The Power of Peers

Nick Craswell[1], Dennis Fetterly[2], and Marc Najork[2]

[1] Microsoft, Bellevue, WA, USA
nickcr@microsoft.com
[2] Microsoft Research Silicon Valley, Mountain View, CA, USA
{fetterly,najork}@microsoft.com

Abstract. We present a study of the contributions of three classes of ranking signals: BM25F, a retrieval function that is based on words in the content of web pages and the anchors that link to them; SALSA, a link-based feature that takes all or part of the result set to a query as input; and matching-anchor count (MAC), a feature that measures precise matches between queries and anchors pointing to result pages. All three features incorporate both link and textual features, but in varying degrees. BM25F is the state-of-the art exponent of Salton's term-vector model, and is based on a solid theoretical foundation; the two other features are somewhat more ad-hoc. We studied the impact of two factors that go into the formation of SALSA's "base" set: whether to use conjunctive or disjunctive query semantics, and how many results to include into the base set. We found that the choice of query semantics has little impact on the effectiveness of SALSA (with conjunctive semantics having a slight edge); more surprisingly, we found that limiting the size of the base set to a few hundred results of high expected quality maximizes performance. Furthermore, we experimented with various linear combinations of BM25F, MAC and SALSA. In doing so, we made a remarkable observation: adding BM25F to a two-way weighted linear combination of MAC and SALSA does not increase performance in any statistically significant way.

1 Introduction

In this work, we compare the ranking performance of linear combinations of three features: BM25F, SALSA-SETR, and MAC. BM25F [8] is a ranking function in the tradition of Salton's term-vector model that is based on a solid theoretical model and that correlates query terms with terms in the title and body of a web page as well as in its URL and any HTML anchors pointing to it. SALSA-SETR [7] is a variant of SALSA [6] which in turn was derived from Kleinberg's HITS [5] algorithm. It projects (some of) the results of a query onto the web graph, extracts (some of) the distance-one neighborhood graph, and computes "authority scores" on that graph. MAC ("matching-anchor count") is a simple heuristic that measures how many anchors pointing to a given result precisely match the query [4]. More precisely, it counts the number of IP subnets containing hosts that serve pages containing one or more matching anchors.

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 497–502, 2011.

Each of the three features incorporates both text and link information: BM25F is predominantly text-based, but it incorporates link information by considering anchor text. SALSA is predominantly link-based, but the "base set" of vertices that is the input to the SALSA algorithm is based on textual matching between query and corpus documents, and in our implementation, is furthermore biased toward documents that have high BM25F scores. MAC incorporates text and link features in equal (and quite simplistic ways), by looking for anchors (which imply links) whose text precisely matches the query.

We conducted our experiments using three data sets: the ClueWeb09 corpus [2], a collection of slightly over one billion web pages (we only use the 503 million English-language pages); the test set used in the TREC 2009 Web Track [1], which contains 50 queries and 27,964 results paired with binary judgments; and an additional test set comprised of the same 50 queries as the previous test set as well as 4,298 binary judgments completed by the authors.

Our evaluation measures are the ones that were used in the diversity task of the TREC 2009 Web Track. In addition we use the measure "IA-P@20 (judged)" where the denominator is the number of judged documents in the top-20 instead of all 20. This addresses the fact that the TREC test set is only partially judged, and that SALSA in particular surfaces a high number of unjudged results.

In order to quantify the impact of unjudged documents on system ranking, we completely judged the result sets for one weighted linear combination of features described in Section 3. We chose a document cutoff value of 20, which is consistent with other evaluations in this paper, which yielded 3,053 query-results pairs. The query-result pairs were grouped by query and then divided roughly into thirds, which were each evaluated by a single judge so that any particular judge would judge all of the results for a single query. Judging was performed using a tool that displayed information relevant to the query, such as the description of the information need, the type of facet of the query, and the information need for that specific facet, as well as the content of the page itself. A proxy server was employed so that stored content from the ClueWeb collection could be returned for document requests, but images, style sheets, and other page content would be requested from the original web site. We were unable to assess 37 query-result pairs, either because the page could not be rendered or because it contained malware. These judgments are available to the research community; please contact the authors to obtain them. The following table relates the TREC judgments (horizontal) to ours (vertical):

	NR	U	R
NR	995	873	154
U	20	13	4
R	261	320	413

For example, we considered 873 of the unjudged TREC results to be relevant. We computed Cohen's Kappa statistic between the TREC 2009 judgments and ours. After flattening the subtopics, we have $\kappa = 0.49$ and a 77% agreement rate.

2 Impact of Base Set on SALSA's Performance

Our first set of experiments studied the impact of the selection and size of the "base set" that serves as the input to SALSA-SETR [7]. For a given query, we produced a "candidate result set" using either disjunctive (t_1 OR t_2) or conjunctive (t_1 AND t_2) semantics for multi-term queries, computed BM25F scores for each candidate, and admitted the k highest-scoring candidates into the "base set" that is the input to any HITS-like ranking algorithm, including SALSA-SETR. Figure 1 shows the results. The two graphs show two different effectiveness measures (α-nDCG@20 and IA-P@20); the horizontal axis plots k; the vertical axis plots effectiveness; and the two curves in each graph show the performance of the two query semantics we consider. We can see that the choice of query semantics does not have a great impact on effectiveness, although AND slightly outperforms OR. More interestingly, we find that SALSA is most effective when given a base set of ten to a few hundred results.

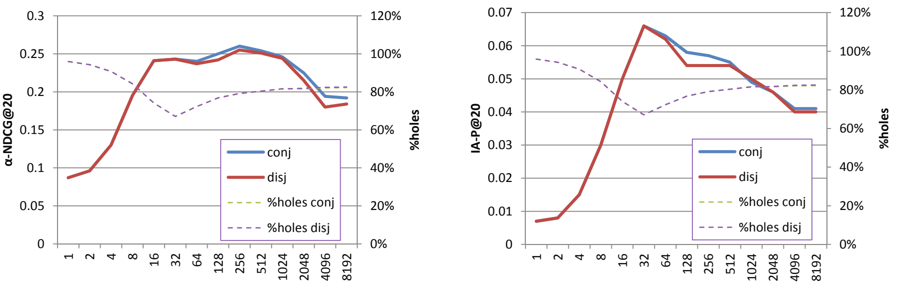

Fig. 1. Impact of the choice and size of the base set on SALSA-SETR's performance. These results use the TREC judgments.

3 Combination of Features

For the remainder of this work, SALSA-SETR is parameterized to use conjunctive query semantics, form the base set out of the 5,000 top candidate results according to BM25F, and to vertex and edge sampling parameters $a = 5$, $b = 6$, $c = 6200$, and $d = 2900$ (see [7] for a detailed description of the algorithm). We studied the performance of the three features in isolation, as well as the three possible pairwise and the one three-wise combination.[1] We applied a log-based transform function to MAC and SALSA when combining them with other features, and we weighted MAC by a factor of 1 and SALSA by a factor of 500. In our experience, increasing SALSA's weight beyond 500 decreases α-nDCG and IA-P, since it substantially increases the number of highly-ranked unlabeled results.

[1] Commercial search engines typically combine hundreds of features, e.g. see [3].

Table 1. Performance of individual features and feature combinations evaluated using the TREC 2009 judgments. Each ranker is a combination of the BM25F score B , the MAC score A, and/or the SALSA-SETR score S. * indicates a significant difference from AS, in a one-tailed t-test, with $p < 0.01$.

	α-nDCG			IA-P			IA-P (judged)		
Ranker	@5	@10	@20	@5	@10	@20	@5	@10	@20
BAS	0.280	0.319	0.365	0.138	0.115	0.109	0.141	0.119	0.113
AS	0.284	0.325	0.363	0.135	0.117	0.094*	0.143	0.135	0.122
BA	0.282	0.311	0.361	0.138	0.113	0.108	0.141	0.116	0.112
A	0.275	0.300	0.350	0.131	0.104	0.090*	0.143	0.121	0.119
BS	0.208*	0.243*	0.281*	0.107	0.098	0.086*	0.129	0.121	0.113
B	0.180*	0.221*	0.253*	0.084*	0.082*	0.072*	0.092*	0.094	0.091*
S	0.143*	0.165*	0.193*	0.058*	0.051*	0.041*	0.109	0.114	0.114

Table 2. Performance of individual features and feature combinations evaluated using the complete judgments performed by the authors. Each ranker is a combination of the BM25F score B , the MAC score A, and/or the SALSA-SETR score S. * indicates a significant difference from AS, in a one-tailed t-test, with $p < 0.01$.

	α-nDCG			IA-P		
Ranker	@5	@10	@20	@5	@10	@20
BAS	0.443	0.491	0.550	0.201	0.174	0.158
AS	0.440	0.502	0.550	0.190	0.170	0.149
BA	0.443	0.487	0.548	0.198	0.171	0.156
A	0.433	0.482	0.548	0.197	0.167	0.149
BS	0.318*	0.370*	0.431*	0.163	0.155	0.135
B	0.262*	0.320*	0.381*	0.121*	0.116*	0.109*
S	0.256*	0.303*	0.363*	0.141	0.141	0.136

We combined the three features (BM25F, MAC, and SALSA) into all seven possible combinations using near-optimal weights, and used each resulting ranker to rank the result sets of each query. Table 1 shows the performance of each of these systems. We can observe that SALSA-SETR performs substantially better under IA-P (judged) than under IA-P or α-nDCG. This is due to the fact that rankers using SALSA surface more unjudged results in the top twenty.

In order to quantify the performance of these systems on a completely judged result set, we pooled the top twenty results of each ranker and judged them using the methodology in Section 1. Table 2 shows the performance of these systems evaluated using the new judgments. MAC is the most efficient single feature, and furthermore, any combination involving MAC does not differ from any other such combination in a statistically significant fashion.

Figure 2 shows the performance of the seven rankers evaluated both the TREC judgments as well as the new judgments. The size of the gap between the curves for TREC IA-P@20 and TREC IA-P@20 (judged) illustrates the fraction of unjudged top 20 results for each ranker. This gap is very small for BAS, indicating

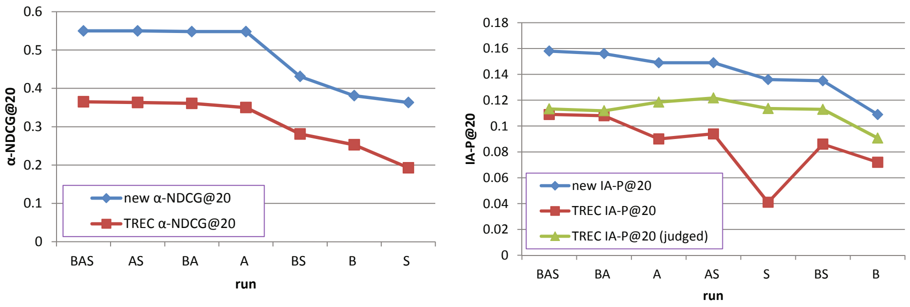

Fig. 2. The system ranking for α-NDCG@20 is similar whether we use TREC judgments or our new judgments. The IA-P@20 system ranking under TREC judgments differs from our fully judged results, and in particular is worse for SALSA.

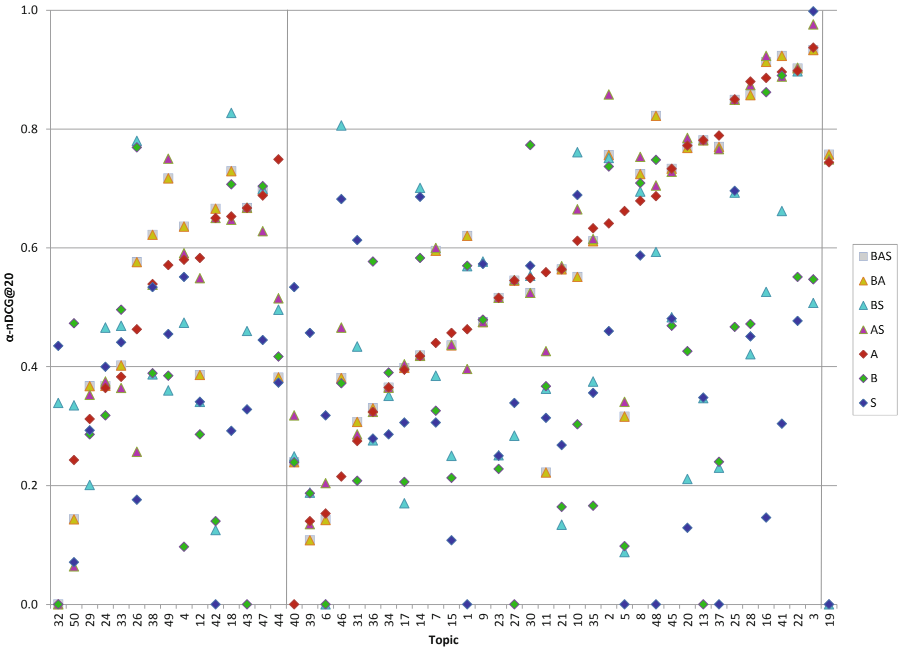

Fig. 3. Performance of the seven rankers in terms of α-nDCG, broken down by topic and grouped by query intent (purely informational, informational+navigational, purely navigational

that the TREC judgments contain virtually no holes for the top-20 results returned by this ranker; conversely it is very large for S. The gap between TREC IA-P@20 (judged) and new IA-P@20 for the BAS ranker indicates that we considered more results to be relevant than the TREC assessors did. For the other rankers, a smaller gap indicates that more of the results that were unjudged in the TREC collection were considered non-relevant by us.

Finally, Figure 3 shows the performance of the seven rankers in terms of α-nDCG, broken down by topic. The left 15 topics are purely informational, the rightmost topic is purely navigational, and the remaining 34 topics have both informational and navigational subtopics.

4 Conclusion

In this work, we studied the effectiveness – in isolation and in combination – of three ranking features that each incorporate both text and link information: BM25F, MAC, and SALSA. We used publicly available data sets and standard measures of retrieval effectiveness to investigate which and how many candidate results to incorporate into the base set of results, and how to combine these features to maximize effectiveness. To our surprise, we found that MAC, a fairly ad-hoc feature, performed better than any other feature, and that any combination of features involving MAC performed equally well in a statistically significant sense.

References

1. Clarke, C., Craswell, N., Soboroff, I.: Report on the TREC 2009 Web Track. In: 18th Text Retrieval Conference (2009)
2. The ClueWeb 09 Dataset, http://boston.lti.cs.cmu.edu/Data/clueweb09/
3. Hansell, S.: Google keeps tweaking its search engine. New York Times (2007), http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html
4. Craswell, N., Fetterly, D., Najork, M., Robertson, S., Yilmaz, E.: Microsoft Research at TREC 2009: Web and relevance feedback tracks. In: 18th Text Retrieval Conference (2009)
5. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms (1998)
6. Lempel, R., Moran, S.: The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In: 9th International World Wide Web Conference (2000)
7. Najork, M., Gollapudi, S., Panigrahy, R.: Less is More: Sampling the neighborhood graph makes SALSA better and faster. In: 2nd ACM International Conference on Web Search and Data Mining (2009)
8. Zaragoza, H., Craswell, N., Taylor, M., Saria, S., Robertson, S.: Microsoft Cambridge at TREC–13: Web and HARD tracks. In: 13th Text Retrieval Conference (2004)