45, AR48, AR49, AR50, AR51, AR55, AR56, AR61, AR69, and AR74. From this run, we learned that title processing can be effective alone, but fails more often than not if it is the sole basis for retrieval. Combining title retrieval (or differently weighting the title words) with other techniques is indicated.

The base run, using all terms (without differential weighting for title terms), yielded a greater number of relevant documents retrieved (730 vs. 121 for title-only) but lesser early precision and weaker precision over all. Exact precision did not suffer as much, presumably due to a smaller number of failed queries. Nevertheless, only 35 out of 50 topics yielded any results, and 15 of those had no relevant documents. Here, we suffered from working exclusively with the exact match Boolean AND of topic terms. The lack of stemming, plus the lack of any query expansion or partial-match ranking, hurt the set of documents that could be considered and ranked for retrieval.

Overall, these results provide a baseline for VSM-style processing of Arabic documents for mono-lingual runs. Obvious features for inclusion for better results include stemming, query expansion, and differential weighting based on document components such as the title.

Web Track

Two runs for the topic distillation task in the Web track were run. As for the Arabic runs, one was title-only and the other used the entire document. The IRTools indexing took about 4 days for the collection (20GB of HTML documents, about 1.2M documents and 6.37M unique terms). Summary results are in Tables 3 and 4.

Table 3: Web Topic Distillation for irtwt title-only

```
Total number of documents over all
queries
    Retrieved:      901
    Relevant:       737
    Rel_ret:         34
Interpolated Recall - Precision
Averages:
    at 0.00       0.2232
    at 0.10       0.0825
    at 0.20       0.0236
    at 0.30       0.0035
    at 0.40       0.0035
    at 0.50       0.0035
    at 0.60       0.0000
    at 0.70       0.0000
    at 0.80       0.0000
    at 0.90       0.0000
    at 1.00       0.0000
```

```
Average precision (non-
interpolated) for all rel
docs(averaged over queries)
                  0.0237
Precision:
  At    5 docs:   0.0741
  At   10 docs:   0.0519
  At   15 docs:   0.0370
  At   20 docs:   0.0333
  At   30 docs:   0.0284
  At  100 docs:   0.0126
  At  200 docs:   0.0063
  At  500 docs:   0.0025
  At 1000 docs:   0.0013
R-Precision (precision after R (=
num_rel for a query) docs
retrieved):
    Exact:        0.0338
```

Table 4: Web Topic Distillation for irtwb (whole document)

```
Total number of documents over all
queries
    Retrieved:     4728
    Relevant:      1574
    Rel_ret:        141
Interpolated Recall - Precision
Averages:
    at 0.00       0.1153
    at 0.10       0.0740
    at 0.20       0.0465
    at 0.30       0.0243
    at 0.40       0.0235
    at 0.50       0.0174
    at 0.60       0.0042
    at 0.70       0.0010
    at 0.80       0.0010
    at 0.90       0.0000
    at 1.00       0.0000
```

```
Average precision (non-
interpolated) for all rel
docs(averaged over queries)
                  0.0222
Precision:
  At    5 docs:   0.0653
  At   10 docs:   0.0429
  At   15 docs:   0.0408
  At   20 docs:   0.0398
  At   30 docs:   0.0381
  At  100 docs:   0.0286
  At  200 docs:   0.0144
  At  500 docs:   0.0058
  At 1000 docs:   0.0029
R-Precision (precision after R (=
num_rel for a query) docs
retrieved):
    Exact:        0.0372
```

The Web results were not good. Some topics (such as 552) had perfect or near-perfect early precision, while others (such as 551) found no relevant documents at all. Analysis of these results indicates that the main problem is not having heuristics in place to identify good distillation pages, instead relying on regular topic-based matching geared towards term matching. Results were marginally better for the title-only run.

Work to improve results will focus on incorporating document structure into results; in particular the title and heading data which might better indicate good candidates for distillation. In addition, heuristics to look at the document URL itself (which was completely ignored) will help, by flagging shorter URLs are potentially more likely to be good distillation candidates.

Interactive Track

For the interactive track, we are comparing two nearly identical systems using a Google-like text-based interface. Both use the same set of documents, and both make an initial set of candidate documents for ranking using a Boolean AND. They use the Web 02 collection (20GB of HTML from .gov). The difference is that one system uses LSI for ranking results, the other uses VSM with Lnu.Ltc ranking. Document summarization is via Perl modules from CPAN.

Our hypothesis is that the differences in ranking will make no difference in the user experience (i.e., results on measured variables will not be significantly different). We intend this as a base study to explore further variations:

- Systems where the ranked set of documents is different, via automatic query expansion
- Systems where result sets are visualized in a 3D fly-through system

Unfortunately, last year's interactive track was not completed (we intended to compare a text list of results to a browseable category hierarchy), primarily because IRTools was not up to the task. This year, however, the systems are up and running and giving reasonable results. Currently, the test interfaces are accessible:

http://underdog.ils.unc.edu/cgi-bin/nph-lsi.cgi (text interface to LSI)

http://underdog.ils.unc.edu/cgi-bin/nph-vsm.cgi (text interface to VSM)

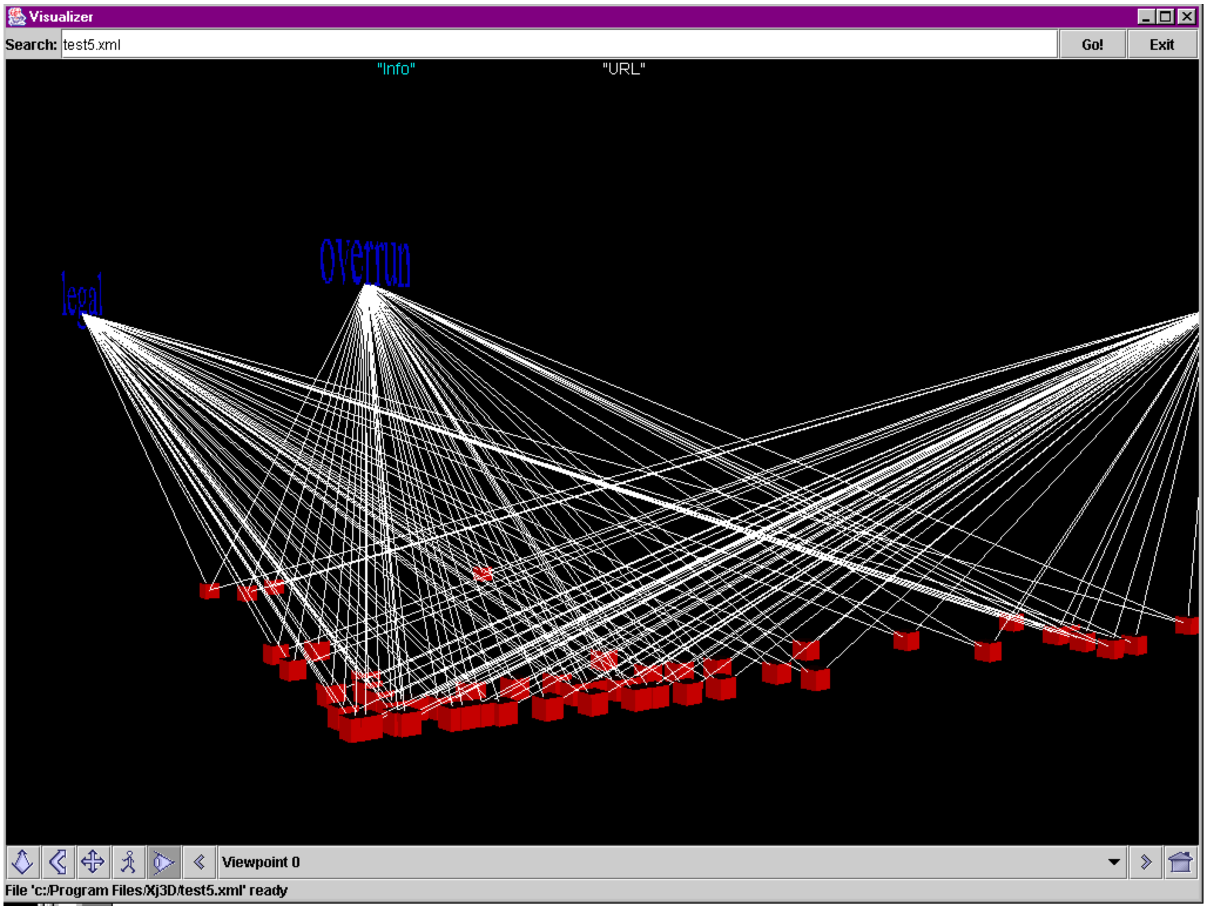
http://underdog.ils.unc.edu/cgi-bin/nph-query.cgi (VSM with database select)

The 3D interface is implemented in Web3D (essentially, a modern VRML '97 implemented over Java3D). This interface runs by accepting user queries, running them against the LSI module of IRTools, then displaying the resulting set of term and document locations and relationships. A simple XML structure is used to communicate between the visualizer and the server.

Note that the LSI applied is only to the Boolean AND of search terms, or a slightly expanded set of search terms. This is done while the user waits (usually within a few seconds, depending on the number of terms and documents being considered). For larger-scale LSI, we intend to construct some very large LSI spaces into which queries may be mapped. But for general visualization of search set results using only documents that contain the query terms, the technique described here seems to work well.

We will evaluate this visual interface in several contexts, and determine whether it is effective in determining relations among documents in post-search result sets.

Figure 1: 3-word query with lines denoting set membership. Clickable documents appear as small cubes.



Search: test5.xml

File 'c:/Program Files/Xj3D/test5.xml' ready

GridIR

Grid computing is an important advance in computational techniques. It has some concepts in common with distributed computing and with massively parallel computing, but many added features. GridIR is IR on the computing grid. The author and his colleagues have worked to form a GridIR working group under the auspices of the Global Grid Forum (http://gridforum.org). We believe that GridIR offers important advantages to IR researchers, and will make experimental and mainstream IR systems more usable and better suited for large-scale research.

Grid computing has a security model built in, making GridIR suitable for publishing partial extranets or implementing security at the query, collection, document or user level. We are currently working on a draft requirements document for the GGF for delivery in Spring 2003, and welcome input and efforts from other IR researchers. Reference systems for GridIR will include IRTools and Amberfish, and we welcome others. Our goal is to develop a set of actual standards for GridIR (under the GGF, following a rulemaking procedure similar to the IETF). We are building on knowledge from Z39.50 and other efforts, and hope to enable a far higher level of interoperability among content maintainers, searchers and IR systems than is now available.

Visit the GridIR Web site to learn more: http://www.gridir.org.

Conclusion

IRTools continues to develop, and despite results being late was able to handle the Web and Arabic tracks with relative ease. Continued work will make IRTools more usable, and integration with the GridIR reference implementation will help to shake out bugs and shape future developments. CLIR continues to be a focus, with new modules for Chinese and Arabic recently added.

IR researchers are urged to consider GridIR as a possible activity. Credibility and buy-in from IR systems developers, vendors, scholars, etc. will help make GridIR as beneficial as possible.

Acknowledgements

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

Many students and employees have worked on the software for this year's results. Their efforts are significant and valued. Some of these fine people include: Li Chen, Nassib Nassar, Mao Ni, Li Wen, and Yuehong Wang. Alan Hudson and Yumetech, Inc. of Seattle have performed work on the visualizer, under a grant from the E.S.P. Das Educational Foundation.