Toward Better Weighting of Anchors

David Hawking
CSIRO ICT Centre
GPO Box 664
Canberra, Australia 2601
david.hawking@csiro.au

Trystan Upstill
Computer Science
Department, ANU
Canberra, Australia 0200
trystan@cs.anu.edu.au

Nick Craswell
CSIRO ICT Centre
GPO Box 664
Canberra, Australia 2601
nick.craswell@csiro.au

ABSTRACT

Okapi BM25 scoring of anchor text surrogate documents has been shown to facilitate effective ranking in navigational search tasks over web data. We hypothesize that even better ranking can be achieved in certain important cases, particularly when anchor scores must be fused with content scores, by avoiding length normalisation and by reducing the attentuation of scores associated with high tf. Preliminary results are presented.

Categories and Subject Descriptors: H.3.3Information Storage and RetrievalInformation Search and Retrieval, Retrieval Models

General Terms: Performance, Experimentation, Theory.

Keywords: Anchor text, Enterprise search, Web search.

1. INTRODUCTION

When the incoming anchor text of links to a particular web page is collected and associated with the target rather than the sources this forms an *anchor text surrogate* for the target. Here, the content of the anchor text surrogate will be called the "anchor text" of the page.

The use of anchor text in Web search was first reported in 1994 [5] and anchor text is believed to make a significant contribution to the effectiveness of the Google search engine [1]. It has also been shown to be very effective on navigational (but not topic relevance) search tasks over enterprise-scale collections such as the TREC .GOV and WT10g collections [2] [3] and in real enterprise intranets [4].

2. OKAPI BM25

Based on consistent results in TREC evaluations, the Okapi BM25 relevance scoring formula [6] can be said to embody a good model of relevance based upon term occurrences within text documents. Here is a simplified version:

$$w_t = tf_d \times \frac{\log(\frac{N-n+0.5}{n+0.5})}{k_1 \times ((1-b) + b \times \frac{dl}{avdl}) + tf_d} \qquad \text{(BM25)}$$

where w_t is the relevance weight assigned to a document due to query term t, tf_d is the number of times t occurs in the document, N is the total number of documents, n is the number of documents containing at least one occurrence of t, dl is the length of the document and $avdl$ is the average document length[1]. $k1 = 2.0, b = 0.75$ are the constants used by [2] and in experiments reported here.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

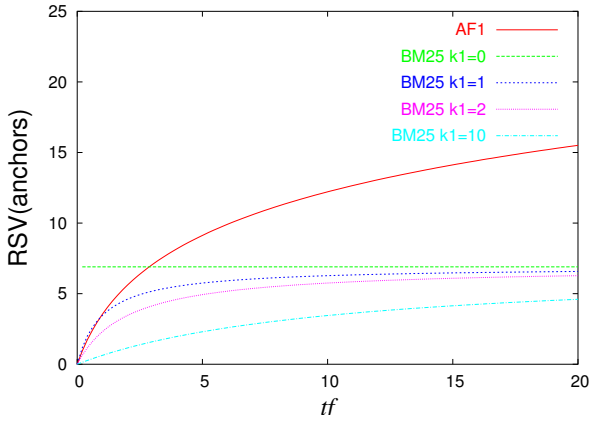

Figure 1: Variation of AF1 and BM25 using several values of k1 with tf assuming a document of average length, $N = 100000, n = 10$

Let us examine the three components of the BM25 formula and consider how applicable they are to anchor text.

Term frequency: Anchor text exhibits very different term frequency distributions to those of document text. We observed that the word 'projects' accounted for 80% of all anchor text for the World Bank projects homepage ($tf = 6798$) but only about 4% of the document content ($tf = 5$).

Figure 1 shows that the curve of Okapi score versus tf is almost flat beyond $tf = 10$. This is not a desirable property when scoring anchor text as each occurrence of a query term in anchortext may be considered to be a separate vote that this page is relevant.

It seems counter-intuitive either that a page with 6000 votes should score almost identically to another page with only 10 such votes (BM25) or that the former should score 600 times as much (linear model). Equation AF1 proposes a simple logarithmic model.

Document Length Normalisation: The document length normalisation in the Okapi formula reflects the fact that, the longer a piece of text, the greater the likelihood that a particular query term will occur by chance.

Again, this seems counter-intuitive when applied to an-

[1]Note: negative values of $\log(\frac{N-n+0.5}{n+0.5})$ are mapped to a small positive constant ϵ in experiments reported here.

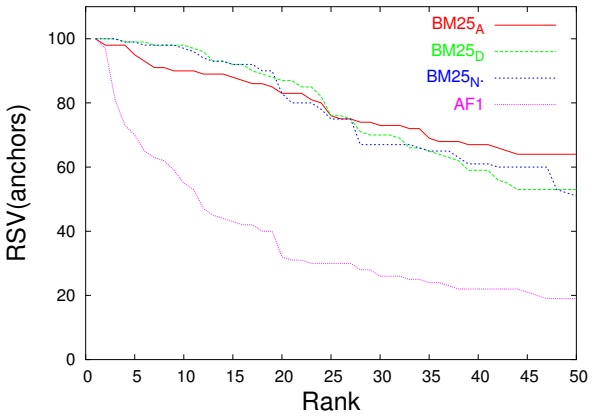

Figure 2: Decline in normalised score with increasing rank for the query 'library' in the University X collection using four different anchor text scoring formulae.

chor text. Greater text length is indicative of more votes (incoming links) for the document and should not be penalised. Equation AF1 eliminates length normalisation altogether. The same effect can be achieved by setting $b = 0$ in Equation BM25.

Inverse document frequency: Here, intuition accords with the Okapi model – it makes sense to give more weight to query terms occurring in fewer anchor documents.

3. AF1 - A FIRST-CUT MODEL

$$w_t = \alpha \log(tf_d + 1) \times \log(\frac{N - n + 0.5}{n + 0.5}) \qquad \text{(AF1)}$$

Equation AF1 presents a first cut at a formula based on the above intuitions. It was used in the experiments described below and also (combined with other evidence) in successful TREC Web Track [3] submissions. In both cases, documents matching all terms in a query were ranked ahead of partial matches.

We hope that the greater AF1 dispersion of scores associated with high anchor *tf*s will facilitate fusion with scores from other types of evidence, such as relevance of the document content.

A more refined version of AF1 is likely to include better weighting of term coordination, weighting of proximity/adjacency of query terms, separate treatment of individual pieces of incoming anchor text and the proportion of anchor text consisting of the query words.

4. RESULTS

Columns 3 and 4 of Table 4 table show ranks and normalised scores for the the best answer in response to the query 'library' when using only anchor text. Subscripts indicate the type of length normalisation used in BM25 runs: A - relative to length of anchor surrogate; D - relative to actual document length; N - no length normalisation.

In $BM25_A$, the correct answer is severely penalised by the length (13,484 words, 262 times the average) of its anchor surrogate, despite a very high *tf* for the query term (tf = 1664). $BM25_D$ places the best answer at rank one but scores

Table 1: Observations and results for 80,000 webpages for University X. Columns three and four show ranks and normalised scores for the best answer to the query 'library'. Columns five and six show results for 332 navigational queries. MRR1 means the mean reciprocal rank of the first correct answer; P@1 is the proportion of queries for which the best answer was returned at rank one.

Formula	Length Normalisation	Library Right Answer		Results over 332 queries	
		Score	Rank	MRR1	P@1
$BM25_A$	anchors	61	62	.6075	.4669
$BM25_D$	document	100	1	.7182	.6265
$BM25_N$	none	100	1	.7002	.6084
AF1	none	100	1	.6909	.5934

it only slightly above many other candidates (only 1% higher than the home page of a minor library whose *tf* is a factor of 7.5 lower), making the ranking vulnerable during fusion with content scores. Figure 2 shows that Okapi scores for the anchor text surrogates decline quite slowly whereas the new formula makes a stronger *tf*-based distinction.

Columns 5 and 6 of Table 4 show results for a set of 332 navigational queries processed over the same collection. Wilcoxon tests showed: $AF1 > BM25_A (p < 10^{-5}), AF1 \approx BM25_N, BM25_D > BM25_N (p < 0.02)$.

5. DISCUSSION AND CONCLUSIONS

We found a 15% deterioration in MRR1 when normalising BM25 scores using anchor text length. We hypothesise that improvement due to normalisation by actual document length is due to the introduction of query independent evidence – home pages are likely to be shorter.

We found no advantage to AF1 over $BM25_N$ in anchor-text-only experiments but hypothesize that the greater differentiation between the AF1 scores of top-ranked documents will permit more effective fusion with other evidence. Testing this is an attractive avenue for future work.

6. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW7*, pages 107–117, 1998.

[2] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, pages 250–257, New Orleans, 2001.

[3] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the trec-2003 web track. In *Proceedings of TREC 2003*, Gaithersburg, Maryland USA, November 2003.

[4] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the workplace web. In *Proceedings of WWW2003*, Budapest, Hungary, May 2003.

[5] O. McBryan. GENVL and WWWW: Tools for taming the web. In *Proceedings of WWW1*, 1994.

[6] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, November 1994.