# Dempster-Shafer theory for a query-biased combination of evidence on the Web

Vassilis Plachouras
*University of Glasgow, Glasgow G12 8QQ, UK*

Iadh Ounis
*University of Glasgow, Glasgow G12 8QQ, UK*

**Abstract.** This paper reports on a large-scale experiment for the evaluation of a formal query-biased combination of evidence mechanism. We use the Dempster-Shafer theory of evidence to combine optimally results obtained by content and link analyses on the Web. The query-biased mechanism is based on the query scope, a measure of the query specificity. The query scope is defined using a probabilistic propagation mechanism on top of the hierarchical structure of concepts provided by WordNet. We use two standard Web test collections and two different link analysis approaches. The results show that the proposed approach could improve the retrieval effectiveness.

**Keywords:** Combination of content and link analysis, Query-Biased combination of evidence, Dempster-Shafer theory of evidence, Web Information Retrieval, Query scope

## 1. Introduction

In classical Information Retrieval (IR), the content analysis of the documents is used to decide whether a document is relevant to a particular query (Van Rijsbergen, 1979). Additionally, on the World Wide Web there is another source of evidence that a Web IR system can explore, namely the hyperlink structure of documents. It has been claimed that hyperlinks can be used, in combination with the content analysis, to detect high quality documents, or what is commonly called the *authority* documents corresponding to a given query (Amento et al., 2000; Kraaij et al., 2002; Silva et al., 2000).

The analysis of the hyperlink structure has received a lot of attention. Kleinberg (1999) proposed the HITS algorithm, where he suggests that Web documents have two qualities; they are hubs and authorities, which ideally form bipartite graphs. The authorities are the relevant documents for a given topic, while the hubs are Web documents which point to the relevant authorities. Moreover, there is a *mutual reinforcement relation* between authorities and hubs: good authorities are pointed by many good hubs and good hubs point to many good authorities. The algorithm works as follows: initially a set of Web documents is retrieved by using a standard search engine and this set is extended by adding documents that are pointed, or that point to the documents in the initial set. The adjacency matrix $A$ of the graph that corresponds to the extended set is created and the principal eigenvectors of the matrices $AA^T$ and $A^TA$ are computed. The component of each document in the principal eigenvector of $AA^T$ corresponds to its hub weight, while its component in the principal eigenvector of $A^TA$ corresponds to its authority value.

Another important contribution is PageRank, introduced by Brin and Page (1998). This algorithm is based on the calculation of the probability of visiting a Web document in the Markov chain induced from the Web graph and returns a global authority score for each indexed Web document. The authority score of a document depends on the authority scores of the Web documents that point to it, and it is calculated by an iterative process. However, the corresponding Markov chain does not always guarantee the convergence of the process. To overcome this limitation, the concept of a rank source is introduced, which replenishes the rank lost in the dangling nodes, or the sinks, that is structures of documents that do not have any outgoing links to other Web documents. This transformation of the Web graph can be interpreted in the following way: a random user that navigates in the Web has two possibilities at each step: either to follow a link from the document that he is currently browsing, or to jump to a randomly selected Web document and continue his navigation from that document. The addition of this element of randomness results into a more stable algorithm (Zheng et al., 2001), and guarantees the existence of the invariant distribution of the corresponding Markov chain.

Extensions and refinements of HITS and PageRank are discussed in (Bharat and Henzinger, 1998; Chakrabarti et al., 1998; Cohn and Chang, 2000; Lempel and Moran, 2000; Kao et al., 2002; Calado et al., 2003) and (Kim and Lee, 2002; Haveliwala, 2002; Richardson and Domingos, 2002; Diligenti et al., 2002) respectively. In all extensions, link analysis is intended to complement content analysis and to improve precision at the top retrieved documents. However, the reported results were not evaluated in a TREC-like laboratory setting. On the contrary, earlier TREC experiments (Hawking and Craswell, 2001; Craswell and Hawking, 2002) suggest that hyperlink analysis does not enhance retrieval effectiveness for ad-hoc retrieval tasks, although the results from TREC12 topic distillation task showed the potential benefit from hyperlink analysis (Craswell et al., 2003).

Most of the proposed approaches have employed an ad-hoc way to combine content and link analyses, without taking into account each individual query. However, queries exhibiting different characteristics require a modified combination of content and link analyses (Plachouras et al., 2003a). Experience on the Web suggests that queries on very specific topics, or on topics not well represented on the Web can hardly benefit from link analysis, because some relevant pages are not popular, as they are dedicated to a specialised audience, and therefore, they are not pointed by many links. On the other hand, it is commonly accepted that the application of link analysis increases precision among the top ranked documents for queries on broad, or popular topics (Kleinberg, 1999). As a consequence, we believe that there is a need for an optimal combination of results obtained from content and link analyses, with respect to the *specificity* of the query.

The notion of specificity has been employed in IR for quantifying the discriminatory power of terms. For example, *idf* (Spärck Jones, 1972) and its extensions (Aizawa, 2000; Wong and Yao, 1992; Rölleke, 2003) weight terms according to the number of documents they appear in. Moreover, Ruthven et al. (2002) define the specificity of a document as the sum of the *idf* of its terms, divided by the length of the document. From a different perspective, Cronen-Townsend et al. (2002), instead of defining a measure of specificity, they model the clarity of a query as the divergence of the query language model from the collection language model.

In this paper, we introduce the notion of query specificity in order to bias the combination of both content and link analyses. Deciding dynamically about the optimal combination demands a method for estimating a measure of how specific a query is. We assume that, given two queries $q_1$ and $q_2$, if $q_1$ is more specific than $q_2$, then the corresponding specificity value $v_1$ for $q_1$ is smaller than the specificity value $v_2$ for $q_2$. We relate the proposed specificity measure, which we call *query scope*, to both the term frequencies in the collection and an approximation of the conceptual content of the query. The latter is achieved by employing an hierarchical structure of concepts, such as WordNet (Miller, 1995; Fellbaum (ed.), 1998), a lexical reference system, in which terms are associated with a set of underlying concepts, and concepts are linked with various types of relations.

We interpret the hierarchical structure provided by WordNet in two ways and, given a particular collection of Web documents, we define two different probabilistic methods for estimating the query scope. Results obtained from content and link analyses are then optimally combined using Dempster-Shafer's theory of evidence (Shafer, 1976). This process can be seen as a dynamic query-biased process, where each source of evidence is assigned a measure of uncertainty, depending on the query characteristics. The focus of this paper is the definition of the query scope and the optimal query-biased combination of evidence.

For the evaluation of the proposed methodology, we use two standard TREC Web test collections, namely the WT10g (Hawking and Craswell, 2001) and the .GOV (Craswell and Hawking, 2002). Moreover, we employ two different link analysis approaches: the well-established PageRank (Brin and Page, 1998) and the Absorbing Model (Plachouras et al., 2003b; Amati et al., 2003), a new well-founded model for link analysis.

The rest of the paper is organised as follows. In Section 2, we describe two probabilistic approaches for defining a measure of the query scope. In Section 3, we present a method for combining different sources of evidence, based on Dempster-Shafer's theory of evidence. The experiments are described in details and the results are presented in Sections 4 and 5 respectively. Section 6 contains a discussion of the proposed approach, and in Section 7, we present our conclusions from this work.

## 2. Query scope

We define the *query scope* as a probabilistic measure of how specific a query is, in a step-wise manner. More specifically, considering the query as a bag of terms, we define the query scope as a function of the *term scope* of its composing terms, that is, a measure of how specific its composing terms are. We compute the scope of a term according to how specific its associated concepts are. The estimation of the term scope is based on defining a probability measure for concepts on top of WordNet's hierarchical structure of concepts (Miller, 1995; Fellbaum (ed.), 1998). For example, part of this hierarchical structure is shown in Figure 1, where each set of terms represents a concept.

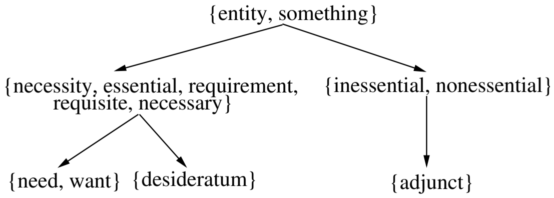
{entity, something}
{necessity, essential, requirement, requisite, necessary}
{inessential, nonessential}
{need, want}
{desideratum}
{adjunct}

*Figure 1.* Part of WordNet's hierarchical structure of concepts.

We interpret WordNet's structure of concepts in two ways. First, we consider it as a lattice, where the probability of a concept propagates to its directly more generic concepts. For example, in Figure 1, the probability of concept {entity, something} is determined by the probabilities of its direct subconcepts, {necessity, essential, requirement, requisite, necessary} and {inessential, nonessential}. The second approach is based on considering the hierarchical structure of concepts as a set of independent concepts. The probability of a concept depends only on its position in WordNet's hierarchical structure. For example, as shown in Figure 1, concept {inessential, nonessential} is one level below the most generic concept {entity, something}.

In Section 2.1, we define the approach based on interpreting WordNet's structure of concepts as a lattice, while in Section 2.2, we present the approach based on interpreting WordNet's structure of concepts as a set of independent concepts. In Section 2.3, we calculate the scope of a term from the probabilities of its associated concepts in WordNet, and then the scope of the query from the scopes of its composing terms.

### 2.1. WordNet considered as a Lattice

Let us consider an arbitrary lattice $\langle T_C, \leq \rangle$. We will first assign an integer value $m(C)$ to each concept $C$, which is interpreted as the frequency of the concept in the document collection, namely the number of documents in which this concept occurs. We recall that the meaning of $C_1 \leq C_2$ in the lattice $\langle T_C, \leq \rangle$ is that any element in the concept $C_1$ is also an element of the concept $C_2$. In First Order Logic (FOL), this is expressed by the formula: $\forall x C_1(x) \rightarrow C_2(x)$.

The problem of assigning weights, or probabilities to FOL formulas depends on whether the formulas are closed or open, that is, if some quantifiers occur or not in the formulas. For example, if $C(x)$ is a concept, we may decide that its probability $Prob(C(x))$ is given by first defining a model of the language and after assigning a probability distribution to the power set of the domain of the model. The probability of the subset of elements satisfying $C(x)$ is then taken as the probability of $C(x)$. On the other hand, if we consider $\forall x C(x)$, the set of elements satisfying $\forall x C(x)$ is either the empty set or the domain of the model so that its probability must be either 0 or 1. Our first assumption here is to treat only open formulas, that is, formulas in which quantifiers do not occur. It is easy to observe that according to this assumption, if $C_1 \leq C_2$ then $Prob(C_1) \leq Prob(C_2)$.

The second assumption we use is that the document collection forms a set of models $M = \langle D, \models \rangle$ (that is a model of modal logic). The semantics is straightforward: if $d \in M$ then $d \models C(a)$ occurs in the document, with $a$ an individual. As noticed above, we have the problem of assigning weights to existential and universal quantified concepts occurring in a document. We suppose to have formulas in prenex normal form, that is, all quantifiers are at the beginning of the formula. In other words, the quantifiers are applied to an open formula. In order to reduce this problem to a probabilistic workable model of FOL, we need to eliminate suitably the existential and universal quantifiers. We assume

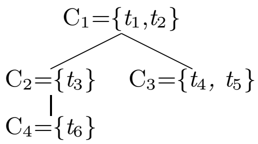

*Figure 2.* Hierarchy of concepts for Example 1

that in indexing our documents, we use only existential formulas. In this case, the process is quite easy. Indeed, we can then introduce for each concept $C$ appearing in the document a *unique* constant, which we call a *witness* and we denote by $p_c$. Then, $d \models C(p_c)$ iff $\exists x \ldots \wedge C(x) \wedge \ldots$ occurs in the document index. This is an important restriction to the formalism of FOL, but it is the assumption usually made to achieve tractable logic-based IR systems (Amati and Ounis, 2000; Ounis, 1998).

After having eliminated both existential and universal quantifiers from the formulas of FOL, we define formally the function $m(C)$ as follows:

**Definition 1** The weight $m(C(x))$ of a concept $C$ is defined recursively as the cardinality of the set $\{d \in D : d \models C'(x), C' \leq C$ for some individual, or witness t$\}$

In other words, $m(C(x))$ is defined recursively as the number of documents in which the concepts $C$ and its direct children occur. Following, we introduce the definition of the probability of a concept $C$:

**Definition 2** The probability $Prob(C(x))$ of a concept $C$ is:

$$Prob(C(x)) = \frac{\sum_{C' \leq C} m(C'(x))}{\sum_{C'} m(C'(x))}. \tag{1}$$

Hence, it turns out that if we consider single $C(a)$ and $C(p)$ as restricted concepts of $C$, that is $C(a) \leq C$ and also $C(a) \leq C(p)$, then $Prob(C(x)) = Prob(C(p))$. Moreover, if $C \leq C'$, then $Prob(C(x)) \leq Prob(C'(x))$. We denote by $Prob(C)$ the probability $Prob(C(x))$.

This is a probability function according to the following interpretation of negation $\neg C$ of $C$: it is the concept which is the union of all concepts not below $C$ in the lattice, that is $\neg C = \cup_{C' \not\leq C} C'$. It is easy then to verify all classical Kolmogorov properties of a probability distribution.

**Example 1** Let the lattice of Figure 2 be the hierarchy of concepts used. Each concept is represented as a set of terms ${t_i}^1$, and the fre-

[1] In Figure 2 the expression $C_1 = \{t_1, t_2\}$ means that concept $C_1$ is associated to terms $t_1$ and $t_2$.

Table I. Frequencies for terms of Example 1.

| Term | Frequency | Concepts |
|---|---|---|
| $t_1$ | 3 | $C_1$ |
| $t_2$ | 4 | $C_1$ |
| $t_3$ | 2 | $C_2$ |
| $t_4$ | 1 | $C_3$ |
| $t_5$ | 2 | $C_3$ |
| $t_6$ | 1 | $C_4$ |

quencies of these terms are shown in Table I. According to Definition 2, for the calculation of the probability of concept $C_2 = \{t_3\}$ we have:

$$Prob(\{t_3\}) = \frac{m(\{t_6\}) + m(\{t_3\})}{m(\{t_1, t_2\}) + m(\{t_3\}) + m(\{t_4, t_5\}) + m(\{t_6\})}.$$

Substituting the values of function $m$, we have that $Prob(\{t_3\}) = 0.23$.

We may also compute the probability of concept $C_1 = \{t_1, t_2\}$ as follows:

$$\begin{aligned} Prob(\{t_1, t_2\}) &= \frac{m(C_2) + m(C_3) + m(\{t_1, t_2\})}{m(\{t_1, t_2\}) + m(\{t_3\}) + m(\{t_4, t_5\}) + m(\{t_6\})} \\ &= \frac{m(\{t_3\}) + m(\{t_6\}) + m(\{t_4, t_5\}) + m(\{t_1, t_2\})}{m(\{t_1, t_2\}) + m(\{t_3\}) + m(\{t_4, t_5\}) + m(\{t_6\})} = 1. \end{aligned}$$

### 2.2. WordNet considered as a Set of Independent Concepts

The second approach proposed is based on the interpretation of the hierarchical structure of concepts of WordNet as a set of independent concepts. Each concept is assigned a weight, or probability, that depends on its position in the hierarchical structure of concepts.

Let $\mathbb{C} = \{C_1, \ldots, C_n\}$ be the set of concepts in the hierarchical structure of WordNet, and each concept $C_i$, where $1 \leq i \leq n$, is at depth $d_i$ in this hierarchical structure. In addition, let $t_k$ be a term which appears in the hierarchical structure of concepts and has a frequency $tf_k$ in the document collection. We also define the sum of all term frequencies as:

$$T = \sum_{t_k} tf_k \tag{2}$$

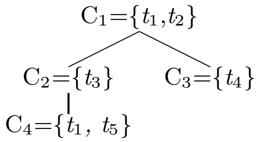

*Figure 3.* Hierarchy of concepts for Example 2

The set $\mathbb{C}_k = \{C_i|$ term $t_k$ is associated to concept $C_i\}$ has $n_k$ elements, its j-th element is denoted by $C_{k,j}$, and the depth of $C_{k,j}$ is denoted by $d_{k,j}$. If there are more than one paths from concept $C_{k,j}$ to the most generic concept in the hierarchy, we consider as the depth $d_{k,j}$ of $C_{k,j}$ the length of the shortest path. Moreover, the maximum depth of concepts $C_{k,j} \in \mathbb{C}_k$ is denoted by $D_k$.

**Definition 3** The *contribution* $a_{k,j}$ of term $t_k$ to concept $C_{k,j}$ is defined by:

$$a_{k,j} = \frac{(D_k + 1) - d_{k,j}}{n_k * (D_k + 1) - \sum_{j=1}^{n_k} d_{k,j}} \tag{3}$$

Then, we define the probability of a concept $C$ as follows:

**Definition 4** The probability of a concept $C \in \mathbb{C}$ is the weighted sum of the term frequency $tf_k$ for each term $t_k$ for which $C \in \mathbb{C}_k$, divided by $T$. The weight of the term frequency for each term $t_k$ is the contribution $a_{k,j}$ of term $t_k$ to concept $C = C_{k,j}$:

$$Prob(C) = \sum_{C = C_{k,j} \in \mathbb{C}_k} a_{k,j} * \frac{tf_k}{T} \tag{4}$$

The probability distribution of the concepts in $\mathbb{C}$ has to satisfy the Kolmogorov properties. It is easy to show that $\forall C\ Prob(C) > 0$ and $\sum_{C \in \mathbb{C}} Prob(C) = 1$, since $\sum_{i=1}^{n_k} a_{k,i} = 1$. In order to calculate the probability of the negation of a concept, we observe that for the calculation of $Prob(C)$, the contributions of a term to the different concepts it belongs to, sum up to one. In that case, it is assumed that the negation $\neg C$ of a concept $C$ is $\mathbb{C} - \{C\}$. Therefore, $Prob(C)$ satisfies the Kolmogorov properties.

**Example 2** Let the lattice of Figure 3 be the hierarchy of concepts used[2]. We calculate the probability of concept $C_4 = \{t_1, t_5\}$ as follows.

[2] Note that the two concepts in Figure 3, namely $C_1$ and $C_4$, are associated with term $t_1$. This is a common situation in the hierarchy of concepts of WordNet, where for example, the term "person" is associated with the concept of a human being and the concept of a grammatical category of pronouns and verb forms.

Table II. Frequencies for terms of Example 2.

| Term | Frequency | Concepts |
|---|---|---|
| $t_1$ | 3 | $C_1, C_4$ |
| $t_2$ | 4 | $C_1$ |
| $t_3$ | 2 | $C_2$ |
| $t_4$ | 1 | $C_3$ |
| $t_5$ | 2 | $C_4$ |

From (2) and the term frequencies of Table II we have: $T = 12$. The term $t_1$ is associated with concepts $C_1$ and $C_4$. Therefore $\mathbb{C}_1 = \{C_1, C_4\}$ and concept $C_4$ is denoted as $C_{1,2}$. According to (3), the contribution $a_{1,2}$ of term $t_1$ to concept $C_{1,2}$ is:

$$a_{1,2} = \frac{(2+1)-2}{2*(2+1)-(0+2)} = 0.25.$$

Similarly, the term $t_5$ is associated with concept $C_4$, so we have $\mathbb{C}_5 = \{C_4\}$ and concept $C_4$ is denoted as $C_{5,1}$. The contribution $a_{5,1}$ is similarly calculated: $a_{5,1} = 1$. From (4), we calculate the probability of concept $C_4$ to be:

$$Prob(C_4) = a_{1,2} * \frac{tf_1}{T} + a_{5,1} * \frac{tf_5}{T} = 0.23$$

### 2.3. Estimation of the query scope

After having defined the probability $prob(C)$ of each concept of WordNet, we calculate the term scope $scope_{t_k}$ for each term $t_k$ using either methods. Let $\mathbb{C}_k$ be the set of concepts associated to term $t_k$.

**Definition 5** The term scope, $scope_{t_k}$, of term $t_k$ is given by:

$$scope_{t_k} = \otimes_{C \in \mathbb{C}_k} Prob(C),$$

where the operator $\otimes$ can be any function such as max, min, *sum*, etc. For example, if we replace $\otimes$ with the function max using either methods for defining the probability $prob(C)$ of concept $C$, we have:

$$scope_{t_k} = \max_{C \in \mathbb{C}_k} Prob(C). \qquad (5)$$

If we consider only the second method proposed in Section 2.2 for defining $Prob(C)$ of concept $C$, we can interpret operator $\otimes$ as the weighted sum of the probabilities of concepts in $\mathbb{C}_k$, where the weight for each concept is the contribution $a_{k,j}$ of term $t_k$ to that concept:

$$scope_{t_k} = \sum_{C=C_{k,j} \in \mathbb{C}_k} a_{k,j} * Prob(C). \tag{6}$$

Once we have defined a measure for the scope of single terms, we need to expand this measure to estimate the scope of a query. Let $q$ be the set of terms for a query.

**Definition 6** The query scope, $scope_q$, of query $q$ is given by:

$$scope_q = \oplus_{t \in q} scope_t,$$

where $\otimes$ is a combination operator.

In this paper, we look into two approaches for the combination operator $\oplus$: the sum and the product of probabilities for single terms, respectively.

**Assumption 1** By taking the sum of values for every term of the query we assume that each term is independent of the others. The contribution of each term's scope is added to the query's scope. Since in this way longer queries would benefit, we normalise by dividing the sum of term scopes by the number of query terms $n$. A measure for the scope of a query $q$ is given by:

$$scope_q = \frac{1}{n} * \sum_{t \in q} scope_t.$$

Alternatively, we can make a different assumption for the independence of terms.

**Assumption 2** We are interested in the scope of the query, in which the terms do not occur independently. Therefore, the co-occurrence of specific terms in the query should contribute more to the overall scope of the query. Again, we need to normalise by multiplying with the number of query terms $n$, since short queries would benefit by this approach. A measure for the scope of a query $q$ is given by:

$$scope_q = n * \prod_{t \in q} scope_t.$$

## 3. Combination of evidence

The combination of two sources of evidence, such as the content analysis and the link analysis, can be modelled using Dempster-Shafer's theory of evidence. This theory introduces the concept of uncertainty in the process of merging different sources of evidence, extending in this way the classical probability theory. Aggregation of different sources of evidence according to a measure of uncertainty is captured by *Dempster's combination rule* (Shafer, 1976). This combination rule is independent of the order in which evidence is gathered.

According to this theory, the set of elements $\Theta = \{\theta_1, \ldots, \theta_n\}$ in which we are interested is called the *frame of discernment.* The measure of uncertainty is based on a *probability mass function* $m$ that assigns zero mass to the empty set, and a value in $[0, 1]$ to each element of $2^\Theta$, the power set of $\Theta$, so that:

$$\sum_{A \subseteq \Theta} m(A) = 1$$

Since we deal with the power set of $\Theta$, which contains not only the base propositions, but all the possible subsets of the set of all base propositions, we can assign the probability mass as we wish, ignoring details we do not know about. Thus, a measure of *uncertainty*, $m(\Theta)$, can be modelled as the probability mass we are unable to assign to any particular subset of $\Theta$. If $A \subseteq \Theta$ and $m(A) > 0$, then $A$ is called a *focal point.* The focal points define a *body of evidence.* Given a body of evidence with a probability mass function $m$, we can compute the total belief given to a subset $A$ of $2^\Theta$ with the *belief function* defined upon $m$:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

When two bodies of evidence are defined in the same frame of discernment, we can combine them using Dempster's combination rule, under the condition that the two bodies are independent of each other. Let $m_1$, $m_2$ be the probability mass functions of the two independent bodies of evidence, defined in the frame of discernment $\Theta$. The probability mass function $m$ defines a new body of evidence in the same frame of discernment $\Theta$ as follows:

$$\begin{aligned} m(A) &= m_1 \oplus m_2(A) \\ &= \frac{\sum_{B \cap C = A} m_1(B) * m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) * m_2(C)} A, B, C \subseteq \Theta \end{aligned} \tag{7}$$

The rule of combination of evidence returns a measure of agreement between two bodies of evidence. The division normalises the new distribution by re-assigning any probability mass which is assigned to the empty set $\emptyset$, by the combination. The corresponding belief function can be easily computed from the mass function $m$.

Coming back to an IR perspective, the frame of discernment $\Theta$ will be the set of Web documents in the collection, i.e. $\Theta = \{d_1, d_2, \ldots, d_n\}$, where $d_i$ is a document of the collection. The scoring functions for the content analysis and the link structure analysis are considered to be the bodies of evidence that will be combined into a single body of evidence in the frame of discernment $\Theta$.

The above combination of evidence is generally computationally expensive. Following (Barnett, 1981; Jose and Harper, 1997; Jose, 1998), we reduce the exponential requirement introduced by the use of the power set $2^\Theta$, to a particular case where we have positive evidence for singleton hypotheses only. Therefore, we assume that the focal elements of our two initial mass functions $m_1$ and $m_2$ are the singleton hypotheses and the frame $\Theta$. In other words, we have positive belief for $\{d_1\}$, $\{d_2\}$, …,$\{d_n\}$ and $\Theta$ only. We denote by $m_1(\Theta)$ and $m_2(\Theta)$ the uncertainties in the bodies of evidence $m_1$ and $m_2$. Hence, the orthogonal sum $m_1 \oplus m_2$, say $m$, can be computed using the combination method in (7):

$$m(\{d_i\}) = \frac{m_1(\{d_i\}) * m_2(\{d_i\}) + m_1(\Theta) * m_2(\{d_i\}) + m_1(\{d_i\}) * m_2(\Theta)}{1 - \sum_{\{d_k\} \cap \{d_j\} = \emptyset} m_1(\{d_k\}) * m_2(\{d_j\})} \tag{8}$$

The denominator in the above equation is a normalising factor and is independent of $\{d_i\}$ (Jose, 1998). Hence, it is not necessary to compute in the ranking process and the above equation can be written as:

$$m(\{d_i\}) \propto \; m_1(\{d_i\}) * m_2(\{d_i\}) + m_1(\Theta) * m_2(\{d_i\}) + m_1(\{d_i\}) * m_2(\Theta) \tag{9}$$

or more simply:

$$m(d) \propto m_1(d) * m_2(d) + m_1(\Theta) * m_2(d) + m_1(d) * m_2(\Theta) \tag{10}$$

We use (10) to compute combined degrees of belief, which is computationally much less expensive than the Dempster's combination rule given in (7).

In our special case, the Dempster's simplified combination rule of (10) for aggregating the ranked lists $l_c$ and $l_l$ obtained from content and link analyses respectively, yields the following formula:

$$m_{c,l}(d) \propto \begin{cases} m_c(d) * m_l(d) + m_c(d) * m_l(\Theta) + m_c(\Theta) * m_l(d) & d \in (l_c \cap l_l) \\ m_c(d) * m_l(\Theta) & d \in (l_c - l_l) \\ m_c(\Theta) * m_l(d) & d \in (l_l - l_c) \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $m_c$, $m_l$ and $m_{c,l}$ denote the bodies of evidence for the content analysis, the link analysis, and the combination of content and link analyses respectively. Since in our combination mechanism, each list $l_c$ and $l_l$ contains all the documents, we can omit the last three cases.

In order to use the Dempster's combination rule of (10), we need to assign to each source of evidence a measure of uncertainty. We propose to use the query scope as an automatically assigned measure. The idea is to optimise the measures of uncertainty $m_c(\Theta)$ and $m_l(\Theta)$, so that we obtain the best combined ranking of the two initial sources of evidence.

As defined in Section 2, the query scope $\mathit{scope}_q$ of a query $q$, is a measure of specificity of $q$. For specific queries, or queries on topics not adequately represented in the collection, $\mathit{scope}_q \leadsto 0$, while for generic queries, or queries on topics well represented in the collection $\mathit{scope}_q \leadsto 1$. Therefore, we set the uncertainty of the content-related body of evidence $m_c$ to be $m_c(\Theta) = \mathit{scope}_q$, while the uncertainty of the link-related body of evidence $m_l$ is set to be $m_l(\Theta) = 1 - \mathit{scope}_q$. The explanation is that for specific queries, i.e. when $m_c(\Theta) = \mathit{scope}_q \leadsto 0$, the content analysis is a more trustful source of evidence than the link structure analysis. Hence, its associated uncertainty is very low, compared to the high $m_l(\Theta) = 1 - \mathit{scope}_q$ uncertainty value associated to the body of evidence $m_l$, and vice-versa.

To summarise the whole dynamic process, the probability mass function $m_{c,l}$ assigned by combining the probability mass functions $m_c$ and $m_l$ could be defined as follows:

$$m_{c,l}(d) \propto m_c(d) * m_l(d) + (1 - \mathit{scope}_q) * m_c(d) + \mathit{scope}_q * m_l(d) \tag{12}$$

Note that if we compute the probability distribution for the WordNet concepts during indexing time, then the query scope $\mathit{scope}_q$ can be computed efficiently as described in Section 2.3, and (12) can be used to compute the final score of a document during query time, with an overall marginal overhead.

## 4. Description of experiments

To evaluate the proposed mechanism for the query-biased combination of evidence, we experiment using two different TREC collections of Web documents. The first, namely the WT10g, consists of 1.64 million documents and it was used for the topic relevance task of TREC10, for which 50 topics have been created (Hawking and Craswell, 2001). The second collection is the .GOV, a crawl from the .gov domain, which consists of 1.25 million Web documents. It was used for the topic distillation task of TREC11, for which 50 topics have been created (Craswell and Hawking, 2002). For indexing the collections, a standard stop word list was used (Van Rijsbergen, 1979) and the stemming algorithm of Porter was applied (Porter, 1980). For both tasks, we used the titles of the provided topics as the queries to be submitted to our retrieval engine, as it is recommended by TREC.

Our retrieval engine consists of two separate modules. The first is the content retrieval module, which is an implementation of the Divergence From Randomness (DFR) probabilistic framework by Amati and Van Rijsbergen (2002). The matching function applied is the $I(n_e)B2$, where the weight of a term $t$ is given by the following formula:

$$weight(t) = \frac{\mathrm{Freq}(t|\mathrm{Collection}) + 1}{\mathrm{doc\_freq} \cdot (tfn + 1)} \left( tfn \cdot \log_2 \frac{N + 1}{n_e + 0.5} \right) \tag{13}$$

where $tfn = \mathrm{term\_freq} \cdot \log_2 \left(1 + c \cdot \frac{\mathrm{average\_document\_length}}{document\_length}\right)$

$N$ is the size of the collection

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

Freq($t$|Collection) is the within-collection term-frequency

term_freq is the within-document term-frequency

doc_freq is the document-frequency of the term

The parameter $c$ is set to 7 for the experiments with the WT10g collection and to 1 for the experiments with the .GOV collection. We choose this specific weighting scheme among the over 50 schemes proposed in DFR, since it is robust for both involved TREC tasks. The weight of each term $t$ is based on the expected inverse document frequency statistics. Then, the weight is adjusted, by modelling the sampling after-effect as Bernoulli trials and by taking into account the document's length (aka *normalisation 2* in (Amati and Van Rijsbergen, 2002)). The content-only retrieval is used as a baseline for all the conducted experiments.

For the link analysis module, two different algorithms are used. The first is a new probabilistic link analysis algorithm, the Absorbing

Model (Plachouras et al., 2003b; Amati et al., 2003) (denoted by AM in the tables). The Absorbing Model is based on modelling the Web graph as an extended Markov chain in the following way: for each state, or node, representing a Web document, we add a virtual node, called the clone node in the Absorbing Model, which corresponds to the event that a user gets absorbed by the specific Web document. The clone node is accessible only from its corresponding original node. The Absorbing Model score for each Web document is the probability of accessing its clone node. The second algorithm is PageRank (Brin and Page, 1998) (denoted by PR in the tables). For both Absorbing Model and PageRank, the link analysis scores for each document are computed during indexing, by taking into account only the hyperlinks between Web documents from different domains. We use only the inter-domain hyperlinks, since they are more likely to convey authority information.

For the calculation of the term scope for every term, we calculate the distribution of the WordNet concepts in the collection. We assume that a term matches a concept in WordNet if this exact term appears in the description of the concept in WordNet. This matching takes place before the query terms are stemmed. At this stage, we do not employ any term disambiguation technique for the matching. The only restriction is that we use only the noun concepts of WordNet, following Brezeale(1999), as nouns tend to have fewer meanings associated with them, while the meanings of verbs tend to depend on the meaning of the surrounding nouns. For the lattice approach (Section 2.1), we consider that each term corresponds to the occurrence of its most generic associated concept, as defined in (5). This may result to an overestimation of the generality of a query but, as we will see, it is not the case for the collections being tested. For the second approach (Section 2.2), each term contributes its occurrences to all of its associated concepts, as defined in (6).

Then, the query scope is computed for each query using two possible ways, and following Definition 6. The first possibility, following Assumption 1 in Section 2.3, corresponds to the sum of the scopes of each query term, which is then divided by the number of query terms (denoted by SUMT in the tables). The second possibility, following Assumption 2, corresponds to the product of the scopes of each query term, which is then multiplied by the number of terms in the query (denoted by PRDT in the tables).

Dempster-Shafer combination of evidence is applied either with fixed measures of uncertainty, as in (10), or using the query scope, as in (12). The constant values of uncertainty used for (10) are 0.50, 0.25 and 0.05 for the content module and 0.50, 0.75 and 0.95 for the link analysis module. We use DSCL to denote the experiments conducted with the constant values, and DSLA and DSIA for the experiments conducted with the lattice and the independent concepts approaches respectively.

Before the application of the formulas, the content and link analysis scores were normalised as follows:

$$m_c(d_i) = \frac{m_c(d_i)}{\sum_j m_c(d_j)}, m_l(d_i) = \frac{m_l(d_i)}{\sum_j m_l(d_j)} \tag{14}$$

The normalisation is necessary because while the link analysis scores are in $[0, 1]$, the scores from the content analysis may be significantly higher, depending on the number of terms in the corresponding query.

Because both TREC tasks do not favour the application of link analysis, as it has been shown in TREC11 (Hawking and Craswell, 2001) and TREC12 conferences (Craswell and Hawking, 2002), we restrict the application of the combination of evidence to the set $B$ of the $|B|$ ranked documents, so that the content-only ranking is not affected significantly. The values we use for $|B|$ are 20, 50 and 1000, respectively.

## 5. Analysis of results

Many factors could affect the results of the conducted experiments and, therefore, the evaluation of the proposed dynamic combination of evidence mechanism. The appropriateness of the tasks for link analysis and the link analysis methods themselves, the chosen methodology for the combination of evidence, and the value of the parameter $|B|$, all influence the retrieval effectiveness of each approach. Under these conditions, we will interpret the effect of each of these factors on the results.

We start by looking into the results of the Dempster-Shafer combination of evidence using constant values. It appears that the average precision of the combination of DSCL(0.50,0.50) with SAM for $|B| = 20, 50, 1000$ for the TREC10 task (0.2069, 0.2066, 0.2041 respectively from Tables III, IV and V) is not greatly affected with respect to the average precision 0.2105 of the content-only baseline. On the other hand, for the TREC11 task, the average precision for the same values of $|B|$ (0.1666, 0.1522, 0.0804 respectively from Tables VI, VII, VIII) drops clearly with respect to the average precision 0.1990 of the baseline. A possible explanation for this is the distribution of link analysis scores. For the WT10g collection, these scores were not as discriminatory as for the .GOV collection, due to the small number of links between documents from different domains. Therefore the results for the WT10g collection are not as much affected as for the .GOV collection.

Comparing the results in Tables III, IV, V and Tables VI, VII, VIII, respectively, we can see the effect of using the two different test collections. Both collections were designed to possess the basic properties of the Web (Bailey et al., 2003). However, it has been noted in TREC10 (Hawking and Craswell, 2001) and TREC11 (Craswell and Hawking, 2002) that using link analysis for the TREC10 topic relevance task and the TREC11 topic distillation task, hardly improves the retrieval effectiveness. Our results show that, while the precision for the TREC11 task is clearly affected by the combination of evidence mechanism, they are more conclusive for the TREC10 task.

The choice of a link analysis algorithm also affects the results. For the TREC10 task, if we compare the columns SAM and PR in Tables III, IV and V, it appears that the Absorbing Model outperforms PageRank in every experiment. Moreover, when using the Absorbing Model, the results are in some cases above the content-only baselines (see the average precision of DSIA-SUMT from Tables III, IV and V w.r.t. the average precision of the baseline). For the TREC11 task, when the dynamic combination of evidence is employed, the Absorbing Model outperforms PageRank for $|B| = 50$ (see Table VII). For the rest of the cases, PageRank performs slightly better than the Absorbing Model, when it is combined with the content analysis by using Dempster-Shafer's theory of evidence (see Tables VI and VIII). These results do not allow us to decide which of the two models is more appropriate as a link analysis method. However, for the TREC10 task, the Absorbing Model proves to be more robust, even if there is a low density of hyperlinks in the WT10g collection.

Comparing the baseline experiment with the ones that employ Dempster-Shafer's combination of evidence with constant values of uncertainty shows that this approach to combination of evidence does not improve precision over the content-only baseline consistently. However, the introduction of the query scope improves the performance in most of the cases over the experiments with the constant level of uncertainty. Moreover, in some of the cases, it results into slightly higher precision in comparison to the baseline. For example, for the TREC10 task, the average precision of the combination of DSIA-SUMT and SAM (0.2111, 0.2112 and 0.2108 from Tables III, IV and V) is higher than the average precision 0.2105 of the baseline, as well as the average precision of the Dempster-Shafer combination with the best constant uncertainties 0.75 and 0.25 (0.2071, 0.2076, 0.2087 from Tables III, IV and V). This shows that our assumption to combine evidence on a per-query basis is valid, since not all queries benefit from a uniform approach.

It is interesting to examine the distribution of the query scope values to see which of the four different approaches, namely DSIA-PRDT, DSIA-SUMT, DSLA-PRDT and DSLA-SUMT, is more suitable in discriminating the specific from the more generic queries. In Figures 4 and 5, the distributions of values for the TREC10 and TREC11 queries are shown respectively. From the figures it appears that the majority of the queries are considered to be specific according to their value of query scope. In other words, the queries are not general enough to favour the application of hyperlink analysis. This is consistent with the results of TREC10 and TREC11, where hyperlink analysis did not lead to improvements, due to the quite specific queries. Therefore, using the query scope biases the combination of evidence mechanism by assigning more importance to the content analysis.

From Figures 4 and 5 it appears that the four variations of query scope are distributed differently. For both collections, the two methods based on the lattice approach, namely DSLA-PRDT and DSLA-SUMT are separated from the two other methods. In addition, the values of the two methods, which are based on the independent concepts approach, namely DSIA-SUMT and DSIA-PRDT, are lower and give even less importance to link analysis. This might suggest that the two methods based on the independent concepts approach, namely DSIA-SUMT and DSIA-PRDT, are more suitable for the tested queries. This is confirmed by the results, especially for the TREC11 queries.

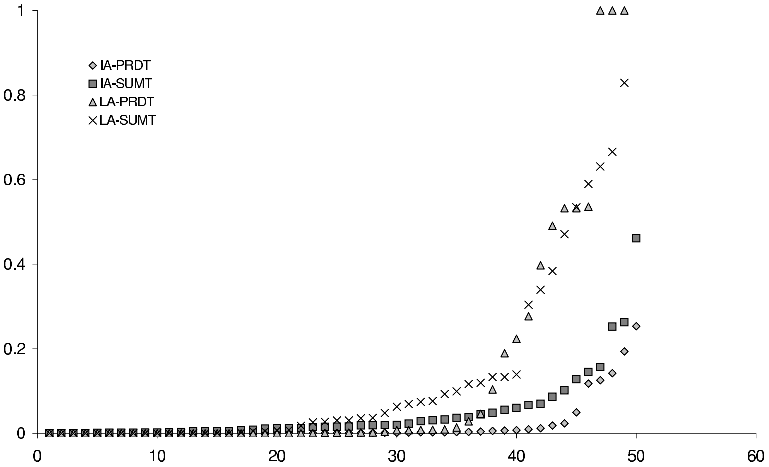

*Figure 4.* The distribution of query scope values by the four different methods for the TREC10 queries.

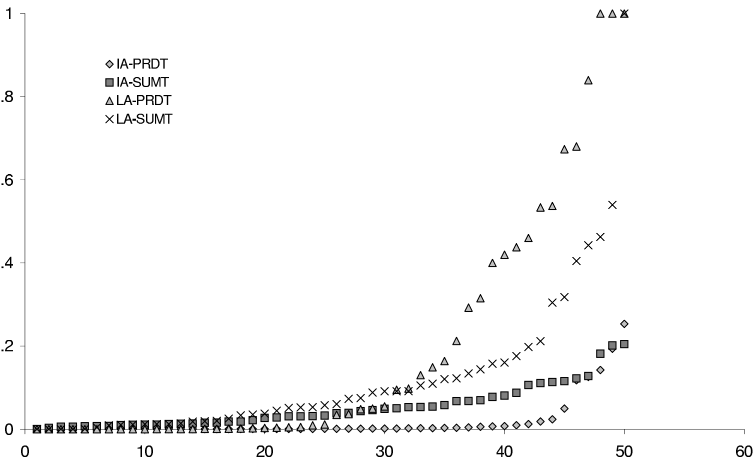

*Figure 5.* The distribution of query scope values by the four different methods for the TREC11 queries.

Table III. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 20$ for WT10g.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.2105 | 0.2105 | 0.4240 | 0.4240 | 0.3720 | 0.3720 | 0.3180 | 0.3180 |
| DSCL(0.50, 0.50) | 0.2069 | 0.2035 | 0.4040 | 0.3840 | 0.3680 | 0.3620 | 0.3170 | 0.3170 |
| DSCL(0.25, 0.75) | 0.2071 | 0.2036 | 0.4040 | 0.3840 | 0.3680 | 0.3620 | 0.3170 | 0.3170 |
| DSCL(0.05, 0.95) | 0.2069 | 0.2063 | 0.4040 | 0.3840 | 0.3680 | 0.3620 | 0.3170 | 0.3170 |
| DSLA-SUMT | 0.2098 | 0.2036 | 0.4120 | 0.3840 | **0.3760** | 0.3620 | 0.3170 | 0.3170 |
| DSLA-PRDT | 0.2102 | 0.2038 | 0.4120 | 0.3840 | **0.3740** | 0.3620 | 0.3170 | 0.3170 |
| DSIA-SUMT | **0.2111** | 0.2037 | 0.4160 | 0.3840 | **0.3780** | 0.3620 | 0.3170 | 0.3170 |
| DSIA-PRDT | **0.2109** | 0.2038 | 0.4160 | 0.3840 | **0.3760** | 0.3620 | 0.3170 | 0.3170 |

## 6. Discussion of the proposed model

The experimental results presented in the previous section show that the effectiveness of our approach is similar to that of the baselines, although a higher precision is sometimes observed, especially for the TREC10 queries. In this section, we discuss how retrieval effectiveness is affected by the specific components we used in our approach. We discuss alternative options for defining the query scope and for combining different sources of evidence.

Table IV. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 50$ for WT10g.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.2105 | 0.2105 | 0.4240 | 0.4240 | 0.3720 | 0.3720 | 0.3180 | 0.3180 |
| DSCL(0.50, 0.50) | 0.2066 | 0.1878 | 0.4040 | 0.3200 | 0.3680 | 0.3240 | 0.3180 | 0.3050 |
| DSCL(0.25, 0.75) | 0.2076 | 0.1885 | 0.4040 | 0.3240 | 0.3680 | 0.3240 | **0.3190** | 0.3050 |
| DSCL(0.05, 0.95) | **0.2113** | 0.1885 | 0.4160 | 0.3200 | **0.3780** | 0.3240 | 0.3180 | 0.3040 |
| DSLA-SUMT | **0.2112** | 0.1906 | 0.4160 | 0.3320 | **0.3740** | 0.3320 | **0.3220** | 0.3050 |
| DSLA-PRDT | **0.2113** | 0.1948 | 0.4160 | 0.3360 | 0.3720 | 0.3380 | **0.3230** | 0.3060 |
| DSIA-SUMT | **0.2112** | 0.1940 | 0.4160 | 0.3320 | **0.3740** | 0.3380 | **0.3190** | 0.3040 |
| DSIA-PRDT | 0.2105 | 0.1936 | 0.4200 | 0.3360 | **0.3740** | 0.3360 | **0.3190** | 0.3050 |

Table V. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 1000$ for WT10g.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.2105 | 0.2105 | 0.4240 | 0.4240 | 0.3720 | 0.3720 | 0.3180 | 0.3180 |
| DSCL(0.50, 0.50) | 0.2041 | 0.0972 | 0.4040 | 0.1240 | **0.3740** | 0.0780 | **0.3190** | 0.0630 |
| DSCL(0.25, 0.75) | 0.2087 | 0.0976 | 0.4000 | 0.1240 | 0.3680 | 0.0780 | **0.3220** | 0.0630 |
| DSCL(0.05, 0.95) | 0.2087 | 0.1163 | 0.4000 | 0.1480 | 0.3680 | 0.1020 | **0.3220** | 0.0880 |
| DSLA-SUMT | **0.2113** | 0.1495 | 0.4160 | 0.2280 | **0.3740** | 0.1820 | **0.3220** | 0.1600 |
| DSLA-PRDT | **0.2109** | 0.1886 | 0.4160 | 0.3440 | **0.3760** | 0.2900 | **0.3200** | 0.2440 |
| DSIA-SUMT | **0.2108** | 0.1712 | **0.4280** | 0.2320 | **0.3740** | 0.2060 | 0.3180 | 0.1920 |
| DSIA-PRDT | 0.2093 | 0.1905 | 0.4120 | 0.3440 | 0.3700 | 0.3000 | 0.3140 | 0.2520 |

The first component of the proposed methodology is the definition of the query scope as a probability distribution on top of WordNet, described in Section 2. Generally, WordNet has been employed in order to disambiguate terms for information retrieval. For example, Voorhees (1994) employed WordNet to disambiguate the query terms, but the effectiveness of this approach was lower than that of stemming. An issue of using external conceptual structures, such as WordNet, is that they may be too specific, or too general for the task, or the collection under consideration. Building hierarchies of concepts on a per-query basis is a different way to overcome these limitations (Sander-

Table VI. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 20$ for .GOV.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.1990 | 0.1990 | 0.3020 | 0.3020 | 0.2408 | 0.2408 | 0.1888 | 0.1888 |
| DSCL(0.50, 0.50) | 0.1666 | 0.1804 | 0.2490 | 0.2612 | 0.2061 | 0.2143 | 0.1888 | 0.1888 |
| DSCL(0.25, 0.75) | 0.1743 | 0.1807 | 0.2612 | 0.2612 | 0.2143 | 0.2204 | 0.1888 | 0.1888 |
| DSCL(0.05, 0.95) | 0.1790 | 0.1811 | 0.2653 | 0.2612 | 0.2184 | 0.2245 | 0.1888 | 0.1888 |
| DSLA-SUMT | 0.1799 | 0.1831 | 0.2694 | 0.2694 | 0.2184 | 0.2286 | 0.1888 | 0.1888 |
| DSLA-PRDT | 0.1726 | 0.1843 | 0.2735 | 0.2735 | 0.2224 | 0.2286 | 0.1888 | 0.1888 |
| DSIA-SUMT | 0.1814 | 0.1830 | 0.2694 | 0.2694 | 0.2143 | 0.2286 | 0.1888 | 0.1888 |
| DSIA-PRDT | 0.1837 | 0.1843 | 0.2898 | 0.2653 | 0.2265 | 0.2286 | 0.1888 | 0.1888 |

Table VII. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 50$ for .GOV.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.1990 | 0.1990 | 0.3020 | 0.3020 | 0.2408 | 0.2408 | 0.1888 | 0.1888 |
| DSCL(0.50, 0.50) | 0.1522 | 0.1487 | 0.2163 | 0.2082 | 0.1878 | 0.1857 | 0.1643 | 0.1755 |
| DSCL(0.25, 0.75) | 0.1619 | 0.1496 | 0.2367 | 0.2082 | 0.1918 | 0.1878 | 0.1684 | 0.1745 |
| DSCL(0.05, 0.95) | 0.1779 | 0.1541 | 0.2694 | 0.2122 | 0.2184 | 0.1878 | **0.1898** | 0.1816 |
| DSLA-SUMT | 0.1743 | 0.1569 | 0.2776 | 0.2245 | 0.2184 | 0.2061 | 0.1765 | 0.1888 |
| DSLA-PRDT | 0.1715 | 0.1590 | 0.2776 | 0.2367 | 0.2102 | 0.2204 | 0.1806 | 0.1888 |
| DSIA-SUMT | 0.1780 | 0.1554 | 0.2735 | 0.2082 | 0.2143 | 0.1898 | 0.1796 | 0.1857 |
| DSIA-PRDT | 0.1779 | 0.1688 | 0.2980 | 0.2408 | 0.2245 | 0.2143 | 0.1867 | **0.1929** |

son and Croft, 1999). Additionally, in the context of Web information retrieval, there are other sources of evidence that can be used to model the specificity of a query, with respect to the statistical characteristics of the set of retrieved documents. It is the latter approach that we have chosen to investigate in TREC12 (Plachouras et al., 2003a), in order to select the most appropriate retrieval approach for each query. This approach is more effective and has lead to important improvements in precision.

The second component we look into is the combination of evidence mechanism. Dempster-Shafer theory of evidence is not effective in sig-

Table VIII. Results for the baseline and Dempster-Shafer combination of evidence with with $|B| = 1000$ for .GOV.

| Experiment | Av. Precision | | Prec. at 5 | | Prec. at 10 | | Prec. at 20 | |
|---|---|---|---|---|---|---|---|---|
| | AM | PR | AM | PR | AM | PR | AM | PR |
| Baseline | 0.1990 | 0.1990 | 0.3020 | 0.3020 | 0.2408 | 0.2408 | 0.1888 | 0.1888 |
| DSCL(0.50, 0.50) | 0.0804 | 0.0744 | 0.0857 | 0.1306 | 0.1000 | 0.1612 | 0.1122 | 0.1286 |
| DSCL(0.25, 0.75) | 0.1104 | 0.0810 | 0.1469 | 0.1510 | 0.1510 | 0.1653 | 0.1449 | 0.1316 |
| DSCL(0.05, 0.95) | 0.1622 | 0.1426 | 0.2327 | 0.2245 | 0.2041 | 0.1939 | 0.1816 | 0.1551 |
| DSLA-SUMT | 0.1531 | 0.1449 | 0.2204 | 0.2408 | 0.1837 | 0.2082 | 0.1612 | 0.1684 |
| DSLA-PRDT | 0.1533 | 0.1584 | 0.2204 | 0.2408 | 0.1796 | 0.2163 | 0.1653 | 0.1796 |
| DSIA-SUMT | 0.1602 | 0.1475 | 0.2204 | 0.2367 | 0.2000 | 0.1878 | 0.1714 | 0.1561 |
| DSIA-PRDT | 0.1869 | 0.1897 | 0.2857 | 0.2980 | 0.2306 | **0.2469** | 0.1837 | **0.1898** |

nificantly improving precision in our experiments, due to either the quality of the sources of evidence, or the appropriateness of the method itself. With respect to the first point, while content-only retrieval is an effective approach for both the tasks we experimented with, hyperlink analysis has not proved to be equally useful. In addition, the normalisation of the scores described in (14) could bias the combination of evidence, since the distribution of hyperlink analysis scores is significantly different from that of the content-only retrieval scores. As for the combination of evidence, there have been several different proposed approaches, based on bayesian networks (Croft and Turtle, 1989; Ribeiro-Neto and Muntz, 1996), or on the propagation of scores across the connections between documents (Dominich, 2002). All these approaches can be used for the combination of evidence, instead of Dempster-Shafer theory.

The last point we note is the effectiveness of hyperlink analysis for the tasks under consideration. The topic relevance task from TREC10, although employing a Web test collection, is an ad-hoc task, where content-only retrieval and query expansion are very effective. On the other hand, the topic distillation task from TREC11 is about finding useful entry points for the query topics and it is focused on Web searching. However, the relevance assessments are more similar to those for a ad-hoc task (Craswell and Hawking, 2002). As a result, content-only retrieval is still the most effective approach for this task, a fact that is also confirmed by our experimental results. As a consequence, combining content and hyperlink analysis is not expected to lead to significant improvements for the specified tasks.

Overall, we have proposed a model for combining evidence on a per-query basis. We have experimented with WordNet's hierarchy of concepts and Dempster-Shafer's combination of evidence, which slightly increased the effectiveness. However, better improvements may be obtained, if we use alternative configurations, based on statistical evidence for the queries, or different mechanisms for the combination of evidence, as shown in our TREC12 experiments (Plachouras et al., 2003a).

## 7. Conclusions

In this paper, we present a query-biased combination of evidence mechanism for the Web. We propose two methods for estimating probabilistically the query scope, that is a measure of the query's specificity. The query scope is related to the query's term frequencies in the document collection and the semantic interpretation of the query according to WordNet. We merge the ranked lists obtained from content and link analyses using Dempster-Shafer's theory of evidence, by assigning to each source of evidence a measure of uncertainty based on the query scope.

We evaluate the effectiveness of the proposed methodologies on two TREC Web tasks, namely the TREC10 topic relevance task (Hawking and Craswell, 2001) and the TREC11 topic distillation task (Craswell and Hawking, 2002). Although both tasks do not favour the application of link analysis, as it has been shown in TREC10 and TREC11, we get a slight improvement for the TREC10 task, when the query-biased combination of evidence mechanism is employed, and for the TREC11 task, our results are close to the levels of precision of the content-only baseline.

The introduction of the query scope improves the retrieval effectiveness when it is compared to Dempster-Shafer combination of evidence with constant values of uncertainty. This shows that the query scope is useful in the sense that it successfully biases the results towards the most appropriate source of evidence for the specific test collections we used. Different approaches to the issue of combination of evidence may prove more effective when query-biased measures, such as the query scope are introduced (Plachouras et al., 2003a).

An issue that needs further investigation is the matching of terms to concepts. Since, no term disambiguation is performed, we expect that at this stage we introduce a level of noise to the computation of the scope of each term in the collection. Therefore, simple mechanisms for term disambiguation might result into improved retrieval effectiveness.

## 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). We would like to thank Gianni Amati for his useful comments on the described work.

## References

Aizawa N (2000) The feature quantity: an information theoretic perspective of Tfidf-like measures. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 104–111.

Amati G and I Ounis (2000) Conceptual Graphs and First Order Logic. The Computer Journal, 43(1):1–12.

Amati G, I Ounis and V Plachouras (2003) The Absorbing Model for the Web. Submitted to Information Processing Letters.

Amati G and CJ van Rijsbergen (2002) Probabilistic models of Information Retrieval based on measuring divergence from randomness. ACM Transactions on Information Systems, 20(4):357–389.

Amento B, L Terveen and W Hill (2000) Does "authority" mean quality? predicting expert quality ratings of Web documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 296–303.

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

Barnett JA (1981) Computational Methods for a Mathematical Theory of Evidence. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81), Vancouver, BC, Canada, pp. 868–875.

Bharat K and MR Henzinger (1998) 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, Melbourne, Australia, pp. 104–111.

Brezeale D (1999) The Organization Of Internet Web Pages Using Wordnet and Self-Organizing Maps. Master Thesis, University of Texas at Arlington, Texas, USA.

Brin S and L Page (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1–7):107–117.

Calado P, B Ribeiro-Neto, N Ziviani, E Moura and I Silva (2003) Local Versus Global Link Information in the Web. ACM Transactions on Information Systems, 21(1):42–63.

Chakrabarti S, B Dom, D Gibson, J Kleinberg, P Raghavan and S Rajagopalan (1998) Automatic Resource List Compilation by Analyzing Hyperlink Structure and Associated Text. Computer Networks and ISDN Systems, 30(1–7):65–74.

Cohn D and H Chang (2000) Learning to Probabilistically Identify Authoritative Documents. In: Proceedings of the 17th International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, pp. 167–174.

Craswell N and D Hawking (2002) Overview of the TREC-2002 Web Track. In: Proceedings of the 11th Text REtrieval Conference (TREC 2002), Gaithersburg, MD, USA, pp. 86–93.

Craswell N, D Hawking, R Wilkinson and M Wu (2003) Overview of the TREC-2003 Web Track. In: Proceedings of the 12th Text REtrieval Conference (TREC 2003), Gaithersburg, MD, USA.

Croft W and H Turtle (1989) A Retrieval Model for Incorporating Hypertext Links. In: Proceedings of the 2nd annual ACM conference on Hypertext (Hypertext'89), Pittsburgh, PA, USA, pp. 213–224.

Cronen-Townsend S, Y Zhou and WB Croft (2002) Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 299–306.

Diligenti M, M Gori and M Maggini (2002) Web page scoring systems for horizontal and vertical search. In: Proceedings of the 11th International World Wide Web Conference (WWW 2002), Honolulu, HI, USA, pp. 508–516.

Dominich S (2002) Connectionist interaction information retrieval. Information Processing & Management, 39(2):167–193.

Fellbaum C, editor (1998) WordNet An Electronic Lexical Database. MIT Press.

Haveliwala TH (2002) Topic-Sensitive PageRank. In: Proceedings of the 11th International World Wide Web Conference (WWW 2002), Honolulu, HI, USA, pp. 517–526.

Hawking D and N Craswell (2001) Overview of the TREC-2001 Web Track. In: Proceedings of the 10th Text REtrieval Conference (TREC 2001), Gaithersburg, MD, USA, pp. 61–67.

Jose JM (1998) An Integrated Approach for Multimedia Information Retrieval. PhD Thesis, The Robert Gordon University, Aberdeen, Scotland.

Jose JM and DJ Harper (1997) A Retrieval Mechanism for Semi-Structured Photographic Collections. In: Proceedings of the Database and Expert Systems Applications, 8th International Conference (DEXA'97), Toulouse, France, pp. 276–292.

Kao H-Y, M-S Chen, S-H Lin and J-M Ho (2002) Entropy-based link analysis for mining web informative structures. In: Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, pp. 574–581.

Kim SJ and SH Lee (2002) Improved Computation of the PageRank Algorithm. In: Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, Glasgow, UK, pp. 73–85.

Kleinberg JM (1999) Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, 46(5):604–632.

Kraaij W, T Westerveld and D Hiemstra (2002) The Importance of Prior Probabilities for Entry Page Search. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 27–34.

Lempel R and S Moran (2000) The stochastic approach for link-structure analysis (SALSA) and the TKC effect. Computer Networks, 33(1-6):387–401.

Miller G (1995) WordNet: A Lexical Database for English. Communications of the ACM, 38(11):39–41.

Zheng AX, Ng AY and MI Jordan (2001) Stable Algorithms for Link Analysis. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, USA, pp. 258–266.

Ounis I (1998) Un modèle d'indexation relationnel pour les graphes conceptuels fondé sur une interprétation logique. PhD Thesis, Université Joseph Fourier, Grenoble, France.

Plachouras V, F Cacheda, I Ounis and CJ van Rijsbergen (2003a) University of Glasgow at the Web track: Dynamic Application of Hyperlink analysis using the Query Scope. In: Proceedings of the 12th Text Retrieval Conference (TREC 2003), Gaithersburg, MD, USA.

Plachouras V, I Ounis and G Amati (2003b) A Utility-oriented Hyperlink Analysis Model for the Web. In: Proceedings of the 1st Latin Web Conference, Santiago, Chile, pp. 123–131.

Porter MF (1980) An algorithm for suffix stripping. Program, 14(3):130–137.

Ribeiro-Neto B and R Muntz (1996) A Belief Network Model for IR. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, pp. 253–260.

Richardson M and P Domingos (2002) The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In: Advances in Neural Information Processing Systems 14 (Neural Information Processing Systems: Natural and Synthetic, NIPS 2001), Vancouver, BC, Canada, pp. 1441–1448.

Rölleke T (2003) A frequency-based and a poisson-based definition of the probability of being informative. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, pp. 227–234.

Ruthven I, M Lalmas and CJ van Rijsbergen (2002) Combining and Selecting Characteristics of Information Use. Journal of the American Society for Information Science and Technology, 53(5):378–396.

Sanderson M and B Croft (1999) Deriving Concept Hierarchies from Text. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, pp. 206–213.

Shafer G (1976) A Mathematical Theory of Evidence. Princeton University Press.

Silva I, B Ribeiro-Neto, P Calado, E Moura and N Ziviani (2000) Link-based and content-based evidential information in a belief network model. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, pp. 96–103.

Spärck Jones K (1972) A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28(1):11–20.

Van Rijsbergen CJ (1979) Information Retrieval, 2nd edition. Buttersworth, London.

Voorhees E (1994) Using WordNet to disambiguate Word Senses for Text Retrieval. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA, USA, pp. 171–180.

Wong SKM and Y Yao (1992) An Information-Theoretic Measure of Term Specificity. Journal of the American Society for Information Science, 43(1):54–61.