Sabir Research at TREC 9

Chris Buckley and Janet Walz

Sabir Research participated in TREC-9 in a somewhat lower key fashion than normal. We participated only in the Main Web Task, and in the Query Track. Most of our interesting work and analysis was done in the Query Track, and is reported in the TREC-9 Query Track Overview. Here we report very briefly on the Main Web Task; briefly because there really isn't much interesting to say this year!

We used the SMART Information Retrieval System Version 13.3.6 for our runs. SMART was developed at Cornell and continues to be developed at Sabir Research. The basic algorithms have been described numerous times in our past TREC papers [1, 2, 3]; we made no major changes this year. They includes blind feedback with query zoning, and looking at correlated terms.

The Web data itself posed no problems for SMART. This was our first Web test collection, but basic indexing and retrieval was straightforward (modulo a forgotten check to ensure no single word exceeded 512 characters in length.) SMART indexes at 3 to 4 GBytes per hour on a cheap PC running Linux.

What does pose a problem is trying to take advantage of the additional Web information available. In our retrospective tests on last year's TREC-8 Web data, and in our tests both before and after TREC-9 submissions, nothing we tried seemed to affect the results much! Experiments with anchor text, links, and trying to emphasize certain parts of the documents all had basically no effect on retrieval results. In most cases they had a minor detrimental effect. Even basic retrieval and indexing variations such as stemming, phrasing, and document length normalization had little effect on the Web results; less effect than we would have expected. Given our results are 10% – 20% under the current top groups, which include other groups that were running SMART like AT&T [4], we obviously need to look at things in more detail.

One known weakness in our current setup is choice of query expansion terms from blind feedback. We haven't yet played around with options here because we have an investigation in the area planned for the near future as we make major changes to SMART. SMART currently offers several choices for expansion, but Sabir has stayed with expanding by terms related to as many top documents as possible. That appears to be non-optimal for recent TREC test collections, as too many expansion terms are not content-bearing terms. In earlier TRECs, with more relevant and near relevant documents per query, we were able to pull in the general terms which described the query content area. We need some method of distinguishing which queries we can draw in these good general

Run	Mean Avg Prec	R-Prec	NumRelRet
Sab9web1	.1265	.1518	1250
Sab9web2	.2122	.2463	1468
Sab9web3	.2159	.2464	1456
Sab9web4	.2091	.2485	1476
Sab9web5	.2018	.2400	1468

Table 1: TREC-9 Main Web Task Results

terms, and which queries we need to target specific terms.

We submitted 5 runs to the Main Web Task. Sab9web1 was run with the title words only from the topic statement. Sab9web2, Sab9web3, and Sab9web4 used the entire topic statement. Sab9web5 used the entire topic statement plus used link information from the Web pages. As can be seen from Table 1, as expected Sab9web1 is significantly worse than the others, but all the variations we tried using the entire topic statement didn't result in any changes. All of our runs are above average for their respective categories, but not in the top group of systems this year.

In conclusion, Sabir Research participated in the Main Web Task and the Query Track. The Query Track investigation is reported elsewhere. In the Web Task, we used the same basic approach as the past few years and got above average, but not top results. We were unable to get any improvement using Web-specific data such as links, anchor texts, and content placement. That doesn't mean the data may not be useful in the future, only that we were unable to take advantage of it here.

References

[1] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. Using clustering and SuperConcepts within SMART: TREC-6. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*, pages 107–124, August 1998. NIST Special Publication 500-240. Electronic version available at http://trec.nist.gov/pubs.html.

[2] Chris Buckley, Mandar Mitra, Janet Walz, and Claire Cardie. SMART high precision: TREC 7. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 285–298, August 1999. NIST Special Publication 500-242. Electronic version available at http://trec.nist.gov/pubs.html.

[3] Chris Buckley and Janet Walz. SMART in TREC 8. In Voorhees and Harman [5]. NIST Special Publication 500-246. Electronic version available at http://trec.nist.gov/pubs.html.

[4] Amit Singhal, Steve Abney, Michiel Bacciani, Michael Collins, Donald Hindle, and Fernando Pereira. AT&T at TREC-8. In Voorhees and Harman [5]. NIST Special Publication 500-246. Electronic version available at http://trec.nist.gov/pubs.html.

[5] E.M. Voorhees and D.K. Harman, editors. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, 2000. NIST Special Publication 500-246. Electronic version available at http://trec.nist.gov/pubs.html.