Retrieval Performance and Visual Dispersion of Query Sets [1,2]

Mark Rorvig [3]
The University of North Texas

Abstract

In the course of eight TREC Conferences, retrieval performance of all systems started high and then declined. This was especially true for conference 5. Only in conferences 7 and 8 have performance levels reached those initially achieved. In this paper, scaling of the corpus of 450 TREC topics is performed. It is observed that as the visual dispersion of a topic set increases, the level of retrieval performance across systems declines for that set. Conversely, as the visual dispersion of topics decreases, system performance rises. In common elements of conferences 2, 5, and 8, this relationship appears to hold despite increases in the number of participating systems in TREC. It is proposed that visual dispersion measures should be used to describe topic set difficulty in addition to measures such as "hardness".

Introduction

In the middle of a wonderful review article of the work of Project Intrex from 1965 to 1973, the authors interject this startling phrase: "Our analysis has shown that *choice of words used in search strategies* has a major influence on retrieval effectiveness" (Overhage and Reintjes, 1974, p. 174).

This phrase startles because it is at once a reduction and an enigma. There is no doubt that word choice is important, but how can it be so important that retrieval performance depends upon it to the exclusion of so many other system and architecture considerations? The query is often the last item considered in IR testing. Usually its study is incorporated in the interaction effects between systems and users; a difficult and fluid arena. The suspicion that queries might establish a system performance limit did not arise in TREC literature until conference 5 (Voorhees and Harman, 1997). However, it has since been recognized as an area for important study, resulting in the establishment of a query track since conference 7 (Buckley, 1998).

It is difficult to quantify the meaning of topic difficulty. Voorhees and Harman (1997) note that it is weakly ($r = 0.33$) correlated with the percent of unique relevant documents for that query. In the same volume, Sparck Jones remarks that "…low levels of performance…in TREC 4 and 5 must be taken as representing a more realistic retrieval situation than TREC 2 and 3…" (Sparck Jones, 1997, p. B-2). Sparck Jones comments further

[1] This study was supported by Intel Corporation.
[2] A color version of this paper is available at <http://www.unt.edu/ir/trec/trec8.htm>.
[3] Correspondence should be sent to the author at mrorvig@unt.edu.

in her review of TREC7 that "Since TREC-7 full topics are shorter than TREC-6, but TREC-7 performance levels are better, the TREC-7 topics are presumably not as hard…However, performance is not as tightly correlated with topic length, and specifically with version, as might be expected…" (Sparck Jones, 1999, p. B-6).

Factors that *cannot* describe query difficulty are: (1) topic components (concepts, narratives, etc.), (2) topic length, (3) and topic construction (creating topics without regard to existing documents vs. the contrary practice). Document uniqueness is the only quantitative measure so far offered. Indeed, topic hardness appears to rest in that zone of phenomena that many can mutually observe, but cannot describe in terms that would eventually permit control.

This paper proposes an additional quantitative measure for query difficulty. The measure is applicable to sets of topics only, but is based on the scaled similarity of documents by text terms. The proposed measure is replicable, and conforms to observed system performance behavior across three representative TREC conferences.

Methodology

TREC Topics were copied from the trec.nist.gov site and parsed into individual documents. A document similarity matrix was created using the cosine vector measure of similarity. The similarity matrix was scaled using maximum likelihood method customary for text data (Rorvig, 1999a) and plotted using a conventional graphics tool.

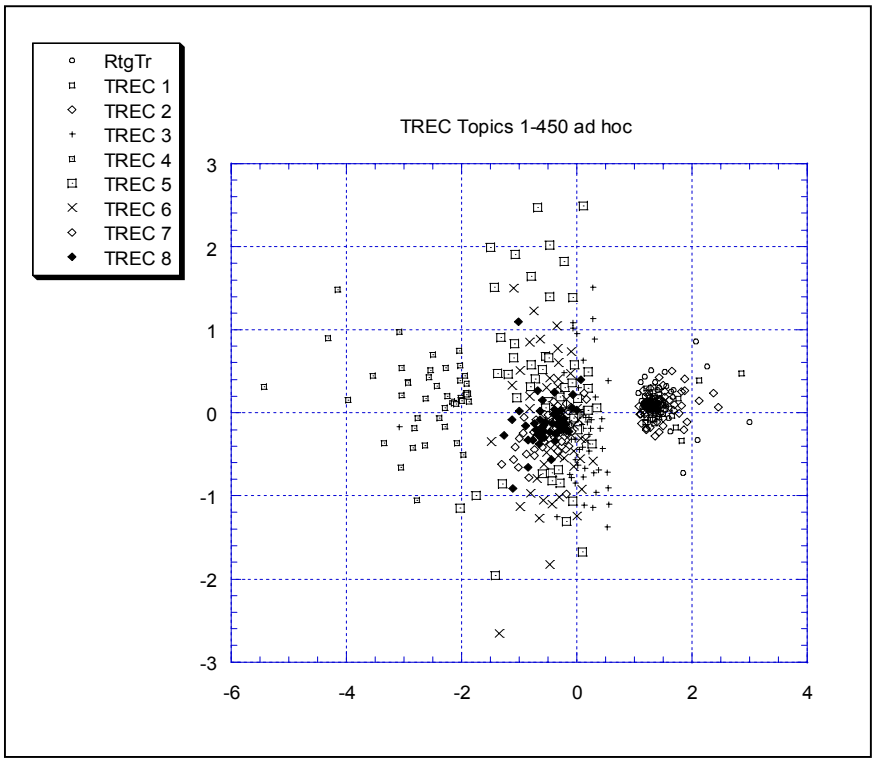

Figure 1: Each dot in the illustration above represents a TREC topic. Arrayed from left to right, topic sets reveal increasing dispersion from topic set 3 onward. This effect does not change until topic sets 7 and 8 appear.

The resulting plot appears as Figure 1. Measures of mean distances among individual documents by set were then taken according to the methodology established in Rorvig and Fitzpatrick (2000). In this method, the centroid of all points in a set is established, and distances of all other points to the centroid calculated. The mean, standard deviation and minimum and maximum point distances were all calculated. These points appear in Figure 2.

TREC	Mean	Std Dev	Minimum	Minimum
routing	0.3111723	0.2805747	0.0747101	1.6268146
1	0.2052631	0.2294708	0.0091761	1.5116826
2	0.3036261	0.1861311	0.0382413	0.9796236
3	0.5627928	0.4018929	0.0406492	1.7142475
4	0.6950999	0.4838904	0.0755262	2.8757636
5	1.0031172	0.6126555	0.1658590	2.4142985
6	0.7522161	0.4572351	0.1520691	2.6426799
7	0.3396361	0.1987803	0.0501366	0.9060473
8	0.3288653	0.2413289	0.0360555	1.3208982

Figure 2: Calculations of interpoint distances of all TREC topic sets after scaling.

As Voorhees and Harman (1997, p. 18) note regarding the reports by Buckley, Singhal, and Mitra, show a comparison of the average precision of the Cornell runs over five TRECs. “Of particular interest here is the fact that the TREC-5 Cornell system performed about 34% worse on the TREC-5 topics than on the TREC-4 topics…most of this difference is due to ‘harder’ topics.” It is an unusual coincidence that the mean dispersion of TREC-5 topics over TREC-4 topics is, if fact about 31% greater.

Because of the differences in various TREC conferences regarding query construction, the three TRECs with the widest variation in dispersion were chosen for further analysis. From published system reports, ad hoc run precision scores of participating systems were recorded at 10% levels of recall and 50% levels of recall including manual systems for TRECs 2, 5, and 8.

Results

Figure 3 shows overall system performance for TRECs 2, 5, and 8 for all systems, the top ten systems, and the top twenty-five systems and precision set at 10% and 50% of recall. Although among the top ten systems at high levels of precision, there is no significant difference, significant differences appear for all systems at high and medium levels of recall, meduim levels of recall for the top ten systems, and for the top twenty-five systems at both high and medium levels of recall.

TREC	p10/all	p50/all	p10/10s	p50/10s	p10/25s	p50/25s
2	0.481	0.385	0.596	0.353	0.559	0.314
5	0.371	0.177	0.556	0.282	0.498	0.250
8	0.447	0.213	0.595	0.319	0.578	0.307

Figure 3: High and medium precision scores for ad hoc runs for three TRECs of all reporting systems, the top ten systems, and the top twenty-five systems.

Discussion

At issue are the various causes of dispersion among the topic documents. Greater dispersion in scaling can be due to a number of factors, among them simple differences in document length, greater heterogeneity in document tokens, or greater heterogeneity in document tokens among some documents but not in others. Investigation of these factors is beyond the scope of this study at this time, however, they are topics of further interesting exploration.

There is also to consider the unique document theory. Although the correlation reported earlier between document hardness and document uniqueness was not high, there is other evidence that high dispersion among topic statements is reflected in wide separation among their associated documents (Rorvig, 1999b). This would appear to support the document uniqueness theory, and upon replication with the qrels document sets for TREC5 suggest other methods by which document uniqueness and document hardness could be calculated.

Finally, as a point of reference, for the next round of TREC, the topic dispersion more than likely will reflect topic hardness. It will make an interesting postscript to this paper to suggest overall system performance for TREC9 merely from introducing the scale of the new topics into the similarity matrix calculated for this study.

Conclusions

This paper is an example of thinking with visualization. A correspondence between topic dispersion in a scaled and visualized space and overall TREC system performance was observed based both on previously published statements of TREC participants and direct observations from printed TREC ad hoc run results. It may be possible to predict overall system performance in TREC9 by scaling the topic set when it becomes available.

Acknowledgements

The author is grateful to David Evans of Claritech for suggesting this study and to Ellen Voorhees of NIST for her helpful comments on data used in this paper.

References

Buckley, C. (1999) “The TREC-7 Query Track,” in the *Seventh Text Retrieval Conference (TREC-7*) (E.M. Voorhees, D.K. Harman, Eds.), NIST Special Publication 500-242, p. 65-72.

Overhage, C.F.J., Reintjes, J.F. (1974) “Project Intrex: A General Review,” *Information Storage and Retrieval*, 10:157-188.

Rorvig, M., (1999a) "Images of Similarity: A Visual Exploration of Optimal Similarity Metrics And Scaling Properties Of TREC Topic-Document Sets," *Journal of the American Society for Information Science*, 50(8): 639-651, 1999.

Rorvig, M., (1999b) "On the Orderliness of TREC Relevance Judgements," *Journal of the American Society for Information Science*, 50(8): 652-660, 1999.

Rorvig, M., Fitzpatrick, S. (2000) "Shape Recovery: A Visual Method for Evaluation of Information Retrieval Experiments," *Journal of the American Society for Information Science*, [in press].

Sparck Jones, K. (1997) "Summary Performance Comparisons TREC-2, TREC-3, TREC-4, TREC-5," in the *Fifth Text Retrieval Conference (TREC-5)* (E.M. Voorhees, D.K. Harman, Eds.), NIST Special Publication 500-238, p. B1-B6.

Sparck Jones, K. (1999) "Summary Performance Comparisons TREC-2 through TREC-7," in the *Seventh Text Retrieval Conference (TREC-7)* (E.M. Voorhees, D.K. Harman, Eds.), NIST Special Publication 500-242, p. B1-B6.

Voorhees, E., Harman, D. (1997) "Overview of the Fifth Text Retreival Conference (TREC5)," in the *Fifth Text Retrieval Conference (TREC-5)* (E.M. Voorhees, D.K. Harman, Eds.), NIST Special Publication 500-238, p. 1-28.