Using Anchor Text for Homepage and Topic Distillation Search Tasks

Mingfang Wu
School of Computer Science and Information Technology, RMIT University, Melbourne, Australia. E-mail: mingfang@gmail.com, mingfang.wu@rmit.edu.au

David Hawking
Funnelback Pty Ltd, Australia and Research School of Computer Science, Australian National University. E-mail: david.hawking@funnelback.com

Andrew Turpin
Department of Computer Science and Software Engineering, University of Melbourne, Australia. E-mail: aturpin@unimelb.edu.au

Falk Scholer
School of Computer Science and Information Technology, RMIT University, Melbourne, Australia. E-mail: falk.scholer@rmit.edu.au

Past work suggests that anchor text is a good source of evidence that can be used to improve web searching. Two approaches for making use of this evidence include fusing search results from an anchor text representation and the original text representation based on a document's relevance score or rank position, and combining term frequency from both representations during the retrieval process. Although these approaches have each been tested and compared against baselines, different evaluations have used different baselines; no consistent work enables rigorous cross-comparison between these methods. The purpose of this work is threefold. First, we survey existing fusion methods of using anchor text in search. Second, we compare these methods with common testbeds and web search tasks, with the aim of identifying the most effective fusion method. Third, we try to correlate search performance with the characteristics of a test collection. Our experimental results show that the best performing method in each category can significantly improve search results over a common baseline. However, there is no single technique that consistently outperforms competing approaches across different collections and search tasks.

Received September 10, 2011; revised December 30, 2011; accepted January 4, 2012

 DOI: 10.1002/asi.22639

Introduction

Pages on the World Wide Web are connected to each other using hyperlinks. This mechanism provides a rich link structure, so that when a web search engine retrieves and ranks web pages that are possibly relevant to a user's query or information need, the search engine can use not only internal evidence from a page itself—for example, word frequency—but also external evidence from other web pages that are connected to the web page through hyperlinks. Retrieval approaches that leverage this extra evidence include the famous PageRank algorithm (Brin & Page, 1998; Page, Brin, Motwani, & Winograd, 1998) and HITS (Kleinberg, 1999), where the credibility of a page is estimated based on the page's indegree (the number of incoming links) and outdegree (the number of outgoing links). Other methods take account of incoming anchor text—the text people click on to follow a link in a browser.

External evidence such as the number of incoming or outgoing links and PageRank are query independent. They have been shown to be useful in the task of homepage finding (Upstill, Craswell, & Hawking, 2003). Other external evidence, such as outgoing links that point to retrieved documents, is query dependent; this type of evidence has been shown to be useful for topic distillation search tasks (Kleinberg, 1999; Wu, Scholer, & Turpin, 2011).

In this article, we focus on one particular piece of external evidence—anchor text. Anchor text is the descriptive textual information that the author of a web page supplies as part of an HTML anchor tag. For example, consider the following HTML fragment that defines a hyperlink: ACM digital library.

Here, the *target* of the hyperlink is the resource http://portal.acm.org. The *anchor text* associated with the link is "ACM digital library"–this is the text that is displayed in a web browser, and serves as a description of the target page for the reader of the current (or *referring*) page.

The rationale behind the use of anchor text for web search is twofold. First, the anchor text is usually an informative description of its target page. Its purpose is to provide an information scent that enables users to predict what they would get if they followed the hyperlink. When a user clicks on a hyperlink, the user would expect to land on a target page that is accurately described by anchor text. This leads to the second point, that a hyperlink acts as a recommendation or citation for the target page. Each link resembles its author's recommendation to the target page, and the more links that point to a target page, the more important that page is likely to be. Critically, different authors of these hyperlinks may use different sets of words to describe the target page, according to the context in which the target page is being referred to. The anchor text gathered from many different links that point to the same target page can therefore provide a rich semantic description of what the target page is about.

The use of anchor text has been shown to be effective for web search as initially shown by Page et al. (1998) and McBryan (1994). Anchor text has also been widely tested for homepage finding, named page finding, topic distillation, and ad hoc search tasks. Experiments with the TREC web track data have concluded that using anchor text on its own can lead to substantial performance gains (Craswell & Hawking, 2004; Craswell, Hawking, & Robertson, 2001; Davison 2000; Koolen & Kamps, 2010). Other approaches have incorporated anchor text with other evidence to improve search quality, for example, by combining search results from an anchor text representation and from an original text collection based on a document relevance score or a document's rank position in a search results list, and combining term frequencies from both anchor text and full text collections at retrieval time. However, these methods were tested under different testbeds and compared to different baselines. As a result, it is difficult to determine which method works better in a given situation. The aims and contributions of this article are threefold: to review and summarize existing representative methods of how anchor text can be used to improve search quality; to evaluate these methods using a common set of testbeds and search tasks, and to relate search performance to the characteristics of a testbed.

The remainder of the article is organized as follows. In the following section, we review existing methods that make use of anchor text. Then we introduce our experimental setting and present our results. We discuss our findings and conclude the study in the remaining sections.

Related Work

Data Fusion

Data fusion entails a range of techniques that combine evidence from multiple sources to improve the effectiveness of information retrieval systems. The sources used by these techniques include different query representations, different document representations, different system representations, and combinations of such.

In an information search, there are three key components that decide the final set of retrieved documents: a query representation, document representations, and a search system that maps a query representation to document representations to produce a result list. Thus, the sources used by fusion techniques include (a) multiple query representations, but with a single retrieval system and single collection (Belkin, Cool, Croft, & Callan, 1993); (b) multiple information sources or collections with a single retrieval system, such as in the context of distributed information retrieval (Larkey, Connell, & Callan, 2000); and (c) multiple search systems with the same collection (Aslam & Montague, 2001; Fox & Shaw, 1993; Lee 1997). Data fusion is also used in situations such as metasearch, where there are multiple information resources or collections, each with their own retrieval systems (Hawking & Thomas 2005).

There are two theories which support the data fusion approach: cognitive poly-representation and inference theory. From a cognitive perspective (Ingwersen, 1996; Larsen, Ingwersen, & Lund, 2009), an information-seeking process involves multiple actors who engage and interact with each other. These actors may include users, authors, and system designers; each group forms their own representation of information objects, and understandings of information systems. For example, users may have different query representations and apply different search strategies with different retrieval systems; authors may present a document differently and give opinions on documents authored by others (for example, through anchor text in a web context, or citations in the context of academic articles), and system designers design a retrieval system according to their own understanding of information objects, search problems in the real world, and the relationship between information objects and a certain search problem. Ingwersen & Järvelin (2005, p.208) argued that "the more interpretation of different cognitive and functional nature, based on information seeking and retrieval situation, that points to a set of objects overlaps, and the more interplay in how they do so, the higher probability that such objects are relevant to a perceived work task/interest to be solved."

Alternatively, information retrieval (IR) can be considered as a problem of relevance inference (Turtle & Croft, 1991): Given an information need and a collection of documents, an IR system infers or estimates the relevance probability that a document satisfies the information need. From this point of view, the greater the number of query formulations and document representations, the more sources of

DOI: 10.1002/asi

evidence an IR system can use to infer the relevance probability in its evidential inference process.

Multiple query representations. In their study of information search behavior of an operational online IR system, Saracevic and Kantor (1988) observed that, given the same search question statement, searchers tend to interpret the question differently; as a result, they formulate different queries and retrieve different set of documents. Empirical studies have shown that combined search results from different Boolean queries (Belkin et al., 1993; Belkin, Kantor, Fox, & Shaw, 1995) or from different query types (such as Boolean queries and natural language queries; Turtle & Croft, 1991) achieved higher precision than any individual search result did. Belkin et al. (1993) summarized a good rule of thumb on combining search results from different query representations: the more, the better.

Kelly, Dollu, and Fu (2005) and Kelly and Fu (2007) investigated the effect of fusing different pieces of information from the same user on retrieval performance (for example, statements of why the user wants to know about the topic, and a description of what they already know about the topic). Their experimental results demonstrated that search effectiveness could be significantly improved by making use of such user responses, above and beyond an initial keyword-based query or automatic query expansion.

Multiple document representations. A document's author uses various strategies to structure and write the document to effectively convey the underlying messages to potential readers. A good title, abstract, and headings attract a reader's attention and deliver the gist of the document, whereas citations and references to other authoritative documents can give readers a sense of reliability. If the document is checked into a large document repository, the document may be assigned metadata for archiving and easy retrieving in the future. All of this intra- and interinformation associated with a document provides a rich human cognitive representation of the document, and richer evidence for the document to be retrieved (Ingwersen 1996). Studies show that, given the same query representation, if there are a substantial number of different document representations that all point toward a document, it is more likely that the document is highly relevant to the query; combining evidence from multiple document representation increases search performance in terms of precision (Skov, Larsen, & Ingwersen, 2008).

Multiple system representations. Search models map a query representation to a document representation to retrieve a set of documents that are likely to be of interest to users. Well-known examples of these models include the Boolean search model, the probabilistic model, the vector space model, and language models. In the probabilistic and vector space models, it is also possible to vary aspects of the models, for example, by applying different term weighting schemes. Each variation of a model and its aspects will result in a different set of documents being retrieved. Larsen et al. (2009) combined search results from the four best-performing retrieval models that participated in the TREC 5 ad hoc task, while Lee (1997) selected six sets of search results from different retrieval systems of the TREC 3 ad hoc track participants. Both studies reported significant improvement in fused results. In particular, Larsen et al. (2009, p.648) observed that "fusions of algorithmically dissimilar models outperform the single constituents and fusions made from similar systems/models." However, unlike the conclusion of Belkin et al. (1993), "the more the better," it was found that more system representations are not always better here. Lund, Schneider, and Ingwersen (2006) and Larsen et al. (2009) observed that the success of fusing search results from different system representations depends on three factors: (a) the degree of dissimilarity between the constituent IR models, (b) how equally the component models perform, and (c) how well the component models perform. On the other hand, Beitzel et al. (2004) experimented with a single information retrieval engine, but with different search strategies to calculate query-document similarity (one vector space model and two probabilistic models) while keeping other system properties (parsers, stemmers, phrase lists and stop words, etc.) the same. Their experimental results indicated that fusing highly effective retrieval strategies (with a single retrieval system) did not guarantee effectiveness improvements, and that the difference between relevant and nonrelevant overlap of component result sets was a poor indicator of the effectiveness of fusion in their testing context.

Data Fusion Methods

Although a wide range of data fusion methods have been proposed, these can be categorized into two broad types: search result fusion and evidence fusion. Search result fusion is carried out after a search to combine several sets of search results arising from different query representations, document representations, or retrieval models. In contrast, evidence fusion is carried out during the search process to combine evidence from different resources during matching and ranking, so that only a single search result list is generated. Thus, evidence fusion methods are usually applied when there are multiple document collections and a single search engine.

Search result fusion. Search result fusion methods use two types of information: a document's relevance score, and the document's rank position.

Fusion-based on document relevance score. Given a query, an information retrieval system typically finds those documents matching at least one query term. If a system adopts a Boolean retrieval model, the system returns a set of formally matching documents, otherwise the system infers a relevance score (RS) for each document and ranks documents in decreasing order of their relevance score. The relevance score represents either the likelihood or probability

DOI: 10.1002/asi

of a document being relevant to the query, or the similarity of the query and document vectors. To combine several candidate lists of search results into one ranked list, a fusion method first finds the union of documents from all candidate lists, calculates a new relevance score for each document in the union list, and then ranks documents accordingly. The two most widely used approaches to calculating the new relevance score are: linear interpolation and a suite of Fox and Shaw's fusion methods (Fox & Shaw, 1993).

Linear interpolation. For any document d_i from candidate lists, its new relevance score ($RS_{new}(d_i)$) in the merged list is simply a linear combination of its original RS from the candidate lists (d_{i1}, d_{i2}, . . . , d_{in}):

$$RS_{new}(d_i) = \alpha_1 * RS(d_{i1}) + \alpha_2 * RS(d_{i2}) + \ldots + \alpha_n RS(d_{in}) \quad (1)$$

where $\alpha_1 + \alpha_2 + \ldots + \alpha_n = 1$.

Parameters $\alpha_j (j = 1,2, \ldots, n)$ adjust the weight given to each candidate list. The relevance score of a document is treated as 0 in any list in which it does not occur.

This linear combination method has shown varying degrees of success (Bartell, Cottrell, & Belew, 2005; Westerveld, Kraaij, & Hiemstra, 2001). Based on empirical and mathematical analysis of linear combination of any two runs from 61 submitted runs (which can be regarded as 61 systems) of the TRECC% adhoc track, Vogt and Cottrell (1999) concluded that a combination would be effective when (a) each single run exhibits good performance, (b) both runs return similar sets of relevant documents, and (c) both return dissimilar sets of irrelevant documents.

Fox and Shaw Fusion methods. This range of methods was first proposed and tested by Fox and Shaw (1993) in their TREC 2 work; we will refer to these methods as "FoxShaw" in this article. The FoxShaw fusion methods come in the following five basic forms:

1. CombMIN: A document's final relevance score is the minimum of its individual relevance scores. This method would minimize the probability that an irrelevant document is highly ranked.
2. CombMAX: A document's final relevance score is the maximum of its individual relevance scores. This method would minimize the probability that a relevant document would receive a low rank.
3. CombSUM: A document's final relevance score is the summation of its individual relevance scores if the document appears in more than one list. Note that a new list generated by the CombSUM method would be equivalent to linear interpolation above with α_i set to 1.
4. CombANZ: A document's final relevance score is its score as calculated by CombSUM *divided* by the number of nonzero relevance scores. This method would punish a document that occurs in multiple lists but all with low scores.
5. CombMNZ: A document's final relevance score is its score as calculated by CombSUM multiplied by the number of nonzero relevance scores. CombMNZ greatly promotes a document that is retrieved in multiple lists.

Unlike the linear interpolation method, which requires training of a set of parameters α_i that would be optimized for the similar type of queries and with a specific collection, the FoxShaw fusion methods combine sets of search results on a query by query basis without any parameter training.

Score Normalization. When several lists of search results are fused based on the relevance score of a document, the original relevance score, as determined by the retrieval system, can be used directly. Alternatively, it can be advantageous to normalize a document's relevance score when the document relevance score from each collection or retrieval system is on a different scale. The three most widely used score normalization methods are

1. *Linear normalization.* The relevance score (RS) of a retrieved document d is normalized by the maximum range of the scores:

$$NRS_d = (RS_d - Min)/(Max - Min)) \quad (2)$$

where *Max* and *Min* represent the maximum and minimum document relevance score of a list. For a ranked list, these are the relevance scores of the first and last documents, respectively. After this normalization, the document's relevance score is constrained within the range [0,1], with the first ranked document having a score of 1 and the last having a score of 0.

2. *Score transformation before linear normaliZation.* Before applying the above linear normalization method, a document score is first transformed by e^{RS_d}, because sometimes logarithms are involved when a document's relevance score is calculated in the first place.
3. *Normalization by maximum relevance score.* Arguably, the above linear normalization is not optimal because for queries where the document representation evidence is not strong, the highest document relevance score will still score 1.0. An alternative approach is therefore to normalize the relevance scores in a search result within the range [0, *MAXPOSS*], where *MAXPOSS* is the maximum possible relevance score of a hypothetical document. For example, in Okapi BM25, *MAXPOSS* corresponds to a zero-length document containing an "infinite" number of occurrences of each query term (Upstill et al., 2003). Thus, if we have different representations of documents, we will have different MAXPOSS values for each collection of document representations.

Fusion based on rank. This approach to aggregation uses the rank position of a retrieved document instead of its relevance score. The family of methods originated from a model of social choice voting (Roberts, 1976). Two representative methods are the positional method and the majority voting method, and these have been instantiated as the Borda Fuse and Condorcet approaches, respectively (Aslam & Montague, 2001; Dwork, Kumar, Naor, & Sivakumar, 2001; Montague & Aslam, 2002).

DOI: 10.1002/asi

Borda Fuse method. This approach uses relative rank position: Each document from a list receives a point that is the reverse to its rank. For example, if a list has n documents, then the first document receives n points, the second one $n-1$ points, and so on. A document's point in the final list can be its accumulated points from all lists, or can be obtained through linear interpolation of the document's points. In the latter case, some lists can be weighted higher than others. The Borda Fuse method rewards documents that have a high rank. For example, consider two lists, each consisting of 1,000 documents. Document a occurs in only one list, at position 1, and so gets a score 1,000. Another document, b, occurs in two lists but at position 500 in one list, and 600 in the other, thus achieving a score 900. Overall, document b will be ranked lower than document a in the new merged list.

A variation of the Borda Fuse method is to use the reciprocal rank (or the harmonic mean), in which a document gets a score of one over its rank, that is, documents from a ranked list of n documents get a score of 1,1/2,1/3, . . . ,1/n. With this method, the importance of lower ranked documents drops dramatically.

Condorcet method. This approach takes the relative order of two documents into account. If a document a is ranked higher than a document b and this order is observed in the majority of lists, then document a will be ranked higher than document b. Thus, it may be inappropriate to apply this method if there are only two lists to fuse, or if the lists being merged overlap only negligibly.

Term frequency merging. In the case where a document has multiple representations, we can regard each document representation as a "field" (or an element of a XML mark-up) of the document. In this case, the above postretrieval relevance score or rank combination approaches resemble field search in structured document retrieval (Myaeng, Jang, Kim, & Zhoo, 1998; Piwowarski & Gallinari, 2003; Wilkinson, 1994). Robertson, Zaragoza, and Taylor (2004) argued that the postretrieval score combination approach could lead to dangerous overestimation of the importance of the term, breaking nonlinear features of term frequency, and causing unstable behavior of document frequency; this can occur because fields may be very different from each other in their length and importance to a query.

One way to overcome the problem with postsearch fusion methods is to combine term frequency from different fields (or document representations) during the retrieval stage, thus building a unified document representation model that allows the generation of a single retrieval list.

For example, the typical term weighting function in the Okapi BM25 model (Robertson, Walker, Hancock-Beaulieu, & Gatford, 1994) is

$$w_t = \frac{(k_1+1)*tf_t}{k_1\left((1-b)+b\dfrac{dl}{avdl}\right)+tf_t} * log\frac{N-df_t+0.5}{df_t+0.5} \quad (3)$$

where N is the total number of documents in a collection, df_t is the document frequency of term t, dl is the document length, $avdl$ is the average document length across the collection, tf_t is term frequency, and $k_1$and b are term weight normalization parameters. A document weight is then obtained by summing the weights of those query terms contained in the document:

$$w(d,q) = \sum_{t\in Q} w_t \quad (4)$$

Robertson et al. (2004) proposed that the term frequency should be calculated as a linear combination of the frequencies of the term in each field. Thus, in Equation (3), tf_t is replaced with:

$$tf_t = \sum_f \alpha_f * tf_{t,f} \quad (5)$$

where α_f is the weight given to each field f, with its value usually set within the range [0,1], and $\sum_f \alpha_f = 1$. The document frequency of a term and document length are calculated as if all fields are merged together. In the same article, Robertson et al. compared this term frequency combination method with a scoring method that uses the linear combination of document weights. Their experimental results showed that the term frequency combination method outperformed the document score combination method for the the Reuters volume I collection (50 topics from the TREC 2002 filtering track) and the .GOV collection (50 topics from 2002 Web track topic distillation task).

The above simple linear combination of term frequency from each field does not take into account that different fields may be of different lengths. In particular, this can be a problem when lengths vary substantially across fields. For example, a term that occurs in a long field five times should not get the same weight as a term that has the same frequency of occurrence in a short field. In this case, the term frequency can be normalized by the field length before it is combined with its frequency from other fields (Zaragoza, Craswell, Taylor, Saria, & Robertson, 2004):

$$\tilde{tf}_t = \sum_f \alpha_f * \frac{tf_{t,f}}{(1-b_f)+b_f\dfrac{dl}{avdl}} \quad (6)$$

Accordingly, a combined term weight is

$$w_t := \frac{\tilde{tf}_t}{k_1+\tilde{tf}_t} * log\frac{N-df_t+0.5}{df_t+0.5} \quad (7)$$

Similarly, in a language model (Ponte & Croft, 1998; Zhai & Lafferty, 2001) documents are ranked by their probability of generating a given query. In a unigram model, a

DOI: 10.1002/asi

document's probability is computed by taking the product over all query terms of the probability of the query term given the document language model θ_d:

$$P(Q \mid \theta_d) = \prod_{i=1}^{|Q|} P(q_i \mid \theta_d) \quad (8)$$

where typically

$$P(q_i \mid \theta_d) = \lambda P(q_i \mid d) + (1-\lambda) P(q_i \mid C) \quad (9)$$

The above language model can be modified to incorporate the linear interpolation of evidence from each document field at the term level (Kraaij, Westerveld, & Hiemstra, 2002; Ogilvie & Callan, 2003a):

$$P(q_i \mid \theta_d) = \sum_f \alpha_f P(q_i \mid \theta_d(f)) \quad (10)$$

where $d(f)$ is the representation of the f^{th} field, α_f is the weight given to the field, and the sum of all α_f scores equals 1.

Experiments by Ogilvie and Callan (2003a) showed that the mixture-based language model gave better results than rank-based fusion algorithms for the homepage finding task from TREC10 and the named page finding task from TREC11.

Anchor Text

Anchor text has been studied and exploited for various purposes, such as document summarization (Amitay & Paris, 2000), query refinement (Kraft & Zien, 2004), and most of all, as a source of evidence to improve search quality. In this section, we review two aspects of anchor text: the analysis of relationships between anchor text, queries and documents, and the actual use of anchor text in improving search quality.

Study of anchor text characteristics. Two important studies by Davison (2000) and Eiron and McCurley (2003) provide some insight into why anchor text could be helpful for some search tasks (and not be helpful for some other tasks).

Davison (2000) studied the topic locality of connected web pages and found that web pages were significantly more likely to be topically related to web pages to which they are linked, as opposed to other randomly selected pages. Similarly, anchor text is significantly similar to the page it references, and less similar to random pages. This suggests that anchor text could be useful in representing its target web page.

Eiron and McCurley (2003) examined relationships among queries, anchor text, and web pages from the IBM intranet. They found that, compared to a page title, the anchor text better resembles queries sent to the intranet search engine, as the anchor text term set was more diverse than the title term set. They also observed that terms that topically occurred in queries were more likely to be repeated in the content than the average anchor text term. Thus searching on anchor text may result in more focused search results than searching on documents, but may give zero results to some queries.

Use of anchor text in search. The approaches for using anchor text in search can be summarized as follows: (a) using anchor text as sole evidence, (b) propagating a document's anchor text to its target document, (c) treating anchor text as another representation of its target page and merging search results from the anchor text collection and original document collection, and (d) building a virtual structured document that treats its aggregated anchor text and the original text as two fields.

Anchor text as sole evidence. For those pages that have incoming links, the anchor text associated with the links are harvested and aggregated, and treated as their target page's representation. Indexing and search are then applied to this anchor-text collection. (We call this collection the *anchor-text* collection, and the original document collection the *original-text* collection.) It has been reported that the use of an anchor-text collection alone can effectively improve results for a home page finding task using either the Okapi retrieval model (Craswell et al., 2001), or language models (Fujii, 2008). However, Chowdhury, Aljlayl, Jensen, and Beitzel (2002) reported that searching original-text significantly outperformed searching only anchor-text for a named page finding task.

Anchor text propagated to its target document. Broder et al. (2000) reported that the distribution of indegree follows a power law: a small number of pages have high indegree, thus are associated with a large amount of anchor text; a large number of pages have low indegree, and thus are associated with little or no anchor text. As a result, searching an anchor-text collection alone would not retrieve pages with zero indegree and may fail to retrieve pages with low indegree. A way to overcome this anchor text sparsity problem is to enrich the anchor text collection by propagating anchor text through a neighborhood link graph (Metzler, Novak, Cui, & Reddy, 2009). We will refer the collection in which a document is represented by its aggregated anchor text as well as its original text as an *extended-text* collection. Searching on an extended-text collection would increase weighting of those terms appearing in anchor text, meanwhile avoids the above sparsity problem.

Anchor text as an alternative representation. Here, the anchor-text and original-text collections are treated as two independent, alternative representations. A query is sent to both representations, search results from each representation are then merged by using a search result fusing method. Sometimes, other external evidence such as a document's indegree and outdegree or URL features are also combined during fusion process (Westerveld et al., 2001).

DOI: 10.1002/asi

Anchor text as a field of its target page. The original-text and anchor-text of a page can be treated as two fields of a super structured document, as follows:

```
<doc>
    <anchor>
        Anchor text from incoming links:
    </anchor>
    <content>
        Original page text:
    </content>
</doc>
```

Field search methods, such as in Equations (7) or (10), can then be applied to incorporate evidence from anchor text during the searching and ranking process.

Anchor text at TREC. The use of anchor text as a source of relevance evidence of its target document was first proposed by Brin and Page (1998) and Chakrabarti et al. (1998). However, with the availability of web document collections at TREC, different ways of exploiting anchor text to improve search quality has been widely explored. TREC had a web track from 1997 to 2003 with the test collections VLC2 (Very Large Collection or WT100G), WT2G, .GOV and .GOV2 (Hawking & Craswell, 2005) give a detailed description of each collection); the track was restarted in 2009 with the ClueWeb collection (Clarke, Craswell, & Soboroff, 2009).

A range of fusion methods have been tested by TREC web track participants to merge search results from searches on original-text representations, anchor-text representations, and other document representations that include specific text extracted from fields in the original document itself, such as <*title*> or <*h2*> (Beitzel et al., 2003; Lu, Hu, & Ma, 2004; Ogilvie & Callan, 2003a, 2003b; Zhou et al., 2004). Lu et al. (2004) found that, for topic distillation queries, linear interpolation produced higher mean average precision than CombMNZ and CombMAX; for homepage finding queries, the performances of CombMNZ and the linear combination were close to each other, as measured by mean reciprocal rank (MRR) and success rate at 10 (S@10)). However, the CombMAX method outperformed CombMNZ and the linear combination method (as measured by S@10). Zhou et al. (2004) also showed that the linear combination method outperformed CombMNZ for mixed search queries. Robertson, Zaragoza & Taylor (2004) compared term frequency merging with linear combination using evidence from anchor text, and reported a significant improvement from term frequency merging for the .GOV-03 topic distillation task.

From 2009, the TREC web track has used the ClueWeb collection, a 25 TB crawl of pages from the web (Clarke et al., 2009). There have been no named page or homepage finding tasks that use this collection to date; therefore, we are unable to report results on this data set, the importance of anchor text for an ad hoc search has been considered in the context of this larger data crawl. Koolen and Kamps (2010) demonstrated that, unlike for smaller TREC collections, an ad hoc search on ClueWeb can benefit strongly from anchor text. This source of evidence led to significantly higher precision in the top 30 rank positions of search results, and a linear mixture model combining full and anchor text evidence significantly outperformed a full text baseline. Moreover, their results showed that the size of a collection can have a substantial impact on the benefits of anchor text, with larger collections showing more substantial increases from this source of evidence. It will be interesting to see if these trends hold for topic distillation tasks, once queries and relevance judgements become available for this type of search task on the ClueWeb collection.

Experimental Framework

As discussed in the second section, a range of fusion methods exploit relevance evidence from anchor text to improve search quality. Previous studies used their own testbeds, baselines, and anchor-text collections—usually compared a fusion method with one baseline at a time. As a result, it is not possible to compare between the methods. The main aim of this study is to compare these fusion methods with the same testbeds and baselines. This section introduces our testbeds, search system, and the fusion methods under comparison.

Testbeds

A testbed includes a set of queries, a collection of documents to search, and relevance assessments that indicate which documents in the collection are relevant to which query. We selected four testbeds from TREC Web tracks, as the majority of the fusion methods were proposed for web searching environments.

Document collections. Our testbeds involve three document collections: an enterprise collection and two web collections.

CERC (CSIRO Enterprise Research Collection). The CERC corpus is a crawl of all public webpages in the .csiro.au domain of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) as they appeared in March 2007. The collection was used for the TREC 2007 and 2008 enterprise tracks, for document search and expert search tasks. The document search task models the problem of finding a set of "key pages" about the topic of interest. The task is similar to a topic distillation task.

.GOV. The .GOV collection is a crawl of websites in the .gov domain from early 2002 (Craswell & Hawking, 2002). This collection was used for the TREC 2002, 2003, and 2004 web track for homepage finding, named page finding, and topic distillation search tasks.

WT10G. The 10 gigabyte WT10G collection is a snapshot of the web in 1997. It was engineered to keep desirable properties of web document collections such as a realistic

DOI: 10.1002/asi

TABLE 1. Properties of three document collections.

	CERC	.GOV	WT10G
Number of pages	370,715	1,247,753	1,692,096
Total number of links	4,670,439	11,164,829	8,063,026
Number of links having anchor text	4,609,687	8,653,291	7,208,776
% of pages with anchor text	78%	55%	87%
Number of documents with a title	323,604 (87%)	984,558 (79%)	1,609,571 (95%)
Average number of words per anchor text	2.15	2.90	2.77

proportion of hyperlinks between pages (Bailey, Craswell, & Hawking, 2003) This collection was used in the TREC 2000 and 2001 web tracks for ad hoc and homepage finding tasks.

These three collections represent the web at different scales; the properties of each collection are shown in Table 1. The three collections have similar percentages of documents that have incoming anchor text. However, the WT10G collection has 8% more documents with a *title* than the CERC collection, which itself has 8% more documents with a title than the .GOV collection. Similar to web queries, an anchor text entry has an average of two to three words. The .GOV collection has slightly more terms from anchor text than the WT10G collection; the CERC collection has the fewest number of terms from anchor text.

Anchor text gathering. To construct the anchor text document representations, the following steps were carried out. First, we processed a collection, one document at a time. In each document, when an HTML anchor tag was encountered, the source URL, target URL, and anchor text were stored in a lookup table. That is, only within-collection links were identified. Second, the URLs in the table were normalized, by reconciling different forms and redirected links. Common default pages were normalized, for example, the URL http://www.csiro.au and http://www.csiro.au/index.html are regarded as the same. Relative URLs were expanded to the full base URL. Third, for each target URL, the anchor text from all source URLs that point to the current target page was concatenated, forming the anchor text representation of that target document. In this step, repeated anchor text entries were preserved.

Search tasks and measures. In this study, we use two representative tasks from the TREC Web track (Hawking & Craswell, 2005): homepage finding and topic distillation.

Homepage finding task. Here the goal is to find the entry page of a website. For example, if a user's query is "California state parks home," it is very likely that the user is looking for the entry page, http://www.parks.ca.gov/. The search quality of this task is typically measured by mean reciprocal rank (MRR) and the success rate at certain cutoff levels. MRR is the reciprocal of the rank of the first relevant entry page in a search results list. Success rate at a certain cutoff (S@n) is the proportion of queries for which a relevant answer is ranked above the cutoff n. For example, if a search engine has a relevant answer appearing at rank 1 for 20 out of 100 queries, then the search engine's success rate at top 1 (S@1) is 0.2. In cases where a website's entry page may have more than one relevant answer, only the first answer is counted. This task was used in the TREC web track from 2001 to 2004 (Craswell & Hawking, 2002, 2004; Craswell, Hawking, Wilkinson, & Wu, 2003; Hawking & Craswell, 2001).

Topic distillation task. This task requires the retrieval of a list of key resource pages, given a broad query, instead of a specific site as in a homepage finding task. A key resource page is a good authority page on the given query and would point users to relevant documents within a site (Craswell et al., 2003). The task was tested from 2002 to 2004 with the .GOV collection (Craswell & Hawking, 2002, 2004; Craswell et al., 2003), and in 2007 with the CSIRO test collection (Bailey, Craswell, Soboroff, & de Vries, 2007).

The performance of systems on this task is measured by mean precision at 10 documents retrieved (P@10); mean R-precision (RP), the precision after R document are retrieved where R is the number of relevant documents for the topic and mean average precision (MAP). Average precision for a single query (AP) is calculated by averaging precision scores at each point in the search results list where a relevant document is found. MAP is then the mean of the AP scores over a set of queries.

Table 2 shows the number of topics used per task and per collection. In total, we consider six testbeds, three for each search task. The table also shows which fields from the topic descriptions were used as queries in our study.

Relevance assessments. For each query in a testbed, there are a small subset of documents from the collection which were assessed as being relevant or irrelevant to the query. If a document is retrieved in response to a query but the document was not assessed, the document is regarded as irrelevant. For the testbeds using the .GOV-03, .GOV-04 and WT10G collections, the relevance assessments are binary. For the CERC testbed, the relevance assessment has three grades: not a key page, possibly a key page, and highly likely to be a key page (Bailey et al., 2007). In our study, we folded the three grades into binary, a document with either of the first two assessments is regarded as "not a key page."

The CERC testbed also has two types of assessment standards: "gold" standard and "bronze" standard. Gold standard relevance assessments were made by CSIRO science communicators who are familiar with CSIRO business, the CERC document collection, and the topic distillation task; for bronze judgments, the assessments were made by groups who participated in the track in 2007, who may understand the topic distillation task, but are unlikely to have been familiar with the document collection. Bailey et al.

DOI: 10.1002/asi

TABLE 2. Topic sets used in our experiments. Numbers in brackets are the number of topics in the corresponding testbed. Words from fields shown in italics are used as a query.

Collection	Field	Example	Average query terms
Homepage Finding			
.GOV-03 (150) and	*Description*	California state parks home	4.19
WT10G (145)			3.08
.GOV-04 (75)	*Title*	Food and Drug Administration	3.63
Topic distillation			
.GOV-03 (50)	*Title*	Wireless Communications	2.26
	Description	Information on existing and planned uses, research/technology, regulations and legislative interest	
.GOV-04 (75)	*Title*	American Music	2.43
CERC (50)	*Query*	Sustainable Agriculture	2.38
	Narrative	General information about CSIRO's activities in sustainable agriculture with links to more specific research areas	

(2008) found that there was a low agreement between the gold standard and bronze standard, and that the difference between the two standards is substantial enough to affect system evaluation. Thus, we use the gold standard in our evaluation for this testbed.

Search Models

Indexing and searching were carried out using the Lemur toolkit (http://www.lemurproject/org). As shown in Table 2, the topic sets from each collection are composed of one or more different field names; the words from the field indicated in italics were used as a bag or words search query. We tested both the Okapi BM25 model and a language model to retrieve initial search results. Each merging method if then executed. We found similar results and trends regardless of which model was used for initial retrieval. We therefore report experimental results only for the Okapi BM25 model. However, we note that the language modeling approach offers an alternative, theoretically clean approach for the incorporation of different sources of evidence, such as combining the original and anchor text representations of documents, using prior probabilities of relevance. Many of the observations in this article support individual findings from the language modeling paradigm, see, for example, Kraaij et al. (2002), Ogilvie and Callan (2003a, 2003b), and Kamps (2005).

During the indexing process, 418 words appearing in the standard INQUERY stoplist were removed (Allan, Callan, Feng, & Malin, 1999), and the remaining words were stemmed using the Porter stemmer. When we apply the Okapi BM25 model (Robertson et al., 1994), we use Equations (3) and (4) for term weights and accumulated document weights, respectively. The term weighting in Formula (3) involves training two parameters k_1 and b for each testbed, where k_1 controls the nonlinear *tf* effect and b controls document length normalization.

For a fielded search, where we treat anchor text and the original document text as two fields of a super document and apply the term frequency combination method [BM25F; see Equations (6) and (7)], we followed the method presented by Robertson, Zaragoza & Taylor (2004) to train five parameters. First, we trained b_f and k_{1f} for each field, then k_1 [Equation (7)] using the previously trained b_f and setting the two field combination weights (α_f) to 1, and finally the two field combination weights α_f.

TABLE 3. Symbols for document representations, merging and normalization methods.

Document representation	Symbol
Original document text	original-text or C
Anchor text	anchor-text or AT
Anchor text plus original document text	extended-text or CAT
Fusion method	**Symbol**
Linear interpolation	Linear
Similarity score-based fusion	CombMAX, CombMIN, CombANZ and CombMNZ
Rank-based fusion	Borda, Reciprocal
Term frequency merging	BM25F
Score normalization method	**Symbol**
No normalization	NNorm
Linear normalization	LNorm
Exponential score transfer and linear normalization	ELNorm
Normalization by maximum score	MNorm

For search result merging, we compare linear interpolation, four fusion methods, and two rank aggregation methods. For each score-based merging method, we compare four methods of score normalization, namely: no normalization (NNorm), linear normalization (LNorm), exponential transfer and linear normalization (ELNorm), and normalization by maximum possible document weight (MNorm). Table 3 lists the symbols used to represent the different fusion and normalization methods.

DOI: 10.1002/asi

TABLE 4. Jaccard similarity between anchor text items and their target pages, averaged across all pairs of items.

	CERC	.GOV	WT10G
Between anchor text entries	0.73	0.68	0.60
Between anchor text entries and titles	0.09	0.31	0.35
Between anchor text entries and their target page	0.01	0.02	0.02

Characteristics of Anchor Text From Each Collection

If a page has multiple anchor text entries, these anchor text entries may come from different referring pages that may be written by different authors; each author may use their own vocabulary to describe the target page. We use the Jaccard similarity coefficient to estimate the vocabulary diversity among multiple anchor text entries of the page. The greater the extent to which the anchor text entries of a page consist of similar words, the closer the Jaccard coefficient score is to 1.

It is also interesting to compare, given a target page, if the authors of anchor text and the author of the target page use different vocabulary to describe the page. A page's author usually summarizes the page in the page's title. If all anchor text entries are very similar to the title of their target page, then we do not need to make the extra effort to harvest anchor text. We use the Jaccard coefficient to measure similarity between two text strings A and B:

$$JS = \frac{|A \cap B|}{|A \cup B|} \tag{11}$$

Table 4 shows the three measures, averaged over all documents in a collection. The numbers from the first two rows of the table show that anchor text entries are more likely to be similar to each other than to the title of their target page. For all three collections, the similarity between title and anchor text is lower than 0.5: This suggests that the two sets of text are not interchangeable. The Jaccard coefficient between anchor text entries and their target page is only 0.01 (for CERC) and 0.2 (for both .GOV and WT10G), this is because the length of an anchor text entry is generally much shorter than the length of the text of its target page. In this case, it is informative to also consider the percentage of anchor text entries that occur in their target page; this is about 0.77 (from CERC), 0.81(from .GOV), and 0.87 (from WT10G). These figures indirectly indicate that anchor text is another set of text, other than the title, that topically describes its target page. It is surprising that the anchor text entries from the CERC collection have the lowest similarity to title and overlap to documents because people from an organization could be expected to share a more common language than people from outside the organization.

Experimental Results

In this section, we describe our experimental investigation into the use of anchor text for homepage finding and topic distillation, aiming to identify the best methods to merge search results from original-text and anchor-text, and the most suitable document weight normalization method.

Homepage Search Task

Baselines. For the homepage search task, we trained the combination of b and k_1 for each collection to optimize S@5, that is, the proportion of queries where at least one of the top five retrieved pages is a good answer. Table 5 shows the results of applying the Okapi BM25 ranking model to each collection. By searching on the original original-text collection alone, S@1 is 16%, 4%, and 20% for the .GOV-03, .GOV-04, and WT10G test collections, respectively. For search on the anchor-text collection, S@1 is 58%, 47%, and 51%, giving a statistically significant improvement for each collection.[1] The rank of the first correct answer (MRR) is also improved significantly. Searching on the extended-text collection can push performance further; however, the performance difference between the anchor-text and extended-text collections is only significant for the WT10G collection.

Merging two lists using document similarity scores. Given two lists, resulting from a search on an anchor-text representation and an original-text representation, respectively, we merge the lists using the four FoxShaw methods and the linear interpolation method. We also investigate the four ways of transforming document relevance scores, as presented in Table 3. The performance of merged lists is shown in Tables 6, 7, and 8.

Among the four fusion approaches, the methods CombMNZ and combMAX clearly show better performance in terms of any evaluation measure and normalization method, followed by CombANZ and CombMIN. For the .GOV-03 and the WT10G collections, CombMAX has a higher S@5 than CombMNX, except when the exponential weight transformation is applied. For the .GOV-04 collection, CombMNZ has the highest S@5 for any weight normalization method.

A similar trend can be seen using the linear combination method: For the collections .GOV-03 and WT10G, the exponential transfer has the highest S@5 and MRR; whereas for the data set .GOV-04, it has lowest S@5 and second highest MRR. Note that the retrieved documents from the original-text representation receive much less weight for the .GOV-03 data ($\alpha = 0.2$), whereas for the .GOV-04 and WT10G collections almost equal weight is assigned for the original-text and anchor-text representations.

[1]Differences in performance metrics presented in this article are tested for statistical significance using the paired Wilcoxon signed-rank test, with $p < .05$ used as a threshold for statistical significance.

DOI: 10.1002/asi

TABLE 5. The performance of three document collections for the homepage finding task with the Okapi BM25 model.

	Representation	MRR	S@1	S@5	S@10	(b, k_1)
.GOV-03	Original-text	0.2441	0.1600	0.3467	0.4600	(0.9, 1.4)
	Anchor-text	0.6790†	0.5800†	0.8067†	0.8200†	(0, 0.5)
	Extended-text	0.6901†	0.5933†	0.8067†	0.8333†	(0.7, 3.7)
.GOV-04	Original-text	0.1638	0.0400	0.3333	0.4533	(0.95, 2.85)
	Anchor-text	0.5759†	0.4667†	0.7333†	0.7733†	(0, 0.35)
	Extended-text	0.5717†	0.4800†	0.7467†	0.8267†	(0.95, 3)
WT10G	Original-text	0.3312	0.2000	0.5172	0.6069	(0.7, 1.3)
	Anchor-text	0.5776†	0.5103†	0.6552†	0.6828	(0.05, 0.5)
	Extended-text	0.6366†‡	0.5448†	0.7724†‡	0.8069†‡	(0.75, 2)

Note. The parameters b and k_1 are optimized for the best performance of S@5. † = a result that is significantly better than the corresponding original-text run; ‡ = a result that is significantly better (or worse) than the corresponding anchor-text run.

TABLE 6. The performance of fusion methods based on similarity scores with the *.GOV-03* collection for the homepage find task.

		MRR	S@1	*S@5**	S@10	α
CombMIN	NNorm	0.2636	0.1867	0.3133	0.4000	
	LNorm	0.3208	0.2067	0.4400	0.5267	
	ELNorm	0.3188	0.2067	**0.4467**	0.5000	
	MNorm	0.2947	0.2133	0.3867	0.4333	
CombMAX	NNorm	0.6590	0.5533	**0.8000**	0.8647	
	LNorm	0.5451	0.3200	**0.8000**	0.8667	
	ELNorm	0.5379	0.3200	0.7733	0.8467	
	MNorm	0.6568	0.5467	**0.8000**	0.8533	
CombANZ	NNorm	0.4404	0.3200	0.5933	0.6933	
	LNorm	0.3761	0.2333	0.5467	0.6467	
	ELNorm	0.4106	0.2400	0.6133	0.7400	
	MNorm	0.4550	0.3333	**0.6200**	0.7133	
CombMNZ	NNorm	0.6514	0.5800	0.7400	0.7733	
	LNorm	0.6078	0.5267	0.6800	0.7667	
	ELNorm	0.6641	0.5667	**0.8000**	0.8600	
	MNorm	0.6399	0.5667	0.7067	0.7867	
Linear	NNorm	0.6780	0.5800	0.8133	0.8133	0.05
	LNorm	0.6913	0.5933	0.8200	0.8200	0.05
	ELNorm	0.7092	0.6067	**0.8533**	0.8867	0.2
	MNorm	0.6945	0.6067	0.8200	0.8333	0.2

Note. Wherever a training parameter is involved, the parameter is set to maximize S@5. $\alpha$is the weight given to the original-text representation.

TABLE 7. The performance of fusion methods that use simiarity scores with the *.GOV-04* collection for the homepage finding task.

		MRR	S@1	*S@5**	S@10	α
CombMIN	NNorm	0.0564	0.0267	0.0667	0.1067	
	LNorm	0.1764	0.0667	0.2667	0.4000	
	ELNorm	0.1623	0.0533	0.2533	0.3600	
	MNorm	0.3936	0.2400	**0.5867**	0.7067	
CombMAX	NNorm	0.2979	0.1467	0.5067	0.6267	
	LNorm	0.4505	0.2533	0.6800	0.8000	
	ELNorm	0.4484	0.2533	0.6933	0.8000	
	MNorm	0.5475	0.4133	**0.7333**	0.8267	
CombANZ	NNorm	0.1022	0.0267	0.1467	0.2800	
	LNorm	0.2388	0.1067	0.3200	0.5467	
	ELNorm	0.2929	0.1067	**0.5200**	0.6933	
	MNorm	0.2020	0.0800	0.2933	0.5067	
CombMNZ	NNorm	0.6343	0.5200	0.7600	0.8400	
	LNorm	0.6073	0.4533	**0.8000**	0.8667	
	ELNorm	0.6205	0.5200	0.7733	0.8533	
	MNorm	0.6142	0.4667	0.7867	0.8267	
Linear	NNorm	0.5958	0.4400	**0.8000**	0.8400	0.6
	LNorm	0.5761	0.4133	**0.8000**	0.8267	0.55
	ELNorm	0.6081	0.4933	0.7600	0.8267	0.05
	MNorm	0.6391	0.5200	0.7867	0.8267	0.3

Note. $\alpha$is the weight given to the original-text representation.

Among the different fusion methods that are based on similarity scores, the best linear combination method achieves equal or higher S@5 than the best performing FoxShaw method, CombMNZ, for all three collections.

Rank aggregation. Results for the two rank-based merging methods, Borda and Reciprocal, are shown in Table 9. The reciprocal method shows higher effectiveness than the Borda method across all collections; however, the difference is only statistically significant for the .GOV-03 and WT10G collections. Note that for the Borda method the α value is 0 for both the .GOV-03 and the .GOV-04 collections, and close to 0 for the WT10G collection. This indicates that search results from the original document representation contribute little to the final merged results. For the reciprocal ranking method, the α values are closer to evenly balanced (0.5) for all three collections, so search results from the anchor text representation and original document representation contribute almost equally to the final merged results.

Term frequency merging. The performance of term frequency merging fusion method, BM25F, is shown in Table 10. For easy comparison, this table also reports the best-performing fusion methods from the similarity score and rank-based groups. Linear combination, reciprocal rank-based merging, and term frequency-based fusion all lead to an improvement over searching the anchor-text representation alone. In most cases, the improvements are statistically significant. However, the largest and most substantial improvement comes from the term frequency merging approach, during a search.

DOI: 10.1002/asi

TABLE 8. The performance of fusion methods based on similarity scores with the *WT10G* collection for the homepage finding task.

		MRR	S@1	*S@5**	S@10	α
CombMIN	NNorm	0.2539	0.1448	0.3862	0.4759	
	LNorm	0.3654	0.2552	**0.5034**	0.5655	
	ELNorm	0.3616	0.2552	0.4897	0.5586	
	MNorm	0.2967	0.1793	0.4621	0.5241	
CombMAX	NNorm	0.5839	0.4828	0.7172	0.7793	
	LNorm	0.5129	0.3034	**0.7310**	0.7931	
	ELNorm	0.5119	0.3034	**0.7310**	0.7931	
	MNorm	0.5414	0.4345	0.6759	0.7517	
CombANZ	NNorm	0.4728	0.3586	0.6207	0.7034	
	LNorm	0.4093	0.2759	0.5931	0.6759	
	ELNorm	0.4605	0.3034	**0.6690**	0.7655	
	MNorm	0.4577	0.3517	0.5931	0.6690	
CombMNZ	NNorm	0.5923	0.5172	0.6897	0.7241	
	LNorm	0.5704	0.4759	0.6828	0.7448	
	ELNorm	0.6177	0.5172	**0.7379**	0.8138	
	MNorm	0.5812	0.4966	0.6759	0.7517	
Linear	NNorm	0.5948	0.4759	0.7379	0.8138	0.7
	LNorm	0.5823	0.4759	0.7310	0.7931	0.5
	ELNorm	0.6177	0.5103	**0.7517**	0.8069	0.4
	MNorm	0.5737	0.4897	0.6828	0.7241	0.45

Note. $\alpha$is the weight given to the original-text representation.

TABLE 9. The performance of rank-based fusion methods for the homepage finding task.

		MRR	S@1	S@5	S@10	α
.GOV-03	Borda	0.6790	0.5800	0.8067	0.8200	0
	Reciprocal	0.6859	0.5800	0.8600†	0.8800†	0.45
.GOV-04	Borda	0.5759	0.4667	0.7333	0.7733	0
	Reciprocal	0.5904	0.4667	0.7467	0.8133	0.3
WT10G	Borda	0.5664	0.4966	0.6621	0.6759	0.05
	Reciprocal	0.6363†	0.5448†	0.7655†	0.8069†	0.35

Note. α is the weight given to the original-text representation.

For further comparison, Table 10 includes the median of the best runs from each group that participated in the corresponding TREC track, and the top run of all submitted runs. We note that it is not straightforward to directly compare between these TREC runs and our results because the TREC runs used a range of different techniques; moreover, some TREC runs used anchor text as a source of evidence, but the anchor text collections used may be different from ours, as harvesting anchor text is a complex task, and there is no single common anchor text collection available to all groups. Nevertheless, the comparison provides a reference for the performance of the fusion methods relative to other techniques in improving searching and ranking performance.

The top 2003 run used a generative language model to include structural information (anchor text, meta tags, document title and character trigrams on URLs) and URL priors (Ogilvie & Callan, 2003b). The top 2004 run linearly combined four relevance scores and a HostRank score (a PageRank-like value). The four relevance scores are searches over the fields of title, body, anchor text, and URL (Song et al., 2004). The top 2001 top run is a linear combination of document scores (resulting from a simple language model), anchor text scores, and URL priors (Westerveld et al. 2001).

From the table it can be seen that we have a good baseline for our fusion methods. Although our search results from the original-text representation has lower performance than the median TREC runs, our search results from the anchor-text representation has higher performance than the median TREC runs. In addition, for any testbed, the BM25F method we tested performs much better than the TREC median in terms of any evaluation measure, but relatively worse than the top run. Note that all three top runs used extra evidences in addition to anchor text.

Topic Distillation Search Task

Baselines. The performance of applying the Okapi BM25 retrieval model on the three document representations of original-text, anchor-text, and extended-text is shown in Table 11. Searching on the anchor-text and extended-text representations results in a substantial improvement over the original-text representation. In most of the cases, the improvements are significant. Unlike the homepage finding task, where searching on the extended-text representation gives better performance than on the anchor-text representation for all test collections, for the topic distillation task this only holds for the CERC test collection.

Merging two lists using document similarity scores. Tables 12, 13, and 14 show the results of this category of merging methods for the collections CERC, .GOV-03 and .GOV-4, respectively. As in the homepage finding task, CombMNZ and CombMAX are the two winners among the four Fox and Shaw fusion methods that use similarity scores for merging: CombMNZ is the best performer for the CERC and the .GOV-04 collections, regardless of which weight normalization scheme is applied. For the .GOV-03 collection, CombMNZ performs better than CombMAX when linear normalization or exponential transformation is applied, whereas CombMAX works better if no normalization or normalization into the maximum document weighting range is applied.

Linear combination performs better than fusion methods, for the CERC and the .GOV-03 collections, in terms of P@10 but not for the MAP measure. For the .GOV-04 collection, the linear combination method and CombMNZ work better for different normalization methods. Overall, for each normalization method, the difference between the linear combination and the best performing fusion method is small—the largest difference occurs in the .GOV-03 collection, where the linear combination is only 4.8% better than CombMNZ with the MAX normalization method.

When the linear combination method is used with the .GOV-03 collection, optimal performance is reached when α

DOI: 10.1002/asi

TABLE 10. The best runs from each type of fusion methods for the homepage finding task.

		MRR	S@1	*S@5**	S@10
.GOV-03	Anchor-text	0.6790	0.5800	0.8067	0.8200
	Linear (Exp)	0.7092† (4.4%)	0.6067	0.8533† (5.8%)	0.8867†
	Reciprocal	0.6859 (1.0%)	0.5800	0.8600† (6.6%)	0.8800†
	BM25F	0.7521† (10.8%)	0.6600†	0.8667† (7.4%)	0.9067†
	(k=11.1, w=7.7)				
Median of TREC runs (19 groups)		0.5840	0.4600	0.7533	0.8200
The top run		0.8070	0.7333	0.9000	0.9200
.GOV-04	Anchor-text	0.5759	0.4447	0.7333	0.7733
	Linear (Max)	0.6391† (11.0%)	0.5200	0.7867 (7.3%)	0.8267
	Reciprocal	0.5904 (2.5%)	0.4667	0.7467 (1.8%)	0.8133
	BM25F	0.6427† (11.6%)	0.5200	0.8267†(12.7%)	0.9067†
	(k = 3.3, w = 11.2)				
Median of TREC runs (18 groups)		0.4730	0.3600	0.6400	0.6800
The top run		0.7290	0.6530	0.8670	0.9070
WT10G	Anchor-text	0.5776	0.5103	0.6552	0.6828
	Linear (Exp)	0.6177† (6.9%)	0.5103	0.7517† (14.7%)	0.8069†
	Reciprocal	0.6363† (10.2%)	0.5448†	0.7655† (16.8%)	0.8069†
	BM25F	0.6835$^{\dagger\ddagger\perp}$ (18.3%)	0.6000†	0.8000†‡ (22.1%)	0.8621†
	(k = 2.8, w = 6.8)				
Median of TREC runs (43)		0.3982	0.3310	0.4897	0.5586
The top run		0.7745	0.7034	0.8552	0.8830

Note. † = a significant improvement over the run with anchor-text; ‡ = significant improvement over the linear interpolation method; $\perp$ = significant improvement over the reciprocal rank. Relative improvements over the anchor-text baseline run are shown in parentheses.

TABLE 11. Performance for the topic distillation task (BM25 model, with b and k_1 tuned for the best performance of P@10).

	Representation	MAP	P@5	*P@10**	R-Prec	(b, k_1)
CERC	Original-text	0.1716	0.1306	0.1082	0.1473	(0.55, 1)
	Anchor-text	0.2962†	0.2000†	0.1245	0.2966†	(0, 0.3~0.5,2.1~2.8)
	Extended-text	0.2760†	0.1878†	0.1429†	0.2760†	(0.4,3.3~4)
GOV-2003	Original-text	0.1145	0.1000	0.0980	0.1152	(0.95,0.2 ~ 0.4)
	Anchor-text	0.1442†	0.1720†	0.1300	0.1511	(0.95,(1.45,1.5,1.55,2))
	Extended-text	0.1414†	0.1400	0.1160	0.1316†	(0.9,2.3 ~ 3))
GOV-2004	Original-text	0.0881	0.1493	0.1227	0.1031	(0.95,1.2 ~ 2)
	Anchor-text	0.1121†	0.2213†	0.1933†‡	0.1599†	(0.65,2.45 ~ 2.5)
	Extended-text	0.1274†	0.2267†‡	0.1893†	0.1590†	(0.9, (2.15,2.2))

equals 0. That means that the search from the original-text representation does not contribute at all to the final merged list. For the other two collections, the α values are higher than the average (0.5), which means that there is a higher contribution from the search on the original-text representation.

Rank aggregation. When two search result lists from the anchor-text and original-text collections are merged using a document's rank information, the reciprocal rank outperforms Borda rank for the CERC and .GOV-04 collections, while the opposite holds for the .GOV-03 collection, as shown in Table 15.

Merging based on term frequency. The performance of the BM25F method is shown in Table 16. The BM25F method performs only slightly better than the similarity score or rank-based after-search merging methods for the .GOV-03 collection. For the other two collections, the BM25F method is slightly worse than the after-search merging methods.

Compared to the top runs from the corresponding TREC submitted runs, searching on anchor text directly and the three best performing merging methods from each category achieve a higher P@10 than the top run of the TREC 2003 web track. However, the opposite result holds for the TREC 2004 data. The top run of the TREC 2003 web track did not use anchor text, but instead used the document title and URL-type information (Tomlinson, 2003). The 2004 top run is from the same group who had the top run for the homepage finding task, and used the same technique of combining four relevance scores and HostRank (Song et al., 2004). For the TREC 2007 enterprise document-search track, we downloaded all submitted runs and applied gold judgments (Bailey et al., 2008). The run with highest P@10 result combined relevance scores from title and body, HostRank, and query expansion with given key pages (Duan et al., 2007).

DOI: 10.1002/asi

TABLE 12. The performance of similarity score-based fusion methods for the topic distillation task with the CERC collection.

Merging method	Normalization	MAP	P@5	*P@10**	R-prec	α
MIN	NNorm	0.1399	0.0857	0.0714	0.1045	
	LNorm	0.1301	0.1020	0.0735	0.1052	
	ELNorm	0.1285	0.1061	**0.0816**	0.1060	
	MNorm	0.1324	0.0735	0.0673	0.0880	
MAX	NNorm	0.2839	0.1837	**0.1347**	0.2711	
	LNorm	0.2538	0.1673	0.1265	0.2362	
	ELNorm	0.2413	0.1551	0.1143	0.2342	
	MNorm	0.2702	0.1837	**0.1347**	0.2599	
ANZ	NNorm	0.2521	0.1469	**0.1184**	0.2059	
	LNorm	0.1741	0.1224	0.0918	0.1571	
	ELNorm	0.1827	0.1347	0.1041	0.1610	
	MNorm	0.2259	0.1388	0.1061	0.1755	
MNZ	NNorm	0.3343	0.2286	**0.1449**	0.3127	
	LNorm	0.3511	0.2367	0.1429	0.3277	
	ELNorm	0.3377	0.2041	0.1347	0.3038	
	MNorm	0.3441	0.2367	0.1408	0.3209	
Linear	NNorm	0.3359	0.2408	**0.1510**	0.3024	0.75
	LNorm	0.3423	0.2204	0.1490	0.3171	0.75
	ELNorm	0.3261	0.1878	0.1388	0.2813	0.3
	MNorm	0.3350	0.2245	0.1490	0.2973	0.8

TABLE 13. The performance of similarity score-based based fusion methods with the .GOV-03 collection for the topic distillation task.

Merging method	Normalization	MAP	P@5	*P@10**	R-prec	α
MIN	NNorm	0.0634	0.0480	0.0340	0.0586	
	LNorm	0.0476	0.0440	0.0560	0.0512	
	ELNorm	0.0366	0.0360	0.0520	0.0493	
	MNorm	0.0696	0.0520	**0.0696**	0.0614	
MAX	NNorm	0.1459	0.1720	**0.1300**	0.1511	
	LNorm	0.1300	0.1280	0.1020	0.1243	
	ELNorm	0.1160	0.1160	0.1000	0.1067	
	MNorm	0.1489	0.1720	**0.1300**	0.1511	
ANZ	NNorm	0.1079	0.1080	**0.0820**	0.1026	
	LNorm	0.0621	0.0560	0.0680	0.0694	
	ELNorm	0.0488	0.0520	0.0660	0.0570	
	MNorm	0.1221	0.1040	**0.0820**	0.1101	
MNZ	NNorm	0.1444	0.1720	0.1220	0.1466	
	LNorm	0.1472	0.1760	0.1180	0.1411	
	ELNorm	0.1435	0.1720	0.1200	0.1452	
	MNorm	0.1498	0.1840	**0.1240**	0.1430	
Linear	NNorm	0.1442	0.1720	**0.1300**	0.1511	0
	LNorm	0.1442	0.1720	**0.1300**	0.1511	0
	ELNorm	0.1442	0.1720	**0.1300**	0.1511	0
	MNorm	0.1442	0.1720	**0.1300**	0.1511	0

Discussion

Summary of Findings

Previous studies show that anchor text is a good descriptor of its target document, and that anchor text provides a good source of evidence that can be used to improve search quality. In this study, we compared the performance of a range of fusion methods on six common testbeds, which involve three document collections (.GOV, WT10G, and CERC) and two search tasks (homepage finding and topic distillation). Our findings are as follows:

1. This study confirms previous studies (Craswell et al., 2003; Craswell & Hawking, 2004), that searching an anchor-text representation alone can give significantly better retrieval performance than searching an original-text representation; this finding holds for all six testbeds.
2. Of the after-search fusion methods that use the similarity scores of retrieved documents, linear interpolation generally gives the best performance. For the score normalization methods, there is no single approach that outperforms the others on the homepage finding task, whereas the simple normalization of relevance scores within the range [0,1] achieves the best performance for the topic distillation task.
3. Generally, the best performing rank-based fusion method performs slightly (but not significantly) better than the best performing similarity score-based fusion method.
4. For the homepage finding task, the term-frequency merging method (BM25F) performs better than the best after-search fusion methods. However, for the topic distillation task, the performance of BM25F is slightly worse than the best after-search fusion methods for the CERC and .GOV-04 testbeds. The performance difference between BM25F and the best performing after-search fusion method is not statistically significant for either search task.
5. The best performing after-search fusion methods and the term-frequency merging method perform better than searching on anchor-text alone, and in most cases the improvement is significant.

DOI: 10.1002/asi

TABLE 14. The performance of similarity score-based based fusion methods with the .GOV-04 collection for the topic distillation task.

Merging method	Normalization	MAP	P@5	*P@10**	R-prec	α
MIN	NNorm	0.0604	0.0960	**0.0880**	0.0838	
	LNorm	0.0546	0.0853	**0.0880**	0.0733	
	ELNorm	0.0415	0.0800	0.0787	0.0624	
	MNorm	0.0572	0.0987	0.0867	0.0812	
MAX	NNorm	0.1263	0.2187	0.1867	0.1636	
	LNorm	0.1072	0.1760	0.1480	0.1348	
	ELNorm	0.0973	0.1760	0.1360	0.1226	
	MNorm	0.1272	0.2240	**0.1920**	0.1629	
ANZ	NNorm	0.1033	0.1760	0.1400	0.1387	
	LNorm	0.0694	0.0987	0.1040	0.0860	
	ELNorm	0.0747	0.1093	0.0973	0.0897	
	MNorm	0.1064	0.1733	**0.1493**	0.1445	
MNZ	NNorm	0.1396	0.2613	0.2053	0.1732	
	LNorm	0.1458	0.2453	0.2067	0.1736	
	ELNorm	0.1198	0.1813	0.1560	0.1456	
	MNorm	0.1436	0.2640	**0.2107**	0.1754	
Linear	NNorm	0.1441	0.2640	0.2067	0.1746	0.9
	LNorm	0.1487	0.2533	**0.2080**	0.1721	0.4
	ELNorm	0.1121	0.2213	0.1933	0.1599	0
	MNorm	0.1348	0.2480	0.2040	0.1714	0.6

TABLE 15. The performance of rank based fusion methods for the topic distillation task.

	Merging method	MAP	P@5	*P@10**	R-prec	α
CERC	Borda	0.2874	0.2082	0.1490	0.2495	0.7
	Reciprocal	0.3170	0.2041	0.1388	0.2977	0.15
GOV-03	Borda	0.1451	0.1720†	0.1300	0.1511	0
	Reciprocal	0.1535	0.1520	0.1340	0.1531	0.35
GOV-04	Borda	0.1335†	0.2667†	0.2093	0.1741†	0.05
	Reciprocal	0.1193	0.2213	0.1947	0.1614	0.05

Difference and Similarity Between Anchor-Text and Original-Text Representations

As was shown in Table 4, most anchor text terms also occur in the page that they refer to (that is, the original-text representation of the document). We therefore investigated whether searching on the anchor-text representations gives a very different set of search results from the original-text representations. To answer this question, we use the Jaccard coefficient to measure the similarity between the retrieved pages from the two collections (that is, the number of pages appearing in both runs divided by the number of unique pages from both runs).

The left-most three columns of Table 17 show the Jaccard similarity for all retrieved documents between any of two document representations, for the homepage finding task. It is apparent that the retrieved list from the anchor-text representation is very different from the original-text representation: The maximum similarity is about 13% for the WT10G representation. Moreover, even though the anchor-text representation is a part of the extended-text representation, the maximum similarity of the two is only around 17%. This is likely to be the case because the extended-text representation has far more words from the original-text representation than the anchor-text representation. Consistent with this explanation, the search result from the original-text representation is very similar to the list from the extended-text representation.

Looking more closely, we find that the above low similarity between sets of search results is largely due to the irrelevant documents retrieved from each set. The right-most three columns of 17 show the Jaccard similarity of only relevant pages retrieved between any two document representations. The numbers indicate that searching any representation delivers a similar set of relevant documents: The Jaccard similarity between two sets of retrieved relevant documents is at least 73%. Figure 1 shows the distribution of the similarity of individual documents for the pairs of collections. This demonstrates that relevant pages tend to be highly similar, whereas the distribution of irrelevant pages is skewed toward low similarity.

A similar trend is also found in the topic distillation task. As shown in Table 18, the overlap of retrieved relevant documents from the anchor-text and original-text representations is high and the overlap of all retrieved documents is still very low. However, Figure 2 shows that the distribution of the similarity between retrieved relevant documents is less skewed than for the homepage finding task. This may be because there are usually only one or two correct answer pages for each query in the homepage finding task, whereas for the topic distillation task there are on average 4, 10, and 21 relevant pages for the CERC, .GOV-03 and .GOV-04 collections, respectively.

The analysis reported in Tables 5 and 11 showed that searching anchor-text and extended-text representations delivers a list of search results that is significantly better than

DOI: 10.1002/asi

TABLE 16. The best runs from each result fusion methods for the topic distillation task.

	BM25	MAP	P@5	*P@10**	R-prec
CERC	Anchor-text	0.2962	0.2000	0.1245	0.2966
	Linear (Norm)	0.3423† (15.6%)	0.2204	0.1490† (19.7%)	0.3171
	Borda	0.2874 (-3.0%)	0.2082	0.1490† (19.7%)	0.2495
	BM25F	0.3180† (7.4%)	0.1878	0.1469† (18.0%)	0.2970
	(K = 9.6, w = 1)				
Median of TREC runs (16 groups)		0.2837	0.2122	0.1571	0.2296
Top TREC run		0.9772	0.6282	0.3531	0.9629
.GOV-03	Anchor-text	0.1442	0.1720	0.1300	0.1511
	Linear (norm)	0.1442 (0.0%)	0.1720	0.1300 (0.0%)	0.1511
	Reciprocal	0.1535 (6.4%)	0.1520	0.1340 (3.1%)	0.1531
	BM25F	0.1505† (4.4%)	0.1800	0.1360 (4.6%)	0.1369
	(K = 27, w = 9.3 ~ 9.6)				
Median of TREC runs (23 groups)		0.1371	0.1560	0.0880	0.1357
Top TREC run		0.1387	0.1440	0.1280	0.1450
.GOV-04	Anchor-text	0.1121	0.2213	0.1933	0.1599
	Linear (Norm)	0.1487† (32.6%)	0.2533†	0.2080 (7.6%)	0.1721
	Borda	0.1335† (19.1%)	0.2667†	0.2093 (8.3%)	0.1741†
	BM25F	0.1405† (25.3%)	0.2320	0.2013 (4.1%)	0.1811
	(K = 20.1, w = 12.1 ~ 12.3, 11.6, 11.7)				
Median of TREC runs (18 groups)		0.1150	0.2453	0.1990	0.1518
Top TREC run		0.1780	0.2960	0.2507	0.2045

TABLE 17. Jaccard similarity between search results from the original-text representation (C), the anchor-text representation (AT), and the combined representation (CAT) for the homepage finding task.

	All retrieved documents			Retrieved relevant documents		
Collection	*C*∩*AT*(%)	*AT*∩*CAT*(%)	*C*∩*CAT*(%)	*C*∩*AT* (%)	*AT*∩*CAT* (%)	*C*∩*CAT*(%)
.GOV-03	6.48	9.15	59.57	73.11	83.3	82.06
.GOV-04	8.80	10.24	92.59	86.67	89.33	92.67
WT10G	12.95	17.01	80.40	72.92	76.72	93.07

the one obtained from the original-text representation. It is also interesting to consider how many queries are actually improved from searching on the anchor-text instead of the original-text. This information is presented in Table 19. Taking the testbed .GOV-03 with the homepage finding task as an example, there are 52 queries that have at least one relevant page in the top 5 search results from the original-text representation (row .GOV-03, columns 1 and 3 (where $C{:}S@5 = 1$), numbers outside parentheses). If we search on the anchor-text representation instead, search performance deteriorates for 12 of them; however, there will be 81 queries for which performance is improved (in the table, this is equivalent to moving from columns 1 and 3 (where $S@5 = 1$ for the original-text representation, to columns 1 and 2 where $S@5 = 1$ for the anchor-text representation). As we can summarize from the table: searching the anchor-text representation improves more queries than it harms, and the homepage finding task gains more benefit from the anchor-text representation than does the topic distillation task.

Table 19 also shows the number of queries that meet performance criteria after applying the best-performing fusion methods from the similarity score, rank, and term frequency families. This shows that the performance improvements achieved by the best-performing fusion methods are likely due to moving the rank positions of the top documents from either document representation. Each column of the table represents queries split into four performance categories: In the first three categories, there is at least one relevant document in the top 5 (for homepage finding) from the original-text, the anchor-text, or both. (For topic distillation, the corresponding analysis is based on a cutoff position of 10, consistent with the nature of the task). In the fourth column, there are no relevant documents above the cutoff from either document representation. We can see that, after the BM25F, ELNorm and reciprocal fusion methods are applied (the respective numbers in parentheses in each cell), there are a similar number of queries in the fourth category for each approach. Again, consider the testbed .GOV-03 with the homepage finding task as an example. Before fusing, there are 40, 81, 12, and 17 queries in each category; after fusing with the ELNorm method, there are 78 (out of 81) and 11 (out of 12) more queries that

DOI: 10.1002/asi

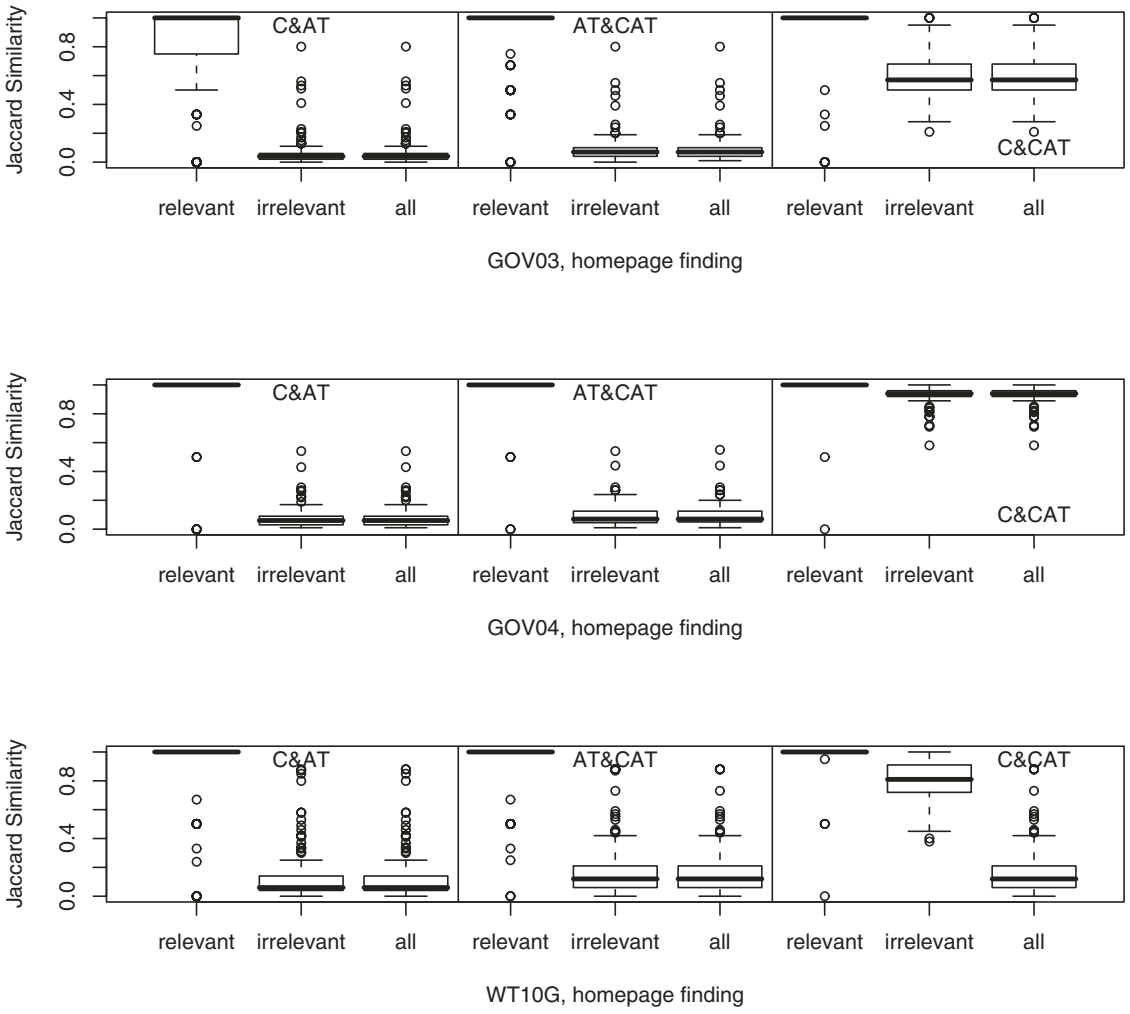

FIG. 1. The overlap between retrieved documents from different document representations, for the homepage finding task. Boxes show the 25th and 75th percentiles, the dark line is the median and whiskers are extreme values.

TABLE 18. Jaccard similarity between search results from the original-text representation (C), the anchor-text representation (AT), and the combined representation (CAT) for the topic distillation task.

	All retrieved documents			Retrieved relevant documents		
	$C \cap AT$	$AT \cap CAT$	$C \cap CAT$	$C \cap AT$	$AT \cap CAT$	$C \cap CAT$
Collection	(%)	(%)	(%)	(%)	(%)	(%)
CERC	10.09	14.15	74.69	58.63	65.20	85.99
.GOV-03	6.67	9.72	66.91	45.30	58.25	75.16
.GOV-04	8.16	10.18	85.61	45.46	50.25	88.45

have a relevant document moved to the top 5 if we take original-text representation and anchor-text representation as a baseline respectively. For those 17 queries that have none of the relevant documents ranked above the cutoffs from the both representations, the number remains similar after the two representations are combined. An example of this behavior at the query level is topic 419. The relevant answer document is at the position 254 and 20 when searching the original-text and anchor-text representations, respectively. The relevant document is moved to the positions 17, 20, and 34 by the fusion methods BM25f, ELNorm and reciprocal rank, respectively. However, none of the fusion methods were able to produce a rank position that is high enough to affect the evaluation metrics.

Conclusion

Anchor test has been shown to be a useful resource in improving search quality. Although many approaches for the use of anchor text have been proposed, these have not been compared across a consistent set of search tasks and document collections. Our contributions in this article are that we compare three classes of fusion methods based on six testbeds (three test collections for two search tasks). We also investigate how search results from anchor text are different from those obtained from original text, and how this difference contributes to the rank improvement through fusion methods.

DOI: 10.1002/asi

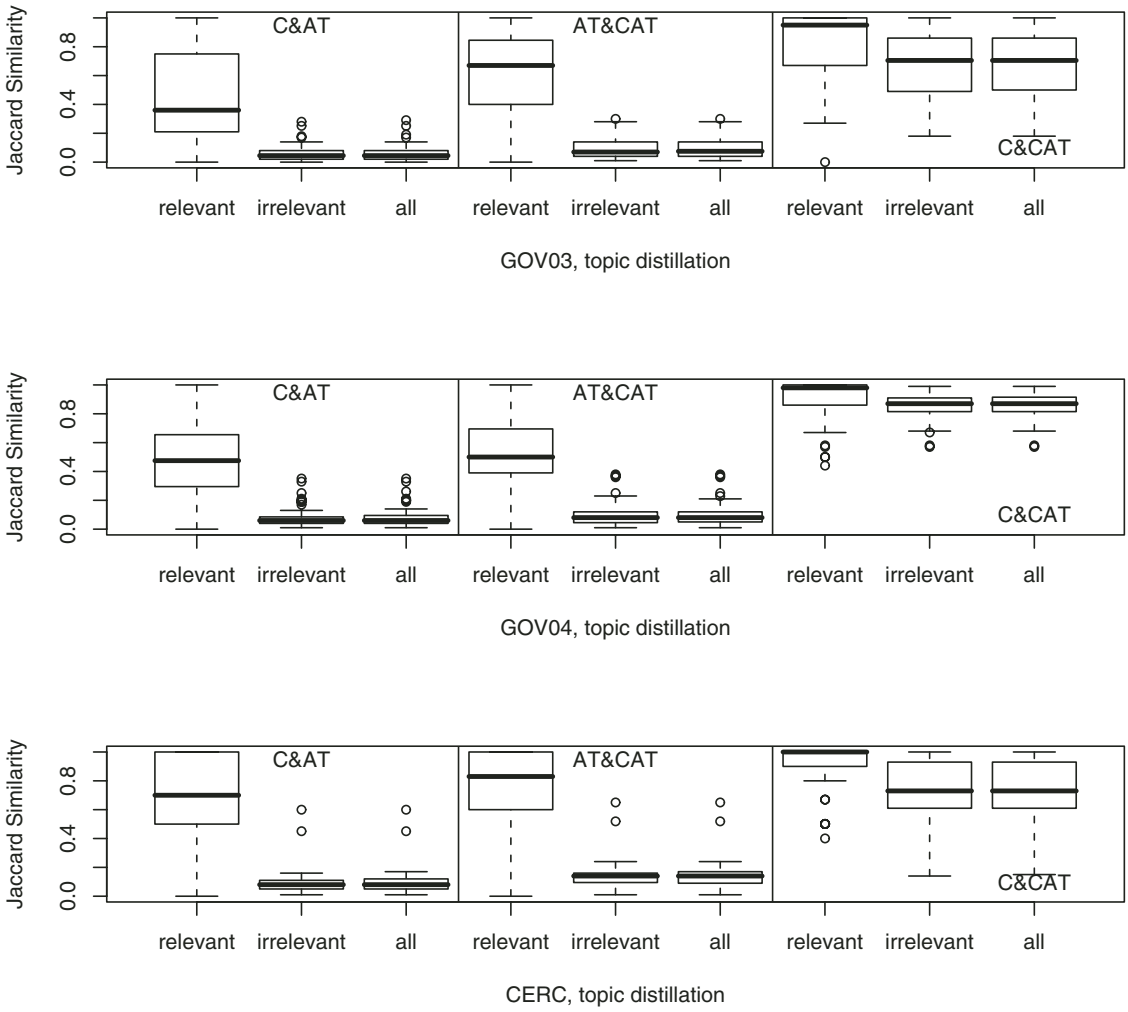

FIG. 2. The overlap between document representations for the topic distillation task.

TABLE 19. The number of queries whose results meet a specified measurement criterion.

Homepage finding task				
Testbeds	AT: S@5 = 1 C: S@5 = 1 (S@5 = 1)	AT: S@5 = 1 C: S@5 = 0 (S@5 = 1)	AT: S@5 = 0 C: S@5 = 1 (S@5 = 1)	AT: S@5 = 0 C: S@5 = 0 (S@5 = 0)
.GOV-03	40 (40,39,40)	81(80,78,80)	12(9,11,9)	17(16,17,17)
.GOV-04	17(17,17,17)	38(35,36,36)	8(7,5,3)	12(9,11,12)
WT10G	51(51,51,51)	44(42,38,43)	24(22,20,17)	26(25,26,26)
Topic distillation task				
Testbeds	AT: P@10 > 0 C: P@10 > 0 (P@10 > 0)	AT: P@10 > 0 C: P@10 = 0 (P@10 > 0)	AT: P@10 = 0 C: P@10 > 0 (P@10 > 0)	AT: P@10 = 0 C: P@10 = 0 (P@10 = 0)
.GOV-03	22(21,22,22)	12(10,12,12)	8(2,8,8)	8(8,8,8)
.GOV-04	37(35,37,37)	19(14,15,18)	9(5,4,3)	10(8,7,7)
CERC	27(27,27,27)	7(7,6,6)	3(3,3,3)	12(12,11,12)

Note. Numbers outside brackets show the count of queries whose search results satisfy both specified measures for the anchor-text and original-text representation (S@5 and P@10 for the homepage and named page finding tasks, respectively). The numbers in brackets indicate the number of queries whose fused search results satisfy the measure, indicated in parentheses in the column heading, after a fusion method is applied (BM25F, linear EZNorm, and Reciprocal/Borda, respectively). See the text for examples.

DOI: 10.1002/asi

Our experimental results indicate that the best performing fusion methods from each class can lead to significant improvements over baselines. Among the similarity score-based fusion methods, the linear interpolation method was the best performer for both search tasks. Between the ranked based fusion methods, reciprocal rank and Borda each performed better for the homepage finding and topic distillation task, respectively. The during-search fusion method, term frequency merge, showed better performance than the best after-search fusion methods for the homepage finding task; however, the performance difference between the best performing fusion methods of each type is not statistically significant.

Our analysis of search results demonstrated that searching on anchor text and original text representations can lead to similar sets of relevant documents, but very different sets of irrelevant documents. When at least one representation already ranks an answer document highly, then the after-search fusion methods can promote the rank of such documents further; however, the fusion approaches are less successful for those documents with a low initial rank from both representations. This explains why similarity score-based fusion and rank-based fusion give very similar results: Both approaches need extra evidence beyond text ranking to bring those low-ranked relevant documents up. We plan to investigate the impact of other evidence such as inlink and outlink information in future work.

In this article, we have focused on the combination of two document representations: anchor text and the original full-text content. It is worth noting that these are only two of many possible rich representative features that could be considered under the polyrepresentation continuum, which includes multiple query, source, and system representations (Larsen, Ingwersen, & Kekalainen, 2006; Lioma, Larsen, Schuetze, & Ingwersen, 2010). It would be interesting to investigate these in future work.

References

Allan, J., Callan, J., Feng, F., & Malin, D. (1999). INQUERY and TREC-8. In Proceedings of the 8th Text Retrieval Conference (TREC 1999). Gaithersburg, MD: National Institute of Standards and Technology.

Amitay, E., & Paris, C. (2000). Automatically summarising Web sites—Is there a way around it? In Proceedings of the 9th ACM International Conference on Information and Knowledge Management (pp. 173–179). New York: ACM Press.

Aslam, J.A., & Montague, M. (2001). Models for metasearch. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 276–284). New York: ACM Press.

Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. Information Processing & Management, 39(6), 853–871.

Bailey, P., Craswell, N., Soboroff, I., & de Vries, A.P. (2007). The CSIRO enterprise search test collection. ACM SIGIR Forum, 41(2), 42–45.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vires, A.P., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter? In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 667–674). New York: ACM Press.

Bartell, B.T., Cottrell, G.W., & Belew, R.K. (2005). Automatic combination of multiple ranked retrieval systems. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 75–82). New York: ACM Press.

Beitzel, S., Jensen, E., Cathey, R., Ma, L., Chowdhury, A., & Pass, G. (2003). IIT at TREC-2003: Task classification and document structure for known-item search. In Proceedings of the 12th Text Retrieval Conference (TREC 2003). Gaithersburg, MD: National Institute of Standards and Technology.

Beitzel, S., Jensen, E., Chowdhury, A., Grossman, D., Frieder, O., & Goharian, N. (2004). Fusion of effective retrieval strategies in the same information retrieval system. Journal of the American Society for Information Science and Technology, 55(10), 859–868.

Belkin, N.J., Cool, C., Croft, W.B., & Callan, J.P. (1993). The effect of multiple query representations on information retrieval performance. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 339–346). New York: ACM Press.

Belkin, N.J., Kantor, P., Fox, E.A., & Shaw, J.A. (1995). Combining the evidence of multiple query representations for information retrieval. Information Processing and Management, 31(3), 431–448.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual web search engine. In Proceedings of the 7th international World Wide Web Conference (pp.107–117). New York: ACM Press

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajageopalan, S., Stata, R., . . . Wiener, J. (2000). Graph structure in the web. Computer Networks, 33(1–6), 309–320.

Chakrabarti, S., Dom, B.E., Raghavan, P., Rajagopalan, S., Gibson, D., & Kleinberg, J.M. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text. In Proceedings of the 7h international World Wide Web Conference (pp. 65–74). New York: ACM Press.

Chowdhury, A., Aljlayl, M., Jensen, E., & Beitzel, S. (2002). IIT at TREC 2002: Linear combinations based on document structure and varied stemming for Arabic retrieval. In Proceedings of the 11th Text Retrieval Conference (TREC 2002). Gaithersburg, MD: National Institute of Standards and Technology.

Clarke, C.L.A., Craswell, N., & Soboroff, I. (2009). Overview of the TREC 2009 Web track. In Proceedings of the 18th Text Retrieval Conference (TREC 2009). Gaithersburg, MD: National Institute of Standards and Technology.

Craswell, N., & Hawking, D. (2002). Overview of the TREC-2002 Web track. In Proceedings of the 11th Text Retrieval Conference (TREC 2002). Gaithersburg, MD: National Institute of Standards and Technology.

Craswell, N. & Hawking, D. (2004). Overview of the TREC-2004 Web track. In Proceedings of the 13th Text Retrieval Conference (TREC 2004). Gaithersburg, MD: National Institute of Standards and Technology.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 250–257). New York: ACM Press.

Craswell, N., Hawking, D., Wilkinson, R., & Wu, M. (2003). Overview of the TREC-2003 web track. In Proceedings of the 12th Text Retrieval Conference (TREC 2003). Gaithersburg, MD: National Institute of Standards and Technology.

Davison, B.D. (2000). Topical locality in the web. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 272–279). New York: ACM Press.

Duan, H., Zhou, Q., Lu, Z., Jin, O., Bao, S., Cao, Y., & Yu, Y. (2007). Research on enterprise track of TREC 2007 at SJTU APEX Lab. In Proceedings of the 16th Text Retrieval Conference (TREC 2007). Gaithersburg, MD: National Institute of Standards and Technology.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In Proceedings of the 10th international World Wide Web Conference (pp. 613–622). New York: ACM Press.

DOI: 10.1002/asi

Eiron, N., & McCurley, K.S. (2003). Analysis of anchor text for web search. In Proceedings of the 26st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 459–460). New York: ACM Press.

Fox, E.A., & Shaw, J.A. (1993). Combination of multiple searches. In Proceedings of the 2nd Text Retrieval Conference (TREC 1993) (pp. 243–252). Gaithersburg, MD: National Institute of Standards and Technology.

Fujii, A. (2008). Modeling anchor text and classifying queries to enhance web document retrieval. In Proceedings of the 17th international World Wide Web Conference (pp. 337–346). New York: ACM Press

Hawking, D., & Craswell, N. (2001). Overview of the TREC-2001 web track. In Proceedings of the 10th Text Retrieval Conference (TREC 2001). Gaithersburg, MD: National Institute of Standards and Technology.

Hawking, D., & Craswell, N. (2005). The very large collection and web tracks. In E.M. Voorhees & D.K. Harman (Eds.), Experiment and evaluation in information retrieval (pp. 199–231). Cambridge, MA: MIT Press.

Hawking, D., & Thomas, P. (2005). Server selection methods in hybrid portal search. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 173–181). New York: ACM Press.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. Journal of Documentation, 52(1), 3–50.

Ingwersen, P., & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Springer-Verlag New York, Inc.

Kamps, J. (2005). Web-centric language models. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 307–308). New York: ACM Press.

Kelly, D., Dollu, V.D., & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 457–464). New York: ACM Press.

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information search systems. Information Processing and Management, 43(1), 30–46.

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5), 604–632.

Koolen, M., & Kamps, J. (2010). The importance of anchor text for ad hoc search revisited. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 122–129). New York: ACM Press.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 27–34). New York: ACM Press.

Kraft, R., & Zien, J. (2004). Mining anchor text for query refinement. In Proceedings of the 13th international World Wide Web Conference (pp. 666–674). New York: ACM Press.

Larkey, L.S., Connell, M.E., & Callan, J. (2000). Collection selection and results merging with topically organized U.S. patents and TREC data. In Proceedings of the 9th International Conference on Information and Knowledge Management (pp. 282–289). New York: ACM Press

Larsen, B., Ingwersen, P., & Lund, B. (2009). Data fusion according to the principle of polyrepresentation. Journal of the American Society for Information Science and Technology, 60(4), 646–654.

Larsen, B., Ingwersen, P., & Kekalainen, J. (2006). The polyrepresentation continuum in IR. In IIiX: Proceedings of the 1st International Conference on Information Interaction in Context (pp. 88–96). New York: ACM Press.

Lee, J.H. (1997). Analysis of multiple evidence combination. In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 267–276). New York: ACM Press.

Lioma, C., Larsen, B., Schuetze, H., & Ingwersen, P. (2010). A subjective logic formalisation of the principle of polyrepresentation for information needs. In IIiX: Proceedings of the 3rd Symposium on Information Interaction in Context (pp. 125–134). New York: ACM Press.

Lu, Y., Hu, J., & Ma, F. (2004). SJTU at TREC 2004: Web track experiments. In Proceedings of the 13th Text Retrieval Conference (TREC 2004). Gaithersburg, MD: National Institute of Standards and Technology. from http://trec.nist.gov/pubs/trec13/t13_proceedings.html

Lund, B.R., Schneider, J.W., & Ingwersen, P. (2006). Impact of relevance intensity in test topics on {IR} performance in polyrepresentative exploratory search systems. In Proceedings of the SIGIR 2006 EESS (Evaluating Exploratory Search Systems) Workshop (pp. 42–46). New York: ACM Press.

McBryan, O.A. (1994). GENVL and WWWW: Tools for taming the web. In Proceedings of the 1st International World Wide Web Conference (pp. 79–90). New York: ACM Press.

Metzler, D., Novak, J., Cui, H., & Reddy, S. (2009). Building enriched document representations using aggregated anchor text. In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 219–226). New York: ACM Press.

Montague, M.H., & Aslam, J.A. (2002). Condorcet fusion for improved retrieval. In Proceedings of the 9th International Conference on Information and Knowledge Management (pp. 538–548). New York: ACM Press.

Myaeng, S.H., Jang, D.H., Kim, M.S., & Zhoo, Z.C. (1998). A flexible model for retrieval of SGML documents. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 138–145). New York: ACM Press.

Ogilvie, P., & Callan, J. (2003a). Combining document representations for known-item search. In Proceedings of the 26st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 143–150). New York: ACM Press.

Ogilvie, P., & Callan, J. (2003b). Language models and structured document retrieval. In Proceedings of the 1st Initiative for the Evaluation of XML Retrieval Workshop (INEX 2003). Saarbrücken, Germany: Saarland University.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web (Technical report). Stanford, CA: Stanford Digital Library Technologies, Stanford University. Retrieved from http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Piwowarski, B., & Gallinari, P. (2003). A machine learning model for information retrieval with structured documents. In Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition (pp. 425–438). Heidelberg, Germany: Springer-Verlag.

Ponte, J.M., & Croft, B.W. (1998). A language modeling approach to information retrieval. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 275–281). New York: ACM Press.

Roberts, F.S. (1976). Discrete mathematic models with applications to social, biological, and environnmental problems. Englewood Cliffs, NJ: Prentice-Hall.

Robertson, S., Walker, S., Hancock-Beaulieu, M.M., & Gatford, M. (1994). Okapi at TREC-3. In Proceedings of the 3rd Text Retrieval Conference (TREC 1994) (pp. 109–126). Gaithersburg, MD: National Institute of Standards and Technology.

Robertson, S., Zaragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (pp. 42–49). New York: ACM Press.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieval III: Searchers, searches and overlap. Journal of the ASIS, 39(3), 197–216.

Skov, M., Larsen, B., & Ingwersen, P. (2008). Inter and intra-document contexts applied in polyrepresentation for best match IR. Information Processing and Management, 44(5), 1673–1683.

Song, S., Wen, J., Shi, S., Xin, G., Liu, T., Qin, T., . . . Ma, W. (2004). Microsoft Research Asia at web track and terabyte track of TREC 2004. In Proceedings of the 13th Text Retrieval Conference (TREC 2004).

DOI: 10.1002/asi

Gaithersburg, MD: National Institute of Standards and Technology. from http://trec.nist.gov/pubs/trec13/t13_proceedings.html

Tomlinson, S. (2003). Robust, web and genomic retrieval with Hummingbird SearchServer at TREC 2003. In Proceedings of the 12th Text Retrieval Conference (TREC 2003) (pp. 254–267). Gaithersburg, MD: National Institute of Standards and Technology.

Turtle, H., & Croft, W.B. (1991). Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, 9(3), 187–222.

Upstill, T., Craswell, N., & Hawking, D. (2003). Query-independent evidence in home page finding. ACM Transactions on Information Systems, 21(3), 286–313.

Vogt, C.C., & Cottrell, G.W. (1999). Fusion via a linear combination of scores. Information Retrieval, 1(3), 151–173.

Westerveld, T., Kraaij, W., & Hiemstra, D. (2001). Retrieving web pages using content, links, URLs and anchors. In Proceedings of the 10th Text Retrieval Conference (TREC 2001) (pp. 663–672). Gaithersburg, MD: National Institute of Standards and Technology.

Wilkinson, R. (1994). Effective retrieval of structured documents. In Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 311–317). New York: ACM Press.

Wu, M., Scholer, F., & Turpin, A. (2011). Topic distillation with query-dependent link connections and page characteristics. ACM Transactions of the Web, 5(2), 6:1–6:25.

Zaragoza, H., Craswell, N., Taylor, M., Saria, S., & Robertson, S. (2004). Microsoft Cambridge at TREC-13: Web and HARD tracks. In Proceedings of the 13th Text Retrieval Conference (TREC 2004). Gaithersburg, MD: National Institute of Standards and Technology. from http://trec.nist.gov/pubs/trec13/t13_proceedings.html

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 334–342). New York: ACM Press.

Zhou, Z., Guo, Y., Wanng, B., Cheng, X., Xu, H., & Zhang, G. (2004). TREC 2004 Web track experiments at CAS-ICT. In Proceedings of the 13thText Retrieval Conference (TREC 2004). Gaithersburg, MD: National Institute of Standards and Technology. from http://trec.nist.gov/pubs/trec13/t13_proceedings.html

DOI: 10.1002/asi