Homepage Finding and Topic Distillation using a Common Retrieval Strategy

Vo Ngoc Anh *Alistair Moffat*

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{*vo,alistair*}*@cs.mu.oz.au*

Abstract: For the *TREC-2002* web track the University of Melbourne experimented with a system designed primarily for topic relevance tasks, and applied it directly to the homepage finding and topic distillation tasks. Our intention was to process queries regardless of their classification, as discriminating information may be unavailable in practice. An integer-valued weighting scheme reported in earlier work was employed, modified to take into account anchor text and many of the metadata fields, but not the URL text, and not the link structure information. Our experiments were carried out using a distributed retrieval system, with data spread across a sixteen node cluster. Indexing and query processing is fast, and the total index size is small.

1 Background

Our goal in participating in the Web track this year is to explore ways in which normal techniques for topic relevance tasks can be applied to the specific tasks of homepage finding and topic distillation. We examined variations of topic relevance systems, hoping to determine the relative performance of this approach in comparison with the mechanisms employed by other participants.

Our approach can be briefly characterized as: (a) using realistic queries; (b) using document structure, metadata and anchor text, but not URL text; (c) not using link structure; (d) using simple stemming with case-folding; (e) using a (modified) vector space model with integer-valued quantized impacts replacing floating-point term weights in the similarity computation; and (f) applying a common retrieval strategy to the homepage finding and topic distillation tasks. Techniques like phrase indexing, locating noun phrases, and query expansion were not considered in our experiments.

A number of considerations led to our selection of runs. First, we believe that our impact transformation technique [Anh and Moffat, 2002] can give good retrieval effectiveness, and we wished to validate that optimism in the *TREC* context. The Web tasks this year do not include direct topic relevance, but we have maintained the main characteristics of the impact transformation technique and have applied it directly to the new tasks. We sought to use a common system for both of the tasks, and also for the traditional topic relevance task. In order to defend this position, the only training we performed was using the 2000 topic relevance task, and we did not then tune against the 2001 homepage task results.

Second, although metadata and anchor text are integrated into the weighting scheme, their inclusion is not at the cost of the "common" system, and they are regarded as being part of the document in question. A good retrieval system should make effective use of all sources of information, but should also be able to retrieve matching documents regardless of whether or not they contain metadata and/or structuring elements. On the other hand, our exclusion of URL details from the indexing process was a programming oversight, and we would expect to include whatever information they contain in a future system.

Third, the link structure is not used in our weighting scheme simply because we still have not found an efficient (and effective) way to do it. It is reasonable to expect that using link structure enhances retrieval effectiveness, but has a non-trivial cost in terms of implementation effort.

Finally, although we have no great expectation that our approach leads to superior effectiveness results, we also believe that producing a universal method for Web searching, regardless of any specific task, is a desirable goal. Even if users are able to tag their queries, it seems improbable that many would.

Level	Description
1	The following fields (and no others) are indexed: text content, outgoing anchors, incoming anchors, titles, headings, keywords, descriptions.
2	Same as level *1*, except with outgoing anchors excluded.
3	Only text content is indexed.
4	Same as level *1*, except with text content excluded.
5	Only incoming anchors are indexed.

Table 1: Information content levels of indexes used in the experiments along with their description.

2 Weighting scheme

The weighting scheme employed in our experiments is based on our impact transformation technique [Anh and Moffat, 2002], with some modifications made to alter the within-document frequencies prior to using them in the similarity score computation.

Weighting document components An important aspect of our experiments is the effect that indexing different document fields has on the performance of the system, in terms of both effectiveness and efficiency. The document fields considered include the text content; any outgoing anchor text; any incoming anchor text; the title; any headings; keywords; and description fields. All other fields are ignored when indexing. The combinations of features used in our experiments are listed in Table 1.

Whenever an indexed term t is parsed in a document d, a certain contribution is added to the within-document frequency $f_{d,t}$. The contribution differs according to the field this occurrence of t is in. As a baseline, the contribution to $f_{d,t}$ is always 1 for term occurrences in the document text. A contribution of 8 is used for outgoing anchor text; 8 for terms in incoming anchor text from a different host; 4 for incoming anchor text from the same host; and 2 for any of the other fields listed in Table 1. Note that the setting of contribution weights is a very coarse way of boosting the importance of metadata fields and anchor text, and to date we have not made any attempt to tune the parameters for better effectiveness.

The sum of the contributions of term t in document d is used as a surrogate for the conventional $f_{d,t}$ value in the ensuing similarity computation.

Vocabulary Every word that appears in one or more of the selected fields is case-folded, slightly stemmed by the removal of "s" and "ed" suffixes, and indexed. The intention of the stemming process is to only remove differences between plural and single nouns, and between variant verb appearances for regular cases.

Similarity score The modified similarity score $S_{d,q}$ between a document d and a query q is represented as:

$$S_{d,q} = \sum_{t \in q \cap d} \omega_{d,t} \cdot \omega_{q,t}$$

where $\omega_{x,t}$ is an integer in the range 1 to 2^b, with (in these experiments) $b = 5$. The value $\omega_{x,t}$ represents the quantized impact of term t in document or query x, and is calculated in two steps.

First, a normal cosine similarity is used to compute $w^*_{d,t}$ and $w^*_{q,t}$:

$$\begin{aligned}
w_{d,t} &= (1 + \log_e f_{d,t}) \\
W_d &= \sqrt{\sum_{t \in d} w^2_{d,t}} \\
W^*_d &= 1/((1-s) + s \cdot W_d / W^a) \\
w^*_{d,t} &= w_{d,t} / W^*_d \\
w^*_{q,t} &= \log_e \left(1 + \frac{f^m}{f_t}\right) \cdot (1 + \log_e f_{q,t})
\end{aligned}$$

where $f_{d,t}$ and $f_{q,t}$ are the term frequency in the document and query calculated as shown above; f_t is the number of documents that contain term t; f^m is the greatest value of f_t in the collection; W_d is document length; W^a is the average value of W_d over the documents in the collection; and W^*_d represents the normalized document length using pivoted normalization [Singhal et al., 1996] with a slope of $s = 0.7$.

Then, a small enough positive value L and a large enough positive value U are chosen such that all of the $w^*_{d,t}$ lie between L and U, thereby allowing the following transformation to be calculated:

Run label	Weighting scheme	Effectiveness		
		Reciprocal rank	% in top 10	% unfound
MU106	*1L*	0.576	78.7	7.3
MU609	*3G*	0.524	73.3	16.7
MU80A	*5G*	0.402	53.3	26.7
MU307	*4L*	0.207	31.3	58.0

Table 2: Description of homepage finding runs and their official results.

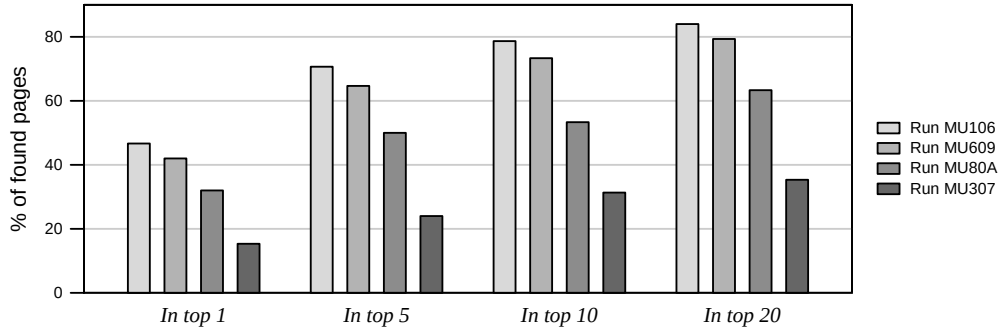

Figure 1: Behavior of weighting schemes for the homepage finding task, percentages of correct homepages found at different number of documents retrieved.

$$\omega_{d,t} = \left\lfloor 2^b \cdot \frac{\log w^*_{d,t} - \log U}{\log U - \log L + \epsilon} \right\rfloor + 1$$

$$\omega_{q,t} = \left\lfloor 2^b \cdot \frac{\log w^*_{q,t} - \log U}{\log U - \log L + \epsilon} \right\rfloor + 1$$

in which $B = (U/L)^{L/(U-L)}$, and ϵ is a small positive quantity, and the impact values are recorded and used as integers.

Our experiments made use of two different types of transformation, characterized by the choice of L and U. In the first, referred to as *global* and denoted by a *G* suffix, the values of L and U respectively are the minimum and maximum values of $w^*_{d,t}$ over the whole database. In the second, referred to as *local* and indicated by an *L* suffix, each document or query x is associated with its own L and U, which are the minimum and maximum among all of the values $w^*_{x,t}$ generated from x. That is, in the local transformation, a value $w^*_{x,t}$ is transformed with respect to the values of L and U of x – the document or query it belongs to.

Weighting scheme notation Two characters are used to denote a weighting scheme. The first character is a digit representing the indexing information level, from 1 to 5 (Table 1). The second is either *G* or *L*, indicating the transformation type.

3 Retrieval effectiveness

Ten runs were submitted for assessment, but an oversight meant that only nine were distinct. Five runs were for the topic distillation task, and four for the homepage finding task.

Homepage finding Our homepage finding runs are summarized in Table 2 and Figure 1. Indexing all components of the documents appears to be better than only indexing parts of each document. To some extent, this conclusion was expected. But the fact that method *1L* performs only slightly better than *3G* indicates that either our crude weighting scheme is contributing little, or that metadata, headings, and titles are relatively unimportant in determining the information content of documents.

When comparing MU609 with MU80A, it is perhaps surprising to see the role of indexing content for this task, and the not-so-good performance of anchor text. Before receiving the results we expected that the use of anchor text alone might give good effectiveness for this task, since anchor text represents a kind of "expert" external view of each document, while content provides a more subjective view. On the other hand, it is also possible that some of the pages sought in these topics were not the target of any other links in the collection.

On the other hand, run MU307 shows that using all metadata plus any anchor text is much worse than using anchor text alone. The use of subjective meta texts (description, keyword, title, heading) by homepage authors may be somewhat arbitrary.

Run label	Weighting scheme	Precision at 10
MU525	*1L*	0.1939
MU111	*2L*	0.1857
MU624	*3G*	0.1694
MU313	*4L*	0.1163
MU212	*4G*	0.1082

Table 3: Description of the topic distillation runs and their official results. For all runs, the title fields of the topics are used unchanged as input queries.

Another point drawn from Figure 1 is the small gaps in performance between the 5, 10 and 20 breakpoints on the number of documents retrieved. If our method is able to find the homepage at all, it appears to rank it highly.

Topic distillation While topic distillation is an interesting and practical task, we were unable to find an effective way of dealing with it using our weighting scheme. Our runs for this task are summarized in Table 3. Part of the poor performance in these runs is accounted for by our lack of use of the URL information that was provided, an omission that with hindsight is now obvious. The key problem in this task, we believe, is to locate good hubs and then discard any child pages connected to those hubs.

The performance of different weighting schemes has the same tendency as in the case of the homepage finding task. Scheme *1L* again perform best. In second place is scheme *2L*, which differs from *1L* only by excluding outgoing anchor text from indexing. Both are significantly better than using content text alone (scheme *3G*) which in turn is better than indexing only metadata and anchor text, or anchor text alone.

The results provide some support for the conjecture that a method that works well for one task can also work well for the other. And while there is no promise of high performance, it does provide hope that building a common system for both of the tasks might be possible.

4 System description and efficiency

Software The experiments were carried out using a modified version of *MG* (see `http://www.cs.mu.oz.au/mg/`). The main feature of *MG* for text retrieval is that it uses compression for document collections as well as for their indexes. This feature is especially appreciated when dealing with large collections like *.GOV*.

Changes have been made to *MG* to suit our needs, in both the indexing and the querying modules. While the changes are already reported in Anh and Moffat [2002], it is worth reiterating that the weighting scheme for document terms is integrated into the index, and that during query processing only a small amount of calculation is required.

Hardware A Beowulf cluster of 16 nodes was used for the experiments. Each of the nodes is an 800 MHz Intel Pentium III with 256 MB RAM, local hard disk of 40 GB, running Debian GNU Linux. The cluster is served by a 933 MHz Intel Pentium III with 1 GB RAM, a 9 GB SCSI disk for system needs, and four 36 GB SCSI disks in a RAID-5 configuration for data. There is a link with the capacity of 100 Mbits/second to and from each node as well as the server with a network switch.

Except for the server, our system uses all nodes for both indexing and querying. The server itself was employed only to deliver jobs to the nodes and to collect the final results. The indicative times reported below are for experiments in which there was no other activity on the hardware.

Data preparation Indexing and query processing was done in a distributed manner. The collection was split randomly across the nodes, with a separate index built for each subcollection. The intention was that each subcollection would be a homogeneous extract of the main collection, and that term frequencies and other collection statistics could be used locally within each subcollection without reference to the global values.

Each query was processed separately against each subcollection using the local index, and a global result listing compiled using the local scores. By assigning documents randomly to collections, we expected minimal degradation of retrieval effectiveness compared to a monolithic system.

Table 4 shows statistics for some of the subcollections and their indexes, built for the weighting scheme *1L*. The figures demonstrate that the subcollections are indeed almost indistinguishable, in terms of their gross statistics, and that the indexes are a small overhead. Their modest size is in part a function of the compression that is used, and in part a consequence of the removal of HTML tags during indexing. Note also that the total in the last row overstates the size of a comparable monolithic index, as there is considerable repetition within the sixteen separate vocabulary files.

Construction of indexes In the inverted list for a term t, each document d containing the term is associated with a quantized impact $\omega_{d,t}$ rather than a conventional $f_{d,t}$ value. Since the number of different quantized values is low, the index structure described by Anh et al. [2001] is appropriate. The pointers in each inverted list are partitioned

Subcollection	Subcollection statistics				Size of index	
	Size (MB)	Documents	Words	Pointers	MB	% of collection size
01	1,095	77,985	672,915	21,849,610	35.82	3.27
06	1,087	77,985	675,313	21,800,614	35.85	3.29
11	1,100	77,984	685,729	22,037,566	36.35	3.30
16	1,096	77,985	631,320	21,924,174	36.64	3.34
Over all subcollections:						
.GOV	17,469	1,247,753	n/a	350,770,758	579.59	3.31

Table 4: Statistics for four of the sixteen subcollections and their respective indexes, built on the 16-node Beowulf cluster for the *.GOV* collection. Document header fields were removed from documents and are not counted in the collection size. Anchor text is added to the corresponding destination document before building subcollections and indexes. Indexes are built for the weighting scheme *1L*, without any pruning, the maximum size among those tested. Sizes includes the cost of the inverted index and the vocabulary file. The last row of the table shows totals over all sixteen subcollections.

into blocks according to their quantized impact. Inside each block, documents are sorted in the increasing order of document numbers. The blocks themselves are arranged in decreasing impact order.

Building such an index requires only a small amount of additional computation compared with building a standard compressed index (described by Witten et al. [1999]). It took around 15 minutes to build the index for the *.GOV* collection and weighting scheme *1L*.

Query processing Query processing is also simple and fast [Anh and Moffat, 2002]. Note in particular that queries are evaluated in a distributed manner on the 16-node cluster. Just a few seconds suffice to run the 50 topic distillation queries and 150 homepage finding queries. The running time does not include time for retrieving the actual documents, and the task is presumed to be completed when the list of answer documents has been created.

5 Conclusion

Our participation in Web track has been limited by a number of simplifications. The link structure within the pages was not investigated, and the weighting scheme and retrieval strategy were not specifically adapted for the topic distillation task. The only feature added to the weighting scheme was the use of high weights for the appearances of words in anchor texts and certain tags. As a result, our results in the two 2002 tasks are not competitive with other systems. Effectiveness on the homepage finding task was better than for the topic distillation task, reflecting the fact that most of the changes to the system were made in favor of homepage finding.

On the other hand, the features we added to our system are not solely driven by this year's tasks, and involve elements that should also be added for the previous topic relevance task. Although the gain in retrieval effectiveness obtained to date is modest, the idea of building a common retrieval strategy for topic relevance, homepage finding and topic distillation seems to be possible.

It seems that a combined task – in which the search systems are not aware of the user's intention – would be a natural next step. That is, no preliminary information about the nature of a query is provided, and the system itself must decide whether it should it be regarded as homepage finding, information finding, or topic distillation. Systems could then approach the task in a number of possible ways, including using a sole system with queries treated equally, or automatically classifying queries into different categories for possible assigning a specific retrieval strategy.

Acknowledgement This work was supported by the Victorian Partnership for Advanced Computing.

References

V. N. Anh, O. de Kretser, and A. Moffat. Vector-space ranking with effective early termination. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–42, New Orleans, LA, September 2001. ACM Press, New York.

V. N. Anh and A. Moffat. Impact transformation: Effective and effi cient web retrieval. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Järvelin, editors, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, Tampere, Finland, August 2002. ACM Press, New York.

A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proc. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29. ACM Press, New York, August 1996.

I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, second edition, 1999.