RMIT University at the TREC 2007 Enterprise Track

Mingfang Wu Falk Scholer Milad Shokouhi
Simon Puglisi Halil Ali

School of Computer Science and IT
RMIT University, GPO Box 2476V
Melbourne 3001, Australia

1 Overview

At TREC 2007, RMIT University participated in the document search task of the enterprise track. Our goals were to investigate:

1. Which sources of external evidence (anchor text, PageRank and Indegree) are useful for improving a document-based ranking scheme for a key page finding task?
2. Should the different source of evidence be used in isolation, or in combination?
3. Can federated search improve performance over single collection search, for example when the collection is divided into discipline or business-function related categories?

In this paper, we discuss our approaches to these three questions and present experimental results.

2 Sources of External Evidence

The 2007 document search task is akin to a topic distillation task, where the search system should identify resource pages that provide links to informational pages that are relevant to a broad topic, aiming to provide a rich information space and comprehensive picture between found documents and the topic. Such a page may or may not exist in the website: if it exists, it would be ideal for a search engine to rank the page highly; otherwise, the search engine should retrieve those pages potentially pointed at by such a resource page, and rank these pages highly.

Anchor text, PageRank, and Indegree have been shown to be useful sources of external evidence for navigational search tasks. We view the topic distillation task (or this year's document search task) as lying somewhere between navigational and informational searches on the spectrum of search tasks. Therefore, as a starting point, we investigate if external sources of evidence such as anchor text are also useful for such a task.

We used the Lemur toolkit [1] for indexing and searching for all of our submitted runs. Answer documents are ranked according to their KL divergence.

2.1 Anchor Text

Anchor text—descriptive text that is included with an HTML anchor tag—often gives a short topical description of its targeted document. Eiron et al. [4] studied pages from the IBM intranet and found

that anchor text resembles real-world queries in terms of its term distribution and length. We hypothesize that the CSIRO collection would have similar characteristics, and in particular that anchor text in such a collection would be more meaningful, and contain less spam, than anchor text form the public web.

In summary, there are following five major ways to use anchor text:

- Concatenate the anchor text of all hyperlinks pointing to a page together and treat them as a surrogate representation of the page. We call this re-constructed collection the *anchor text* collection (and the original collection the *content* collection). Queries are then run on this anchor text collection only. Craswell et al. showed that searching this anchor text collection alone can effectively improve the entry page finding task [3].

- Treat the anchor text collection and content collection separately. When a query is run, two lists of retrieved pages are returned, one from each collection. These two lists are then merged together, and a page's score (*dw*) in the merged list is an interpolation of its scores from each collection: $dw_{mergedlist} = \alpha \cdot dw_{content} + (1 - \alpha) \cdot dw_{anchor}$. Westerveld et al. tested this method (with $\alpha = 0.9$) and found that this combination of two lists also leads to improved results for the entry page finding task [13].

- Combine the anchor text model and content model by using a unified language model to obtain a single result list. The previous interpolation method is a kind of meta-search approach where the anchor text and the content text provide two very different textual representations of a page. In the unified model, the two representations are combined into a mixture model to estimate a query on a term by term basis [6, 8].

- Treat the anchor text and full text content of a page as two different fields of the page, then apply structured document retrieval techniques [7, 9]. The retrieved pages would be ranked on a combination of field scores.

- Extend each page from the content collection to include its all anchor text as suggested in [10], we call this collection the *extended collection*. To weight anchor text higher than content text (for example, 5 times higher), we could simply repeat the anchor text 5 times during the merging process. In this way, we get an integrated model of the anchor text and document, and a field is up (or down) weighted on a term by term basis.

There are 370,715 documents in the CSIRO collection, and we detected 5,233,862 links. For each page that has incoming links from pages other than itself, we extracted 4,686,442 pieces of anchor text (we call each anchor text an entry). We ignored those hyperlinks that are images — thus we ignored any text such as those from the *alt* attribute of HTML image tags. On average, there are 12.64 anchor text entries per page. Figure 1 shows the occurrence of anchor text across pages. It can be seen that the distribution is skewed; around 76% of documents have three or less associated pieces of anchor text.

2.2 Indegree and PageRank

Indegree and PageRank have also been explored and are used mainly for entry page finding task. These methods includes:

- Use Indegree or PageRank as a document prior in a language model [6],

- Use Indegree or PageRank to re-rank a list of documents retrieved from some content based collection. The re-ranking can be done by using Indegree or PageRank directly to rank retrieved pages above a certain cut-off [5, 12], or

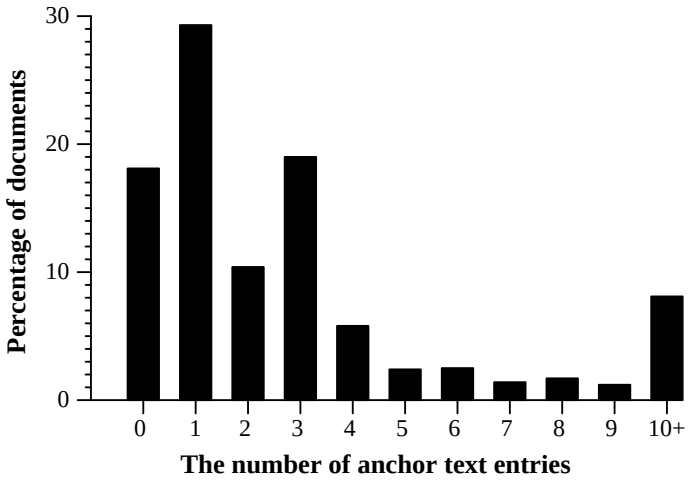

Figure 1: The distribution of anchor text entries per page.

- Combine Indegree or PageRank with a document score through interpolation [13], we tried: $\alpha \cdot dw_{content} + (1 - \alpha) \cdot \log(indegree + 1)$.

2.3 Federated Search

CSIRO has 17 divisions and wide ranges of research areas. Considering that a domain name is more or less related to a function area of CSIRO, we divide the CSIRO collection into sub-collections. All pages with a same domain name are assigned to the same sub-collection. Each sub-collection is indexed separately, and documents are retrieved and ranked in a federated manner. In total, we had 256 sub-collections.

We used CORI [2] for collection selection, and SSL single-model [11] for result merging. For each query, the top five collections ranked by CORI are selected. Selected collections receive the query and return their top 1,000 answers to the broker. The broker then uses SSL to merge the results. The federated search environment used in our experiment is assumed to be *cooperative*. That is, the broker has access to the term frequency information of documents in all collections.

3 Results

For the document search task, a system/run is evaluated on its capability to retrieve the key pages, i.e. those pages which have relevance judgement score of 2. Table 1 summarises our runs and their associated results measured by MAP, P@5 and P@20. Among these runs, *RmitQ*, *RmitQAnc*, *RmitQAncIndg* and *RmitQFir* are our submitted runs; and *RmitQ* is the baseline. The three runs RmitQAnc, RmitQAncIndg and RmitQFir were chosen to submit, as the initial evaluation by using the key pages from the topic description indicated that the RmitQAnc and RmitQAncIndg significantly improved over the baseline RmitQ in terms of MAP and P@10. RmitQFir also showed an improvement although that was not significant.

Anchor text: We tried three methods to use anchor text: search anchor text collection only; search extended collection; and, combine content and anchor text runs through interpolation. None of these methods show improvement over the baseline in terms of the three evaluation measures. In fact, as anchor text gets more weight (either in the extended or interpolation method), the performance deteriorates.

Run	Description	MAP	P@5	P@20
content (*RmitQ*)	content collection only	0.388	0.612	0.471
extended (*RmitQAnc*)	extended collection	0.387	0.604	0.466
anchor	anchor text collection only	0.098	0.441	0.242
anchor-content-merge-75	interpolation of content run and anchor text run ($\alpha = 0.75$)	0.366	0.600	0.458
anchor-content-merge-50	as above ($\alpha = 0.50$)	0.346	0.560	0.437
anchor-content-merge-25	as above ($\alpha = 0.25$)	0.334	0.560	0.435
content-indegree-95	interpolation of the content run and Indegree ($\alpha = 0.95$)	0.390	0.632	0.475
content-indegree-90	as above ($\alpha = 0.90$)	0.386	0.612	0.462
content-indegree-80	as above ($\alpha = 0.80$)	0.346	0.516	0.405
content-indegree-rerank	using Indegree to re-rank top 20 pages from the content run	0.388	0.612	0.471
extended-indegree-0.95	interpolation of the extended run and Indegree ($\alpha = 0.95$)	0.389	0.612	0.465
extended-indegree-0.90	as above ($\alpha = 0.90$)	0.387	0.600	0.461
extended-indegree-0.80	as above ($\alpha = 0.80$)	0.371	0.576	0.434
extended-indegree-0.70 (*RmitQAncIndg*)	as above ($\alpha = 0.70$)	0.324	0.460	0.404
extended-indegree-rerank	using Indegree to re-rank top 20 pages from the extended run	0.386	0.604	0.466
content-pagerank-0.95	interpolation of content run and PageRank ($\alpha = 0.95$)	0.387	0.628	0.472
content-pagerank-0.90	as above ($\alpha = 0.90$)	0.384	0.624	0.469
content-pagerank-0.80	as above ($\alpha = 0.80$)	0.364	0.564	0.434
content-pagerank-rerank	using PageRank to re-rank top 20 pages from the content run	0.387	0.604	0.466
extended-pagerank-95	interpolation of the extended run and PageRank ($\alpha = 0.95$)	0.386	0.600	0.465
extended-pagerank-90	as above $\alpha = 0.90$)	0.384	0.596	0.461
extended-pagerank-80	as above $\alpha = 0.80$)	0.377	0.576	0.450
extended-pagerank-70	as above $\alpha = 0.70$)	0.358	0.524	0.422
extended-pagerank-rerank	using PageRank to re-rank top 20 pages from the extended run	0.387	0.604	0.466
FIR: separate the collection into sub-collections and apply the federated search				
FIR-05 (*RmitQFir*)	5 collections selected	0.265	0.524	0.395
FIR-20	20 collections selected	0.283	0.484	0.374
FIR-50	50 collections selected	0.266	0.448	0.343

Table 1: *Summary of our runs.*

Indegree: Indegree was used to re-rank search results from the content run and extended run in two ways: one is to use Indegree to re-rank the retrieved pages above certain cutoffs, another is to re-rank the whole list through interpolation. We tried the first method at cutoffs: 10, 20, 30, 40, 50 and 100, and we didn't observe any difference in terms of the three measures even though the orders of the pages above the cut-offs are different. Table 1 shows that re-ranking the top 20 retrieved pages of either the content or extended run doesn't result in the improvement over their originated runs. By interpolating Indegree into the content run and the extended run, P@5 is slightly improved (when $\alpha = 0.95$).

PageRank: PageRank was tested in the same ways as in Indegree and led to similar results.

FIR: We also compared the results of three federated search runs with different collection selection cutoff (CO) values (CO$\in \{5, 20, 50\}$). We varied the number of collections that are selected per query, and investigated the impacts on different evaluation metrics. For $CO = 5$ the performance of federated search is better than the other methods; however it is still poorer than the baseline.

4 Discussion

Unlike the entry page finding task in which the use of anchor text has significantly improved the search results over using the content collection only, A similar result was not achieved here for the key document search task. Using an interpolation of Indegree/PageRank and the content run provides a small but not significant improvement in precision.

We observed that the web pages from the CSIRO collection follow a certain template: global navigation bar at the top of a page, local navigational bar on the left, related link area on the right, copyright bar at the bottom, and content area in the middle. We observed that most links come from the non-content area, this may provide an explanation why anchor text, Indegree and PageRank may be more helpful for the entry page or named page finding task than for the key document finding task - which may require the authored links from the content and/or related link area. We are doing a post analysis of this issue.

References

[1] Lemur project. *http://www.lemurproject.org*.

[2] J. Callan, Z. Lu, and W. B. Croft. Search distributed collections with inference networds. In *Proceedings of the 18th International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 21–28, Seattle, Washington, 1995.

[3] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 250–257, New Orleans, Louisiana, USA, September 9-12 2001.

[4] N. Eiron and K. S. McCurley. Analysis of anchor text for web search. In J. Callan, D. Hawking, and A. Smeaton, editors, *Proceedings of the 26st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 459–460, Toronto, Canada, July 28 - August 01 2003.

[5] R. Fagin, R. Kumar, and K. McCurley. Searching the workplace web. In Y. R. Bhen, L. Kovacs, and S. Lawrence, editors, *Proceeding of the 12th International Conference on on World Wide Web*, pages 366–375, Budapest, Hungary, May 20-24 2003.

[6] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the 25st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34, Tampere, Finland, August 11-15 2002.

[7] S. H. Myaeng, D. H. Jang, M. S. Kim, and Z. C. Zhoo. A flexible model for retrieval of sgml documents. In *Proceedings of the 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 138–145, Melbourne Australia, 1998.

[8] P. Ogilvie and J. Callan. Combining document representations for known-item search. In *Proceedings of the 26st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 143–150, Toronto, Canada, 2003. A mixture-based language model gives better results than applying meta-search techniques. Rank-based fusion algorithm did not perform as well as the score-based algorithms.

[9] P. Ogilvie and J. Callan. Language models and structured document retrieval. In *Proceedings of the the first INEX workshop*, 2003.

[10] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *Proceedings of CIKM'04*, Washington, DC, USA, 2004. compare the field weight, up weight a term of a field at a document level is logcal reasonable and performs better than weight a term at a field level (take a field as a document collection).

[11] L. Si and J. Callan. A semisupervised learning method to merge search engine results. *ACM Transactions on Information Systems (TOIS)*, 21(4):457–491, 2003.

[12] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286–313, 2003.

[13] T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, urls and anchors. In *Proceedings of the Tenth Text Retrieval Conference, TREC2001, NIST Special publication 500-250*, pages 663–672, 2002.