PERGAMON

Information Processing and Management 39 (2003) 853–871

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Engineering a multi-purpose test collection for Web retrieval experiments

Peter Bailey [a,*], Nick Craswell [b], David Hawking [b]

[a] *Department of Computer Science, The Australian National University, Canberra, ACT 0200, Australia*
[b] *CSIRO Mathematics and Information Sciences, GPO Box 664, Canberra, ACT 2601, Australia*

Received 9 April 2002; accepted 9 September 2002

Abstract

Past research into text retrieval methods for the Web has been restricted by the lack of a test collection capable of supporting experiments which are both realistic and reproducible. The 1.69 million document WT10g collection is proposed as a multi-purpose testbed for experiments with these attributes, in distributed IR, hyperlink algorithms and conventional ad hoc retrieval.

WT10g was constructed by selecting from a superset of documents in such a way that desirable corpus properties were preserved or optimised. These properties include: a high degree of inter-server connectivity, integrity of server holdings, inclusion of documents related to a very wide spread of likely queries, and a realistic distribution of server holding sizes. We confirm that WT10g contains exploitable link information using a site (homepage) finding experiment. Our results show that, on this task, Okapi BM25 works better on propagated link anchor text than on full text.

WT10g was used in TREC-9 and TREC-2000 and both topic relevance and homepage finding queries and judgments are available.

Keywords: Web retrieval; Link-based ranking; Distributed information retrieval; Test collections

1. Introduction

An IR test collection is a laboratory model of a class of real world searching. It generally consists of a corpus of documents, a set of encapsulated information needs (topics), and a sufficiently complete set of relevance judgments.

* Corresponding author.
E-mail addresses: peter@synop.com (P. Bailey), nick.craswell@csiro.au (N. Craswell), david.hawking@csiro.au (D. Hawking).

0306-4573/03/$ - see front matter
doi:10.1016/S0306-4573(02)00084-5

A test collection for Web retrieval research must simulate the salient properties of real Web search. Experimental work with past Web datasets in TRECs 7 and 8 (Hawking, Craswell, & Thistlewaite, 1998; Hawking, Voorhees, Craswell, & Bailey, 1999) has highlighted the need for a new Web test collection which:

- models real Web search, by means of:
  - a sufficiently large and representative document set,
  - a large set of representative Web queries, and
  - a corresponding set of "sufficiently complete" relevance judgments.
- enables meaningful evaluation of hyperlink-based retrieval methods.
- supports experimentation with server selection and result merging algorithms for distributed information retrieval.
- is neither too large nor too "messy" to discourage its use. Large scale use of the corpus is necessary for success of the pooling method (NIST, 2001) in building up reusable relevance judgments.

Here we concentrate on the engineering of a new corpus, to be known as WT10g, which supports as many as possible of the above properties. This task required both characterising the Web and reflecting its properties in a sample collection.

Accurately characterising the *whole Web* poses a formidable challenge, because the Web is huge, complex, highly dynamic and systematically repetitious due to aliasing and mirroring. Even worse, the Web contains vast and unknowable quantities of *dark matter* (Bailey, Craswell, & Hawking, 2000) due to Web page access restrictions, to lack of connectivity in the link graph and to the existence of automatic Web page generators.

Properties of the whole Web such as size and dynamism obviously cannot be reflected in a small fixed corpus. Deciding which properties to preserve or optimise is crucial to creating a successful collection.

Here we present a new methodology for the engineering of test collections to exhibit multiple desired properties and describe its application to the construction of WT10g. We envisage that the same methodology may be applied again in the future.

Here we also document for the benefit of the many researchers working with WT10g, [1] the properties of the collection and the strategy by which it was selected. Finally, we present results of an initial experiment which suggests that the corpus is sufficiently extensive to demonstrate the utility of link based methods, given an appropriate task.

2. Related work

The present work is related to past research in the areas of Web structure and sampling, evaluation of Web search engines "in the wild", and test collection construction.

[1] At the time of writing, more than 40 copies of WT10g have been distributed to research groups around the world and more requests are in the pipeline.

Several authors, e.g. (Lawrence & Giles, 1999; O'Neill, Lavoie, & McClain, 1998), have sampled the Web by probing port 80 (normally the Web server port) of IP (Internet Protocol) addresses chosen at random from among the 2^{32} possible values. Others, e.g. (Henzinger, Heydon, Mitzenmacher, & Najork, 1999; Henzinger, Heydon, Mitzenmacher, & Najork, 2000) have attempted to achieve uniform Web page sampling using random walks on the Web link graph. Characteristics of the Web graph structure, growth rate and server/page size distribution have been investigated by many authors, including (Broder et al., 2000; Faloutsos, Faloutsos, & Faloutsos, 1999; Huberman & Adamic, 1999; Lawrence & Giles, 1999).

Various attempts have been made to measure the retrieval performance of commercial search engines, e.g. (Gordon & Pathak, 1999; Hawking, Craswell, Bailey, & Griffiths, 2001). These studies treat the whole of the Web as a document collection and necessarily evaluate the combination of crawling [2] and document ranking. These studies are unfortunately not repeatable due to changes in the Web and to the inability to obtain complete relevance judgments.

A number of standard test collections for information retrieval have been developed to support repeatable experiments, of which the most widely used are those of TREC (Harman, 1997; NIST, 2001). The TREC ad hoc collections contain more than half a million documents each and represent an increase in data size by three orders of magnitude over earlier collections such as Cranfield (Cleverdon, 1997), CACM and NPL.

Three generally accessible Web test collections are now available (CSIRO, 2001; Hawking et al., 1998; Hawking & Thistlewaite, 1997; Hawking et al., 1999). They are summarised in Table 1. As noted, the source data for VLC2, WT2g and WT10g was the same 1997 crawl (Internet Archive, 1997). Unfortunately, too few relevance judgments are available for VLC2 to support the pooling assumption that unjudged documents can be safely considered to be irrelevant. WT2g addressed this limitation but is very small, and contains very few inter-server links. Possibly as a result of this, TREC-8 participants investigating link-based algorithms such as Kleinberg's HITS (Kleinberg, 1997) and PageRank (Brin & Page, 1998) were unable to demonstrate benefits from using these algorithms in TREC-style ad hoc retrieval tasks.

3. Motivations

The small size of WT2g and its lack of inter-server links were the key motivating factors for creating a new collection but a number of other perceived deficiencies also played a part:

Binary "documents". WT2g includes binary documents incorrectly reported by their Web server as being of type `text/html`. These documents add to the data size, but contribute nothing to the collection.

Duplicate documents. Many server holdings included in WT2g contain duplicate documents (due to name aliasing and/or broken links leading to error pages), and repeated documents at the same URL.

Non-English documents. A number of non-English language documents were included in VLC2 and WT2g. While such documents are interesting and contribute to a realistic representation of

[2] Crawling is the process of identifying and fetching Web pages, usually for indexing, by traversing the Web link graph from a set of starting points.

Table 1
Web test collections

Collection	Documents	Size (GB)	Year distribution	Composition	Queries with judgments
VLC	7.5M	20	1997	For scalability experiments (only 4.5% Web documents)	TREC-6 - Not reusable
VLC2	18.5M	100	1998	From 1997 crawl by Internet Archive	TREC-7,8,9 - Not reusable
WT2g	0.25M	2	1999	Subset of VLC2	TREC-8 topics 401–450 - NIST generated
WT10g	1.7M	10	2000	Subset of VLC2	TREC-9 topics 451–500 - From search logs, word misspellings TREC-2001 topics 501–550 - From search logs TREC-2001 EP1–EP145 - Homepage finding - NIST generated

Relevance judments have been made for many topics with respect to VLC and VLC2, but they are incomplete and therefore difficult to reuse.

the Web, their presence increases the difficulty of indexing without bringing any benefit to the performance of an English-language task. Furthermore, small numbers of documents representing an uncontrolled spread of non-English languages would not support useful cross-language experiments.

Missing homepages. Several servers represented within WT2g are missing their homepage. This is a potentially serious shortcoming as far as link based methods are concerned.

Absence of metadata. WT2g pages include very little useful metadata.

We set out to create a corpus which would remedy as many as possible of the identified shortcomings of WT2g. There were however, other, conflicting requirements. The desire for representativeness of the general Web argues for a large collection but the requirement for "sufficiently complete" relevance judgments argues against it.

We compromised by targetting a corpus size of 10 GB, representing about 1.5 million Web documents. Although the raw data size is five times that of earlier TREC ad hoc and WT2g corpora, the number of documents is only a little more than double that of some of the ad hoc collections. By eliminating most of the documents which would cause unnecessary indexing difficulty, we hoped to facilitate high levels of participation in the TREC tasks using the new collection, which in turn would contribute to more complete judgments.

In summary, we wished to construct a 10 GB corpus which:

- was broadly representative of Web data in general (to improve the chance that conclusions drawn from WT10g experiments will be applicable on the real Web),
- contained many inter-server links (to permit meaningful experimentation with hyperlinks),
- contained all available pages (always including homepages) from a set of servers (to permit meaningful distributed IR and hyperlink experiments),

- contained an interesting set of metadata (to permit experimentation on retrieval methods based on content and metadata), and
- contained few binary, duplicate or non-English documents (to minimise difficulty of use).

4. Corpus construction methodology

We selected a superset of documents which we hoped would be large enough to represent the whole Web, and measured its properties. We then chose a subset of the desired size using a heuristic algorithm designed to preserve or optimise those properties. Finally, we compared this subset (WT10g) with hypothetical subsets resulting from competing selection strategies to confirm that WT10g best suited the purposes for which the collection was intended.

We rejected the approach of crawling the target quantity of data directly from the Web, even though this would have provided more recent documents and a large degree of control over what was included. Without a process of large scale measurement and careful selection, there could be no guarantee that such a corpus would have satisfied the requirements outlined above.

5. Constucting the WT10g corpus

Construction of WT10g proceeded in four phases.

5.1. Phase 1: choice of superset

Having chosen to use a large archive of Web data as a starting point, we considered using the full 320 GB of the 1997 Internet Archive data, or alternatively obtaining a more recent large crawl from the Internet Archive or from another source. Proximity to the deadline for distribution of WT10g, coupled with the time consuming nature of certain phases of the analysis (such as link connectivity) compelled us to work with the 100 GB VLC2 as our superset.

5.2. Phase 2: rejection of unwanted data

5.2.1. Binary and non-English language data

Although there are standard HTTP/HTML [3] mechanisms for indicating the character set used within a document (and hence which documents are likely not to be English language), these mechanisms are not uniformly or reliably used, particularly with data from 1997.

Accordingly, we developed a heuristic tool based on the PADRE retrieval system (Hawking, Thistlewaite, & Craswell, 1997) to analyse VLC2 documents and accept only those:

- shorter than 75 indexed words, or
- which included a common English word (“and”, “the”, “for”, “you”, “not”, “this”) occurring at least twice, or

[3] HTTP (HyperText Transmission Protocol) and HTML (HyperText Markup Language) are the two standards on which the current Web is based.

- which contained at least 30 words (from `/usr/dict/words`), or
- which contained at least one word more than three times.

More sophisticated systems for language detection, such as those of (Alis Technology Inc., 2001) are available commercially. Unfortunately, we lacked the time necessary to obtain one or to develop an equivalent.

We hypothesised that the removal of non-English documents would lead to a greater decrease in vocabulary size than would the removal of an equivalent number of randomly chosen documents. We expected that non-English documents would contain large numbers of unusual letter sequences which would add to the vocabulary.

We tested this using the WT2g collection and the PADRE retrieval system. The initial vocabulary size was 1 047 075. Removing a few thousand documents identified as non-English from the 247 491 in WT2g caused the vocabulary size to drop by 6.7%, while removing the same number of documents chosen at random resulted in a 4.2% drop. This suggests that while our hypothesis may be correct, its effect is small.

We also tested how accurately our heuristic identified non-English documents. We did this by randomly selecting 100 of the rejected documents from the VLC2 collection and manually classifying their contents. (We also used this technique during tuning of the heuristic.) While the sample is small, the resulting categories are instructive. The results are reported in Table 2. The absence of any binary documents in this sample might be because there are relatively few binary documents or because our heuristic is failing to identify them correctly. The majority of the remaining documents, while English language, would be unlikely to be retrieved for the majority of topic relevance queries as they contain few English language words. The removal of these documents, though unintended, is therefore considered to be of little consequence.

5.2.2. Avoiding other undesirable corpus properties

An initial draft selection of WT10g was analysed by the present authors and by Allison Powell and Jim French at the University of Virginia. Some additional problems were identified.

Table 2
Document categories for 100 documents rejected by binary and non-English heuristic

Quantity	Description
81	Non-English
0	Binary
4	Lists of URL
3	Directory indexes
3	Sports scores
2	Email archives
2	Homepages
2	Advertisements
2	HTTP server statistics
1	Table of statistics

The VLC2 contains a number of sets of documents whose members share the same URL. [4] Regardless of whether the content was identical or not, only the first document instance for any given URL from the VLC2 was considered acceptable.

Name aliasing can lead to HTTP servers returning the same document for different URLs. Also, servers may contain broken links which lead to apparently unique and valid pages returning the same error page, instead of an HTTP 404 error code. Although not a completely reliable heuristic (see (Bharat, Broder, Dean, & Henzinger, 1999) for a more thorough treatment of the identification of duplicate documents in Web data), a 64-bit CRC64 checksum provides a reasonable signature of identity. The probability that CRC64 will falsely signal a duplicate in this size collection is estimated to be of the order of 10^{-7}. The main limitation of CRC methods is their inability to detect "near-duplicates", i.e. pages which differ only in some small detail.

We eliminated all documents sharing the same checksum as another document on the same server.

A final problem was the existence of large quantities of generated material and/or non-HTML documents. Although the W3C (The World Wide Web Consortium, 1998) recommends that any URL, whatever the filename extension, is acceptable as an HTML document, in practice, most valid HTML documents from 1997 end in some variant of `.html` or `.htm`. We made a decision that only documents whose URL ended with a `.html` (and variants) or `.txt` extension would be acceptable. We also specifically eliminated any identifiably generated documents. If there was no extension, then we considered the file acceptable (such URLs frequently refer implicitly to the default document for a Web directory).

A summary of the VLC2 documents rejected by our heuristics is given in Table 3.

5.3. Phase 3: characterisation of servers

Unlike a classic document collection, the Web possesses additional structure arising from the grouping of documents by HTTP server and by directory and from the hyperlinks which interconnect documents. HTTP servers are analogous to the publishers, such as the Wall Street Journal, of documents in the standard TREC collections. However, HTTP servers are more numerous and vary more widely in the number of documents they serve.

We considered it important to preserve server structure in the subset and to include all the available pages from each selected server for two reasons:

1. Servers would constitute the basic building blocks in a distributed search simulation, and
2. Server holdings correspond to top level sites on the Web. Web searchers frequently use search engines to assist them to navigate to the entry page of a top level site, such as `www.sony.com` or `www.anu.edu.au`. Once there, they may browse the site by following links.

[4] For example, there are several occurrences of `http://dns.prox.it:80/`. We are unable to specify exactly why this is the case as we have been unable to obtain any information about the Internet Archive's crawler. It is possible that the URLs requested were actually different but that the crawler canonicalized them when it recorded them or alternatively that the crawler recorded the final target of HTTP redirections without recognizing that the target had already been crawled.

Table 3
VLC2 properties: URLs and servers, unacceptable documents by category

Quantity	Description
18 571 671	Total documents
117 101	Servers (no hostname alias detection)
1 739 552	Binary and non-English documents
862 410	Duplicate documents at the same URL
1 691 342	Identical checksums on the same server
3 563 775	Generated and rejected URL extensions
7 781 408	Total eliminated

A subset selected without regard to server holdings would be very unnatural in Web terms.

Having identified and recorded documents that were unacceptable for various reasons during Phase 2, we measured the relevant properties of each server in the VLC2 collection.

During the process, we realised that too few VLC2 pages contain useful metadata to permit a 10 GB subset capable of supporting interesting metadata experiments. Accordingly, metadata content was abandoned as a factor in our selections.

For each server we recorded:

- total number of documents,
- number of documents rejected in Phase 2
- existence of a homepage
- general relevance score
- link counts

The latter two properties require some explanation.

The general relevance score for each server was the proportion of a large set of Web queries for which at least one document from the server was ranked in the top 500 documents returned by PADRE. The queries were the 10 000 queries from the TREC-8 Web Track large task. The intent of this arbitrary measure was to establish a level of general relevance to many topics. This metric is biased in favour of larger servers, but Phase 4 (below) imposes constraints on the distribution of server sizes selected.

Link counts include only inter-server links whose source and target both lie within VLC2. The server inlink count includes all links referencing any page in a server's collection which originate from another server. The outlink count was computed in an analogous way. A server–server sparse matrix recorded the number of links between each pair of servers.

5.4. Phase 4: selection of WT10g servers

5.4.1. Size representativeness

A representative distribution of server sizes was a very important goal for WT10g. Huberman and Adamic (1999) approximate the distribution of Web server sizes using the power law

$$P(n_s) = Cn_s^{-\beta}$$

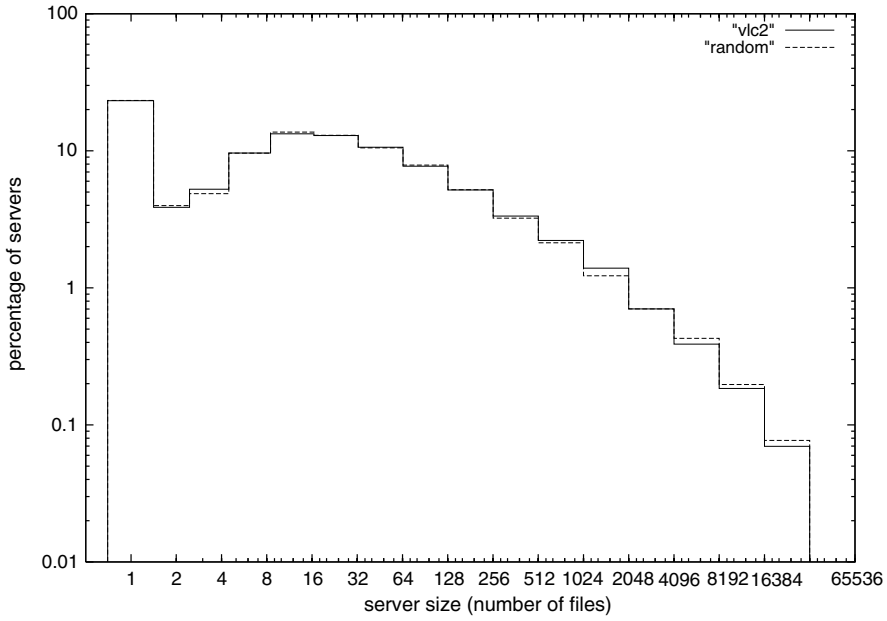

Fig. 1. The server size distribution for VLC2, quantised into $\log_2$ scaled buckets. Also shown is the corresponding distribution for a random (approximately 10%) sample.

where $P(n_s)$ is the probability that a Web site has n_s pages and C and $\beta \geqslant 1$ are constants. This means that there are a few very large servers, and many very small ones.

Fig. 1 shows the server size distribution of VLC2 bucketed according to powers of two. Thus the buckets consisted of server sizes as follows: 1, 2, 3–4, 5–8, 9–16,... etc. We expressed the bucket size as a percentage of the total number of servers, allowing us to compare selections of different sizes. The second histogram shows the server size distribution for a randomly chosen selection of the same number of servers present in the final WT10g. This confirms that random sampling does not affect the server size distribution.

To enable direct comparison with the Huberman and Adamic model, Fig. 2 shows the probability that a randomly selected VLC2 server has a particular size. Visual comparison with the Alexa plot in Fig. 3 of Huberman and Adamic [5] shows a very similar shape, although the dip at small server sizes ($1 < n_s < 8$) appears more pronounced. As a result, the slope of the line of best fit obtained using a linear regression of log–log data suggests $\beta = 1.54$ which is lower than the range of $1.65 \ldots 1.85$ reported by Huberman and Adamic.

The fact that WT10g contains none of the multi-million page servers such as Geocities is not a problem as the probability of occurrence of these servers is very low. Of more concern were the number of single page servers. Often they contained an error page or an "under-construction" message. Neither of these kinds of page are interesting for a Web collection. Many other singleton pages we examined contained links to same-server pages not represented within VLC2. We could

[5] The Internet Archive is operated by the Alexa Corporation. It is possible that Huberman and Adamic were in fact using the same (full) crawl from which the VLC was selected.

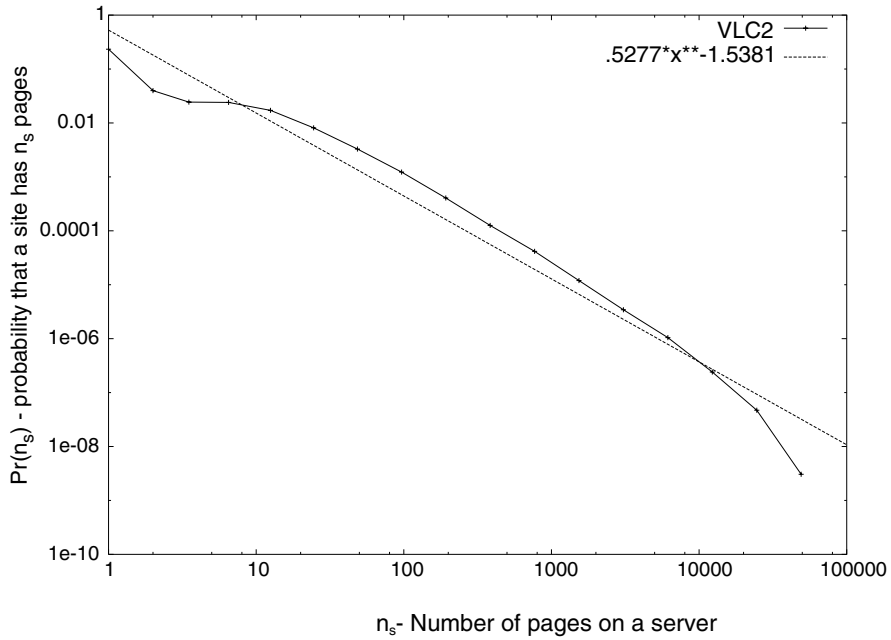

Fig. 2. Fig. 1 data, presented to enable direct comparison with Fig. 3 of Huberman and Adamic and showing the least-squares line of best fit.

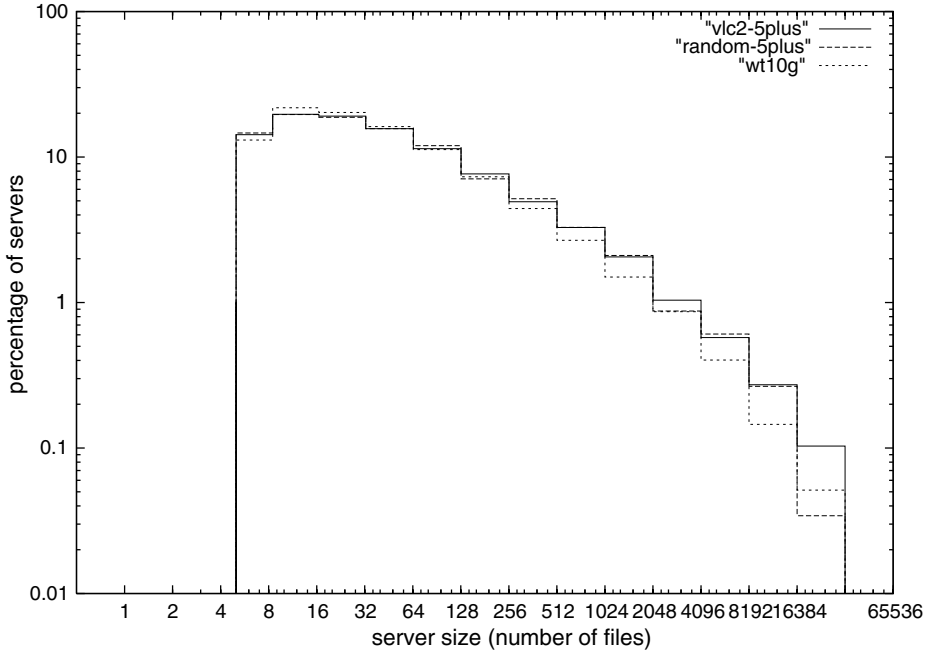

Fig. 3. Size distribution of the VLC2 and WT10g collections, split into $\log_2$ scaled buckets, and only for servers with five or more pages.

not tell whether they appeared later in the crawl, or whether they were missed by the crawler for some reason such as network outage or server downtime.

After further consideration, we decided to select only servers represented by at least five pages.

Fig. 3 shows the final WT10g server size distribution together with distributions for a random server selection of the same size and for the VLC2 minus small servers. The small WT10g deviations are discussed in the next section, however the general similarity of the three plots is reassuring.

5.4.2. Selection algorithm

We used an obvious technique to preserve the server size distribution in the final collection. We ordered the VLC2 servers into their respective size buckets and selected from each bucket according to the proportion it represented of the overall count. Each time a bucket was selected, our algorithm sorted its servers according to a ranking formula described below, to determine which servers should be taken.

We set an upper limit on the number of servers to be selected by estimating how many servers would be required to make up 10 GB. We made a series of passes through the complete set of buckets in ascending frequency order. We started by selecting one server from the bucket with the fewest entries (containing servers with the largest number of documents). From each successive bucket we selected more servers, in proportion to the number of servers they contained. It turned out that two passes through the buckets were more than sufficient, gathering 10% more data than we needed. To trim back to our desired target corpus size, we arbitrarily eliminated servers selected in the second pass which ranked lowest in their buckets. We excluded the first two buckets since we wanted to preserve large servers.

In ordering servers within each bucket, certain properties were given priority. Thus, given two servers A and B, A is chosen over B if:

- A's homepage is present and B's is not
- A has no rejected documents and B does
- A's homepage is not rejected and B's is

The net result of this is that the top servers contain unrejected homepages, those containing some rejected documents ranking below those with none. After this come servers with rejected homepages, and finally those with no homepages. In fact, all servers selected for WT10g contained unrejected homepages. This is considered to be a desirable property because:

1. Homepages are normally the primary source of links by which the rest of the site can be accessed, and the primary target of incoming links to the site.
2. Homepages are often the targets of navigational queries submitted to Web search engines with the object of locating the entry page to a site. Examination of sample Web search logs and anecdotal evidence from search engine companies suggests that this type of query typically constitutes around 15–20% of queries submitted.

Within tiers of servers determined by the above ordering, servers s were ranked according to scores computed as follows:

$$\text{Score}_s = 0.25\text{il} + 0.25\text{ol} + 0.3\text{rel} + 0.2\text{acc}$$

where il is the inlink score, ol is the outlink score, rel is the relevance score, and acc is the acceptable page ratio for s.

The inlink and outlink scores were calculated dynamically as each bucket was encountered, with respect to the servers already selected. Thus for an individual server, its inlink score was determined only from servers already in the selected pool. Similarly, its outlink score was determined only by links it contained to servers already in the selected pool. For the very first (large) server, these inlink and outlink scores were instead the statically computed scores for the whole VLC2.

The intent of this dynamic recalculation was to maximise the inter-server link count within the collection. An even more effective strategy would have been to reorder the servers within a bucket after every new server was chosen. However, this method would have been computationally prohibitive. (As it was, with two passes and 13 buckets, there were only 26 recalculations required. Buckets with larger numbers of servers naturally took longer to reorder. Even so, the total selection time, dominated by the dynamic inlink and outlink calculations, took approximately 2 h. Reordering buckets after every new server was selected would have increased the time taken by three orders of magnitude.)

It is possible, but by no means certain, that the use of link count rather than link density, may have biased the selection toward larger servers, but again, the overall server size distribution was constrained.

6. Analysis of the final WT10g selection and possible alternatives

Table 4 reports statistics for the final WT10g selection.

While developing the selection methodology for WT10g, we created a tool to analyse the properties of any given (mooted) selection. We used it to compare a number of alternative selections arising from varying the constants in the ranking formula Score_s, while holding the server size distribution constant. These alternate selection policies are listed below. We also analysed the

Table 4
WT10g properties

Quantity	Value
Documents	1 692 096
Servers	11 680
Average documents per server	144
Inter-server links (within WT10g)	171 740
Servers with inter-server inlinks	9988
Servers with inter-server outlinks	8999
Documents with outlinks	1 295 841
Documents with inlinks	1 532 012
Servers without a homepage	0

result of choosing servers on the basis of static VLC2 inlink and outlink counts rather than dynamic counts within the growing WT10g collection.

- *dynamic_by_inoutlink*: Score_s modified to weight il at 0.5 and ol at 0.5; (0 for both acc and rel).
- *dynamic_by_inlink*: Score_s modified to weight il at 1 and 0 for ol, acc, and rel.
- *static_by_inlink*: the same as *dynamic_by_inlink*, but using the static VLC2 inlink score for bucket ordering.
- *dynamic_by_outlink*: Score_s modified to weight ol at 1 and 0 for il, acc, and rel.
- *static_by_outlink*: the same as *dynamic_by_outlink*, but using the static VLC2 outlink score for bucket ordering.
- *static_by_relevance*: Score_s modified to weight rel at 1 and 0 for il, ol, and acc (in other words, links are ignored).
- *static_by_acceptable_ratio*: Score_s modified to weight acc at 1 and 0 for il, ol, and rel.

Finally, we analysed a number of other selection policies in which the server size distribution properties was not constrained:

- *vlc2_all_buckets*: the supercollection from which WT10g is derived.
- *vlc2*: all VLC2 servers with five or more pages.
- *wt2g*: the 2 GB collection from TREC-8's Web Track, also a subcollection of the VLC2.
- *random_all_buckets*: randomly selected servers of any size; same total number of servers as WT10g.
- *random*: randomly selected servers having five or more pages but without bucketing them; same total number of servers as WT10g.
- *dynamic_two_pass*: two complete passes using Score_s, without removal of servers to match the size limits of WT10g's selection (a total of 610 extra servers more than WT10g).
- *static_two_pass*: two complete passes using Score_s modified to use the static VLC2 inlink and outlink scores, but the same weighting for il, ol, acc, and rel.

The corpus properties reported by the tool are listed below. Their values for various hypothetical selections are compared (with appropriate scaling) in Fig. 4.

- *average inout*: the average number of inlinks per server (which is the same as the average number of outlinks per server since all inter-server links are measured from within the collection).
- *average relevance*: the per-server relevance score, averaged over the 10 000 queries, scaled by a factor of 5000.
- *average score*: the per-server average of Score_s, as modified for the *static_two_pass* selection, scaled by 100.
- *server outlink*: the percentage of servers in the selection containing any outlinks.
- *server inlink*: the percentage of servers in the selection containing any inlinks.
- *server relevance*: the percentage of servers in the selection containing any relevant pages.
- *server good page ratio*: the percentage of servers without rejected pages (in the VLC2 collection). Note that rejected pages are not included in the final collection.

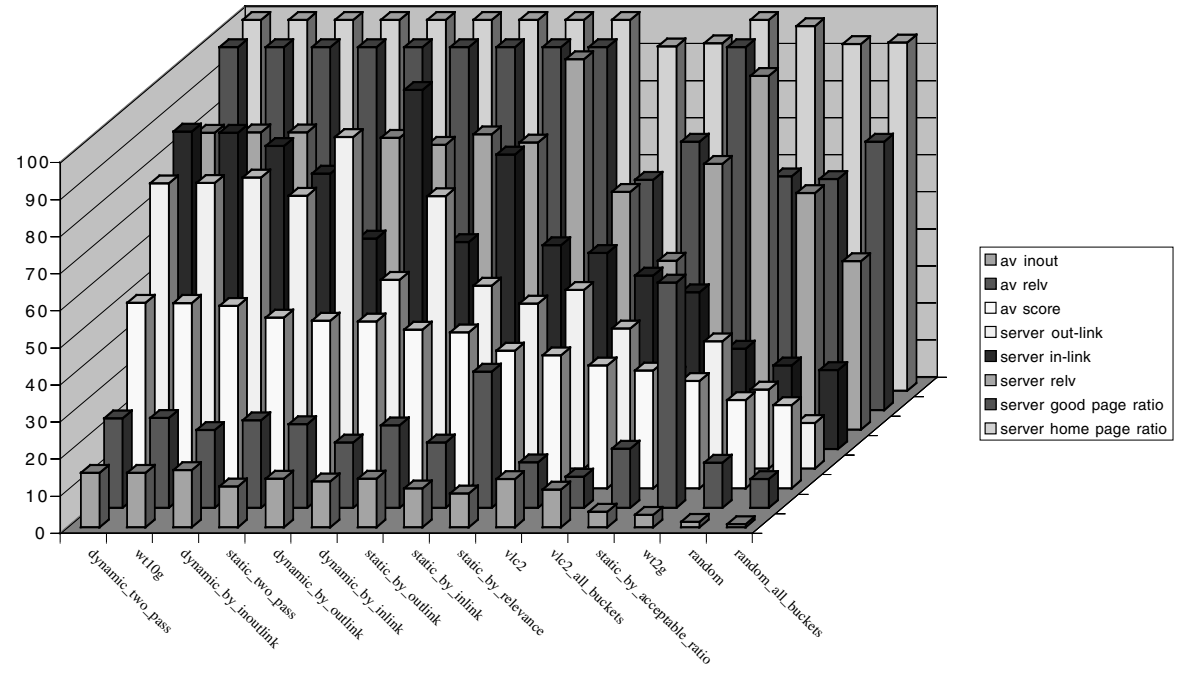

Fig. 4. Comparison of alternative selection strategies.

- *server homepage ratio*: the percentage ratio of servers with homepages to the total number of servers in the selection.

In examining Fig. 4, we see that the *wt10g* selection is close to the best overall performing selection. (The ordering from left to right of the selections is based on the value of *average score* as this is the best indicative metric of all the properties.) The *dynamic_two_pass* selection is marginally better for its *average inout* metric, because servers were not subsequently removed from the selection to meet the 10 GB target.

Individual selections chosen explicitly to optimise a particular metric outperform *wt10g* on those metrics, but are outperformed by *wt10g* on most of the other metrics. For example, the *dynamic_by_outlink* selection has a significantly higher *server outlink* ratio. Some other points worth noting are that both the *vlc2* and the *random* selections (containing only servers with five or more pages) perform better than their all page counterparts *vlc2_all_buckets* and *random_all_buckets* across all metrics other than the *server good page ratio*. Also worth noting is that *wt2g* performs substantially better on the relevance metrics *average relevance* and *server relevance* than any of the other selections. Only the *static_by_relevance* selection comes close.

This suggests strongly that the *average relevance* metric is biased by the number of servers in the selection. Intuitively this is obvious, since we applied a fixed document cutoff of top 500 documents per query. With fewer servers, it is far more likely that any individual server is likely to get a document into the top 500, particularly when averaged over 10 000 topics. Confirming this suspicion is that both *vlc2* selections have relatively poor values for the two *relevance* metrics.

6.1. TREC experience with the WT10g collection

The WT10g corpus was used in the Web Track of both TREC-9 and TREC-2001 and consequently it is now possible to comment on the extent to which design goals were achieved in practice. Detailed descriptions of the TREC evaluations are to be found in Hawking and Craswell (2001) and Hawking, Voorhees, Bailey, and Craswell (2000).

The hurdle of indexing 10 GB of data was overcome by 23 groups in TREC-9 and 30 in TREC-2001. TREC-9 participants included a roughly equal number of commercial organisations and universities plus a government research agency and an individual student. It seems safe to conclude that WT10g is neither too large nor too "messy" for use even by organisations with relatively modest hardware resources.

As reported in Section 8 below a large set of realistic queries and judgments was created in these TREC evaluations. The topic relevance queries were taken from real search engine logs and the homepage finding queries are representative of a type of query which is common on the Web, but not meaningful in search of collections comprising newspaper data, reports, transcripts or email.

TREC exercises to date have not explicitly addressed the evaluation of distributed IR techniques in the context of WT10g. It is fair to say that the potential of WT10g to support meaningful server selection and merging experiments remains to be confirmed. However, the Université de Neuchâtel (Savoy & Rasolofo, 2000; Savoy & Rasolofo, 2001) report merging experiments based on a four-way division of the data.

Results described in Section 7 below have been confirmed in the TREC-2001 entry page task. Anchor text [6] (purely link based evidence) was effectively exploited on this task by various participants. The Twenty-One group (Westerveld, Kraaij, & Hiemstra, 2001) also showed that inlink count and link-based hub and authority scores could be used to improve a content-only method on this task. They found the evidence supplied by URL structure, which is another Web-specific feature, to be even more useful.

It is clear that WT10g does contain exploitable link evidence. However, it comprises less than 0.1% of the pages reportedly indexed by major search engines such as Google (www.google.com) and FAST (www.alltheweb.com) in August 2002. Consequently, there are severe limits on the extent to which WT10g results can be extrapolated to the Web. On the other hand, WT10g findings may be applicable to organisational Webs, many of which are smaller than WT10g.

7. A link experiment using WT10g

Prior to the distribution of the WT10g collection in early 2000, we conducted a small experiment to determine whether there is sufficient link information within WT10g to enhance homepage finding effectiveness. [7] Homepage finding is an important Web search problem in which link information has proven useful in previous experiments (Craswell, Hawking, & Robertson, 2001) and in commercial search engines.

[6] Anchor text comprises the browser-highlighted words which a person clicks on to follow a link. In the HTML link <AHREF="http://www.cdsnet.net/vidiot/">Vidiot Home</A> the words "Vidiot Home" constitute the anchor text.

[7] A homepage is the main entry page to a Web site.

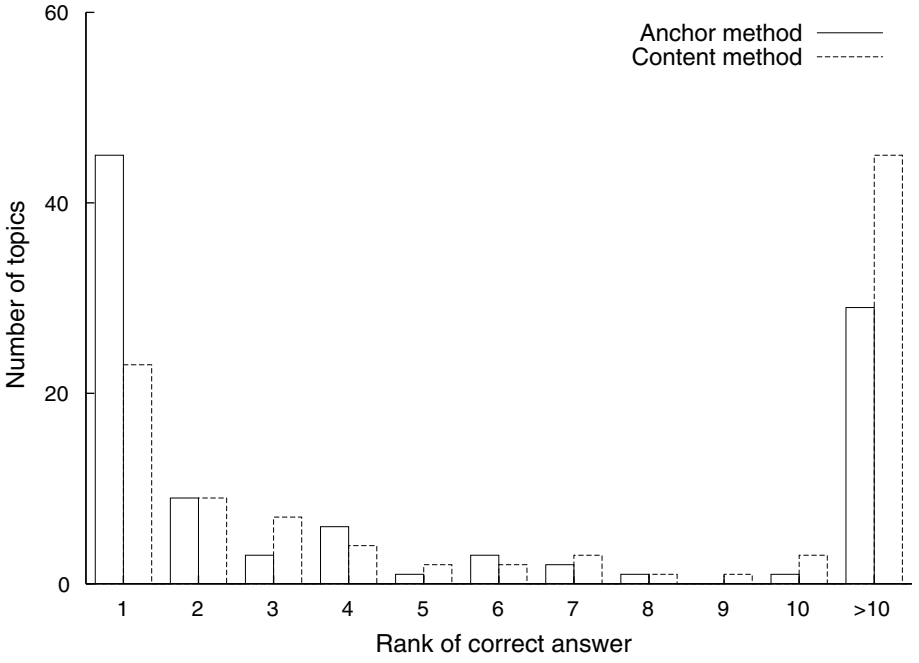

Fig. 5. Relative effectiveness of link and content methods on a homepage finding task.

We randomly chose 100 site homepages from WT10g, and manually generated a query (a name for the site) for each. For example, the query "Vidiot" was generated for the homepage http://www.cdsnet.net/vidiot/. The query processing task demands that the required homepage be returned as close to the top of the results list as possible.

Two runs were performed and compared. In the content-only run, PADRE searched WT10g using the Okapi BM25 ranking algorithm (Robertson, Walker, Hancock-Beaulieu, & Gatford, 1994). In the second, PADRE searched document surrogates called *anchor documents*. An anchor document for a page with incoming links, contains all (and only) the anchor texts of the page's incoming links. A page without any incoming links has no anchor document and cannot be retrieved by this means. PADRE indexed the document surrogates and applied exactly the same ranking algorithm as in the content method. Since the link documents were based solely on link (anchor and target) information, this is a pure link-based ranking method.

For 56 out of the 100 queries, the link method retrieved the homepage at a higher rank than the content method. Fig. 5 presents a histogram of the ranks at which the right answer was obtained. The methods were equal on 17 queries and content was superior on 27. A sign test (Hays, 1963, p. 625) shows that the differences are significant. $p < 0.01$.

8. Relevance judgments

A useful test collection requires extensive sets of requests with corresponding judgments. A substantial quantity have been built up for WT10g in the course of TREC-9 and TREC-2001 Web

Track evaluations (Hawking & Craswell, 2001; Hawking et al., 2000). Details of available relevance judgments are listed in Table 1. They include 100 judgments for ad hoc relevance topics taken from search engine logs and 245 for home (entry) page finding tasks. They are all available from the TREC Web site (NIST, 2001) except for the 100 homepage finding queries used in the experiment reported in Section 7, which are available from the TREC Web Track Web site (CSIRO, 2001).

Note that judgments for the ad hoc relevance topics were ternary (irrelevant, relevant, highly relevant) (Voorhees, 2001) and based on pooling. These topics were reverse engineered by NIST from Web search engine logs and include the original Web query in the `title` field. Topics 451–500 include a number of misspelled words (from the original queries) but topics 501–550 do not.

9. Conclusions

The WT10g corpus [8] is proposed as a resource for repeatable retrieval experiments which model Web search better than possible with any alternative test collection. Its small size necessarily means that certain Web properties cannot be accurately modelled, but we have demonstrated that it does contain exploitable link information. We expect WT10g to be useful for a number of different types of retrieval task, corresponding to different types of information need, and supporting distributed as well as centralized techniques. We have partially confirmed these expectations within the TREC Web Track.

A substantial number of topic relevance and homepage finding judgments are now readily available.

The methodology used in engineering WT10g should be readily applicable to the construction of future Web corpora. Such corpora might be based on larger and more current supersets and/or different selection criteria.

Acknowledgements

The authors wish to acknowledge that this work was partly carried out within the Cooperative Research Centre for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

Allison Powell and Jim French at the University of Virginia provided valuable feedback in discussions about desirable (and undesirable) properties of WT10g during its construction. Ellen Voorhees, TREC Program Manager at NIST, provided advice and assistance with respect to the TREC requirements of the collection. Isao Namba of Fujitsu Laboratories provided early advice on approaches to non-English language detection.

[8] WT10g and other Web test collections are now distributed by CSIRO, which requires that recipients sign a usage agreement and charges a fee to cover its costs. WT10g (compressed with `gzip`) is distributed on 5 CD-ROMS and a variety of additional information, including link connectivity matrix is included. Full details are available at `http://www.ted.cmis.csiro.au/TRECWeb/`.

References

Alis Technology Inc. (2001). *?'qué? system for identification of language and character encoding*. Available: www.alis.com/castil/silc/. Accessed 20 Aug 2002.

Bailey, P., Craswell, N., & Hawking, D. (2000). Dark matter on the web. In *WWW9 Poster Proceedings*. Available: www.www9.org/final-posters/poster30.html. Accessed 20 Aug 2002.

Bharat, K., Broder, A., Dean, J., & Henzinger, M. (1999). A comparison of techniques to find mirrored hosts on the WWW. In *Proceedings of the ACM Digital Libraries '99 Workshop on Organizing Web Space (WOWS)*, Berkeley, CA. Available: research.microsoft.com/research/db/debull/A00dec/bharat.ps. Accessed 20 Aug 2002.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the seventh international world wide web conference*, pp. 107–118.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2000). Graph structure in the web. In *Proceedings of WWW9*, Amsterdam. Available: www9.org/w9cdrom/contents.html#CHARACTERIZATION. Accessed 20 Aug 2002.

Cleverdon, C. (1997). The Cranfield tests on index language devices. In K. S. Jones & P. Willett (Eds.), *Readings in information retrieval* (pp. 47–59). San Francisco: Morgan Kauffman, Reprinted from Aslib Proceedings, vol. 19, pp. 173–192.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information. In *Proceedings of ACM SIGIR 2001*, New Orleans, pp. 250–257. Available: www.ted.cmis.csiro.au/~nickc/pubs/sigir01.pdf.

CSIRO (2001). *TREC Web Tracks homepage*. Available: www.ted.cmis.csiro.au/TRECWeb/.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Proceedings of ACM SIGCOMM'99*, pp. 251–262.

Gordon, M., & Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management, 35*(2), 141–180.

Harman, D. (1997). The TREC Conferences. In K. Sparck-Jones, & P. Willett (Eds.), *Readings in information retrieval* (pp. 247–256). San Francisco: Morgan Kaufmann.

Hawking, D., & Craswell, N. (2001). Overview of the TREC-2001 Web Track. In *Proceedings of TREC-2001*, Gaithersburg MD. NIST special publication 500-250. Available: trec.nist.gov.

Hawking, D., Craswell, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval, 4*(1), 33–59, Preprint at www.ted.cmis.csiro.au/~dave/INRT83-00.ps.gz.

Hawking, D., Craswell, N., & Thistlewaite, P. (1998). Overview of the TREC-7 very large collection track. In *Proceedings of TREC-7*, Gaithersburg, MD, NIST, pp. 91–104.

Hawking, D., & Thistlewaite, P. (1997). Overview of the TREC-6 very large collection track. In *Proceedings of TREC-6*, Gaithersburg, MD. NIST.

Hawking, D., Thistlewaite, P., & Craswell, N. (1997). ANU/ACSys TREC-6 experiments. In *Proceedings of TREC-6*, Gaithersburg, MD, NIST, pp. 275–290.

Hawking, D., Voorhees, E., Bailey, P., & Craswell, N. (2000). Overview of the TREC-9 web track. In *Proceedings of TREC-9*, Gaithersburg MD, NIST special publication 500-249. Available: trec.nist.gov.

Hawking, D., Voorhees, E., Craswell, N., & Bailey, P. (1999). Overview of the TREC-8 web track. In *Proceedings of TREC-8*, Gaithersburg, MD, NIST, pp. 131–150.

Hays, W. L. (1963). *Statistics*. London: Holt, Rinehart and Winston.

Henzinger, M.R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999). Measuring index quality using random walks on the web. In *Proceedings of WWW8*, Toronto. Available: www8.org/w8-papers/2c-search-discover/measuring/measuring.html.

Henzinger, M.R., Heydon, A., Mitzenmacher, M., & Najork, M. (2000). On near-uniform url sampling. In *Proceedings of WWW9*, Amsterdam. Available: www9.org/w9cdrom/contents.html#Characterization. Accessed 20 Aug 2002.

Huberman, B.A., & Adamic, L.A. (1999). *Evolutionary dynamics of the World Wide Web*. Available: www.hpl.hp.com/shl/papers/webgrowth/. Accessed 20 Aug 2002.

Internet Archive (1997). *Building a digital library for the future*. Available: www.archive.org/. Accessed 20 Aug 2002.

Kleinberg, J. (1997). *Authoritative sources in a hyperlinked environment*. Technical Report RJ 10076, IBM.

Lawrence, S., & Giles, C. L. (1999). Accessiblity and distribution of information on the Web. *Nature, 400*, 107–109.

NIST (2001). *Trec home page*. Available: trec.nist.gov/. Accessed 20 Aug 2002.

O'Neill, E., Lavoie, B., & McClain, P. (1998). OCLC web characterization project. In *Report of the W3C web characterization group conference*, Cambridge, MA. Available: www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm. Accessed 20 Aug 2002.

Robertson, S.E., Walker, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of TREC-3*, Gaithersburg, MD, NIST.

Savoy, J., & Rasolofo, Y. (2000). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In *Proceedings of TREC-9*, Gaithersburg, MD, NIST, pp. 579–588. Available: trec.nist.gov.

Savoy, J., & Rasolofo, Y. (2001). Report on the TREC-10 experiment: Distributed collections and entrypage searching. In *Proceedings of TREC-2001*, Gaithersburg, MD, NIST, pp. 586–595. Available: trec.nist.gov.

The World Wide Web Consortium (1998). Available: www.w3c.org.. Accessed 20 Aug 2002.

Voorhees, E. (2001). Evaluation by highly relevant documents. In *Proceedings of SIGIR'01*, New Orleans, LA, pp. 74–82.

Westerveld, T., Kraaij, W., & Hiemstra, D. (2001). Retrieving web pages using content, links, urls and anchors. In *Proceedings of TREC-2001*, Gaithersburg, MD, NIST, pages 663–672. Available: trec.nist.gov.