

Usefulness of Hyperlink Structure for Query-Biased Topic Distillation

Vassilis Plachouras
University of Glasgow
Glasgow, G12 8QQ, U.K.
vassilis@dcs.gla.ac.uk

Iadh Ounis
University of Glasgow
Glasgow, G12 8QQ, U.K.
ounis@dcs.gla.ac.uk

ABSTRACT

In this paper, we introduce an information theoretic method for estimating the usefulness of the hyperlink structure induced from the set of retrieved documents. We evaluate the effectiveness of this method in the context of an optimal Bayesian decision mechanism, which selects the most appropriate retrieval approaches on a per-query basis for two TREC tasks. The estimation of the hyperlink structure's usefulness is stable when we use different weighting schemes, or when we employ sampling of documents to reduce the computational overhead. Next, we evaluate the effectiveness of the hyperlink structure's usefulness in a realistic setting, by setting the thresholds of a decision mechanism automatically. Our results show that improvements over the baselines are obtained.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms

Experimentation, Algorithms, Theory

Keywords

Web information retrieval, Bayesian decision theory, information theory, link analysis

1. INTRODUCTION

Web Information Retrieval (IR) can exploit a number of different retrieval approaches, taking into account various sources of information, in addition to the textual content of documents. For example, we can use evidence from the document structure, or hyperlink structure, to enhance retrieval effectiveness. However, not all queries benefit equally from using the same sources of evidence, or from applying the same retrieval approach. Indeed, hyperlink analysis tends to be more effective for broader queries than it is for specific ones. In other words, the linkage information may not be

adequate to improve effectiveness for specific queries, due to the weaker nature of evidence obtained from hyperlinks [10].

We believe that the optimal retrieval effectiveness could be obtained by applying the most appropriate retrieval approaches on a per-query basis. A necessary step towards this direction is quantifying how much information is obtained by considering evidence other than the content of documents. In this paper, we focus on evidence from the hyperlink structure of the content-based retrieved set of documents and how it can be employed to select the most appropriate retrieval approaches on a per-query basis.

We propose to estimate the usefulness of the retrieved documents' hyperlink structure for each query, by employing Information Theory and measuring the divergence between two score distributions. The first one is the content analysis score distribution, which serves as a baseline. The second distribution depends on both the content analysis scores and the hyperlinks within the set of retrieved documents. The divergence indicates the usefulness of the hyperlink structure by estimating whether there are any patterns in the distribution of hyperlinks for a given query. If the divergence is low, then there is no apparent pattern in the distribution of hyperlinks. On the other hand, if it is high, we expect to find some non-random patterns of hyperlinks, resulting in clusters of related documents. We choose to restrict our approach to the set of retrieved documents, in order to avoid the topic drift effect [5]. Next, we introduce a decision mechanism, which takes as input the estimate of the hyperlink structure's usefulness, and selects the most appropriate retrieval approaches from a set of candidate ones.

The proposed methodology can be seen as a two-step process. First, the usefulness of the hyperlink structure is estimated, without indicating anything about the final document ranking. In other words, the score distributions, whose divergence we measure, are not necessarily used for ranking the retrieved documents. In the second step, the most appropriate retrieval approaches are selected for document ranking. We introduce the second step in order to decouple the estimation of the hyperlink structure usefulness from document ranking. This enables us to use any retrieval approach in the second step, thus making our methodology more general.

We experiment with the data of the topic distillation tasks from TREC11 [7] and TREC12 [9], and our results show that improvements over the baselines are obtained from a selective application of the most appropriate retrieval approaches on a per-query basis. To evaluate the estimation of the hyperlink structure usefulness in the first step, we assume the existence of relevance information, and use Bayesian decision theory [12] to assign a retrieval approach to each query optimally. We evaluate the usefulness measures by using different content analysis weighting schemes for defining the score distributions in the first step. Moreover, in order to reduce the computational overhead of the first step, we experiment with

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
Copyright 2004 ACM 1-58113-881-4/04/0007 ...\$5.00.

using various subsets of top-ranked documents. In both cases, we find that the effectiveness of the usefulness measures remains stable. Next, we evaluate the proposed methodology without relevance information. We assume that if we find useful patterns in the distribution of hyperlinks, then employing evidence from the hyperlink structure, or the URLs, to detect useful entry points, will be more beneficial to the ranking. In order to set the thresholds for the decision mechanism without relevance information, we employ an approach similar to that used by Cronen-Townsend et al. [11].

The remainder of the paper is organised as follows. In Section 2, we present the measures of the hyperlink structure's usefulness. In Section 3, we describe the experimental setting and evaluate the introduced measures in the context of a Bayesian decision mechanism with relevance information. Sections 4 and 5 contain the experiments, where different weighting schemes and sampling of documents are used for estimating the usefulness of the hyperlink structure. In Section 6, we set the thresholds of the decision mechanism automatically, without relevance information. Section 7 contains a brief presentation of the related work, and in Section 8 we present some concluding remarks from this work.

2. USEFULNESS OF HYPERLINK STRUCTURE

We define the usefulness of the hyperlink structure of the set of retrieved documents D as the information theoretic divergence between two probability distributions. The first one is the distribution S of the content analysis scores s_i for the documents $d_i \in D$. The second distribution is constructed so as to favour the relevant documents that link to other relevant documents. This is a desired property of the second distribution, in the sense that it is more useful for a user, who is browsing a top ranked document, to navigate to other relevant but not necessarily highly ranked documents. The new score u_i for document $d_i \in D$ depends on its content-based score s_i , as well as on the content-based scores s_j of all retrieved documents d_j , which are pointed by d_i :

$$u_i = s_i + \sum_{d_i \rightarrow d_j} s_j, \quad d_i, d_j \in D \quad (1)$$

where $d_i \rightarrow d_j$ means that there is a hyperlink from d_i to d_j . We normalise both distributions to probabilities as follows:

$$sn_i = \frac{s_i}{\sum_{d_j \in D} s_j} \quad un_i = \frac{u_i}{\sum_{d_j \in D} u_j} \quad (2)$$

Having defined the distributions $S_n = \{sn_i\}$ and $U_n = \{un_i\}$, we employ their Kullback-Leibler symmetric divergence J [19]¹, as a measure of the usefulness of the hyperlink structure:

$$J(S_n, U_n) = \sum_{d_i \in D} un_i \log_2 \frac{un_i}{sn_i} + \sum_{d_i \in D} sn_i \log_2 \frac{sn_i}{un_i} \quad (3)$$

Before continuing, we should note two issues. First, $J(S_n, U_n)$ is almost positive definite, that is $J(S_n, U_n) \geq 0$, with equality if and only if the distributions S_n and U_n are equivalent. However, there is no upper bound for $J(S_n, U_n)$. This is addressed by employing the Jensen-Shannon symmetric divergence L , also called the total divergence to the average [20]:

$$L(S_n, U_n) = \sum_{d_i \in D} un_i \log_2 \frac{un_i}{\frac{un_i}{2} + \frac{sn_i}{2}} + \sum_{d_i \in D} sn_i \log_2 \frac{sn_i}{\frac{un_i}{2} + \frac{sn_i}{2}} \quad (4)$$

¹The notation for the divergence measures is taken from [19, 20].

One of the properties of the Jensen-Shannon divergence is that it is upper bounded: $L(S_n, U_n) \leq 2$ [20].

The second issue with respect to $J(S_n, U_n)$ is that the distributions S_n and U_n should be mutually absolutely continuous. In other words, we should have $sn_i = 0$ for all i for which $un_i = 0$ and *vice versa*. In our case, this condition is satisfied because in (1), we have $u_i \geq s_i$ and $s_i > 0$ for all $d_i \in D$. This condition is not necessary for the Jensen-Shannon divergence. Thus, the divergence $L(S_n, U_n)$ is defined even if $sn_i > 0$ but $un_i = 0$, or *vice versa*. We consider this property of the Jensen-Shannon divergence to define the distributions $U' = \{u'_i\}$ and $U'_n = \{un'_i\}$ as follows:

$$u'_i = \sum_{d_i \rightarrow d_j} s_j \quad un'_i = \frac{u'_i}{\sum_{d_j \in D} u'_j} \quad (5)$$

The distribution $\{u'_i\}$ differs from $\{u_i\}$ in the sense that the dependence $u_i \geq s_i$, which satisfied the condition of mutual absolute continuity in Kullback-Leibler divergence, is removed. For the distribution $\{u'_i\}$, it is easy to verify that if a document d_i does not have outgoing links, then $u'_i = 0$. Moreover, if d_i points to documents with low scores, it may be the case that $0 < u'_i < s_i$. Therefore, the third divergence measure we use is $L(S_n, U'_n)$.

We assume that when the divergence is high, there exist non-random patterns in the distribution of the hyperlinks. On the other hand, when the divergence is low, we assume that there is no apparent pattern in the way hyperlinks are distributed among the retrieved documents for a query. We continue with an example to show how the divergence between S_n and U_n or U'_n depends on the hyperlink structure of the retrieved documents. Suppose that we have a ranked list of six documents, with distribution $\{s_i\} = \{0.5, 0.4, 0.2, 0.2, 0.1, 0.1\}$, i.e. the score of document 1 is 0.5, the score of document 2 is 0.4, etc. We will compute $J(S_n, U_n)$ for four different arrangements of hyperlinks among the documents.

In Figure 1(a), the first arrangement corresponds to a case, where there is no apparent pattern in the distribution of hyperlinks. After computing the distributions S_n and U_n , we have $J(S_n, U_n) = 0.2738$. In Figure 1(b), the second arrangement corresponds to a case, where the top ranked documents are strongly connected and we find that $J(S_n, U_n) = 0.4070$. For the third arrangement, shown in Figure 1(c), there is a group of documents that are strongly connected, without all of them being highly ranked. In this case we find that $J(S_n, U_n) = 0.7227$. For the last case, suppose that the graph of the example is complete, that is it contains one hyperlink between any ordered pair of documents. The divergence $J(S_n, U_n)$ for this arrangement is equal to 0.5203.

If we look at the computed values for $J(S_n, U_n)$ and the corresponding structure of the hyperlinks for each of the four cases described above, we can see that $J(S_n, U_n)$ has the lowest value when there is no pattern in the way hyperlinks are distributed, and increases when there is a connected group of documents. The increase is higher if the documents from the connected group are ranked lower in the list of documents. In this case, we assume that the information from the hyperlink structure is more useful. The measures $L(S_n, U_n)$ and $L(S_n, U'_n)$ give similar results. We aim to use these measures to detect queries for which there are some patterns in the distribution of hyperlinks. Consequently, we would be able to apply an appropriate retrieval approach for each query.

So far, we have defined the three measures of usefulness of the hyperlink structure, without imposing any constraint on the retrieval approaches that could be used in document ranking. In the next section, we evaluate the effectiveness of each measure in the context of a decision mechanism for selecting optimally the most appropriate retrieval approaches for each query.

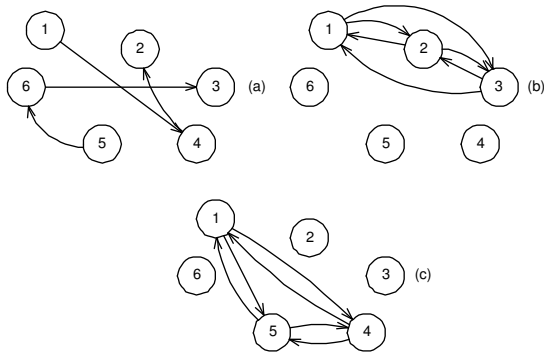


Figure 1: The hyperlink graphs of the ranked documents, corresponding to the first three cases described in the example.

3. EVALUATING THE USEFULNESS MEASURES

For the second step of our methodology, we introduce a decision mechanism, based on the usefulness measures, for selecting the most appropriate retrieval approach on a per-query basis. For each task, we form a set of candidate approaches, corresponding to the most effective retrieval approaches (Section 3.1). Then, we use Bayesian decision theory to select the most appropriate candidate approach for each query (Section 3.2). The evaluation of the usefulness measures is based on the overall effectiveness of the decision mechanism, assuming that relevance information is available.

3.1 Selection of candidate approaches

There are many different retrieval approaches, in addition to content analysis of documents, which can be applied for a query. For example, we can employ hyperlink structure analysis [17, 6], the anchor text of incoming links [8], or use information from the URL of documents [18].

We consider the following retrieval approaches: content-only retrieval (C), content and anchor text retrieval (CA), content with anchor text retrieval and URL length (CAU). For all three approaches, we employ the weighting scheme $PL2^2$ from Amati and Van Rijsbergen’s probabilistic framework of Divergence from Randomness (DFR) [2], where the only parameter of the system is automatically set to $c = 1.28$ [15]. For CA, we extend documents with the anchor text of their incoming links and perform content analysis. For CAU, we re-rank the top 1000 documents retrieved by CA, by taking into account the length of their URL as shown below:

$$s_i = \frac{sca_i}{\log_2(urlpath_len_i + 1)} \quad (6)$$

where sca_i is the score of the content and anchor text analysis, and $urlpath_len_i$ is the length in characters of d_i ’s URL path. We did not use query expansion in any of the above retrieval approaches.

We experiment with the data from the topic distillation tasks of TREC11 [7] and TREC12 [9] on the .GOV collection (for indexing the collection, we removed standard stop words and applied stemming). Both tasks involve 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 to be homepages of relevant sites. This definition resulted in a lower number of relevant documents for the TREC12 task, affecting the stability of precision at 10 documents, which was the evaluation measure for

²The formula of $PL2$ is shown in the appendix of the paper.

the TREC11 task. For TREC12, the Web track organisers chose R-Precision (precision after R documents have been retrieved, where R is the number of relevant documents for a query). For our analysis, we use precision at 10 documents for both TREC tasks and also R-Precision for the TREC12 task.

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

Table 1: Evaluation results of C, CA and CAU for TREC11 and TREC12 topic distillation tasks, respectively.

In this paper, for simplicity and in order to overcome the issue of sparsity of data, we choose the two most effective retrieval approaches as candidates for each task we test, based on relevance information (Table 1)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 are the most effective according to both precision at 10 and R-Precision measures. We have also used the distributions U_n and U'_n , defined in Section 2, for document ranking, but precision was significantly lower than that of C, CA, or CAU.

If we combine manually C and CA for TREC11 into 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 MAX(CA, CAU) of CA and CAU 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.

3.2 Bayesian decision mechanism

After forming the set of candidate retrieval approaches, we proceed to the evaluation of the usefulness measures in the context of a simple decision mechanism. We make a binary decision and select one of the two candidate approaches, using thresholds for the usefulness of the retrieved documents’ hyperlink structure, similarly to the approach taken by Cronen-Townsend et al. [11]. For computing $S = \{s_i\}$, needed for the divergence measures of the first step, we employ the weighting model $PL2$ from the DFR framework.

More specifically, we first group the queries according to which of the candidate approaches is more effective, with respect to precision at 10. We consider only the queries, for which there is a difference in precision at 10 between the candidate approaches, since we are more interested in the top ranked documents, as usually the case in Web IR. For TREC11, there are 9 queries for which C outperforms CA and 8 queries for which CA outperforms C. For TREC12, there are 19 queries for which CAU outperforms CA, and 11 queries where CA results in higher precision at 10 than CAU.

³The best submitted run to TREC11 topic distillation task achieved 0.2510 precision at 10 [7]. The highest R-Precision achieved by the submitted runs to TREC12 topic distillation task was 0.1636, while the highest precision at 10 was 0.1280 (these figures correspond to two different runs) [9].

Then, we compute the prior probabilities of a given retrieval approach being more appropriate for a query (Table 2).

TREC11	TREC12
P(C) 0.5294	P(CA) 0.3667
P(CA) 0.4706	P(CAU) 0.6333

Table 2: The prior probabilities of a retrieval approach being more appropriate for a query (with respect to precision at 10), for both TREC11 and TREC12 topic distillation tasks.

After grouping queries into two classes for each task (the classes are C and CA for TREC11, while for TREC12 they are CA and CAU), we proceed with estimating the conditional probability densities of the divergence measures, given the most appropriate retrieval approach on a per-query basis. We employ kernel density estimation with Gaussian kernels and automatic setting of the bandwidth, using Silverman’s *rule of thumb* [23]. The products of the conditional probability densities for $J(S_n, U_n)$, $L(S_n, U_n)$ and $L(S_n, U'_n)$, with the prior probabilities from Table 2, are shown in Figures 2 and 3 for TREC11 and TREC12, respectively. For example, we can see that when $J(S_n, U_n)$ is used for TREC11, C is more appropriate for queries associated with a low value of $J(S_n, U_n)$, while CA is more appropriate for queries with a higher value of $J(S_n, U_n)$. The results are similar for TREC12 queries, where we select either CA or CAU (see Figure 3).

Next, we set the thresholds for the decision mechanism, equal to the values corresponding to the intersection points of the product of the conditional probability densities and the prior probabilities. In this way, for each query, we apply the retrieval approach with the highest probability of being the most effective. This is optimal in the sense that it minimises the error rate, given that the cost of any wrong decision is the same.

We can extend the above decision mechanism in order to select the most appropriate retrieval approach from any number of candidates. For example, we could compute the prior and conditional probabilities for C, CA and CAU, and then select the most appropriate one. Alternatively, we could compute the divergence between the score distributions of any two retrieval approaches and use this in the decision mechanism. Even though similar improvements are obtained in this way, a restriction is that we can only select between two retrieval approaches for each query.

TREC11	Pr. at 10	Thresholds
C	0.2694	-
MAX(C, CA)	0.2898	-
$J(S_n, U_n)$	0.2796	2.02, 2.12, 2.24, 2.64
$L(S_n, U_n)$	0.2755	0.45, 0.58
$L(S_n, U'_n)$	0.2796	0.79, 1.17

Table 3: Optimal selective application of C and CA for TREC11 topic distillation, using the three measures of the hyperlink structure usefulness and the thresholds, which correspond to the intersection points of the probability densities in Figure 2. C is the best performing baseline.

The evaluation results and the thresholds of the decision mechanism described above are shown in Tables 3 and 4 for TREC11 and TREC12, respectively. We can see that precision increases over the best performing baseline, using any of the divergence measures from Section 2. More specifically, for TREC11, using $J(S_n, U_n)$ or $L(S_n, U'_n)$, for selecting between C and CA, results in the highest precision at 10. However, the selection based on $L(S_n, U'_n)$ is

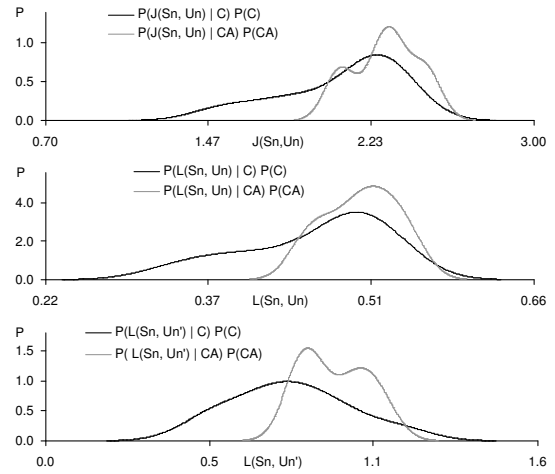


Figure 2: The products of the conditional probability densities and the prior probabilities for the divergence measures of TREC11 queries. The candidate approaches are C and CA.

TREC12	Pr. at 10	R-Pr.	Thresholds
CAU	0.1400	0.1369	-
MAX(CA, CAU)	0.1680	0.1881	-
$J(S_n, U_n)$	0.1540	0.1671	1.79, 2.87
$L(S_n, U_n)$	0.1540	0.1671	0.39, 0.59
$L(S_n, U'_n)$	0.1500	0.1442	0.72, 1.30

Table 4: Optimal selective application of CA and CAU for TREC12 topic distillation, using the three measures of the hyperlink structure usefulness and the thresholds, which correspond to the intersection points of the probability densities in Figure 3. CAU is the best performing baseline.

preferable due to the lower number of thresholds. For TREC12, employing $J(S_n, U_n)$ or $L(S_n, U_n)$ to select either CA or CAU, performs equally well, using two thresholds. Note that the improvements in retrieval effectiveness for TREC12 are reflected in the increased value of both precision at 10 and R-Precision.

Summing up, all three usefulness measures, when employed in the context of an optimal decision mechanism, result in improvements in precision, compared to the results from Table 1. Moreover, the small differences in their performance do not allow us to state that one is more effective than the others.

4. ALTERNATIVE WEIGHTING SCHEMES

For the first step of the proposed methodology, the measures of the hyperlink structure’s usefulness depend on both a content analysis score distribution $S = \{s_i\}$ and the hyperlink structure of the retrieved documents. In this section, we examine how the choice of a weighting scheme for computing the distribution S influences the effectiveness of the usefulness measures, in the context of the second step’s decision mechanism.

For computing the usefulness measures in the experiments of Section 3, the distribution S (and consequently the normalised distribution S_n) was based on the weighting scheme *PL2*. In the following experiments we replace *PL2*, first with the well-known *BM25* formula [22], and then with $I(n_e)C2^4$, a variant of the weighting scheme $I(n_e)B2$ from the DFR framework [2], which

⁴The formula of $I(n_e)C2$ is shown in the appendix of the paper.

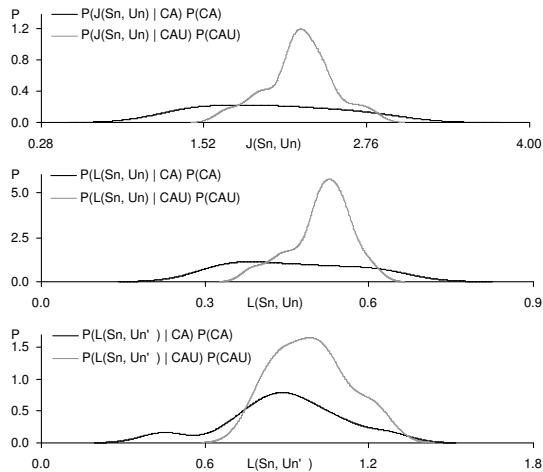


Figure 3: The products of the conditional probability densities and the prior probabilities for the divergence measures of TREC12 queries. The candidate approaches are CA and CAU.

has been used in TREC successfully [1]. The employed weighting schemes are based on different statistical models. Additionally, since the combination of content and anchor text is an effective retrieval approach in the Web context [9], we employ it to estimate the usefulness of the hyperlink structure. We compute the distribution $S = \{s_i\}$ by using *PL2* with content and anchor text, and examine how this affects the retrieval effectiveness of the decision mechanism. The evaluation of the three different weighting schemes, and the combination of content with anchor text for both TREC11 and TREC12 is shown in Table 5. We can see that all schemes perform well. *PL2* is the most effective for TREC11, while *PL2* with anchor text is the most effective for TREC12.

TREC11	Aver. Pr.	Pr. at 10	R-Pr.
<i>PL2</i>	0.2053	0.2694	0.2362
<i>BM25</i>	0.1919	0.2408	0.2067
<i>I(n_e)C2</i>	0.1984	0.2490	0.2229
<i>PL2</i> + anchor	0.1953	0.2551	0.2237
TREC12			
<i>PL2</i>	0.0886	0.0680	0.0730
<i>BM25</i>	0.0999	0.0720	0.0934
<i>I(n_e)C2</i>	0.0832	0.0680	0.0733
<i>PL2</i> + anchor	0.1273	0.1020	0.1325

Table 5: Evaluation results of *PL2* (equivalent to approach C), *BM25*, *I(n_e)C2* and *PL2* with content and anchor text of documents (equivalent to approach CA) for TREC11 and TREC12 topic distillation tasks, respectively.

The evaluation of the usefulness measures in the context of a decision mechanism is carried out in the same way as in Section 3.2, using the Bayesian decision mechanism. We only use the different weighting schemes and the anchor text to compute the divergence values, independently of the document ranking. In all cases, the candidate retrieval approaches are based on the weighting scheme *PL2*. We adopt this approach in order to focus our evaluation on the effectiveness of the usefulness measures. The results are shown in Tables 6 and 7 for TREC11 and TREC12, respectively.

Overall, we can see that using different weighting schemes does not affect the overall performance of the decision mechanism sig-

nificantly. Both average effectiveness measures and thresholds remain stable, with some exceptions. For example, it is more effective for TREC11 to use $J(S_n_{BM25}, U_n)$ than $J(S_n, U_n)$. The results show that improvements over the baseline C (Table 1) are obtained, irrespectively of the weighting scheme used for computing the divergences. This lack of strong dependence between the choice of a weighting scheme for computing the divergence and the effectiveness of the decision mechanism is a desirable property for the hyperlink structure information value measures, since they should depend primarily on the hyperlinks between the retrieved documents.

TREC11	Pr. at 10	Thresholds
$J(S_n, U_n)$	0.2796	2.02, 2.12, 2.24, 2.64
$L(S_n, U_n)$	0.2755	0.45, 0.58
$L(S_n, U'_n)$	0.2796	0.79, 1.17
$J(S_n_{BM25}, U_n)$	0.2837	1.84, 2.06, 2.28, 2.51
$L(S_n_{BM25}, U_n)$	0.2612	0.40, 0.46, 0.50, 0.55
$L(S_n_{BM25}, U'_n)$	0.2796	0.83, 1.18
$J(S_n_{I(n_e)C2}, U_n)$	0.2796	1.92, 2.11, 2.34
$L(S_n_{I(n_e)C2}, U_n)$	0.2796	0.42, 0.46, 0.52
$L(S_n_{I(n_e)C2}, U'_n)$	0.2776	0.79, 1.17
$J(S_n_{anchor}, U_n)$	0.2714	1.75, 1.97, 2.27
$L(S_n_{anchor}, U_n)$	0.2776	0.38, 0.45, 0.52
$L(S_n_{anchor}, U'_n)$	0.2612	0.85, 1.18

Table 6: Optimal selective application of C and CA for TREC11 topic distillation, using different weighting schemes for the computation of the divergences. The baseline is C from Table 1. Unless otherwise noted, S_n is based on *PL2*.

TREC12	Pr. at 10	R-Pr.	Thresholds
$J(S_n, U_n)$	0.1540	0.1671	1.79, 2.87
$L(S_n, U_n)$	0.1540	0.1671	0.39, 0.59
$L(S_n, U'_n)$	0.1500	0.1442	0.72, 1.30
$J(S_n_{BM25}, U_n)$	0.1520	0.1642	1.77, 2.84
$L(S_n_{BM25}, U_n)$	0.1520	0.1642	0.39, 0.60
$L(S_n_{BM25}, U'_n)$	0.1500	0.1442	0.72, 1.30
$J(S_n_{I(n_e)C2}, U_n)$	0.1520	0.1642	1.79, 2.83
$L(S_n_{I(n_e)C2}, U_n)$	0.1520	0.1642	0.39, 0.59
$L(S_n_{I(n_e)C2}, U'_n)$	0.1500	0.1442	0.75, 1.30
$J(S_n_{anchor}, U_n)$	0.1520	0.1648	1.85, 2.77
$L(S_n_{anchor}, U_n)$	0.1540	0.1698	0.40, 0.58
$L(S_n_{anchor}, U'_n)$	0.1480	0.1442	0.78, 1.33

Table 7: Optimal selective application of CA and CAU for TREC12 topic distillation, using different weighting schemes for the computation of the divergences. The baseline is CAU from Table 1. Unless otherwise noted, S_n is based on *PL2*.

5. SAMPLING DOCUMENTS

Any approach that aims to optimise the retrieval effectiveness of a system should not affect its responsiveness significantly. From this perspective, the topic distillation queries retrieve on average several tens of thousands of documents from the .GOV collection, and computing the divergence measures may result in processing a significant number of hyperlinks for the propagation of scores. In order to reduce the computational overhead of the first step, we examine the effect of sampling a subset of top ranked documents, instead of the whole set of retrieved documents.

We select the subset D^k of the top k ranked documents from the set D , with respect to the score distribution $\{s_i\}$ and compute the distributions of scores from these documents, following hyperlinks to documents in the whole set of retrieved documents D . We employ this approach, because if we considered only hyperlinks between documents in D^k , we would restrict the number of processed hyperlinks significantly. This results in the subsequent modification of the distributions $\{u_i\}$ and $\{u'_i\}$:

$$u_i = s_i + \sum_{d_i \rightarrow d_j} s_j \quad u'_i = \sum_{d_i \rightarrow d_j} s_j, \quad d_i \in D^k, d_j \in D \quad (7)$$

For the experiments, we set k equal to 100, 1000 and 10000 documents for computing $J(S_n, U_n)$, $L(S_n, U_n)$ and $L(S_n, U'_n)$, respectively. Using the decision mechanism described in Section 3.2, we compute the thresholds for each divergence and we employ them to select the most appropriate retrieval approach on a per-query basis. The baseline for these experiments is the case where all retrieved documents are used to compute the divergence measures. Results for TREC11 topic distillation are shown in Table 8, where we can see that sampling of documents does not really affect precision for $J(S_n, U_n)$ and $L(S_n, U_n)$. For the measure $L(S_n, U'_n)$, we even see that there is a slight increase in precision at 10 when only 1000 or 10000 documents are sampled.

Docs.	$J(S_n, U_n)$	$L(S_n, U_n)$	$L(S_n, U'_n)$
	Pr. at 10	Pr. at 10	Pr. at 10
100	0.2735	0.2673	0.2776
1000	0.2776	0.2755	0.2816
10000	0.2735	0.2735	0.2816
All	0.2796	0.2755	0.2796

Table 8: Optimal selective application of C and CA for TREC11 topic distillation, with sampling of documents for computing the divergences. The last row, 'All', corresponds to the baseline, where we use all retrieved documents to compute the divergences.

The experiments for TREC12, shown in Table 9, result in similar effectiveness as in the case of using the whole set of retrieved documents D . If we consider precision at 10, we can see that the measures $J(S_n, U_n)$ and $L(S_n, U_n)$ are the most effective for $k = 10000$ documents. With respect to R-Precision, using document sampling with $L(S_n, U'_n)$ does not really affect precision. For $J(S_n, U_n)$ and $L(S_n, U_n)$, we notice an improvement when $k = 100$, but when k increases, R-Precision is lower. Another interesting observation is that for $J(S_n, U_n)$ and $L(S_n, U_n)$, there is a difference between the most appropriate k value for precision at 10 and for R-Precision. The highest R-Precision results from using $k = 100$ documents for sampling, while we obtain the highest precision at 10 when $k = 10000$. Therefore, depending on the used evaluation measure, we could set k to a lower or a higher value.

Overall, reducing the number of documents used for computing the divergence measures, does not really affect the precision of the decision mechanism. Also, this result may indicate that the information obtained from the outgoing hyperlinks of the top-ranked documents is more useful than that obtained from the outgoing hyperlinks of lower-ranked documents. In terms of timing, there is a 92%, 90% and 79% average decrease in processing time when 100, 1000 and 10000 documents are sampled, respectively. The improved timings and the quite stable precision suggest that document sampling can be used in an operational environment to reduce the computational overhead of computing the divergence measures. We will use this conclusion in the next section.

Docs.	$J(S_n, U_n)$		$L(S_n, U_n)$		$L(S_n, U'_n)$	
	Pr. 10	R-Pr.	Pr. 10	R-Pr.	Pr. 10	R-Pr.
100	0.1440	0.1686	0.1420	0.1686	0.1460	0.1426
1000	0.1420	0.1432	0.1420	0.1499	0.1460	0.1398
10000	0.1580	0.1533	0.1580	0.1655	0.1460	0.1398
All	0.1540	0.1671	0.1540	0.1671	0.1500	0.1442

Table 9: Optimal selective application of CA and CAU for TREC12 topic distillation, with sampling of documents for computing the divergences. The last row, 'All', corresponds to the baseline, where we use all retrieved documents to compute the divergences.

6. DECISION MECHANISM WITHOUT RELEVANCE INFORMATION

The experiments of all previous sections are based on assuming that relevance information exists. Indeed, we can choose the best two retrieval approaches as candidates for each task, and select one of them optimally, using Bayesian decision theory. However, in an operational environment, it is not likely that we will have relevance information readily available. Therefore, it is important to have a method for selecting the candidate approaches and setting the threshold values of the decision mechanism automatically.

For selecting the candidate approaches, it is reasonable to assume that we can choose the two most effective retrieval approaches, based on a training process. The potential for improvement is higher if each of the two retrieval approaches outperforms the other on a large subset of the queries. In this section's experiments, we will employ the candidates C and CA for TREC11, and CA and CAU for TREC12, similarly to the previous experiments.

For setting the thresholds, we employ a similar method to that of Cronen-Townsend et al. [11]. We compute the divergence measures by considering a sample of terms from the lexicon as one-term queries. We ignore the terms that appear in less than 1000 documents in the collection, since those terms would retrieve few documents with a small number of hyperlinks, biasing the computed divergence measures. Considering the results from the previous section, and in order to minimise the training time, we use the top 100 retrieved documents to compute the divergence measures. The estimated probability densities are shown in Figure 4. We can see that the probability densities of $J(S_n, U_n)$ and $L(S_n, U_n)$ are very similar, while $L(S_n, U'_n)$ is more symmetric. We select three threshold values, for which 30%, 50%, or 70% of the computed divergence values are below the threshold.

We introduce a decision mechanism that uses evidence from the hyperlink structure, or the URLs, when we detect useful patterns in the distribution of hyperlinks. Hence, if the divergence measure is lower than a threshold, we apply C for the associated TREC11 query (CA for a TREC12 query), otherwise we apply CA (CAU for a TREC12 query). As a consequence, for both tasks, when the divergence measure is higher than the threshold, we employ more evidence from the hyperlink structure, or the URLs, to detect the useful entry points of the retrieved documents.

The results are shown in Table 10. For TREC11, we obtain a slight improvement over the best performing baseline C (Table 1) when we use $L(S_n, U'_n)$ with the 30% and 50% thresholds. For $J(S_n, U_n)$ and $L(S_n, U_n)$, we can see that the performance for the 30% and 50% thresholds is equal to that of CA, and increases when we use the 70% threshold. We believe that we only obtained improvements for the 70% threshold, because the sampling of terms, to set the threshold values for $J(S_n, U_n)$ and $L(S_n, U_n)$ automatically, resulted in lower divergence values than those of the queries.

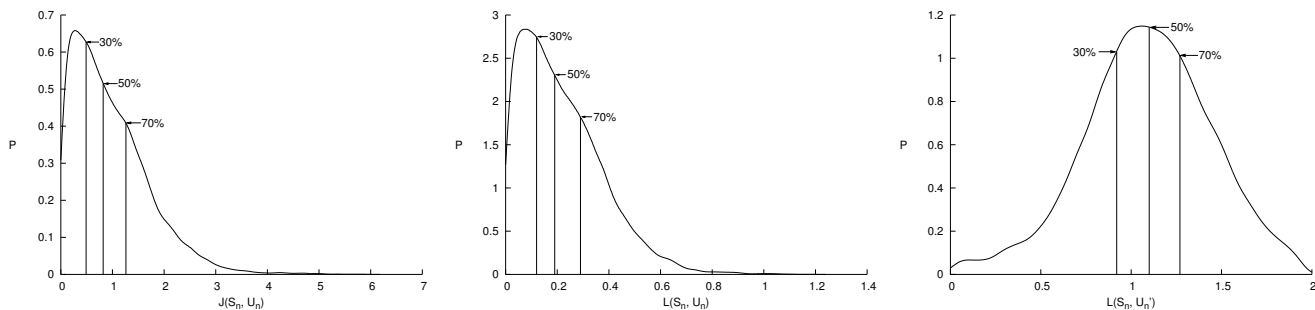


Figure 4: The estimated probability densities from the sampling of terms as queries.

	Threshold	TREC11		TREC12	
		Pr. 10	Pr. 10	R-Pr.	
$J(S_n, U_n)$	0.49 (30%)	0.2551	0.1400	0.1419	
$L(S_n, U_n)$	0.12 (30%)	0.2551	0.1400	0.1369	
$L(S_n, U'_n)$	0.92 (30%)	0.2714	0.1440	0.1619	
$J(S_n, U_n)$	0.82 (50%)	0.2551	0.1420	0.1419	
$L(S_n, U_n)$	0.19 (50%)	0.2551	0.1400	0.1369	
$L(S_n, U'_n)$	1.10 (50%)	0.2735	0.1340	0.1666	
$J(S_n, U_n)$	1.26 (70%)	0.2571	0.1340	0.1644	
$L(S_n, U_n)$	0.29 (70%)	0.2592	0.1400	0.1369	
$L(S_n, U'_n)$	1.27 (70%)	0.2694	0.1340	0.1644	

Table 10: Evaluation of setting the thresholds, without relevance information. The baselines are shown in Table 1.

For TREC12, we achieve similar improvements, depending on the evaluation measure we use. $L(S_n, U'_n)$ and the 30% threshold give the highest precision at 10, while $L(S_n, U'_n)$ and the 50% threshold result in the highest R-Precision. $L(S_n, U_n)$ performs on average the same as CAU (see Table 1), while $J(S_n, U_n)$ results in improved R-Precision. Looking at the results for both TREC tasks, we see that the most reliable measure to use with an automatically set threshold is $L(S_n, U'_n)$.

If we compare the results from Table 10 with the first rows of Tables 8 and 9, where we sample the top 100 documents, we can see that the decision mechanism with the automatically set thresholds, performs nearly as well as the Bayesian decision mechanism, especially when we use $L(S_n, U'_n)$.

7. RELATED WORK

The combination of different sources of evidence has been studied in IR, in order to increase retrieval effectiveness. 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 [14, 13], or on the characteristics of the set of retrieved documents for each query [4, 3, 21]. In all these approaches, the application of hyperlink analysis depends on the assessment of the hyperlink structure, in contrast to our approach, where the hyperlink structure’s usefulness is independent of the document ranking. Moreover, Kang and Kim [16] tried to identify the type of each query from a mixture of topic relevance and home page finding queries, and then applied a retrieval approach for each query type.

Our approach is related to the work of Cronen-Townsend et al. [11], in the sense that we employ a similar way to compute the thresholds of the decision mechanism, with or without relevance information. On the other hand, the information theoretic measures differ from query clarity in that we use the hyperlink structure of the retrieved

documents. Additionally, computing the usefulness of the hyperlink structure is not directly related to document ranking.

8. CONCLUSIONS

In this paper, we present a set of three information theoretic measures that estimate whether there are useful patterns in the distribution of hyperlinks of the retrieved documents. More specifically, we measure the divergence between the content analysis score distribution and a distribution of scores, modified by a single-step propagation of scores through the hyperlinks. All three measures estimate the degree to which the hyperlink structure is random or not, independently of the document ranking. The higher values of the measures signify that the distribution of hyperlinks among the retrieved documents contains some patterns, which may correspond to clusters of related documents.

We evaluate the information theoretic measures in the context of a methodology with two independent steps. First, we compute the divergence measures, independently of the document ranking. Then, we use the computed measures in a decision mechanism based on Bayesian decision theory, where the most appropriate approach, from a set of candidate ones, is used for retrieval on a per-query basis. Having two steps enables us to employ any retrieval approach for the second step, thus making our methodology more general. We assume that there exists relevance information and we find the optimal setting for the decision mechanism. In this case, we obtain improvements over the best baseline for both TREC11 and TREC12 topic distillation tasks. Moreover, similar improvements are obtained when we compute the divergences using alternative weighting schemes, or when we use sampling of documents to reduce the computational overhead.

We employ sampling of documents, in order to experiment and evaluate a decision mechanism, where the thresholds are set automatically, without relevance information. The experiments show that the automatic setting can lead to improvements over the most effective baseline (C for TREC11 and CAU for TREC12), especially when we use the divergence measure $L(S_n, U'_n)$.

There are several interesting directions in which we can refine this work. The first is forming the set of the candidate approaches, and the evaluation of selecting an approach from more than two candidates, in the context of the Bayesian decision theory. A second direction is looking into refined ways to compute the thresholds without relevance information. In addition, we plan to test our approach on other appropriate Web test collections, as they become available.

9. ACKNOWLEDGEMENTS

This work 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>).

10. REFERENCES

- [1] G. Amati, C. Carpineto, and G. Romano. FUB at TREC 10 web track: a probabilistic framework for topic relevance term weighting. In *NIST Special Publication: SP 500-250 The Tenth Text Retrieval Conference (TREC 2001)*, Gaithersburg, MD, 2001.
- [2] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.
- [3] E. Amitay, D. Carmel, A. Darlow, M. Herscovici, R. Lempel, and A. Soffer. Juru at TREC 2003 - Topic Distillation Using Query-Sensitive Tuning and Cohesiveness Filtering. In *NIST Special Publication: SP 500-255 The Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD, 2003.
- [4] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. Topic Distillation with Knowledge Agents. In *NIST Special Publication: SP 500-251 The Eleventh Text Retrieval Conference (TREC 2002)*, Gaithersburg, MD, 2002.
- [5] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 104–111. ACM Press, 1998.
- [6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
- [7] N. Craswell and D. Hawking. Overview of the TREC-2002 Web Track. In *NIST Special Publication: 500-251 The Eleventh Text REtrieval Conference (TREC 2002)*, Gaithersburg, MD, 2002.
- [8] N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 250–257. ACM Press, 2001.
- [9] N. Craswell, D. Hawking, R. Wilkinson, and M. Wu. Overview of the TREC-2003 Web Track. In *NIST Special Publication: 500-255 The Twelfth Text REtrieval Conference (TREC 2003)*, Gaithersburg, MD, 2003.
- [10] W. B. Croft. Combining approaches to information retrieval. In W. B. Croft, editor, *Advances in Information Retrieval from the Center for Intelligent Information Retrieval*, pages 1–36. Kluwer Academic, 2000.
- [11] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 299–306. ACM Press, 2002.
- [12] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, USA, 1973.
- [13] M. Fisher and R. Everson. When Are Links Useful? Experiments in Text Classification. In *Advances in Information Retrieval: 25th European Conference on IR Research*, pages 41–56. Springer-Verlag, 2003.
- [14] C. Gurrin and A. F. Smeaton. Improving the Evaluation of Web Search Systems. In *Advances in Information Retrieval: 25th European Conference on IR Research*, pages 25–40. Springer-Verlag, 2003.
- [15] B. He and I. Ounis. A study of parameter tuning for term frequency normalization. In *Proceedings of the 12th international Conference on Information and Knowledge Management (CIKM)*, pages 10–16. ACM Press, 2003.
- [16] I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 64–71. ACM Press, 2003.
- [17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [18] W. Kraaij, T. Westerveld, and D. Hiemstra. 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*, pages 27–34. ACM Press, 2002.
- [19] S. Kullback. *Information Theory and Statistics*. Jown Wiley & Sons, New York, USA, 1959.
- [20] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37:145–151, Jan. 1991.
- [21] V. Plachouras, F. Cacheda, I. Ounis, and C. J. van Rijsbergen. University of Glasgow at the Web track: Dynamic Application of Hyperlink analysis using the Query Scope. In *NIST Special Publication: SP 500-255 The Twelfth Text Retrieval Conference (TREC 2003)*, Gaithersburg, MD, 2003.
- [22] S. E. Robertson and S. Walker. 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*, pages 232–241. Springer-Verlag New York, Inc., 1994.
- [23] B. Silverman. *Density Estimation*. Chapman & Hall, London, 1986.

APPENDIX

The formulae of the weighting schemes $PL2$ and $I(n_e)C2$ (a variant of $I(n_e)B2$), which we employed from Amati and Van Rijsbergen’s DFR framework [2], are shown below:

$$w_{PL2}(t) = \left(tfn_1 \cdot \log_2 \frac{tfn_1}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tfn_1} - tfn_1 \right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tfn_1) \right) \cdot \frac{1}{tfn_1 + 1}$$

$$w_{I(n_e)C2}(t) = \frac{\text{Freq}(t|\text{Collection}) + 1}{\text{doc_freq} \cdot (tfn_2 + 1)} \left(tfn_2 \cdot \ln \frac{N + 1}{n_e + 0.5} \right)$$

where:

λ is the mean and variance of a Poisson distribution,
 $tfn_1 = \text{term_freq} \cdot \log_2 \left(1 + c \cdot \frac{\text{average_document_length}}{\text{document_length}} \right)$,

$tfn_2 = \text{term_freq} \cdot \ln \left(1 + c \cdot \frac{\text{average_document_length}}{\text{document_length}} \right)$,

N is the size of the collection,

$n_e = N \cdot \left(1 - \left(\frac{1}{N} \right)^{\text{Freq}(t|\text{Collection})} \right)$,

$\text{Freq}(t|\text{Collection})$ is the within-collection term-frequency,
 term_freq is the within-document term-frequency and
 doc_freq is the document-frequency of the term.