Integrating Link Structure and Content Information for Ranking Web Documents

Tapas Kanungo and Jason Y. Zien
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
Email: kanungo,jasonz@almaden.ibm.com

SUMMARY

In this article we describe the results of our experiments on combining the two sources of information: The web link structure and the document content. We developed a new scoring function that combines TF*IDF scoring with the document rank and show that it is particularly effective in the Home Page Finding task.

1 The Indexer

The IBM Almaden system consists of three components: the indexer, the DocRanker, and the query engine.

- The indexer tokenizes the documents and records the presence of various attributes: capitalization, presence in title or bulleted lists, color, etc. Prior to pushing the attributed token into the index, the token is stemmed using the Porter stemmer. Document frequency of tokens and other global statistics are also recorded on the fly.
- The DocRanker ranks documents by extracting the link structure of the crawled pages, then computing the SameSite function [2], and finally computing the PageRank [3], on the graph. Our PageRank calculations used a weighted graph where the weight of a link was 1.0 if the link was composed of two pages from different sites, and 0.0001 if they were on the same site, to deemphasize self-references.
- The query engine retrieves the filtered set of documents that contain the the rarest query term. Then, all the filtered documents are scored using all of the query terms (both stemmed and unstemmed versions). These pages are ranked using the integrated ranking function and then sorted result is returned.

2 Scoring Function Details

The first part in our scoring algorithm is deciding which documents to filter out. Our scoring algorithm uses the rarest term to drive the query evaluation. Only documents which contain the rarest term are examined during scoring.

Let $w_{d,t}$ be the score of a document with respect to term t. The general form of a TF*IDF scoring function looks like this:

$$w_{d,t} = r_{d,t} \cdot w_t \tag{1}$$

where $r_{d,t}$ (the TF component) is a function based on the frequency of a term t in document d, and w_t (the IDF component) is the weight of a term in the corpus.

2.1 Inquery/Okapi

The Inquery scoring function is a variation of the well-known Okapi scoring function. Let A be the average length of a document in terms (not bytes), and $|D_d|$ be the length of document d in terms (not bytes), and N be the number of documents in the collection. Let $f_{d,t}$ be the number of occurrences of t in a document d and f_t be the number of documents in which t occurs. Give a query Q, the Inquery scoring formula is:

$$score = \sum_{t \in Q \cap D_d} TF \cdot IDF \tag{2}$$

$$score = \sum_{t \in Q \cap D_d} r_{d,t} \cdot w_t \tag{3}$$

$$r_{d,t} = 0.4 + 0.6 \frac{f_{d,t}}{f_{d,t} + 0.5 + 1.5 \frac{|D_d|}{A}} \tag{4}$$

$$w_t = \frac{log_e(\frac{N+0.5}{f_t})}{log_e N + 1} \tag{5}$$

In order to take advantage of document contextual information, we modified $f_{d,t}$ so that it is not just the number of occurrences of t in d, but instead, is a weighted sum of occurrences of t in d. The weighting of occurrences in titles and headings, was configurable. Also of note, our $f_{d,t}$ includes both stemmed and unstemmed occurrences of a term. The effect of this is to essentially boost the score for pages that have an exact match, while also giving a chance for pages that have only the stemmed term to appear in the result set.

2.2 Incorporating DocRank into the Scoring Function

Link analysis methods may be used to obtain an ordered ranking of documents, for instance, through a PageRank calculation. Although PageRank provides useful information for scoring, it is unclear how this information should be combined into a page-content-based scoring function. We propose a new scoring function that blends document ranking with a TF*IDF formulation. Let us examine the Inquery/Okapi function's TF component in detail. The component $1.5\frac{|D_d|}{A}$ is the only portion of the function that contains document-related information. This component provides a bias to the score based on the importance of a document. In this case, importance is defined by the size of a document. When a document is large, this score component is large (and hence since this is in the denominator, the overall score is reduced). We propose incorporating a document rank (DocRank) ρ_d into this component. ρ_d is the scaled ordinal rank of a page. For instance, if the document is the the 3^{rd} ranked document in a collection of N documents, $\rho_d = 3/N$. Note that in particular, the DocRank is not the actual PageRank value. Rather, the use of ordinal rank provides a smoother, more gradually changing value than the actual PageRank. Also, it is obvious that any algorithm that can generate an ordering of documents could be used in place of PageRank for our purposes. There are two straightforward ways of combining the DocRank with the document component of the score — multiplicative and additive. An example of the multiplicative form is:

$$1.5\rho_d \frac{|D_d|}{A}. \tag{6}$$

However, after experimentation, the form that we settled on is the additive form:

$$\alpha\rho_d + 1.5\frac{|D_d|}{A}, \tag{7}$$

where α is a user-specified constant. The α term allows the user to tune the relative importance of the document ranking component in the scoring function. For the Home Page Finding task, we used $\alpha = 10.0$. For the Ad Hoc task, we used $\alpha = 1.5$

Our final modified scoring formula is:

$$score = \sum_{t \in Q \cap D_d} TF \cdot IDF \tag{8}$$

$$r_{d,t} = 0.4 + 0.6\frac{f_{d,t}}{f_{d,t} + 0.5 + \alpha\rho_d + 1.5\frac{|D_d|}{A}} \quad (9)$$

$$w_t = \frac{log_e(\frac{N+0.5}{f_t})}{log_e N + 1} \quad (10)$$

3 Results at TREC 2001

We participated in the two Web tracks at TREC 2001: Ad Hoc, and Home Page Finding. Our contribution is a method of incorporating a DocRank term into a TF*IDF cost function that allows us to control the relative contribution of a document's rank to that of text content. The ranking function itself is based on the Inquery variant of the Okapi ranking function [1].

Also, to take advantage of contextual cues on a page, we made use of heading and title information by giving more weight to term occurrences in those contexts.

Results for the Home Page Finding task:

Metric	Rank not used	Rank used
Average reciprocal rank over 145 topics	0.382	0.611
Number of topics for which entry pages found in top 10	90 (62.1%)	113 (77.9%)
Number of topics for which no entry pages was found	17 (11.7%)	15 (10.3%)

The Home Page Finding task shows clearly that when our DocRank scoring is used, both the average reciprocal rank and the top ten scoring method showed substantial improvement. Using linkage information was a clear win with Home Page Finding.

Results for the Ad Hoc task:

Metric	Rank not used	Rank used
Precision at 5 docs	0.40	0.20
Precision at 10 docs	0.20	0.20
Precision at 15 docs	0.133	0.20

For Ad Hoc queries, link information did not improve the results.

4 Conclusion

We introduced a novel new scoring function that combines TF*IDF scoring with link-based ranking. Our experiments showed that this combined scoring method was exceptionally well-suited to Home Page Finding.

References

[1] J. Allan, M. Connell, W. B. Croft, F. F Feng, D. Fisher, and X. Li. INQUERY and TREC-9. 2000.

[2] Soumen Chakrabarti, Byron Dom, David Gibson, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Spectral filtering for resource discovery. In *Proceedings of the SIGIR 98 Workshop on Hypertext Information Retrieval for the Web*, August 1998.

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA, November 1999. http://dbpubs.stanford.edu/pub/1999-66.