

# Selective Combination of Evidence for Topic Distillation using Document and Aggregate-level Information

Vassilis Plachouras<sup>1</sup>, Iadh Ounis<sup>1</sup> & Fidel Cacheda<sup>2</sup>

<sup>1</sup> University of Glasgow  
Glasgow G12 8QQ, UK  
{vassilis, ounis}@dcs.gla.ac.uk

<sup>2</sup> University of A Coruña  
15071 A Coruña, Spain  
fidel@udc.es

## Abstract

The combination of evidence for Information Retrieval has been studied extensively in order to increase effectiveness. In this paper, we study the selective application of different retrieval approaches on a per-query basis for Web Information Retrieval. Our methodology is based on the assumption that not all queries benefit from a uniform retrieval approach. In order to select the most appropriate retrieval approaches for each query, we use evidence from the hyperlink structure of the retrieved documents and also from the distribution of aggregates, that is the groups of documents from the same domain. Our experimental results show that it is possible to obtain important improvements in retrieval effectiveness by using simple statistical decision mechanisms on a per-query basis.

## 1 Introduction

It has been recognised that the combination of different sources of evidence can improve the effectiveness of Information Retrieval (IR) (Croft, 2000). In the context of Web IR, in addition to the textual content of documents, there is another source of evidence, namely the hyperlink structure, which can be used 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 (e.g. specific vs. broad queries) (Kleinberg, 1999) suggest that combining content and hyperlink analysis in a uniform way, independently from the queries, would not result in optimal retrieval effectiveness (Plachouras *et al.*, 2003). For example, if we apply hyperlink analysis for a specific query in the same way as for a broad query, then the benefit should be higher for the broad query. Indeed, evidence from the hyperlink information contained in the set of the retrieved documents for a specific query would be weaker than the corresponding evidence for a broad query.

We investigate a different approach for Web IR, in which we select the most appropriate retrieval approaches from a set of candidate approaches, on a *per-query* basis. This is achieved by quantifying different aspects of the *query scope*, a composite measure of how broad or specific a query is. Each aspect, or component, of the query scope models a feature of broad or specific queries. For example, we expect that broad queries would result in retrieving a high number of documents. Apart from examining the distribution of retrieved documents, we also group them in aggregates, that is sets of related documents. We employ evidence from the distribution of aggregates for each query, in order to select the most appropriate retrieval approaches for the considered query. As an example, we expect that broader queries would result in larger aggregates. After computing the values of the query scope components, we use them in a decision mechanism that selects the most appropriate retrieval approaches for each query.

We evaluate the proposed methodology with a TREC standard Web test collection, the .GOV, and the queries from the topic distillation tasks of TREC11 (Craswell & Hawking, 2002) and TREC12 (Craswell *et al.*, 2003). Both tasks and the corresponding query sets are about finding useful entry points for the query topics. The results show that important improvements over the baselines are possible, even

when we employ a simple selection mechanism to dynamically assign a retrieval approach to each query, depending on its computed query scope.

The remaining of the paper is organised as follows. In Section 2, we introduce the query scope components and the decision mechanism. Section 3 contains a detailed description of our experiments, the results of which we analyse in Section 4. Section 5 contains a brief overview of related work. Finally, we provide some concluding remarks in Section 6.

## 2 Query Scope for Topic Distillation

Assuming that not all queries benefit from the same retrieval approach, we need to find which of the available approaches are most appropriate for a specific query. Hence, we introduce a selection/decision mechanism that will associate the most appropriate retrieval approaches to each query. For example, for specific queries we employ a content-only retrieval, while for more broad queries, we use evidence from the hyperlink structure, or from the URLs of documents. The selection mechanism employs a composite measure, the query scope, which addresses important statistical aspects of the set of retrieved documents.

Furthermore, apart from employing statistics from the document-level only, we take a different perspective, considering 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 topics. Within sites, documents are usually organised in a hierarchical directory structure. Moreover, 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 & McCurley, 2003). There have been different efforts towards an automatic identification of aggregates of hypertext, or Web pages. Botafogo and Shneiderman (1991) have employed graph theoretic approaches in order to identify aggregates. Alternatively, Eiron and McCurley (2003) and Li *et al.* (2000) have defined 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). Similarly, for simplicity, in this paper we define the aggregates to be groups of documents from the same domain.

In the remainder of this section, we present the components of the query scope. Section 2.1 contains the definitions of the components that use evidence from the distribution of retrieved documents (document-level information), and Section 2.2 contains the definitions of the components that depend on the distribution of aggregates (aggregate-level information). In each of these sections, we define two components: one depending on the number of retrieved documents, and another depending on the hyperlinks among the retrieved documents. In Section 2.3, we present the selection mechanism that employs the query scope, in order to dynamically select a retrieval approach for each query.

### 2.1 Document-level Information

The first component of the query scope is related to the number of retrieved documents. We assume that for the broader queries, there will be many documents, which contain all the query terms. In this case, 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.

We introduce the *query\_extent*, the number of retrieved documents that contain all the query terms, normalised between 0 and 1 by dividing with a given fraction  $\alpha$  of the total number of documents in the test collection:

$$query\_extent = \min \left( \frac{|\{d_i \mid d_i \text{ contains all query terms}\}|}{\alpha}, 1 \right) \quad (1)$$

The normalisation is introduced as most of the queries tend to retrieve only a small fraction of documents from the collection and therefore, dividing by  $\alpha$  leads to a better-distributed measure<sup>1</sup>.

The second component is related to the *cohesiveness* of the retrieved documents. We assume that for queries on topics covered by either whole sites, or groups of documents, there will be more hyperlinks among the retrieved documents. We expect that these queries will benefit from retrieval approaches where evidence from the hyperlink structure is used.

In order to measure the cohesiveness of the retrieved documents  $D = \{d_i\}$ , we employ a percolation threshold that estimates the stability of a directed network (Schwartz *et al.*, 2002):

$$q_c(\{d_i\}) = \frac{\langle outdegree_i \rangle}{\langle outdegree_i \cdot indegree_i \rangle} \quad (2)$$

where  $\langle outdegree_i \rangle$  stands for the average of the outdegree distribution of the retrieved documents and  $\langle outdegree_i \cdot indegree_i \rangle$  corresponds to the average of the product of the outdegree with the indegree for each retrieved document. The used hyperlinks are only those within the set of retrieved documents, i.e. the hyperlinks for which the source and destination documents have been retrieved. If  $\langle outdegree_i \rangle = 0$ , then  $q_c(\{d_i\})$  is undefined. If  $q_c(\{d_i\}) \rightarrow +\infty$ , we assume that there is no correlation between the incoming and outgoing links in  $\{d_i\}$ . On the other hand, if  $q_c(\{d_i\}) \rightarrow 0$  for the set of retrieved documents  $\{d_i\}$ , we expect that the existence of incoming and outgoing links in  $\{d_i\}$  is correlated, and that the considered documents are well connected. In this case, we assume that the corresponding query would benefit from using evidence from the hyperlink structure of documents.

## 2.2 Aggregate-level Information

In addition to the document-level, we also use information from the aggregate-level. In this work, we form an aggregate from the documents belonging to the same domain. Let  $D = \{d_i\}$  be the set of retrieved documents. Then, for each unique domain that appears in  $D$ , we create one separate aggregate  $a_j$ . We define two different components that use the aggregate-level information.

The first component is related to the average size of the aggregates formed for a given query. We assume that if the formed aggregates are relatively large for a given query, then there exist clusters of documents on the query topic. In this case, using evidence from hyperlink analysis might be beneficial to find the entry points to each cluster. If the size of aggregate  $a_j$  is  $size(a_j)$ , then we compute the average aggregate size  $\langle size(a_j) \rangle$ .

The second component we consider is the number of aggregates, for which there is some correlation between the outgoing and incoming links. If such a correlation exists for a high number of aggregates, then employing additional evidence from hyperlink analysis may be quite appropriate, in order to detect the entry points of the aggregates. For each aggregate  $a_j$ , we compute the percolation threshold  $q_c(a_j)$ , considering only the hyperlinks within  $a_j$ , since we want to measure the cohesiveness of the specific aggregate. We compute the number of aggregates for which  $q_c(a_j)$  is defined and it is finite:

$$dfperc(\{a_j\}) = \left| \{a_j \mid q_c(a_j) \text{ is defined and } q_c(a_j) < +\infty\} \right| \quad (3)$$

## 2.3 Selection Mechanism for Topic Distillation

We introduce a mechanism for selecting the most appropriate retrieval approaches for a query  $q$ , based on evidence from the query scope components. In order to evaluate the effectiveness of each component separately, we employ a simple approach, where the value of one of the components, computed for the retrieved documents, is compared to a threshold. According to the result of this comparison, we assign

---

<sup>1</sup> In our experiments, we normalised the query extent by dividing with 1% of the number of documents in the collection, after looking at the average number of retrieved documents from the .GOV collection, for the TREC11 and TREC12 queries.

one of the candidate retrieval approaches to the query  $q$ . In this work, we consider two candidate approaches, which we select by using relevance information. Note that we can extend this selection mechanism, in order to use more thresholds and candidate approaches.

**Algorithm** SELECTRETRIEVALAPPROACH

Input : The query  $q$  under consideration,  
the candidate retrieval approaches  $A1$  and  $A2$ ,  
the query scope component  $Comp$  and  
a threshold value  $t$ .

Output: The set of retrieved documents, ranked according  
to the selected retrieval approach.

Method:

1. if  $Comp(q) \leq t$  then
2.     apply approach  $A1$  for query  $q$
3. else
4.     apply approach  $A2$  for query  $q$

Figure 1: Algorithm SELECTRETRIEVALAPPROACH.

Figure 1 shows a description of the selection mechanism. For each query, we compute the value of the query scope component  $Comp$ , given the set of retrieved documents and then we compare its value to the threshold  $t$ . If  $Comp(q) \leq t$ , we select  $A1$  as the most appropriate approach for query  $q$ , otherwise we select  $A2$ . Note that the order in which  $A1$  and  $A2$  are assigned to specific retrieval approaches may significantly affect the effectiveness of the selection mechanism. This order should be consistent with the basic assumptions underlying the employed query scope component and each of the involved retrieval approaches. If there is an inconsistency, then we should expect a detrimental effect in the retrieval effectiveness.

### 3 Experiments

In this section, we investigate the potential benefit from using the query scope components and the decision mechanism, as defined in Section 2. For the evaluation of the proposed methodology, we experiment with a standard TREC collection, the .GOV (for the indexing of documents, standard stop words were removed and stemming was applied), and the topics from the topic distillation tasks of TREC11 (Craswell & Hawking, 2002) and TREC12 (Craswell *et al.*, 2003). Both tasks are about finding key resources, or entry points for the topics. However, a difference between the two tasks is that the relevant documents for the TREC12 task are restricted only to the homepages of sites about the query topics. This difference is reflected in the choice of the evaluation measure, proposed by the TREC Web track organisers. For TREC11 topic distillation, the precision at 10 documents is employed, while for TREC12 task, R-Precision (precision after  $R$  documents have been retrieved, where  $R$  is the number of relevant documents for a query) is used, due to the lower number of relevant documents, which affects the stability of precision at 10 documents (Craswell *et al.*, 2003). For our analysis, we will use precision at 10 documents for both TREC tasks and also, R-Precision for the TREC12 task.

The selection mechanism described in Section 2.3 requires a set of candidate retrieval approaches. We consider three different retrieval approaches. The first one is content-only retrieval (C). For the second one, we extend each document by adding the anchor text of its incoming links (CA), and perform retrieval as in the case of the first approach. We employ the weighting scheme  $PL2$  from Amati and Van Rijsbergen's (2003) Divergence From Randomness (DFR) probabilistic framework. The formula of  $PL2$  can be found in the appendix of the paper. We select this weighting scheme from the available schemes in the DFR framework, since it is very robust for both TREC tasks we are testing (Plachouras *et al.*, 2002; Plachouras *et al.*, 2003). Moreover, the only parameter of the system is automatically set equal to  $c = 1.28$  (He & Ounis, 2003). The third retrieval approach takes into account both content and

anchor text of documents, and modifies the scores according to the length of the document's URL (CAU), as follows:

$$score_i = \frac{sc_i}{\log_2(urlpath\_len_i + 1)} \quad (4)$$

where  $sc_i$  is the content analysis score for document  $d_i$  and  $urlpath\_len_i$  is the number of characters of  $d_i$ 's URL path. In order not to boost non-relevant documents, we apply this approach only to the top 1000 retrieved documents, where content and anchor text is used for retrieval.

The evaluation of the candidate approaches for both TREC11 and TREC12 is shown in Table 1. Following Section 2.3, for TREC11, we choose the approaches C and CA, which are the most effective. For TREC12, we select the approaches CA and CAU respectively, which correspond to both the highest precision at 10 and R-Precision measures. Note that the selection of the best retrieval approaches for each TREC task is independent of the evaluation measure used, since the ordering of the approaches is the same.

If we manually combine C and CA for TREC11 in  $\text{MAX}(C, CA)$ , so that the most effective approach for each individual query is used, we get 0.2898 average precision at 10. For TREC12, the ideal combination  $\text{MAX}(CA, CAU)$  of CAU and CA results in 0.1680 precision at 10, or 0.1881 average R-Precision. From both ideal combinations, we can see that there is room for improvement between the uniform application of a retrieval approach on all queries and the ideal combination, where the most effective approach is used on a per-query basis.

TREC11	Aver. Prec.	Prec. at 10	R-Prec.
<b>C</b>	0.2053	0.2694	0.2362
<b>CA</b>	0.1953	0.2551	0.2237
CAU	0.1001	0.1367	0.1400
<b>MAX(C, CA)</b>	<b>0.2161</b>	<b>0.2898</b>	<b>0.2465</b>
TREC12			
C	0.0886	0.0680	0.0730
<b>CA</b>	0.1273	0.1020	0.1325
<b>CAU</b>	0.1428	0.1400	0.1369
<b>MAX(CA, CAU)</b>	<b>0.1874</b>	<b>0.1680</b>	<b>0.1881</b>

Table 1: The evaluation results of C, CA and CAU for TREC11 and TREC12 topic distillation tasks. The retrieval approaches in bold are the candidate approaches for each TREC task.

The next step of our experiments involves the evaluation of the different query scope components, in the context of the selection mechanism  $\text{SELECTRETRIEVALAPPROACH}$ , or SRA for brevity (see Figure 1). We use each of the components separately for different values of the threshold  $t$ , ranging from each component's minimum value  $\min(\text{Comp}(q))$  to its maximum value  $\max(\text{Comp}(q))$ , as computed for the queries, with a step of 1% of this range. For the analysis and presentation of the results, we normalise the threshold values between 0 and 1, in order to compare the different components. This linear transformation of the threshold values does not affect the results of the analysis. We should note that because the percolation threshold  $q_c(\{d_i\})$  is lower for queries, where there are more hyperlinks, we swap the retrieval approaches C and CA, so that for TREC11 we apply CA when  $\text{Comp}(q) \leq t$  and C otherwise (the same swapping of the retrieval approaches applies to the experiments for TREC12). In addition, for efficiency, we compute  $q_c(\{d_i\})$ ,  $\langle \text{size}(a_j) \rangle$  and  $\text{dfperc}(\{a_j\})$  from the set of the top 20000 documents. For computing the values of these three query scope components, we employ content-only retrieval<sup>2</sup>.

<sup>2</sup> At this stage, we could compute the values of the query scope components by using different weighting schemes, or content with anchor text. It would be interesting to investigate the impact of such alternatives on the components' values.

For TREC11 topic distillation, the results for a range of threshold values are presented in Figure 2, while Table 2 contains the highest precision values for each of the four components and the corresponding threshold values. It is worth noting that the best official run submitted for TREC11 topic distillation task achieved 0.2510 precision at 10 (Craswell & Hawking, 2002). We can see that the most effective component of query scope is the number of aggregates with a positive and finite percolation threshold  $dfperc(\{a_j\})$ , which results in 0.2796 precision at 10 for  $t \in [0.52, 0.54]$ . This result is quite close to the ideal combination  $\text{MAX}(C, CA)$  of  $C$  and  $CA$ , which yields 0.2878 precision at 10. The second most effective component is the percolation threshold  $q_c(\{d_i\})$ , for which the highest precision at 10 is 0.2755 for  $t \in [0.06, 0.07]$ . The two other components do not outperform the content-only baseline, which corresponds to the middle straight line in Figure 2. We can see that using either document, or aggregate-level information to measure the cohesiveness of the retrieved documents, is effective for the TREC11 experiments.

Component	Prec. At 10	Thres. range
C Baseline	0.2694	—
CA Baseline	0.2551	—
$\text{MAX}(C, CA)$	0.2898	—
$\text{SRA}(q, C, CA, query\_extent(\{d_i\}), t)$	0.2633	[0.65, 0.75]
$\text{SRA}(q, CA, C, q_c(\{d_i\}), t)$	<b>0.2755</b>	[0.06, 0.07]
$\text{SRA}(q, C, CA, \langle size(a_j) \rangle, t)$	0.2694	{1.00}
$\text{SRA}(q, C, CA, dfperc(\{a_j\}), t)$	<b>0.2796</b>	[0.52, 0.54]

Table 2: The highest values for each component for TREC11 topic distillation and the corresponding threshold values.

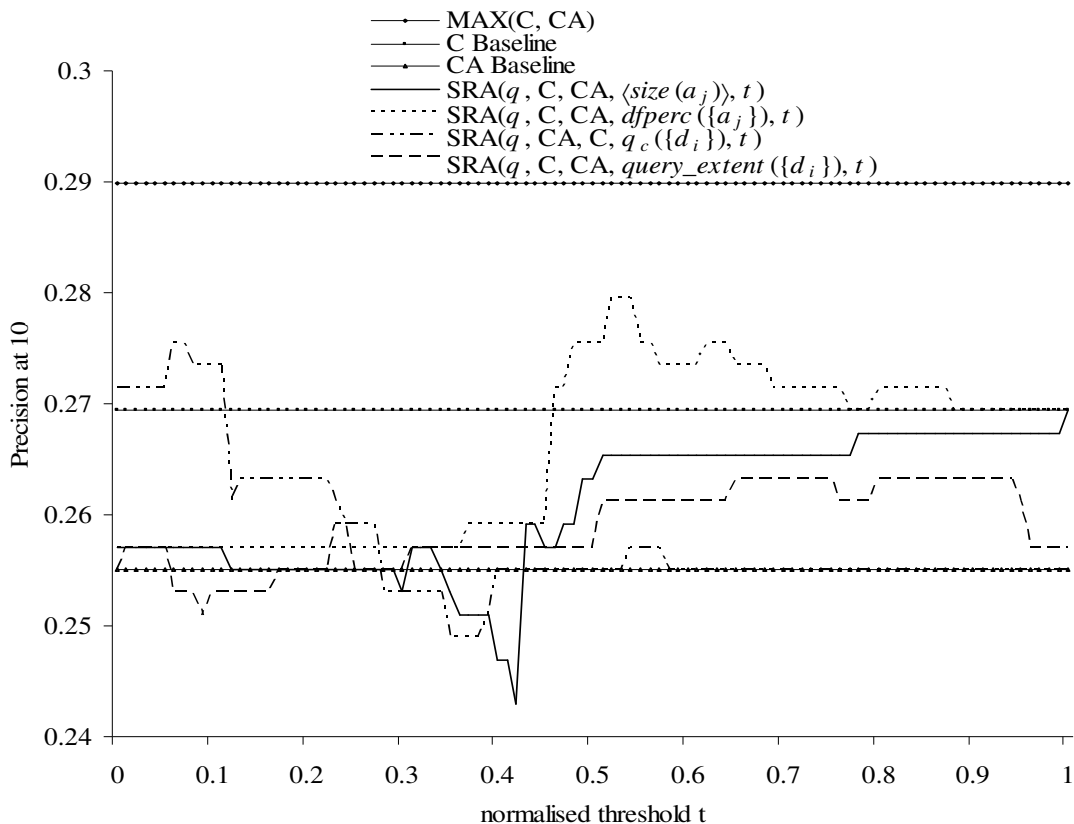


Figure 2: Evaluation of the decision mechanism SRA with  $C$  and  $CA$  for TREC11. The evaluation measure used is Precision at 10.

Important improvements are also obtained from the selective application of different retrieval approaches for TREC12 topic distillation queries. We use both precision at 10 documents (Figure 3) and R-Precision (Figure 4). Moreover, the highest precision values and their corresponding threshold values are shown in Table 3 for both evaluation measures. It is worth noting that the highest R-Precision achieved by the submitted official runs in TREC12 topic distillation task was 0.1636, while the highest precision at 10 was 0.1280 (these figures correspond to two different runs) (Craswell *et al.*, 2003). All four components result in improvements over the CAU baseline. More specifically, the most effective component, with respect to both evaluation measures, is  $\langle size(a_j) \rangle$ , as shown in Table 3. For precision at 10,  $dfperc(\{a_j\})$  is equally effective for a wider range of threshold values. However, when we look at R-Precision, the second most effective component is  $query\_extent(\{d_i\})$ , which achieves 0.1701 R-Precision for  $t \in [0.37, 0.38]$ . Overall, although all the components perform similarly well, we can see that the aggregate-level information leads to higher retrieval effectiveness for the TREC12 experiments.

Component	Prec. at 10	Thres. range	R-Prec.	Thres. range
CA Baseline	0.1020	—	0.1325	—
CAU Baseline	0.1400	—	0.1369	—
MAX(CA, CAU)	0.1680	—	0.1881	—
$SRA(q, C, CA, query\_extent(\{d_i\}), t)$	0.1460	[0.09, 0.13]	<b>0.1701</b>	[0.37, 0.38]
$SRA(q, CA, C, qc(\{d_i\}), t)$	0.1500	[0.47, 0.52]	0.1682	[0.42, 0.43]
$SRA(q, C, CA, \langle size(a_j) \rangle, t)$	<b>0.1520</b>	[0.24, 0.26]	<b>0.1766</b>	[0.36, 0.38]
$SRA(q, C, CA, dfperc(\{a_j\}), t)$	<b>0.1520</b>	$\{0.20\} \cap [0.22, 0.32]$	0.1663	$\{0.20\}$

Table 3: The highest precision at 10 and R-Precision values for each component for TREC12 topic distillation and the corresponding threshold values.

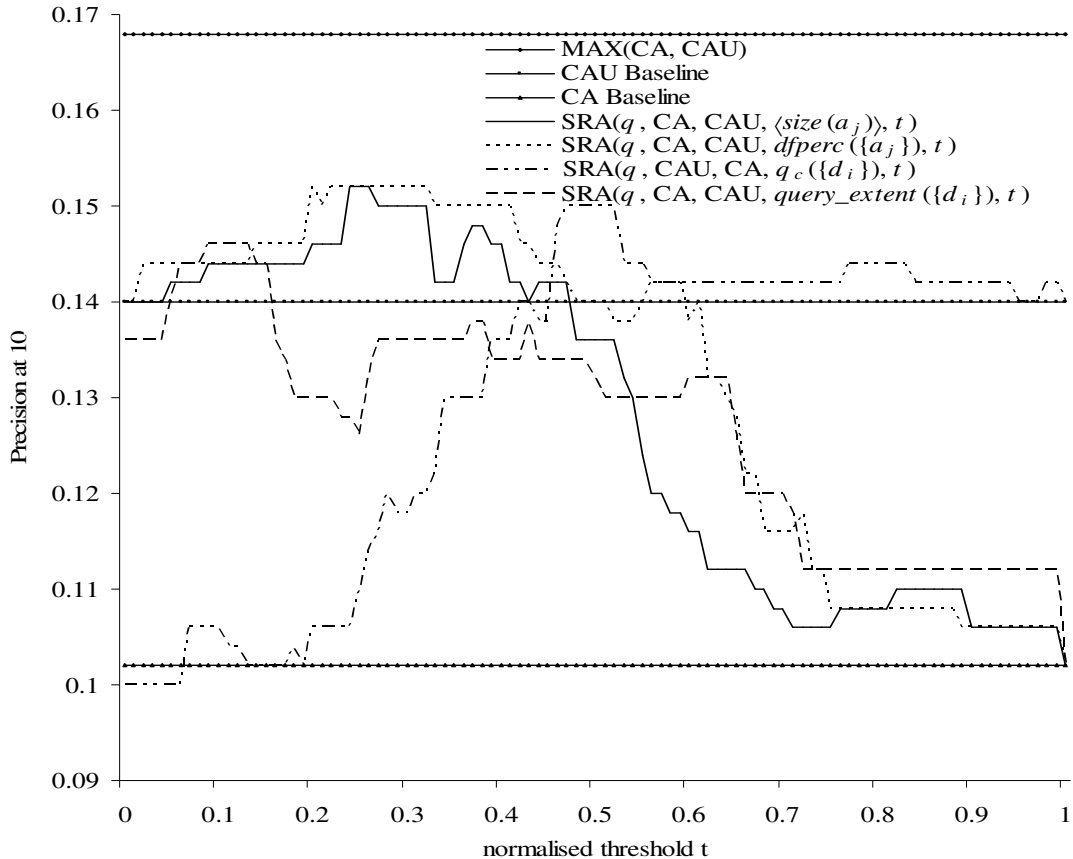


Figure 3: Evaluation of the decision mechanism SRA with CA and CAU for TREC12. The evaluation measure is Precision at 10.

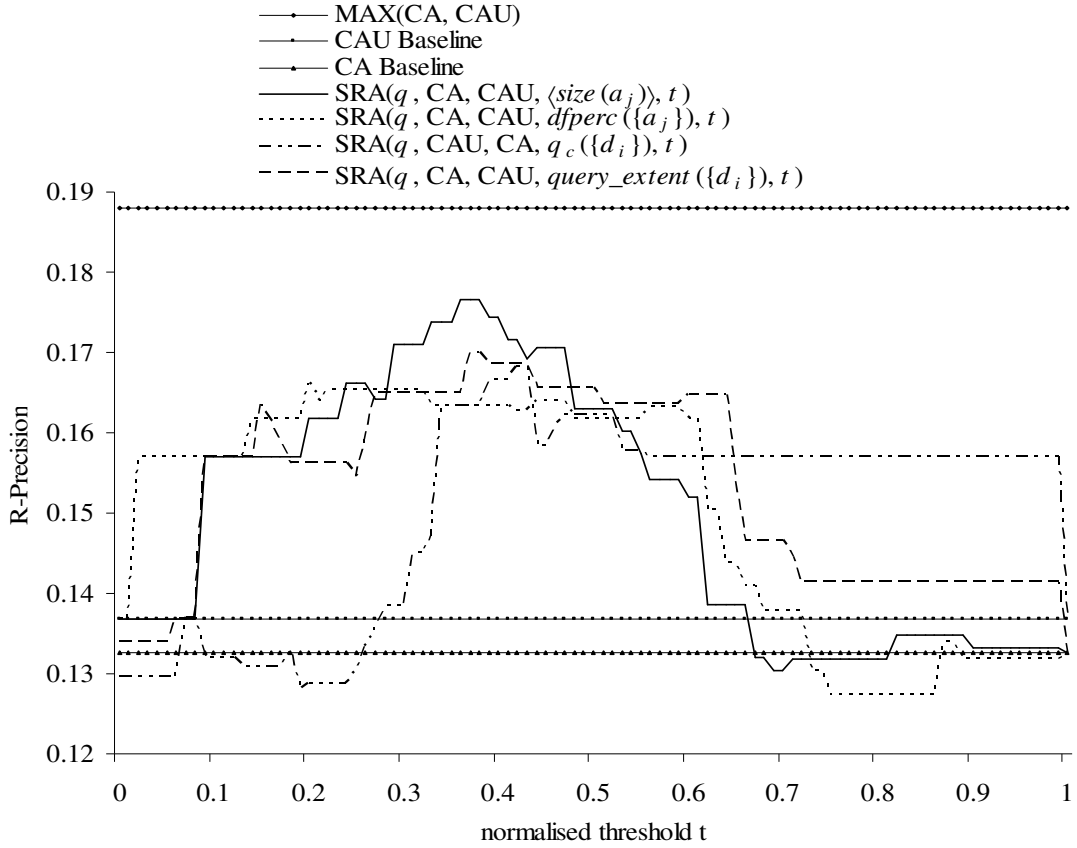


Figure 4: Evaluation of the decision mechanism SRA with CA and CAU for TREC12. The evaluation measure used is R-Precision.

#### 4 Discussion

The experiments have shown that it is possible to improve retrieval effectiveness by applying different retrieval approaches on a per-query basis, based on evidence from either the document-level, or the aggregate-level information. In this section, we analyse the results from different perspectives.

First, we look at the implications of choosing the candidate retrieval approaches. The selective application of a retrieval approach requires that each one of the candidate approaches should outperform the other candidates at least for a subset of the queries. The potential improvement from the selective application of different retrieval approaches depends on the relative size of these subsets of queries. For example, for TREC11 topic distillation, C outperforms CA for 9 queries, while CA is more effective than C for 8 out of the 49 queries. For TREC12, the size of the subsets of queries is higher. Indeed, CAU outperforms CA for 16 queries and CA outperforms CAU for 9 out of the 50 queries. Thus, it should be expected that the potential improvement for TREC12 is higher than for TREC11, and our experimental results confirm this fact. Moreover, in this work, the selection of the candidate retrieval approaches is based on relevance information. In an operational environment, where relevance information is not available, the selection of the candidate retrieval approaches could be based on a training process, where a set of retrieval approaches are evaluated on a set of training queries.

Even if we select the candidate retrieval approaches successfully, we must be able to identify an effective query scope component for assigning the most appropriate approaches to each query. In other words, the values of a query scope component should be correlated with the differences between the retrieval effectiveness of the candidate approaches. If there is such a correlation, then it should result in increased precision. On the other hand, if the assumptions underlying a feature are not consistent with

the actual data, then we should expect the retrieval effectiveness to be close to, or lower than that of the baselines. As an example of the first case, we could look at the percolation threshold  $q_c(\{d_i\})$ , which quantifies the cohesiveness of the retrieved documents. When it is used in the selection mechanism SRA, it results in improved average precision, reflecting the fact that evidence from hyperlink analysis is beneficial for queries that retrieve a more cohesive set of documents. Therefore, the percolation threshold is an effective query scope component for assigning the most appropriate retrieval approaches to a query.

Component	Prec. at 10	Threshold range
C Baseline	0.2694	—
MAX(C, CA) Baseline	0.2898	—
SRA( $q$ , CA, C, $query\_extent(\{d_i\})$ , $t$ )	0.2735	{0.09}
SRA( $q$ , CA, C, $\langle size(a_j) \rangle$ , $t$ )	0.2816	{0.42}

Table 4: The highest values for  $query\_extent(\{d_i\})$  and  $\langle size(a_j) \rangle$  and the corresponding threshold values, after swapping the retrieval approaches, for TREC11 topic distillation.

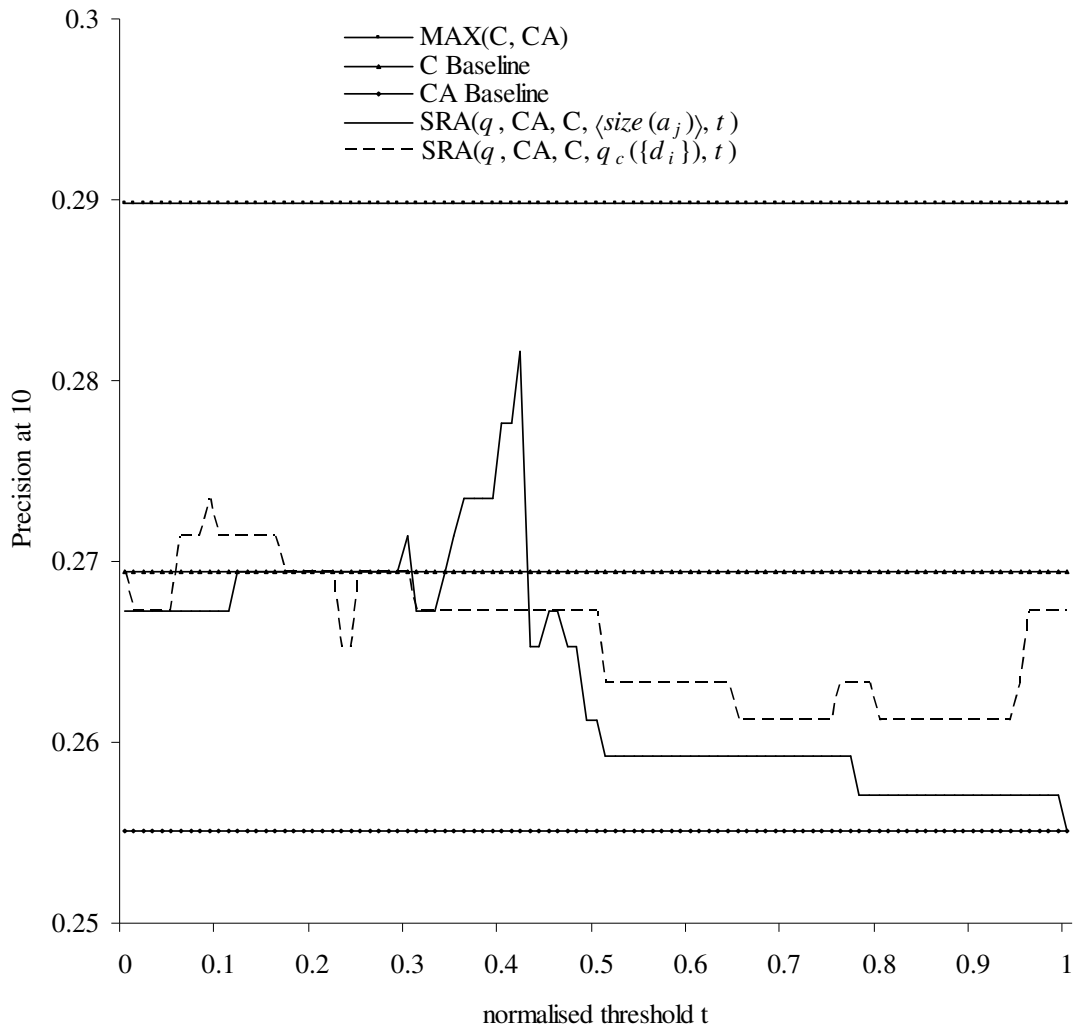


Figure 5: Evaluation of the decision mechanism SRA, using the swapped assignment of C and CA, for TREC11. The query scope components used are  $query\_extent(\{d_i\})$  and  $\langle size(a_j) \rangle$ .

As an example of the second case, we will attempt to explain why  $\langle size(a_j) \rangle$  and  $query\_extent(\{d_i\})$  did not result in improvements for the TREC11 experiments. Both these components, when employed as shown in Table 2, result in lower or equal precision, compared to the content-only baseline. This is an indication that the assignment of the retrieval approaches by the selection mechanism was not appropriate. Indeed, if we swap the approaches in the selection mechanism SRA, so that CA is used when  $Comp(q) \leq t$ , otherwise C is employed, we get the results shown in Figure 5 and Table 4. Thus, the fact that there are improvements over the content baseline when we swap the retrieval approaches can be explained as an inconsistency between the assumptions underlying each component and the actual results. For example, the assumption for  $\langle size(a_j) \rangle$  is that more evidence from hyperlink structure should be used for queries, where a large average aggregate size is observed. However, TREC11 results show that queries benefit from using anchor text when there are smaller aggregates. The same conclusion can be drawn for the query extent as well. We should note that these inconsistencies appear only for TREC11 and not for TREC12 experiments, a fact that underlines the importance of the real difference between the definitions of the two tasks.

It is interesting to link our findings with respect to the validity of our assumptions for TREC11, and the effectiveness of C and CA for another retrieval task, the TREC11 named-page finding task (Craswell & Hawking, 2002). In this task, there is only one relevant document per topic and generally the queries are longer and fewer documents are retrieved. Results from TREC11 have shown that CA is significantly more effective than C for named-page finding. Thus, we could say that when we have fewer relevant documents for a topic, then content and anchor text tends to be more effective than content-only retrieval, independently from the task. As an illustration of this finding, for the TREC12 topic distillation task, the number of relevant documents is quite low (Craswell *et al.*, 2003), and our experimental results confirm that CA is more effective than C.

An important property of the query scope components is the stability in the improvements over the baselines. A component, for which we get improvements for a wide range of threshold values, is preferable to a component, which results in improvements for only few threshold values. In Table 5, we present the percent of the threshold values for each component, ranging from 0 to 1 with a step of 0.01, for which the selective application of different retrieval approaches outperforms the best uniform approach. We should note that the values in parentheses for  $\langle size(a_j) \rangle$  and  $query\_extent(\{d_i\})$  refer to the set of experiments, where we swapped the application of C and CA (Figure 5 and Table 4). The most stable component for TREC11 is  $dfperc(\{a_j\})$ . On the other hand,  $q_c(\{d_i\})$  is not so stable, since it overcomes the baseline for a smaller fraction of the threshold values.

The situation is significantly different for TREC12, where all the components result in improvements for more than 50% of the threshold values, when considering R-Precision. Moreover, when we use precision at 10, all the components except for  $query\_extent(\{d_i\})$  outperform CAU, the baseline with the highest precision, for more than 40% of the threshold values. The property of stability for a component is crucial, if we want to introduce a mechanism for predicting an appropriate threshold value. Indeed, in an operational environment, the probability of selecting a safe threshold will be higher for more stable query scope components.

Component	TREC11 Prec. at 10 threshold values %	TREC12 Prec. at 10 threshold values %	TREC12 R-Precision Threshold values %
$query\_extent(\{d_i\})$	0.00 (10.89)	9.90	93.07
$q_c(\{d_i\})$	11.88	50.50	73.27
$\langle size(a_j) \rangle$	0.00 (8.91)	41.58	57.43
$dfperc(\{a_j\})$	45.54	49.51	70.30

Table 5: Percent of threshold values for which precision is over precision of C for TREC11, or CAU for TREC12.

A different way to look at the results is to consider the kind of information used by the query scope components. The components  $q_c(\{d_i\})$  and  $dfperc(\{a_j\})$  that are based on measuring the cohesiveness of the retrieved documents, outperform the other two components  $\langle size(a_j) \rangle$  and  $query\_extent(\{d_i\})$  for TREC11. The assumptions for the first two components are consistent with the results, while this does not necessarily hold for the other two components. On the other hand,  $\langle size(a_j) \rangle$  and  $query\_extent(\{d_i\})$ , which are based on counting the number of retrieved documents, are more effective for TREC12. We could attribute this difference between the two tasks to the way the tasks are defined. Finally, we should note that  $dfperc(\{a_j\})$  is the most effective query scope component for TREC11, while  $\langle size(a_j) \rangle$  is the most effective component for TREC12. Both these components are based on information obtained from the aggregate level. Thus, the aggregate-level information proves to be an effective source of evidence for defining the query scope components, for both TREC11 and TREC12.

## 5 Related Work

The combination of evidence from different sources has been employed in various ways in order to increase the 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 estimates of the individuals. Instead of combining linearly the scores of different retrieval approaches, 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.

In the context of Web IR, recent research has focused on detecting when to employ evidence from the hyperlink structure. Approaches proposed towards this end were based on the density of the links in a collection (Gurrin & Smeaton, 2003; Fisher & Everson, 2003), or on the characteristics of the set of retrieved documents for each query (Amitay *et al.*, 2002).

Our approach can be seen as a linear combination of multiple retrieval approaches with binary coefficients, where there is only one non-zero coefficient for each query. Additionally, while Amitay *et al.* (2002) use the number of documents added to the root set, during the expansion step, to adapt the contribution of HITS and SALSA for each query, our approach aims to explicitly assign the most appropriate retrieval approach for each query. 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.

## 6 Conclusions

In this paper, we have presented a method for a selective application of different retrieval approaches on a per-query basis. We define the query scope by employing evidence from either the distribution of retrieved documents  $\{d_i\}$  for a query ( $query\_extent(\{d_i\})$  and  $q_c(\{d_i\})$ ), or from the distribution of aggregates  $\{a_j\}$  ( $\langle size(a_j) \rangle$  and  $dfperc(\{a_j\})$ ). Additionally, the components  $query\_extent(\{d_i\})$  and  $\langle size(a_j) \rangle$  are based on counting the number of retrieved documents, while the other two, i.e.  $q_c(\{d_i\})$  and  $dfperc(\{a_j\})$ , measure the cohesiveness of the retrieved documents. The query scope, in combination with a selection mechanism, is applied in order to find the most appropriate retrieval approaches for each query, from a set of candidate approaches. The choice of the candidate approaches in this paper is based on relevance information. An important extension to our methodology would be the investigation of different ways to form the set of candidate approaches, independently of the relevance information.

Our experiments with a standard TREC Web test collection and two query sets from the topic distillation tasks of TREC11 and TREC12 show that important improvements over the baselines are obtained with our approach. In particular, the improvements for TREC12 are stable over a wide range of

threshold values. In addition, the query scope components that employ evidence from the aggregate-level result in the highest precision for both query sets, a result that confirms the importance of the aggregates.

In conclusion, we have shown that by employing simple statistical mechanisms, it is possible to improve the retrieval effectiveness for Web IR, by dynamically combining evidence from content and hyperlink analysis. Currently, we are working on the automatic setting of the threshold values, assuming that there is no relevance information readily available, and the results are promising. Interesting extensions of this work will be a principled way to combine the different query scope components, and the application of this approach in the context of different retrieval tasks, such as filtering (Hull, 1998), where relevance information becomes incrementally available.

## 7 Acknowledgements

The work of the first and second 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 third author has been partially supported by the Comisión Interministerial de Ciencia y Tecnología (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).

## 8 Bibliographical References

- Amati, G. & 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., Lempel, R. & Soffer, A. (2002). Topic Distillation with Knowledge Agents. In *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC)*. Gaithersburg, MD: National Institute of Standards and Technology.
- Aslam, J.A. & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 276—284). ACM Press.
- Bartell, B.T., Cottrell, G.W. & Belew, R.K. (1994). Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 173—181). ACM Press.
- Belkin, N.J., Cool, C., Croft, B.W. & Callan, J.P. (1993). The effect of multiple query representations on information retrieval system performance. In *Proceedings of the 16<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 339—346). ACM Press.
- Botafogo, R.A. & Shneiderman, B. (1991). Identifying aggregates in hypertext structures. In *Proceedings of the 3<sup>rd</sup> Annual ACM Conference on Hypertext* (pp. 63—74). ACM Press.
- Craswell, N. & Hawking, D. (2002). Overview of the TREC-2002 Web Track. In *NIST Special Publication: 500-251 The Eleventh Text Retrieval Conference (TREC)* (pp. 86—93). Gaithersburg, MD: National Institute of Standards and Technology.
- Craswell, N., Hawking, D., Wilkinson, R. & Wu, M. (2003). Overview of the TREC-2003 Web Track. In *Proceedings of The Twelfth Text Retrieval Conference (TREC)*. Gaithersburg, MD: National Institute of Standards and Technology.
- 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* (pp. 1—36). Boston, MA: Kluwer Academic Publishers.
- Eiron, N. & McCurley, K. (2003). Untangling compound documents on the web. In *Proceedings of the 14<sup>th</sup> ACM Conference on Hypertext and Hypermedia* (pp. 85—94). ACM Press.
- Fisher, M. & Everson, R. (2003). When Are Links Useful? Experiments in Text Classification. In F. Sebastiani (ed.) *Advances in Information Retrieval: 25<sup>th</sup> European Conference on IR Research* (pp. 41—56). Heidelberg: Springer-Verlag.
- Gurrin, C. & Smeaton A.F. (2003). Improving the Evaluation of Web Search Systems. In F. Sebastiani (ed.) *Advances in Information Retrieval: 25<sup>th</sup> European Conference on IR Research* (pp. 25—40). Heidelberg: Springer-Verlag.

- He, B. & Ounis, I. (2003). A study of parameter tuning for term frequency normalization. In *Proceedings of the 12<sup>th</sup> International ACM CIKM Conference on Information and Knowledge Management* (pp. 10—16). ACM Press.
- Hull, D.A. (1998). The TREC-7 Filtering Track: Description and Analysis. In *NIST Special Publication: 500-242 The Seventh Text Retrieval Conference (TREC)* (pp. 33—56). Gaithersburg, MD: National Institute of Standards and Technology.
- Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604—632.
- Kwok, K.L., Deng, P., Dinstl, N. & Chan, M. (2002). TREC2002 Web, Novelty and Filtering Track Experiments using PIRCS. In *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC)* (pp. 520—528). Gaithersburg, MD: National Institute of Standards and Technology.
- Li, W.-S., Kolak, O., Vu, Q. & Takano, H. (2000). Defining logical domains in a web site. In *Proceedings of the 11<sup>th</sup> ACM Conference on Hypertext and Hypermedia* (pp. 123—132). ACM Press.
- Plachouras, V., CACHEDA, F., Ounis, I. & van Rijsbergen, C.J. (2003). University of Glasgow at the Web track: Dynamic Application of Hyperlink analysis using the Query Scope. In *The Twelfth Text Retrieval Conference (TREC)*. Gaithersburg, MD: National Institute of Standards and Technology.
- Plachouras, V., Ounis, I., Amati, G. & van Rijsbergen, C.J. (2002). University of Glasgow at the Web track of TREC 2002. In *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC)* (pp. 645—651). , MD: National Institute of Standards and Technology.
- Ribeiro-Neto, B. & Muntz, R. (1996). A belief network model for IR. In *Proceedings of the 19<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 253—260). ACM Press.
- Schwartz, N., Cohen, R., ben-Avraham, D., Barabási, A.-L. & Havlin, S. (2002). Percolation in directed scale-free networks. *Physical Review E* 66, 015104(R).
- Turtle, H. & W.B. Croft (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187—222.

## 9 Appendix

In our experiments, we employed the weighting scheme *PL2* from Amati and Van Rijsbergen's (2002) DFR framework. According to this weighting scheme, the weight of a term *t* is given by the following formula:

$$weight_{PL2}(t) = \left( tfn \cdot \log_2 \frac{tfn}{\lambda} + \left( \lambda + \frac{1}{12 \cdot tfn} - tfn \right) \cdot \log_2 e + 0.5 \cdot \log_2 (2\pi \cdot tfn) \right) \cdot \frac{1}{tfn + 1}$$

where:

- $\lambda$  is the mean and variance of a Poisson distribution,
- $tfn = term\_freq \cdot \log_2 \left( 1 + c \cdot \frac{average\_document\_length}{document\_length} \right)$ ,
- $c$  is a parameter used for term frequency normalisation,
- $term\_freq$  is the within-document term frequency.