A Decision Mechanism for the Selective Combination of Evidence in Topic Distillation

Vassilis Plachouras[1], Fidel Cacheda[2], and Iadh Ounis[1]

[1]University of Glasgow, G12 8QQ Glasgow, UK
[2]University of A Coruña, 15071 A Coruña, Spain

Abstract

The combination of evidence can increase retrieval effectiveness. In this paper, we investigate the effectiveness of a decision mechanism for the selective combination of evidence for Web Information Retrieval and particularly for topic distillation. We introduce two measures of a query's broadness and use them to select an appropriate combination of evidence for each query. The results from our experiments show that there is a statistically significant association between the output of the decision mechanism and the relative effectiveness of the different combinations of evidence. Moreover, we show that the proposed methodology can be applied in an operational setting, where relevance information is not available, by setting the decision mechanism's thresholds automatically.

Keywords Web information retrieval, Topic distillation, decision mechanism, selective combination of evidence, query scope, aggregates

1 Introduction

It has been recognised that the combination of different sources of evidence can improve the effectiveness of Information Retrieval (IR) systems (Croft 2000). In the context of Web IR, and more specifically topic distillation, in addition to the textual content of documents, there is another source of evidence, namely the hyperlink structure, which can be employed to refine the content-based retrieval and increase the precision among the top ranked documents. However, the weaker evidence provided from the hyperlink structure and the different types of queries, i.e. specific or broad queries (Kleinberg 1998), or navigational versus informational queries (Broder 2002), suggest that combining content and hyperlink analysis in a uniform way, independently from the queries, does not necessarily lead to optimal retrieval effectiveness (Plachouras et al. 2003, Plachouras and Ounis 2004).

We propose a decision mechanism for the selective combination of evidence. This decision mechanism is based on the *query scope*, which estimates two important statistical aspects of the set of retrieved documents. The first aspect, *query extent*, estimates how broad a query is, by counting the number of documents that contain all the query terms. Query topics that correspond to high query extent are well covered in the documents collection. Therefore, more evidence from hyperlink analysis can be used to detect documents of high quality. The second aspect, *result extent*, counts the number of domains or directories that contain a high number of retrieved documents. This aspect of the set of retrieved documents indicates whether there are whole or large parts of sites about the query topic. If this is the case, then we can employ evidence from the hyperlink or the document structure and retrieve the entry points of these sites.

The decision mechanism uses the query scope aspects to select an appropriate combination of evidence for each query. For the evaluation of the decision mechanism, we experiment with the .GOV TREC Web test collection and the associated topics from the topic distillation tasks of TREC11 (Craswell and Hawking 2002) and TREC12 (Craswell et al. 2003), respectively. The evaluation of the decision mechanism takes place in two steps.

In the first step, we use relevance information and evaluate each of the query extent and result extent independently, by testing a range of different threshold values. We show that improvements are obtained over the uniform combination of evidence for all queries. Moreover, we find that there is a statistically significant association between the values of the query scope aspects and the relative effectiveness of the combinations of evidence. For example, we show that content retrieval is more effective for the queries that correspond to a low query or result extent values. For the second step of the evaluation, we consider a more realistic setting, where relevance information is not readily available for tuning the thresholds of the decision mechanism. We sample terms from the document collection and use them as single-term queries, an approach similar to that proposed by Cronen-Townsend et al. (2002). After computing the values of query extent and result extent for the engineered queries, we set the thresholds of the decision mechanism, so that a given percent of each aspect's values are below the corresponding threshold. The results show that we can still obtain improvements in retrieval effectiveness by setting the thresholds automatically.

The remainder of the paper is organised as follows. In Section 2, we look at the potential benefit from the selective combination of evidence on a per-query basis, compared to a uniform combination of evidence. Section 3 contains the description of the proposed decision mechanism for selecting the most appropriate combination of evidence for each query. In Sections 4 and 5, we evaluate the decision mechanism in different settings, using relevance information. Next, in Section 6, we set the decision mechanism's thresholds automatically, without relevance information. We continue in Section 7 with a discussion of interesting points from the experiments. Section 8 contains a brief overview of the related work, and we close with concluding remarks in Section 9.

2 Combination of Evidence for Topic Distillation

In this section, we investigate the potential benefit from a dynamic selection of sources of evidence for topic distillation. We evaluate different uniform combinations of evidence, where the same sources of evidence are used for all queries. Following, we assume that there exists a decision mechanism, which selects the most appropriate combination of evidence for each query, and we show that there is room for improvement in retrieval effectiveness over the uniform combination of evidence.

We experiment with the TREC .GOV collection, a crawl of approximately 1.25 million Web documents from the .gov domain. For indexing the collection, we removed stop-words and applied Porter's stemming algorithm. We use the topics[1] and the corresponding relevance information from the topic distillation tasks of TREC11 (Craswell and Hawking 2002) (49 topics[2]) and TREC12 (Craswell et al. 2003) (50 topics).

Both used tasks involve finding useful entry points to sites that are relevant to the query topics. However, a difference between the two tasks is that the relevant documents for the TREC12 topics were restricted to be homepages of relevant sites. This resulted in a lower number of relevant documents, less than 10 relevant documents for many topics. Thus, it would not be theoretically possible to obtain 100% precision at 10 documents, which was the evaluation measure for TREC11. For this reason, the TREC Web track organisers chose the R-Precision (precision after R documents have been retrieved, where R is the number of relevant documents for the topic) as the official evaluation measure for the TREC12 topic distillation task. We will use precision at 10 for both TREC11 and TREC12 tasks, and we will also report R-Precision for the TREC12 experiments.

There are various sources of evidence that can be used for Web IR, in addition to the textual content

[1] The queries are created from the title of the topics.

[2] Originally, the task involved 50 topics, but there were no relevant documents for one of them.

of documents. In this paper, we focus on three different sources of evidence. The first is the textual content of documents, denoted by C. The second is the anchor text associated with the incoming links of a document, denoted by A. It has been shown that the anchor text is an effective source of evidence for topic distillation (Craswell et al. 2003), as well as for other tasks, such as finding homepages (Craswell et al. 2001). The third source of evidence is the document's URL (Kraaij et al. 2002, Upstill et al. 2003), and more specifically the length in characters of its path, since it is likely that homepages of sites will have URLs with shorter paths. We denote this last source of evidence as U.

From all possible combinations of the sources of evidence mentioned above (C, A, U, CA, CU, AU, CAU), we will restrict our analysis to three of them, which have proved to be effective for the TREC tasks we experiment with. More specifically, for the first combination we use only the textual content of documents C. For the second combination of evidence, we extend each document with the anchor text of its incoming links (CA). For the third combination of evidence, we combine CA with the length of the document's URL (CAU), as follows:

$$score_i = \frac{sc_i}{\log_2(urlpath\_len_i + 1)} \tag{1}$$

where sc_i is the content analysis score for document d_i, based on CA, and $urlpath\_len_i$ is the number of characters in the URL path if d_i. In order to avoid boosting non-relevant documents, we apply this approach only to the top 1000 retrieved documents, where content and anchor text is used for retrieval.

For the content analysis, we consider three diverse and statistically independent weighting schemes. The first one is the well-known BM25 weighting scheme (Robertson and Walker 1994), with b set empirically to 0.72. The other two are from Amati and Van Rijsbergen's (2002) Divergence From Randomness (DFR) probabilistic framework, namely PL2 and I(n_e)C2. The weight of a query term t that occurs in a document is given by the following formulae:

$$\begin{aligned} weight_{\mathrm{PL2}}(t) &= \Big(tfn_1 \cdot \log_2 \tfrac{tfn_1}{\lambda} + (\lambda + \tfrac{1}{12 \cdot tfn_1} - tfn_1) \cdot \log_2 e \\ &\quad +0.5 \cdot \log_2(2\pi \cdot tfn_1)\Big) \cdot \tfrac{1}{tfn_1 + 1} \\ weight_{\mathrm{I}(n_e)\mathrm{C2}}(t) &= \tfrac{col\_freq+1}{doc\_freq \cdot (tfn_2 + 1)} \cdot \Big(tfn_2 \cdot \ln \tfrac{N+1}{n_e + 0.5}\Big) \end{aligned}$$

where:

- $tfn_1 = term\_freq \cdot \log_2 \Big(1 + c \cdot \frac{average\_document\_length}{document\_length}\Big)$,
- $tfn_2 = term\_freq \cdot \ln \Big(1 + c \cdot \frac{average\_document\_length}{document\_length}\Big)$,

- $n_e = N \cdot \left(1 - \left(\frac{1}{N}\right)^{col\_freq}\right)$,
- N is the size of the collection,
- col_freq is the within-collection term-frequency,
- $term\_freq$ is the within-document term-frequency,
- doc_freq is the document-frequency of the term and
- $\lambda = col\_freq/N$, which is the mean and variance of a Poisson distribution, with $col\_freq \ll N$.

The only parameter of the system is automatically set to $c = 1.28$, using the approach proposed by He and Ounis (2003). The final score of a document d corresponds to the sum of the weights of all terms that occur in the query q and in the document:

$$weight_x(d|q) = \sum_{t \in q \wedge t \in d} weight_x(t) \tag{2}$$

where x corresponds to BM25, PL2, or I(n_e)C2.

In Table 1, we present the evaluation of the three selected combinations of evidence, when applied for all queries uniformly. We can see that for TREC11, using the textual content of documents is sufficiently effective and any of the two other combinations yields lower precision at 10. The results for TREC12 show that when the relevant documents are only homepages of sites, then the most effective combination of evidence uses all three sources of evidence. For both TREC11 and TREC12, employing content and anchor text (CA) is the second most effective combination of evidence. Before continuing, it is worth noting that the best official run submitted for the TREC11 topic distillation task achieved 0.2510 precision at 10 (Craswell and Hawking 2002). The highest precision at 10 and R-Precision, achieved by the official runs submitted for the TREC12 topic distillation task, were 0.1280 and 0.1636 respectively (these two numbers correspond to two different runs) (Craswell et al. 2003).

In what follows, we investigate whether the selective application of a combination of evidence on a per-query basis can lead to improvements in retrieval effectiveness. We assume that there is a decision mechanism MAX(X, Y,. . .) that will select on a per-query basis the most effective combination of evidence from X, Y,. . ., where each of X, Y,. . . corresponds to one of the combinations of evidence C, CA, or CAU. In this way, we would obtain the highest possible retrieval effectiveness on average with the given setting. Table 2 contains the results of the decision mechanism MAX(X, Y,. . .) for different weighting schemes and all possible pairs of the three combinations of evidence C, CA and CAU. For example, we can see that MAX(C_{PL2}, CA_{PL2}), which selects between C and CA, using the weighting scheme PL2,

Table 1. The evaluation of the uniform application of C, CA and CAU for the topic distillation tasks of TREC11 and TREC12.

	TREC11	TREC12	
Run	Precision at 10	Precision at 10	R-Precision
$\mathrm{C}_{\mathrm{PL2}}$	**0.2694**	0.0680	0.0730
$\mathrm{CA}_{\mathrm{PL2}}$	0.2551	0.1020	0.1325
$\mathrm{CAU}_{\mathrm{PL2}}$	0.1367	**0.1400**	**0.1369**
$\mathrm{C}_{\mathrm{BM25}}$	**0.2408**	0.0720	0.0934
$\mathrm{CA}_{\mathrm{BM25}}$	**0.2408**	0.1020	**0.1293**
$\mathrm{CAU}_{\mathrm{BM25}}$	0.1000	**0.1180**	0.1216
$\mathrm{C}_{\mathrm{I}(n_e)\mathrm{C2}}$	**0.2490**	0.0680	0.0732
$\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$	0.2388	0.0940	0.1222
$\mathrm{CAU}_{\mathrm{I}(n_e)\mathrm{C2}}$	0.1061	**0.1160**	**0.1284**

results in 0.2898 average precision at 10 for TREC11. In addition, we can see that $\mathrm{C}_{\mathrm{PL2}}$ performs better for 9 queries, while $\mathrm{CA}_{\mathrm{PL2}}$ is more effective for 8 out of 49 queries in total. From the same table, we can see that for TREC11, the decision mechanism MAX is most effective, when we use $\mathrm{C}_{\mathrm{PL2}}$ and $\mathrm{CAU}_{\mathrm{PL2}}$, resulting in 0.2959 average precision at 10. Similarly for TREC12, the decision mechanism MAX is most effective when we employ $\mathrm{CA}_{\mathrm{PL2}}$ and $\mathrm{CAU}_{\mathrm{PL2}}$, achieving 0.1680 average precision at 10 and 0.1880 average R-Precision. For the two other weighting schemes, namely BM25 and $\mathrm{I}(n_e)\mathrm{C2}$, we observe that improvements are also obtained over the uniform combination of evidence for all queries.

Table 2. The evaluation of the decision mechanism MAX that selects the most effective combination of evidence on a per-query basis for TREC11 and TREC12. The first (second) number in parentheses denotes the number of queries for which the combination of evidence X (Y) in MAX(X,Y) outperforms Y (X).

	TREC11		TREC12			
Hypothetical Run	Precision at 10		Precision at 10		R-Precision	
MAX($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$)	0.2898	(9, 8)	0.1120	(5, 18)	0.1400	(2, 17)
MAX($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$)	**0.2959**	(28, 8)	0.1600	(8, 24)	0.1699	(7, 23)
MAX($\mathrm{CA}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$)	0.2673	(29, 5)	**0.1680**	(11, 19)	**0.1880**	(9, 16)
MAX($\mathrm{C}_{\mathrm{BM25}}$, $\mathrm{CA}_{\mathrm{BM25}}$)	0.2531	(5, 5)	0.1080	(3, 13)	0.1321	(1, 13)
MAX($\mathrm{C}_{\mathrm{BM25}}$, $\mathrm{CAU}_{\mathrm{BM25}}$)	**0.2571**	(31, 5)	0.1500	(11, 22)	0.1714	(7, 18)
MAX($\mathrm{CA}_{\mathrm{BM25}}$, $\mathrm{CAU}_{\mathrm{BM25}}$)	**0.2571**	(32, 5)	**0.1620**	(16, 19)	**0.1796**	(10, 15)
MAX($\mathrm{C}_{\mathrm{I}(n_e)\mathrm{C2}}$, $\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$)	**0.2612**	(9, 5)	0.1000	(3, 13)	0.1303	(3, 14)
MAX($\mathrm{C}_{\mathrm{I}(n_e)\mathrm{C2}}$, $\mathrm{CAU}_{\mathrm{I}(n_e)\mathrm{C2}}$)	**0.2612**	(29, 5)	0.1460	(10, 22)	0.1597	(7, 22)
MAX($\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$, $\mathrm{CAU}_{\mathrm{I}(n_e)\mathrm{C2}}$)	0.2551	(29, 5)	**0.1580**	(15, 19)	**0.1830**	(10, 17)

For a more realistic decision mechanism, we would have to choose the combinations of evidence to use among all the possible ones. We could employ two approaches to perform this selection. First, we could use the average retrieval effectiveness of the decision mechanism MAX, and select the combinations of evidence that would result in the highest retrieval effectiveness. As we can see from Table 2, for the

TREC11 topic distillation task, MAX($\mathrm{C_{PL2}}$, $\mathrm{CAU_{PL2}}$) outperforms MAX($\mathrm{C_{PL2}}$, $\mathrm{CA_{PL2}}$), even though $\mathrm{CAU_{PL2}}$ is less effective than $\mathrm{CA_{PL2}}$ (0.1367 with respect to 0.2551 average precision at 10, from Table 1). This suggests that employing the average retrieval effectiveness of MAX for selecting the combinations of evidence to use is prone to over-fitting. The second approach is based on selecting those combinations of evidence that result in the highest average retrieval effectiveness. For the remainder of the paper, we will employ the second approach and select the two combinations of evidence with the highest average precision at 10 retrieved documents, as shown in Table 1. Therefore, for TREC11, we will employ $\mathrm{C_{PL2}}$ and $\mathrm{CA_{PL2}}$, while for TREC12, we will use $\mathrm{CA_{PL2}}$ and $\mathrm{CAU_{PL2}}$.

So far, we have looked at the potential improvements from the decision mechanism MAX, where the combinations of evidence use the same weighting scheme. In Table 3, we consider some cases, where a different weighting scheme is used for each of the two combinations of evidence. For example, MAX($\mathrm{C_{PL2}}$, $\mathrm{CA_{I(n_e)C2}}$) corresponds to the case, where the decision mechanism MAX selects either content-only retrieval with PL2, or content and anchor text retrieval with I(n_e)C2. Among the results shown in Table 3, the most effective pair of combinations of evidence for TREC11 is $\mathrm{C_{PL2}}$ and $\mathrm{CA_{BM25}}$, which results in 0.3041 average precision at 10. For TREC12, the highest R-Precision is obtained when $\mathrm{CA_{PL2}}$ and $\mathrm{CAU_{I(n_e)C2}}$ are used.

Table 3. The evaluation of the decision mechanism MAX that selects the most effective combination of evidence on a per-query basis for TREC11 and TREC12 topic distillation tasks. The combinations of evidence are based on different weighting schemes.

	TREC11		TREC12			
Hypothetical Run	Precision at 10		Precision at 10		R-Precision	
MAX($\mathrm{C_{PL2}}$, $\mathrm{CA_{I(n_e)C2}}$)	0.2938	(15, 8)	0.1100	(7, 17)	0.1342	(3, 16)
MAX($\mathrm{C_{PL2}}$, $\mathrm{CA_{BM25}}$)	**0.3041**	(15, 11)	**0.1120**	(5, 17)	0.1362	(3, 16)
MAX($\mathrm{CA_{PL2}}$, $\mathrm{CAU_{I(n_e)C2}}$)	**0.2694**	(31, 5)	**0.1620**	(15, 18)	**0.1901**	(12, 17)
MAX($\mathrm{CA_{PL2}}$, $\mathrm{CAU_{BM25}}$)	0.2673	(31, 4)	**0.1620**	(16, 16)	0.1817	(12, 15)
MAX($\mathrm{CA_{I(n_e)C2}}$, $\mathrm{CAU_{PL2}}$)	0.2551	(26, 7)	0.1660	(10, 21)	0.1825	(8, 17)
MAX($\mathrm{CA_{BM25}}$, $\mathrm{CAU_{PL2}}$)	**0.2633**	(26, 8)	**0.1680**	(11, 19)	**0.1875**	(9, 16)

By comparing the results in Tables 1 and 2, we can see that there is room for improvement between the uniform combination of evidence and an appropriate combination of sources of evidence on a per-query basis. Based on this conclusion, we will introduce in the next section a simple decision mechanism for selecting an appropriate combination of evidence on a per-query basis. In our experiments, we will employ the weighting scheme PL2, which has been shown to be effective for both used TREC topic distillation tasks. We will also report results from experiments with different weighting schemes.

3 Decision Mechanism

We have shown in the previous section that not all queries benefit equally from the same combination of evidence. In this section, we introduce a decision mechanism that aims to apply an appropriate combination of evidence for each query. For example, we employ content-only retrieval for specific queries, while we use evidence from the hyperlinks, or the URLs, for more generic queries. The decision mechanism is based on the query scope, which addresses two important statistical aspects of the set of retrieved documents.

The first aspect, *query_extent*, is related to the number of retrieved documents. We assume that for the more generic queries, there will be many documents that contain all the query terms. In these cases, the queries address a topic that is widely covered in the collection. Therefore, evidence from hyperlink analysis may be more useful in detecting high quality documents, or homepages of relevant sites. The *query_extent* is the number of retrieved documents $\{d_i\}$ that contain all the query terms, normalised between 0 and 1 by dividing with a given percentage α of the total number of documents in the test collection:

$$query\_extent = \min\left(\frac{|\{d_i | d_i \text{ contains all query terms}\}|}{\alpha \cdot N}, 1\right) \tag{3}$$

where N is the number of documents in the collection ($N = 1,247,553$ for the .GOV collection). The *query_extent* does not depend on the relevance scores assigned to documents. The normalisation is introduced as most of the queries retrieve only a small fraction of documents from the collection and therefore, dividing by α leads to a better distribution of the query scope values. The parameter α can take values from the range $(0, 1]$. Because for each query from both TREC11 and TREC12 topic distillation tasks, the retrieved documents that contain all the query terms are on average fewer than 1% of the collection size, we will only test α values less or equal to 0.01. For the remainder of the paper, the *query_extent* will be writen as $query\_extent_\alpha$, in order to denote the used value of α clearly. A similar measure, without the normalisation, has been introduced by Amitay et al. (2003), simultaneously to that defined by Plachouras et al. (2003).

For the second query scope aspect, *result_extent*, we take a different perspective and consider additional structural information. Indeed, hypertext and the Web encourage authors to organise documents in several different ways. First, documents are grouped in sites, where most of the documents cover either a specific topic, or a series of related topics. Within sites, documents are usually organised in a hierarchical directory structure. In addition, a document may correspond to more than one Web page. This is different from classical IR document collections, where each physical document constitutes a logical document (Eiron and McCurley 2003*a*). There have been different efforts towards the automatic identification of aggregates of hypertext, or Web pages. Botafogo and Shneiderman (1991) have employed a graph theoretic approach in order to identify aggregates. Differently, Eiron and McCurley (2003*a*), and Li et al. (2000), define heuristics based on observations of the structure of sites. In the context of TREC experiments, grouping documents according to their domain has been employed in order to limit the redundancy of retrieving many documents from a given site (Kwok et al. 2002).

We define *result_extent*, so that it indicates whether there are aggregates of documents, devoted to the query's topic. If there exist aggregates with a high number of documents containing all the query terms, we expect that their entry points will be more useful than other documents. We denote by $size_j$ the number of documents from the aggregate a_j. In addition, let μ_{size} and σ_{size} be the average and the standard deviation respectively of $size_j$ for $j \in [1, n]$, where n is the number of aggregates formed from the set of retrieved documents that contain at least one query term. We define the *result_extent* as the number of aggregates a_j for which $size_j$ is higher than $\mu_{size} + 2 \cdot \sigma_{size}$:

$$result\_extent = |\{a_j | size_j > \mu_{size} + 2 \cdot \sigma_{size}\}| \quad (4)$$

The *result_extent* does not depend on the relevance scores assigned to documents. We form the aggregates in two different ways. First, a coarse-grained approach is adopted, where an aggregate is formed by grouping all the retrieved documents that belong to the same domain. In this case, *result_extent* is denoted as $result\_extent(domains)$. Alternatively, we take a more granular approach and define aggregates as groups of documents from the same directory. In this case, *result_extent* is denoted as $result\_extent(directories)$.

After defining *query_extent* and *result_extent*, we introduce a simple decision mechanism, which uses either of the two query scope aspects, to select an appropriate combination of evidence on a per-query basis. In order to evaluate the effectiveness of each measure separately, we employ a simple approach, where the value of one of the measures, computed for the retrieved documents, is compared to a threshold. According to the result of this comparison, we assign one of the available combinations of

Table 4. Algorithm SelectCombinationOfEvidence.

```
Algorithm SelectCombinationOfEvidence
Input:  The query $q$ under consideration,
        the combinations of evidence $E1$ and $E2$,
        the query scope aspect $qscope$ of the query and
        a threshold value $t$
Output: The set of retrieved documents, ranked according
        to the selected approach
Method:
if $qscope(q) \leq t$ then
    apply combination of evidence $E1$ for query $q$
else
    apply combination of evidence $E2$ for query $q$
```

evidence to the query. In this work, we consider two possible combinations of evidence, the selection of which is initially based on thresholds obtained by using relevance information. It should be noted that this selection mechanism can be extended, in order to use more query scope aspects and combinations of evidence. For example, we can select from more than two combinations of evidence, or employ both query scope aspects in the selection process. The selection mechanism is shown in Table 4.

For each query, we compute the value of the query scope aspect *qscope*, given the set of retrieved documents and then, we compare its value to the threshold t. If $qscope(q) \leq t$ then we select $E1$ as the most appropriate combination of evidence for query q, otherwise we select $E2$. Note that the order in which specific combinations are assigned to $E1$ and $E2$ may significantly affect the effectiveness of the selection mechanism. This order should be consistent with the basic assumptions underlying the employed query scope aspect and each of the combinations of evidence. If there is an inconsistency, then we should expect a detrimental effect in the retrieval effectiveness.

In the remainder of the paper, we experiment and evaluate the effectiveness of the decision mechanism SelectCombinationOfEvidence, or SCE for brevity, in two steps. For the first step, Sections 4 and 5 contain the results from experiments, where we test *query_extent* and *result_extent* for different threshold values, under the assumption that there exists relevance information. The second step involves a method for setting the thresholds automatically, without relevance information (Section 6).

4 Evaluation of the Decision Mechanism

In this section, we aim to evaluate how effective the decision mechanism and each of *query_extent* and *result_extent* are in identifying appropriate combinations of evidence for each query. For each TREC topic distillation task, we use the two most effective combinations of evidence, according to the evaluation presented in Table 1. For TREC11, we employ $\mathrm{C}_{\mathrm{PL2}}$ and $\mathrm{CA}_{\mathrm{PL2}}$, while for TREC12, we use $\mathrm{CA}_{\mathrm{PL2}}$ and $\mathrm{CAU}_{\mathrm{PL2}}$. We test each of $query\_extent_{0.01}$, $result\_extent(domains)$ and $result\_extent(directories)$ for a range of 100 threshold values, from each measure's minimum value, to its maximum value, as computed for the queries[3]. Because the values of the different query scope aspects are not in the same range, the threshold values in subsequent figures will be normalised between 0 and 1. This linear transformation of the threshold values does not affect the results. The normalised values are also reported in all corresponding tables, in order to facilitate reading the figures.

The results from the experiments on TREC11 data are shown in Figure 1 and Table 5. Both measures based on the *result_extent* are more effective, with $result\_extent(domains)$ achieving 0.2796 average precision at 10, when the threshold values are in the range $[20.04, 23.88]$. The threshold values in this range, which result in the highest average precision at 10, correspond to the 11% of the tested threshold values, showing that $result\_extent(domains)$ can give stable improvements. According to Fisher's exact test, which determines if there is any non-random association between categorical variables, when the threshold t is in the range $[20.04, 23.88]$, the decision mechanism that uses $result\_extent(domains)$, selects the most appropriate combination of evidence for a statistically significant number of queries ($p = 0.0294$), for which there is a difference in C and CA's performance. This means that the queries for which $\mathrm{C}_{\mathrm{PL2}}$ is more effective than $\mathrm{CA}_{\mathrm{PL2}}$ are more likely to correspond to low values of $result\_extent(domains)$ and *vice versa*, enabling us to use $result\_extent(domain)$ to select an appropriate combination of evidence. In Figure 1, and in subsequent figures, the range of threshold values, where there is such a significant association, is marked with a bold line.

From Figure 1 and Table 5, we can also see that the decision mechanism, which uses $query\_extent_{0.01}$, performs as well as $\mathrm{C}_{\mathrm{PL2}}$ only when the threshold $t = 1$. We believe that this is due to an inconsistency between the assumption underlying *query_extent* and the two combinations of evidence, $\mathrm{C}_{\mathrm{PL2}}$ and $\mathrm{CA}_{\mathrm{PL2}}$. If we swap the combinations of evidence in the decision mechanism, so that $\mathrm{CA}_{\mathrm{PL2}}$ is applied

[3] If the minimum and maximum values of a query scope aspect, as computed for the queries, are *min* and *max* respectively, the n-th tested threshold value is given by $min + (n-1)\frac{max-min}{99}$.

when $qscope(q) \leq t$, otherwise $\mathrm{C_{PL2}}$ is applied, we are able to improve precision at 10 over that of $\mathrm{C_{PL2}}$, as discussed by Plachouras et al. (2004).

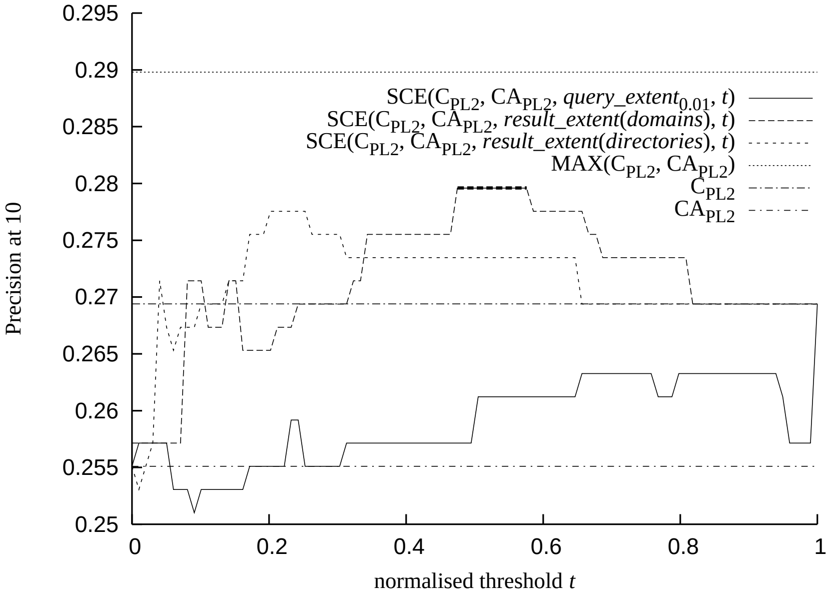

Figure 1. Evaluation of the decision mechanism for the TREC11 topic distillation task. The evaluation measure is precision at 10.

Table 5. The highest precision at 10, the corresponding thresholds for the TREC11 topic distillation task, as shown in Figure 1, and the % of threshold values, for which we obtain the highest improvement over the baseline. The decision mechanism is SCE(q, $\mathrm{C_{PL2}}$, $\mathrm{CA_{PL2}}$, $extent$, t).

Query Scope	Precision at 10	Norm. Threshold	Threshold	%
$\mathrm{C_{PL2}}$ Baseline	0.2694	–	–	
MAX($\mathrm{C_{PL2}}$, $\mathrm{CA_{PL2}}$)	0.2898	–	–	
$query\_extent_{0.01}$	0.2694	1.000	1.000	1%
$result\_extent(domains)$	0.2796	[0.475, 0.576]	[20.04, 23.88]	11%
$result\_extent(directories)$	0.2776	[0.202, 0.253]	[31.49, 38.87]	6%

We obtain similar improvements, when we experiment with the topic distillation data from TREC12 (see Figure 2 and Table 6). More specifically, $query\_extent_{0.01}$ outperforms the baseline CAU and results in 0.1460 average precision at 10 for the threshold range $[0.097, 0.137]$ (5% of the tested threshold values). The corresponding decision mechanism selects the most effective combination of evidence for a statistically significant number of queries, for the threshold ranges $[0.097, 0.137]$ ($p = 0.0472$), $[0.438, 0.498]$ ($p \leq 0.0231$) and $[0.599, 0.639]$ ($p = 0.0140$), as we can see from the bold lines in Figure 2, even though the last two ranges of thresholds do not correspond to the highest precision at 10. In addition, when we use *result_extent*(*domains*), the highest average precision at 10 is 0.1480 for the threshold ranges $[8.17, 9.61]$ and $[12.48, 13.91]$. These ranges correspond to the 6% of the tested threshold values. From Figure 2, Fisher's exact test shows that for the threshold range $[7.45, 13.91]$ the output of the decision mechanism is statistically significant, with $p \leq 0.0419$. This range of threshold values includes values, which do not result in the highest precision at 10. When *result_extent*(*directories*) is used, the highest average precision at 10 is 0.1400, which is equal to that of CAU.

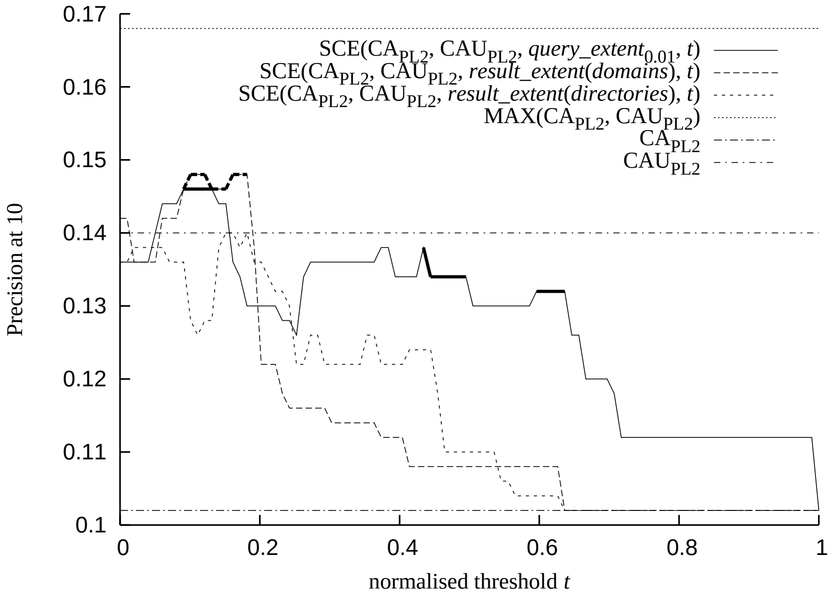

Figure 2. Evaluation of the decision mechanism for the TREC12 topic distillation task. The evaluation measure is precision at 10.

Table 6. The highest precision at 10, the corresponding thresholds for the TREC12 topic distillation task, as shown in Figure 2, and the % of threshold values, for which we obtain the highest improvement over the baseline. The decision mechanism is SCE(q, CA$_{\text{PL2}}$, CAU$_{\text{PL2}}$, *extent*, t).

Query Scope	Precision at 10	Norm. Threshold	Threshold	%
CAU$_{\text{PL2}}$ Baseline	0.1400	–	–	
MAX(CA$_{\text{PL2}}$, CAU$_{\text{PL2}}$)	0.1680	–	–	
query_extent$_{0.01}$	0.1460	[0.091, 0.131]	[0.097, 0.137]	5%
result_extent(*domains*)	0.1480	[0.101, 0.121], [0.161, 0.181]	[8.17, 9.61], [12.48, 13.91]	6%
result_extent(*directories*)	0.1400	[0.152, 0.162]	[19.33, 20.56]	3%

For the TREC12 experiments, we also consider R-Precision as the evaluation measure (Figure 3 and

Table 7). All three query scope aspects result in improvements over the baseline CAU for a wide range of threshold values. The highest average R-Precision is 0.1701, obtained with $query\_extent_{0.01}$ for the threshold range $[0.378, 0.388]$. Following, $result\_extent(domains)$ results in 0.1612 R-Precision for the threshold range $[8.17, 8.89]$. In both cases, the association between the relative performance of CA and CAU, and the output of the decision mechanism that uses $query\_extent_{0.01}$ or $result\_extent(domains)$, is non-random, according to Fisher's exact test, with p equal to 0.0414 and 0.0403 respectively. These results were obtained for 2% of the tested threshold values (Figure 3). When $result\_extent(directories)$ was employed, we obtained 0.1539 R-Precision for only 1% of the tested threshold values.

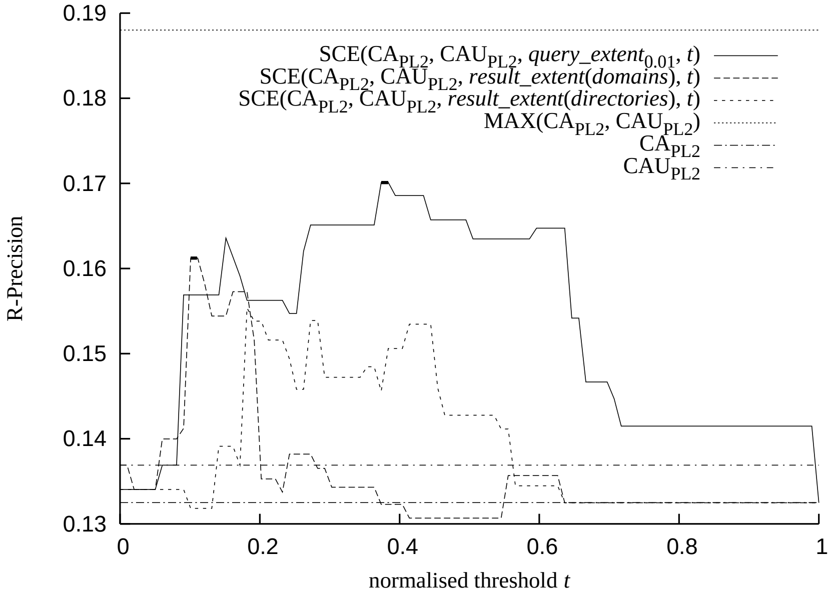

Figure 3. Evaluation of the decision mechanism for the TREC12 topic distillation task. The evaluation measure is R-Precision.

Table 7. The highest R-Precision, the corresponding thresholds for the TREC12 topic distillation task, as shown in Figure 3, and the % of threshold values, for which there is an improvement over the baseline. The decision mechanism is SCE(q, CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, $extent$, t).

Query Scope	R-Precision	Norm. Threshold	Threshold	%
CAU$_{\mathrm{PL2}}$ Baseline	0.1369	–	–	
MAX(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$)	0.1880	–	–	
$query\_extent_{0.01}$	0.1701	[0.374, 0.384]	[0.378, 0.388]	2%
$result\_extent(domains)$	0.1612	[0.101, 0.111]	[8.17, 8.89]	2%
$result\_extent(directories)$	0.1539	[0.273, 0.283]	[34.00, 35.22]	1%

In this section, we have investigated the effectiveness of both *query_extent* and *result_extent* for the TREC11 and TREC12 topic distillation tasks. We have shown that improvements are obtained for wide ranges of thresholds. Moreover, according to Fisher's exact test, we have found that there is a statistically significant association between the output of the decision mechanism, which employs either *query_extent* or *result_extent*, and the relative effectiveness of the employed combinations of evidence. Thus, the two query scope aspects are successful in capturing how broad a query is, and the decision mechanism can effectively identify an appropriate combination of evidence for each query.

5 Additional experiments with the Decision Mechanism

So far, we have shown that the decision mechanism that employs either *query_extent* or *result_extent* improves the retrieval effectiveness. In this section, we focus our analysis on how different parameters of the decision mechanism, such as the normalisation of the *query_extent*, or employing different weighting schemes, affect its effectiveness. The remainder of this section is organised as follows. Section 5.1 contains experiments, where we examine the effect of using different α values for normalising *query_extent*. In Section 5.2, we test the effect of using different weighting schemes for each combination of evidence. Section 5.3 contains experiments, where we show how the query scope aspect values change when we use both content and anchor text for their calculation.

5.1 Normalisation of query extent

In this section, we focus on *query_extent* and the effect of the normalising parameter α. So far, we have set $\alpha = 0.01$. In the following experiments, we set α equal to 0.001, 0.002, 0.005 in addition to 0.01, and observe the effect on the performance of our decision mechanism. Figure 4 and Table 8 show the results for the TREC11 topics. Moreover, Figure 5 and Table 9 contain the results for the TREC12 topics, when precision at 10 is employed for the evaluation, and Figure 6 and Table 10 contain the results for the TREC12 topics, considering R-Precision as the evaluation measure.

For the TREC11 experiments (Table 8), we can see that, for all the tested values of α, the highest obtained average precision at 10 is 0.2694 for the threshold $t = 1$. However, the effect of changing the value of α is demonstrated in Figure 4 by the *shifting* of the curves corresponding to different α values. More specifically, when we decrease the value of α, the ratio $\frac{|\{d_i | d_i \text{ contains all query terms}\}|}{\alpha \cdot N}$ in Equation (3)

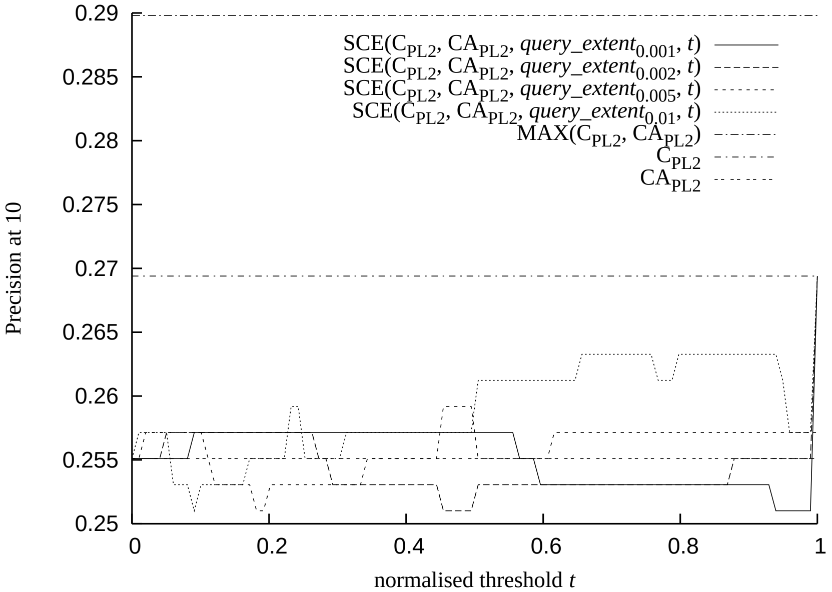

Figure 4. Evaluation of the decision mechanism and *query_extent* for the TREC11 topic distillation task, using different values for the normalising parameter α. The evaluation measure is precision at 10.

increases and therefore, *query_extent* becomes more sensitive to specific queries.

This is more evident in the TREC12 experiments, for both precision at 10 (Figure 5 and Table 9) and R-Precision (Figure 6 and Table 10). We observe that the shift of the optimal thresholds is approximately reciprocal to the ratio between the different α values used. Indeed, we can see from Table 5 that when $\alpha = 0.01, 0.005$ and 0.002 respectively, the middle values of the threshold ranges are 0.117, 0.233 and 0.585, from which we have:

$$\frac{0.01}{0.005} = 2 \approx \frac{0.233}{0.117}, \qquad \frac{0.01}{0.002} = 5 \approx \frac{0.585}{0.117}$$

The shifting is also reflected on the threshold ranges, for which there is a statistically significant association between the relative effectiveness of CA and CAU, and the output of SCE(q, CA$_{\mathrm{PL2}}$, *query_extent*$_\alpha$, t), as shown by the bold parts of the lines in Figure 5. This is expected because α is in the denominator of the ratio $\frac{|\{d_i | d_i \text{ contains all query terms}\}|}{\alpha \cdot N}$ in Equation (3).

Similar results are obtained when we use R-Precision for the evaluation (Figure 6 and Table 10). Additionally, we can see the effect of selecting unsuitable parameter values. As α decreases, the highest

Table 8. The highest precision at 10, the corresponding thresholds for the TREC11 topic distillation task, as shown in Figure 4, and the % of threshold values, for which there is an improvement over the baseline. The decision mechanism used is SCE(q, $\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$, $query\_extent_{\alpha}$, t). The last four rows of the table are equivalent because we report the highest obtained precision at 10.

Query Scope	Precision at 10	Norm. Threshold	Threshold	%
$\mathrm{C}_{\mathrm{PL2}}$ Baseline	0.2694	–	–	
MAX($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$)	0.2898	–	–	
$query\_extent_{0.001}$	0.2694	1.000	1.000	0%
$query\_extent_{0.002}$	0.2694	1.000	1.000	0%
$query\_extent_{0.005}$	0.2694	1.000	1.000	0%
$query\_extent_{0.01}$	0.2694	1.000	1.000	0%

Table 9. The highest precision at 10, the corresponding thresholds for the TREC12 topic distillation task, as shown in Figure 5, and the % of threshold values, for which there is an improvement over the baseline. The decision mechanism is SCE(q, $\mathrm{CA}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$, $query\_extent_{\alpha}$, t).

Query Scope	Precision at 10	Norm. Threshold	Threshold	%
$\mathrm{CAU}_{\mathrm{PL2}}$ Baseline	0.1400	–	–	
MAX($\mathrm{CA}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$)	0.1680	–	–	
$query\_extent_{0.001}$	0.1460	[0.949, 0.990]	[0.953, 0.991]	5%
$query\_extent_{0.002}$	0.1460	[0.465, 0.677]	[0.482, 0.688]	22%
$query\_extent_{0.005}$	0.1460	[0.182, 0.192]	[0.193, 0.272]	9%
$query\_extent_{0.01}$	0.1460	[0.091 ,0.131]	[0.097, 0.137]	5%

precision decreases as well, because *query_extent* becomes more sensitive to specific queries, and the decision mechanism applies $\mathrm{CAU}_{\mathrm{PL2}}$ for more queries, even though $\mathrm{CA}_{\mathrm{PL2}}$ may be more effective.

Table 10. The highest R-Precision, the corresponding thresholds for the TREC12 topic distillation task, as shown in Figure 6, and the % of threshold values, for which there is an improvement over the baseline. The decision mechanism is SCE(q, $\mathrm{CA}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$, $query\_extent_{\alpha}$, t).

Query Scope	R-Precision	Norm. Threshold	Threshold	%
$\mathrm{CAU}_{\mathrm{PL2}}$ Baseline	0.1369	–	–	
MAX($\mathrm{CA}_{\mathrm{PL2}}$, $\mathrm{CAU}_{\mathrm{PL2}}$)	0.1880	–	–	
$query\_extent_{0.001}$	0.1569	[0.949, 0.990]	[0.953, 0.991]	5%
$query\_extent_{0.002}$	0.1635	[0.758, 0.778]	[0.766, 0.785]	3%
$query\_extent_{0.005}$	0.1701	[0.747, 0.768]	[0.751, 0.771]	3%
$query\_extent_{0.01}$	0.1701	[0.374, 0.384]	[0.378, 0.388]	2%

Overall, we have seen that the normalisation of *query_extent* by dividing with α affects the decision mechanism's effectiveness. Employing an inappropriate value may lead to a decision mechanism with reduced potential for improvement. Therefore, in an operational environment, it is important to select an appropriate value for α, so that the highest possible retrieval effectiveness is obtained within the threshold range and the decision mechanism is effective for a wider range of threshold values. We believe

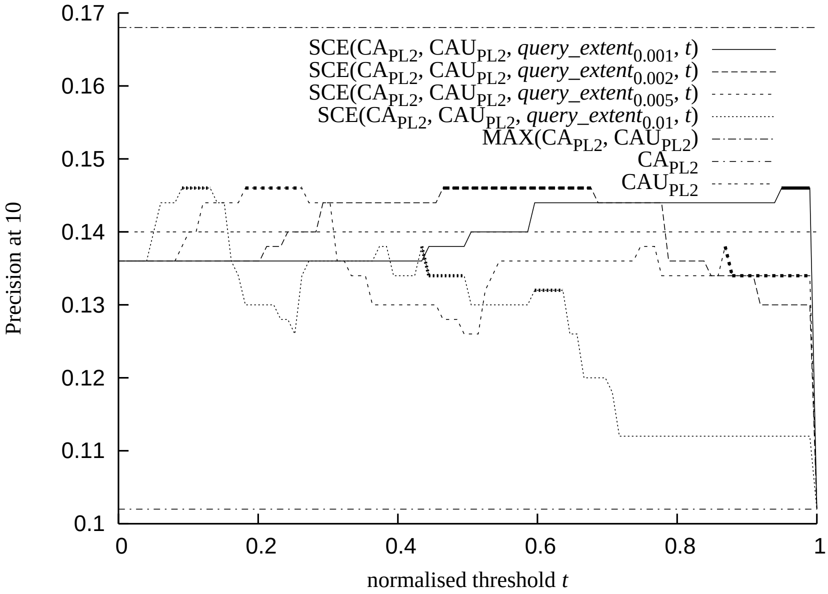

Figure 5. Evaluation of the decision mechanism and *query_extent* for the TREC12 topic distillation task, using different values for the normalising parameter α. The evaluation measure is precision at 10.

that a reasonable value for the normalising parameter should not be significantly higher than the average number of documents containing all query terms.

5.2 Employing different weighting schemes

In our experiments, we have used the same weighting schemes for the different combinations of evidence. However, in Section 2, we saw that using different weighting schemes for each combination of evidence may result in even higher retrieval effectiveness (see Table 3).

In the following experiments, we employ different weighting schemes for both TREC11 and TREC12. More specifically, we show that if we use $\mathrm{C}_{\mathrm{PL2}}$ and $\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$, with *result_extent*(*domains*) for TREC11, we get improvements over the highest average precision obtained when we employ only PL2 for both C and CA (0.2837, as shown in Table 11, with respect to 0.2796 average precision at 10, as shown in Table 5). Both *query_extent*$_{0.01}$ and *result_extent*(*directories*) are also more effective than when they are used with $\mathrm{C}_{\mathrm{PL2}}$ and $\mathrm{CA}_{\mathrm{PL2}}$. When we employed other combinations of weighting schemes, for both tested tasks, the correspondence between the query scope aspect values and the relative effectiveness of the combinations of evidence was weak, and the decision mechanism did not perform better than the

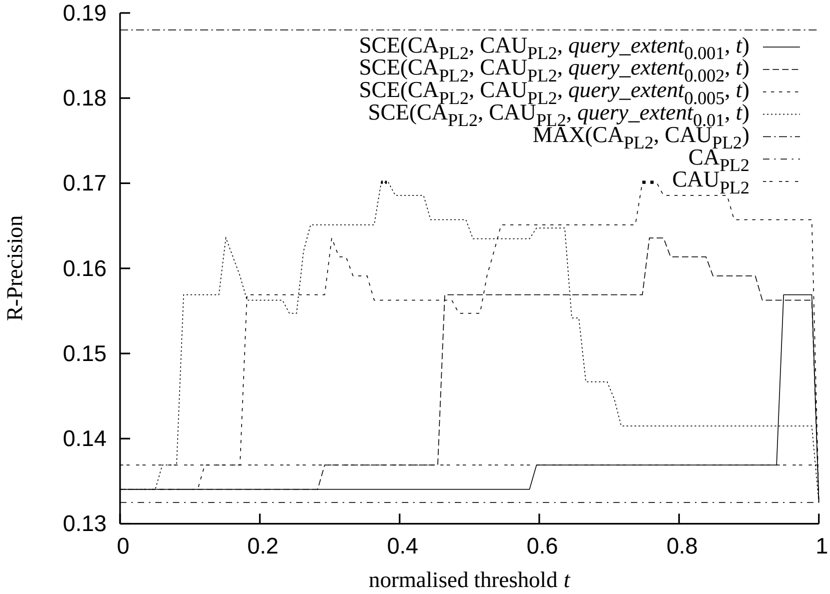

Figure 6. Evaluation of the decision mechanism and *query_extent* for the TREC12 topic distillation task, using different values for the normalising parameter α. The evaluation measure is R-Precision.

baselines.

Table 11. The highest precision at 10, the corresponding thresholds for the TREC11 topic distillation task, as shown in Figure 7, and the % of threshold values, for which there is an improvement over the baseline. The decision mechanism used is SCE(q, C_{PL2}, $CA_{I(n_e)C2}$, *extent*, t).

Query Scope	Precision at 10	Norm. Threshold	Threshold	%
C_{PL2} Baseline	0.2694	–	–	
MAX(C_{PL2}, $CA_{I(n_e)C2}$)	0.2938	–	–	
query_extent$_{0.01}$	0.2714	[0.232, 0.242]	[0.237, 0.247]	2%
result_extent(*domains*)	**0.2837**	[0.475, 0.576]	[20.04, 23.88]	11%
result_extent(*directories*)	0.2796	[0.263, 0.303]	[40.34, 46.24]	5%

These experiments, where we employed different weighting schemes, show that significant improvements can be obtained when we use the weighting schemes PL2 and I(n_e)C2. Further experiments are required in order to establish a relation between the independence of weighting schemes and their effectiveness in the context of the decision mechanism.

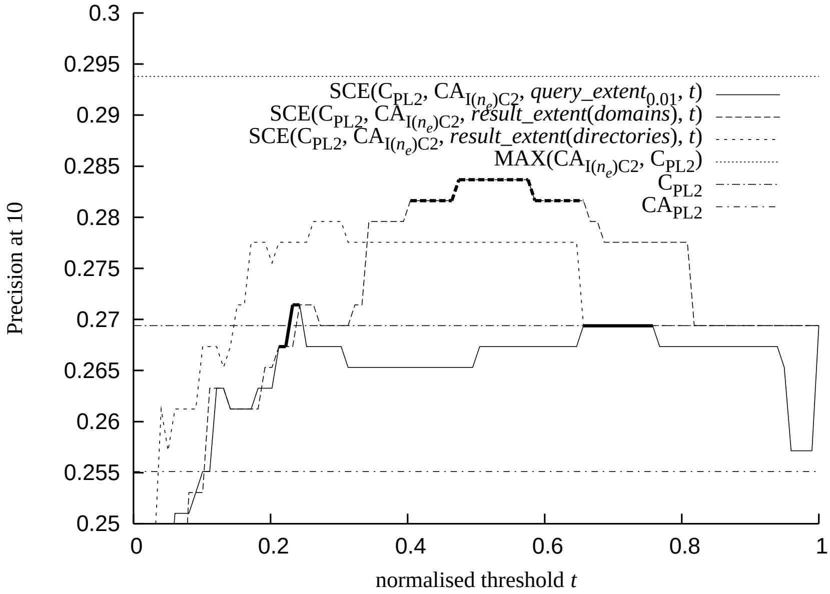

Figure 7. Evaluation of the decision mechanism for the TREC11 topic distillation task, using different weighting schemes. The evaluation measure is precision at 10.

5.3 Using content and anchor text for computing the query scope

The values of the query scope aspects depend only on the retrieved documents for each query. For our last experiments, we investigate how the values of the query scope aspects change if we use a different document representation, where the document is extended by adding the anchor text of its incoming hyperlinks. Indeed, employing the anchor text is an effective retrieval approach for finding entry points of sites (Craswell et al. 2001, Craswell et al. 2003).

In Table 12, we show that the mean values of the query scope aspects for both TREC11 and TREC12 tasks, computed using either document representations, are similar. According to the p values of Wilcoxon's signed rank test for paired samples, which determines if the medians of two distributions are different, there is no significant change in the values of $result\_extent(domains)$ for the two tested document representations. On the other hand, the values of $query\_extent_{0.01}$ are significantly different when anchor text is used, in addition to the content of documents. Moreover, the values of $result\_extent(directories)$ for the two document representations are significantly different only for the TREC12 topic distillation queries. These differences are due to the fact that when we employ content

Table 12. The effect of using different document representations on the values of the query scope aspects and the % of thresholds, for which there is a difference in the average retrieval effectiveness of the decision mechanism. The level of the difference is shown in the parenthesis.

	Mean		p values	% thresholds	
	C	CA		Precision at 10	R-Precision
			TREC11		
$query\_extent_{0.01}$	0.3386	0.3390	$p < 10^{-4}$	0%	-
$result\_extent(domains)$	13.43	13.43	$p = 1.000$	0%	-
$result\_extent(directories)$	27.10	27.68	$p < 10^{-4}$	1% ($\leq$ 0.0143)	-
			TREC12		
$query\_extent_{0.01}$	0.3827	0.3832	$p < 10^{-4}$	3%($\leq$ 0.0020)	3%($\leq$ 0.0029)
$result\_extent(domains)$	14.96	14.98	$p = 0.750$	3%($\leq$ 0.0060)	3%($\leq$ 0.0044)
$result\_extent(directories)$	30.08	30.22	$p = 0.031$	6%($\leq$ 0.0060)	5%($\leq$ 0.0075)

and anchor text, the number of documents with all query terms can only increase.

We compare the effect of the different document representations on the decision mechanism. The setting of the decision mechanism is the same as in Section 4. Moreover, we test the same threshold values for the query scope aspects, for both document representations. We have found that the highest obtained retrieval effectiveness remains the same, independently of the used document representations (the highest obtained values are the same as in Tables 5, 6 and 7). From the last two columns of Table 12, we can see that the average retrieval effectiveness of the decision mechanism is the same when we use $query\_extent_{0.01}$ or $result\_extent(domains)$ for TREC11, and it changes for 3% of the tested thresholds for TREC12. When we use $result\_extent(directories)$ for TREC11, there is a difference for 1% of the tested thresholds. For TREC12, the average precision at 10 of the decision mechanism with $result\_extent(directories)$ changes for 6% of the tested thresholds. If we consider R-Precision for TREC12, there is a change in the average effectiveness of the decision mechanism for 5% of the tested thresholds.

These small changes occur because there are only few documents for which the query terms do not appear in their body, but appear only in the anchor text of their incoming hyperlinks. For the used .GOV collection, one possible explanation is that this collection contains high quality documents from controlled sites. If we use a less controlled Web test collection, where there is greater variability in the anchor text, we may find that using the content of documents and their anchor text for computing $query\_extent$ and $result\_extent$ might have an impact on their values.

6 Setting thresholds without relevance information

In the first step of the decision mechanism's evaluation, we have employed relevance information, in order to choose the two most effective combinations of evidence for each task, and to select the threshold ranges for which we obtain improvements in effectiveness. However, in an operational environment, where relevance information is not readily available, we should be able to select the combinations of evidence $E1$ and $E2$ used in the decision mechanism of Table 4, as well as automatically set the thresholds to values, which would result in improved retrieval effectiveness.

For selecting the most effective combinations of evidence to use in the decision mechanism, it is reasonable to assume that we can perform experiments with a set of training queries. From the results, we could obtain the more appropriate combinations of evidence for the queries that users submit to our retrieval system.

In order to predict appropriate threshold values, we employ an approach, based on the sampling method introduced by Cronen-Townsend et al. (2002). We approximate the process of querying, by sampling terms from the vocabulary of the test collection, and submitting them to our retrieval system as single-term queries. The intuition for sampling single-terms to set the decision mechanism's threshold for multi-term queries is that in both cases we use the same vocabulary for querying and therefore, we expect to obtain a similar distribution of query scope aspect values. From the set of retrieved documents, we compute the values of $query\_extent$ and $result\_extent$. In our experiments, we have sampled the terms with document frequency in the range $[500, 40000]$, in order to obtain approximately the number of retrieved documents with all query terms for the TREC11 and TREC12 topic distillation queries.

After sampling the terms from the vocabulary, we estimate the probability densities for the number of documents containing a term, $result\_extent(domains)$ and $result\_extent(directories)$. We employ kernel density estimation with Gaussian kernels and automatic setting of the bandwidth (Scott 1992). The estimated probability densities for the number of documents containing all the query terms (in this case, this corresponds to the probability density of the document frequency), $result\_extent(domains)$ and $result\_extent(directories)$ are shown in Figure 8. Because $query\_extent$, as defined in Equation (3), is not a continuous function of the number of documents that contain all the query terms, the estimation of its probability density would not be precise for values close to 1. Therefore, we have estimated the probability density for the document frequency and we compute the values of $query\_extent$ from the number of documents containing a term by using Equation (3). We set $\alpha = 0.01$, which is not significantly higher than the average percentage of documents containing all the query terms for the TREC11 and TREC12 topic distillation queries.

From the estimated probability densities, we compute the threshold values at the points where n percent of the values are lower than the threshold, for $n = 25\%, 50\%, 75\%$. The values for the estimates $query\_extent_{0.01}$, and $result\_extent$ for both domains and directories are shown in Table 13.

Table 13. The thresholds at given points of the estimated probability densities, shown in Figure 8.

Percent	Documents	$query\_extent_{0.01}$	$result\_extent(domains)$	$result\_extent(directories)$
25%	795.66	0.064	4.61	8.01
50%	1407.70	0.113	9.02	15.54
75%	3549.86	0.285	17.09	31.57

Next, we employ the estimated threshold values in the decision mechanism. Table 14 contains the results for the TREC11 topics, while Table 15 contains the results for the TREC12 topics, using either precision at 10 or R-Precision for the evaluation. We can see that improvements over the baselines (C for TREC11 and CAU for TREC12) are obtained, when the thresholds are set automatically. Furthermore, the entry in bold shows that the association between the relative effectiveness of the different combinations of evidence and the output of the decision mechanism for the corresponding thresholds is statistically significant.

More specifically, we can see from Table 14 that improvements are obtained for all query scope aspects, and especially for the higher threshold values of $result\_extent(domains)$ and $result\_extent(directories)$. According to Fisher's exact test, the decision mechanism SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$, $result\_extent(domains)$, t) with the corresponding 75% threshold, selects the most effective combination of evidence for a statistically significant number of queries ($p = 0.0329$).

Table 14. TREC11 Precision at 10 with automatically set thresholds. Entries in bold correspond to the cases where Fisher test is statistically significant.

	Precision at 10		
	25%	50%	75%
SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$, $query\_extent_{0.01}$, t)	0.2530	0.2530	0.2551
SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$, $result\_extent(domains)$, t)	0.2571	0.2653	0.2755
SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{PL2}}$, $result\_extent(directories)$, t)	0.2673	0.2694	0.2776
SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$, $result\_extent(domains)$, t)	0.2388	0.2653	**0.2816**
SCE($\mathrm{C}_{\mathrm{PL2}}$, $\mathrm{CA}_{\mathrm{I}(n_e)\mathrm{C2}}$, $result\_extent(directories)$, t)	0.2571	0.2673	0.2775

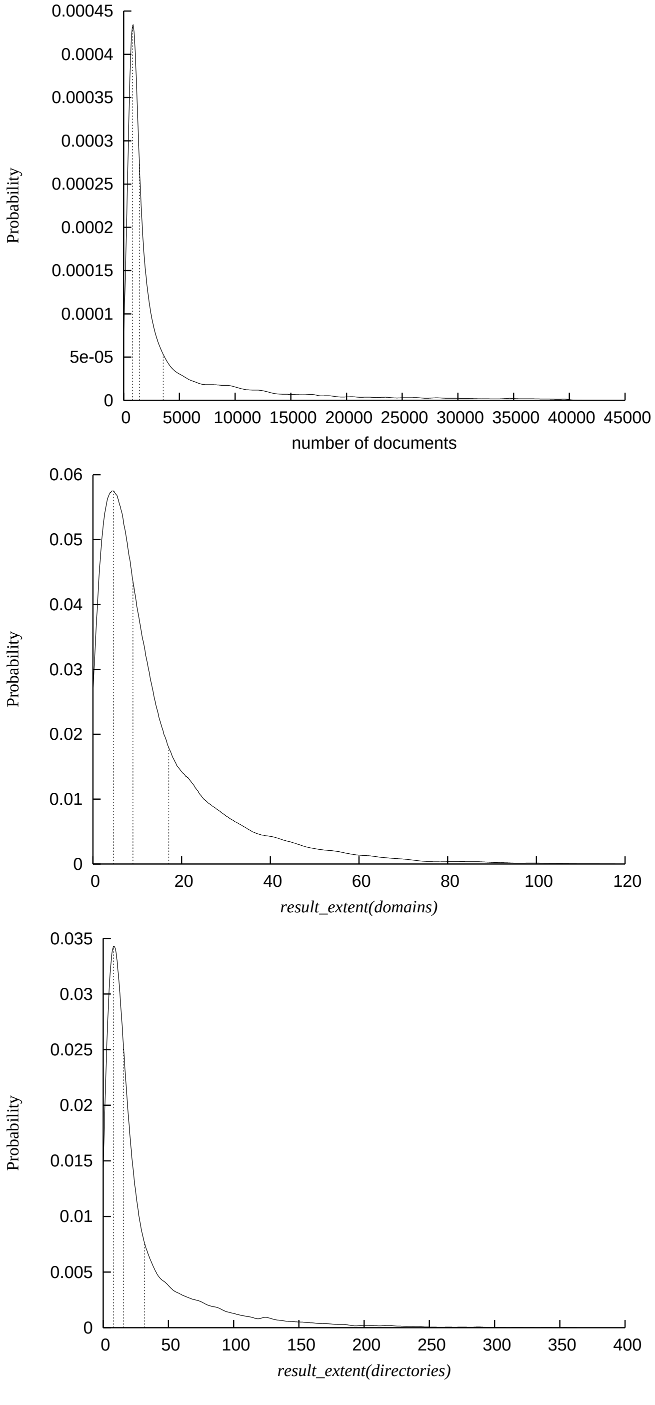

Figure 8. The estimated probability densities from the sampling of terms as queries.

The results for the TREC12 queries are similar. As shown in Table 15, we obtain improvements in precision at 10 over the baseline CAU for both *query_extent*$_{0.01}$ and *result_extent*(*domains*). According to Fisher's exact test, the decision mechanism that uses either *query_extent*$_{0.01}$ or *result_extent*(*domains*) with the 50% threshold, selects the most effective combination of evidence for a statistically significant number of queries ($p = 0.0472$ and $p = 0.0086$ respectively). We obtain the highest R-Precision when we use *result_extent*(*domains*) and the 50% threshold. Thus, *result_extent*(*domains*) proves to be more effective and stable than *query_extent* or *result_extent*(*directories*).

Table 15. TREC12 results with automatically set thresholds. Entries in bold correspond to the cases where Fisher test is statistically significant.

	Precision at 10		
	25%	50%	75%
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *query_extent*$_{0.01}$, t)	0.1440	**0.1460**	0.1360
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *result_extent*(*domains*), t)	0.1360	**0.1480**	0.1160
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *result_extent*(*directories*), t)	0.1380	0.1280	0.1220
	R-Precision		
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *query_extent*$_{0.01}$, t)	0.1369	0.1569	0.1651
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *result_extent*(*domains*), t)	0.1340	0.1582	0.1382
SCE(CA$_{\mathrm{PL2}}$, CAU$_{\mathrm{PL2}}$, *result_extent*(*directories*), t)	0.1340	0.1318	0.1472

Overall, we can say that the sampling of terms from the collection's vocabulary, in order to set the decision mechanism's thresholds automatically, without relevance information, is an effective approach. These results show that the selective combination of evidence is a method applicable in an operational environment. From all three query scope aspects we tested, we found that *result_extent*(*domains*) is the most effective one in the context of the decision mechanism with the automatically set thresholds.

For the experiments in this section, we have considered single terms from the document collection as queries. Even though we set the thresholds to a value, so that 75% of the estimated values are lower, this threshold value is quite low. Especially in the case of *query_extent*$_{0.01}$, the estimated threshold values are lower than the ones resulting in optimal precision (Table 7 and Table 13). We believe that sampling sets of related terms from the collection's vocabulary may result in a better approximation of the querying process. Such an approach would require measuring the association between terms (Silverstein et al. 1999), or using the anchor text of documents (Eiron and McCurley 2003*b*).

7 Discussion

We have shown that the decision mechanism from Section 3, when used with the proposed estimates of how specific or broad a query is, increases retrieval effectiveness for both TREC11 and TREC12 topic distillation tasks.

Comparing the effectiveness of $result\_extent(domains)$ with $result\_extent(directories)$, we can see that the decision mechanism is more effective for both TREC tasks, when domains are used for forming the aggregates. A reason for this is that there may exist many directories that contain few documents, thus biasing the average directory size towards lower values, and resulting in a relatively high value for $result\_extent(directories)$. Further research is needed in order to employ more granular approaches for detecting aggregates of documents (Eiron and McCurley 2003*a*, Li et al. 2000).

We have found that $result\_extent(domains)$ has other useful properties. First, its values do not change significantly when they are computed using a different document representation, such as content and anchor text. In addition, the average retrieval effectiveness of the decision mechanism changes only for a small percentage of thresholds when we use $result\_extent(domains)$ with either document representations. When we employ it in the decision mechanism, we obtain the highest precision for a wide range of consecutive thresholds. In addition, when the decision mechanism's thresholds are set automatically, it is effective for both tested TREC tasks. Therefore, we believe that $result\_extent(domains)$ is a robust query scope aspect, which can be used in a realistic setting.

An interesting generalisation of our methodology would be to investigate the automatic selection of specific combinations of evidence to use with each query scope aspect. Suppose that there is no relevance information available. Instead of comparing the relative retrieval effectiveness between two different combinations of evidence, we could compute a measure of distance between the corresponding rankings of documents. If the distance between the rankings is correlated with the difference between the query scope aspect value for a query, and the threshold value of the decision mechanism, then we would expect that this particular setting of the decision mechanism can be used effectively. In this way, we would obtain an estimation of how effectively we can use a query scope aspect and two different combinations of evidence, in the context of the decision mechanism.

8 Related Work

The combination of evidence from different sources has been employed in various ways in order to increase retrieval precision (Croft 2000). Belkin et al. (1993) used different query representations, while Turtle and Croft (1991) proposed a Bayesian inference network in order to combine different query and document representations. Similarly, Ribeiro-Neto and Muntz (1996) proposed a belief network model for the combination of different sources of evidence, including the hyperlink structure.

From a different perspective, Bartell et al. (1994) investigated the automatic combination of multiple retrieval approaches. They model the combination of evidence from individual "experts" as the linear combination of the individuals' estimates. Instead of combining the scores of different retrieval approaches linearly, Aslam and Montague (2001) proposed a method for fusing ranked lists of documents, an approach with applications in meta-search engines, where the scores of documents are not available. From the perspective of model selection, He and Ounis (2004) consider applying a different weighting model per query, by clustering a set of training queries and applying the same weighting model for all queries in the same cluster.

In the context of Web IR, recent research has focused on detecting when to employ additional evidence to the textual content of documents. Ogilvie and Callan (2003), and Fagin et al. (2003), investigate a variety of combinations of evidence in different settings. Other approaches focus on the relation between the effectiveness of hyperlink analysis and the density of hyperlinks in a test collection (Gurrin and Smeaton 2003, Fisher and Everson 2003).

There have also been proposed approaches, where the weight of additional evidence from either hyperlink analysis algorithms HITS or SALSA (Amitay et al. 2002), or other sources of evidence such as the anchor text and the number of incoming hyperlinks (Amitay et al. 2003), is set on a per-query basis, according to characteristics of the set of retrieved documents. Differently, our decision mechanism enables or disables a source of evidence on a per-query basis. This methodology can be seen as a linear combination of multiple retrieval approaches, where only one coefficient is non-zero. As a result, the query scope aspect values and the relative effectiveness of the combinations of evidence do not need to be strongly correlated. A second difference is that we employ evidence from the distribution of aggregates of documents, in order to model how broad or specific the queries are.

In addition, Kang and Kim (2003) propose a query type classification method. The query type corresponds to the retrieval task, and a different combination of evidence is applied for each query type. On the other hand, our methodology does not consider the query type explicitly, but aims to apply an appropriate combination of evidence on a per-query basis.

9 Conclusions

In this paper, we have presented a decision mechanism that enables the selective combination of evidence on a per-query basis, for Web IR, and more specifically, for topic distillation. The decision mechanism is based on two query scope aspects. The first one is the *query_extent*, which corresponds to the number of documents in the collection that contain all the query terms. The second aspect is the *result_extent*, which corresponds to the number of large aggregates in the set of retrieved documents. An aggregate of documents is a group of related documents, belonging to the same domain (*result_extent*(*domains*)), or stored in the same directory (*result_extent*(*directories*)).

We have shown that the decision mechanism and the query scope aspects can increase the retrieval effectiveness, compared to the uniform combination of evidence, irrespectively of the queries. Improvements over the baselines are obtained for the TREC11 topic distillation task, when we use the combinations C and CA with *result_extent*(*domains*), as well as for the TREC12 topic distillation task, when we employ CA and CAU with either *query_extent* or *result_extent*(*domains*). We have found that using domain-based aggregates is more effective than directory-based aggregates. The improvements in retrieval effectiveness correspond to the cases where there is a statistically significant association between the relative effectiveness of the different combinations of evidence and the output of the decision mechanism that uses the query scope aspects. Following, we have tested our methodology by automatically setting the threshold values, without using relevance information. Overall the most effective query scope aspect was found to be *result_extent*(*domains*), and improvements over the baselines were obtained, proving that our approach can be applied in an operational setting, where relevance information is not readily available.

In the future, we would like to experiment with larger test collections, such as the .GOV2, or with other, less controlled test collections, as they become available. Additionally, we would like to generalise our approach, by introducing a measure that estimates the effectiveness of the decision mechanism without relevance information. Experimentation in the context of different retrieval tasks, such as the filtering task, where relevance information becomes incrementally available, is another possible application of our approach.

Acknowledgements

The work of the first and third authors is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) project grant, number GR/R90543/01. The project funds the development of the Terrier Information Retrieval framework (URL: http://ir.dcs.gla.ac.uk/terrier).

The work of the second author has been partially supported by the Comisión Interministerial de Ciencia y Tecnologı́a (CICYT) of the Spanish government, under project TIC2001-0547 and by the Fundación Caixa Galicia (Beca curso 2002/2003 para Estudios de Postgrado en Universidades y en Centros de Investigación de Excelencia Académica).

References

Amati, G. and van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems* **20**(4), 357–389.

Amitay, E., Carmel, D., Darlow, A., Herscovici, M., Lempel, R., Soffer, A., Kraft, R. and Zien, J. (2003). Juru at TREC 2003 - Topic Distillation Using Query-Sensitive Tuning and Cohesiveness Filtering. In NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003).

Amitay, E., Carmel, D., Darlow, A., Lempel, R. and Soffer, A. (2002). Topic Distillation with Knowledge Agents. In NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC 2002).

Aslam, J. A. and Montague, M. (2001). Models for metasearch. In Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 276–284.

Bartell, B. T., Cottrell, G. W. and Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval. pp. 173–181.

Belkin, N. J., Cool, C., Croft, B. and Callan, J. P. (1993). The effect of multiple query representations on information retrieval system performance. In Proceedings of the 16th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 339–346.

Botafogo, R. A. and Shneiderman, B. (1991). Identifying aggregates in hypertext structures. In Proceedings of the third annual ACM conference on Hypertext. ACM Press. pp. 63–74.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum* **36**(2), 3–10.

Craswell, N. and Hawking, D. (2002). Overview of the TREC-2002 Web Track. In NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002). pp. 86–93.

Craswell, N., Hawking, D. and 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. ACM Press. pp. 250–257.

Craswell, N., Hawking, D., Wilkinson, R. and Wu, M. (2003). Overview of the TREC-2003 Web Track. In NIST Special Publication 500-255: The Twelfth Text REtrieval Conference (TREC 2003).

Croft, W. B. (2000). Combining approaches to information retrieval. In W. B. Croft, ed., Advances in Information Retrieval from the Center for Intelligent Information Retrieval. Kluwer Academic. pp. 1–36.

Cronen-Townsend, S., Zhou, Y. and Croft, W. B. (2002). Predicting query performance. In Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 299–306.

Eiron, N. and McCurley, K. (2003*a*). Untangling compound documents on the web. In Proceedings of the fourteenth ACM conference on Hypertext and Hypermedia. ACM Press. pp. 85–94.

Eiron, N. and McCurley, K. S. (2003*b*). Analysis of anchor text for web search. In Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 459–460.

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A. and Williamson, D. P. (2003). Searching the workplace web. In Proceedings of the twelfth international conference on World Wide Web. ACM Press. pp. 366–375.

Fisher, M. and Everson, R. (2003). When Are Links Useful? Experiments in Text Classification. In Advances in Information Retrieval: 25th European Conference on IR Research. Springer-Verlag. pp. 41–56.

Gurrin, C. and Smeaton, A. F. (2003). Improving the Evaluation of Web Search Systems. In Advances in Information Retrieval: 25th European Conference on IR Research. Springer-Verlag. pp. 25–40.

He, B. and Ounis, I. (2003). A study of parameter tuning for term frequency normalization. In Proceedings of the 12th international Conference on Information and Knowledge Management (CIKM). ACM Press. pp. 10–16.

He, B. and Ounis, I. (2004). A Query-based Pre-retrieval Model Selection Approach to Information Retrieval. In Proceedings of RIAO 2004 (Recherche d'Information Assistee par Ordinateur - Computer assisted information retrieval) on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval. CID. pp. 706–719.

Kang, I.-H. and Kim, G. (2003). Query type classification for web document retrieval. In Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 64–71.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms. ACM Press. pp. 668–677.

Kraaij, W., Westerveld, T. and 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. ACM Press. pp. 27–34.

Kwok, K. L., Deng, P., Dinstl, N. and Chan, M. (2002). TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS. In NIST Special Publication 500-251: The Eleventh Text Retrieval Conference (TREC 2002).

Li, W.-S., Kolak, O., Vu, Q. and Takano, H. (2000). Defining logical domains in a web site. In Proceedings of the eleventh ACM on Hypertext and hypermedia. ACM Press. pp. 123–132.

Ogilvie, P. and Callan, J. (2003). Combining document representations for known-item search. In Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 143–150.

Plachouras, V. and Ounis, I. (2004). Usefulness of hyperlink structure for query-biased topic distillation. In Proceedings of the 27th annual international ACM SIGIR conference on Research and Developement in Information Retrieval. ACM Press. pp. 448–455.

Plachouras, V., Cacheda, F., Ounis, I. and van Rijsbergen, C. J. (2003). University of Glasgow at the Web track: Dynamic Application of Hyperlink analysis using the Query Scope. In NIST Special Publication 500-255: The Twelfth Text Retrieval Conference (TREC 2003).

Plachouras, V., Ounis, I. and Cacheda, F. (2004). Selective Combination of Evidence for Topic Distillation using Document and Aggregate-level Information. In Proceedings of RIAO 2004 (Recherche d'Information Assistee par Ordinateur - Computer assisted information retrieval) on Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval. CID. pp. 610–622.

Ribeiro, B. and Muntz, R. (1996). A belief network model for ir. In Proceedings of the 19th annual international ACM SIGIR conference on Research and Development in Information Retrieval. ACM Press. pp. 253–260.

Robertson, S. E. and Walker, S. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Proceedings of the 17th annual international ACM-SIGIR conference on Research and Development in Information Retrieval. Springer-Verlag New York, Inc. pp. 232–241.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization.* John Wiley and Sons, New York, USA.

Silverstein, C., Marais, H., Henzinger, M. and Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum* **33**(1), 6–12.

Turtle, H. and 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. and Hawking, D. (2003). Query-independent evidence in home page finding. *ACM Transactions on Information Systems* **21**, 286–313.