QUERY-BIASED COMBINATION OF EVIDENCE ON THE WEB

Vassilis Plachouras
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
vassilis@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
ounis@dcs.gla.ac.uk

Abstract

Recent research has shown that the hyperlink structure of the World Wide Web can be a rich source of information about the content of the environment, provided that there are effective means of understanding it. However, experience also suggests that while hyperlink structure analysis improves the overall precision of the top relevant documents for broad and generic queries, it usually fails to address the expectations of the users when the information need is precise and specific. Indeed, the authoritative measure of an information source, underpinning the hyperlink structure analysis, cannot be used as a substitute for relevance in the case of specific queries.

We propose a query-biased combination of evidence mechanism that optimally merges results obtained by content and link analyses, based on the query characteristics. The combination mechanism is modelled using Dempster-Shafer's theory of evidence, and the query characteristics are defined using a probabilistic propagation mechanism on top of the hierarchical structure of concepts provided by WordNet. Our evaluation experiments indicate that the proposed query-biased merging technique yields some interesting conclusions.

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 [21]. Additionally, on the World Wide Web there is another source of evidence that an IR system can explore, namely the analysis of the hyperlink structure of the documents. It is claimed that links within documents can be used, in combination with the content analysis, to detect high quality documents, or what is commonly called the *authoritative* documents corresponding to a given query.

Some recent approaches have shown how IR on the Web can benefit from the combination of link analysis and content analysis. Kleinberg in [15] proposes an algorithm which, given a query, finds authority and hub documents in the Web. The input to the algorithm is a set of documents retrieved using a conventional content-based search engine, where traditional IR techniques are employed. Brin and Page in [7] present the architecture of Google, a search engine in which for every web page indexed, a score based on the page's popularity is precomputed. At query time, a set of documents relevant to the query is formed and sorted according to their score obtained from the content and link analyses. Extensions and refinements of these two algorithms are discussed in [5, 8] and [14, 11, 18] respectively. In both cases, link analysis complements content analysis by improving precision in the top retrieved documents.

However, queries exhibiting different characteristics require alternate combination of content and link analyses. 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 specialized audience, and therefore, they are not highly connected. On the contrary, it is commonly accepted that increased precision among the top ranked documents is observed for queries on a broad or popular topic, when link analysis is performed [15].

As a consequence, there is a need for optimal combination of results obtained from content and link analyses with respect to the query. Intuitively, the contribution of link analysis should be higher for generic queries and lower for specific ones. Deciding dynamically about the optimal combination demands a method for estimating the *query scope*, that is, a measure of how specific or generic a query is. This measure should depend on the term frequencies in the collection as well as the semantic content of the query. An hierarchical structure of concepts can be used for estimating the query scope. Such an hierarchical structure is provided by WordNet [16, 9], a lexical reference system where 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 [19]. 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.

The rest of the paper is organized 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 performed are described in detail and the results are presented in Section 4. In the closing Section 5, we discuss the results and the possible further refinements of this work.

2. QUERY SCOPE

We consider the query as a set of terms. We define the query scope as a probabilistic measure of how specific or generic a query is. The query scope is a function of the *term scope* of its composing terms, that is, a measure of how specific or generic its composing terms are. We compute the term scope according to how specific or generic its associated concepts are. We propose to estimate the term scope by defining a probability measure for concepts on top of the hierarchical structure of concepts provided by WordNet [16, 9]. For example, part of this hierarchical structure is shown in Figure 1, where each set of terms represents a concept.

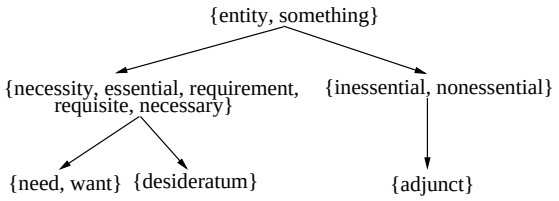

Figure - 1. Part of WordNet's hierarchical structure of concepts.

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 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 $< T_C, \leq >$. 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 $< T_C, \leq >$ 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 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 $M =< D, \models >$ (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, that is, they 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 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 [2, 17].

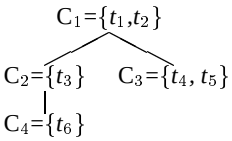

Figure - 2. Hierarchy of concepts for Example 1

Table - 1. 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

Definition 1 *The weight $m(C(x))$ of a concept C is the cardinality of the set $\{d \in D : d \models m(C(x))$ for some individual or witness $t\}$*

It is easy to see that $m(C(x))$ is the number of documents in which the concept C occurs.

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

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

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 Kolmogrov 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 frequencies of these terms are shown in Table 1. According to Definition 2, for the calculation of

[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.

the probability of concept $C_2 = \{t_3\}$ we have:

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

Substituting the values of function m, we have that $Prob(\{t_3\}) = 0.077$.

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:

$$T = \sum_{t_k} tf_k \tag{1}$$

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:*

$$a_{k,j} = \frac{(D_k + 1) - d_{k,j}}{n_k * (D_k + 1) - \sum_{j=1}^{n_k} d_{k,j}} \tag{2}$$

Then, we define the probability of a concept C.

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 weights of the term frequency for each term* t_k *is the contribution* $a_{k,j}$ *of term* t_k *to concept* $C = C_{k,j}$*:*

$$prob(C) = \sum_{C = C_{k,j} \in \mathbb{C}_k} a_{k,j} * \frac{tf_k}{T} \tag{3}$$

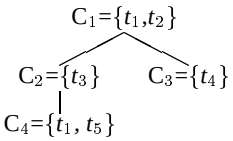

Figure - 3. Hierarchy of concepts for Example 2

Table - 2. 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

The probability distribution of the concepts in $\mathbb{C}$ has to satisfy 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 $prob(C)$, terms are considered only once and are assigned to a specific concept. 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.

From Equation 1 and the term frequencies of Table 1 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 Equation 2, 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 Equation 3, we calculate the probability of concept C_2 to be:

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

[2]We could note in Figure 3 that two concepts, 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.

2.3 Estimation of the Query Scope

After having defined the probability $prob(C)$ of each concept of WordNet, we calculate $scope_{t_k}$ 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:*

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

For example, if we replace $\otimes$ with function $\max$ using either methods for defining the probability $prob(C)$ of concept C, we have:

$$scope_{t_k} = \max_{C \in \mathbb{C}_k} prob(C).$$

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:

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

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

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

$$scope_q = \oplus_{t \in Q} scope_t$$

We look into two approaches for the combination, the sum and the product of probabilities for single terms.

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 normalize 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 normalize 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* [19]. 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^Θ, the power set of Θ, so that:

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

Since we deal with the power set of Θ, 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 Θ. 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^Θ 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 Θ. The probability mass function m defines a new body of evidence in the same frame of discernment Θ as follows:

$$\begin{aligned} m(A) &= m_1 \oplus m_2(A) \\ &= \frac{\sum_{B \cap C = A} m_1(B) * m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) * m_2(C)} A, B, C \subseteq \Theta \end{aligned} \tag{4}$$

The rule of combination of evidence returns a measure of agreement between two bodies of evidence. The division normalizes 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 Θ 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 Θ.

The above combination of evidence is generally computationally expensive. Following [4, 13, 12], we reduce the exponential requirement introduced by the use of the power set 2^{Θ}, to a particular case where we have positive evidence for singleton hypotheses only. Hence, we assume that the focal elements of our two initial mass functions m_1 and m_2 are the singleton hypotheses and the frame Θ, that is, we have positive belief for $\{d_1\}$, $\{d_2\}$, $\ldots$,$\{d_n\}$ and Θ 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 Equation 4:

$$\begin{aligned} &m(\{d_i\}) = \\ &\frac{m_1(\{d_i\}) * m_2(\{d_i\}) + m_1(\Theta) * m_2(\{d_i\}) + m_1(\{d_i\}) * m_2(\Theta)}{1 - \sum_{\{d_k\} \cap \{d_j\} = \emptyset} m_1(\{d_k\}) * m_2(\{d_j\})} \end{aligned} \tag{5}$$

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

$$\begin{aligned} m(\{d_i\}) \propto\ & m_1(\{d_i\}) * m_2(\{d_i\}) + m_1(\Theta) * m_2(\{d_i\}) + \\ & m_1(\{d_i\}) * m_2(\Theta) \end{aligned} \tag{6}$$

or more simply:

$$m(d) \propto m_1(d) * m_2(d) + m_1(\Theta) * m_2(d) + m_1(d) * m_2(\Theta) \qquad (7)$$

Equation 7 is used to compute combined degrees of belief and is computationally much less expensive than the Dempster's combination rule given in Equation 4.

In our special case, the Dempster's simplified combination rule of Equation 7 for aggregating the ranked lists l_c and l_l obtained from content and link analyses respectively yields the following formula:

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

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 Equation 7, we need to assign to each source of evidence a measure of uncertainty. We propose to assign this measure automatically based on the scope of the query. 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 scope of a query q, $scope_q$ is a measure of specificity or generality of the query q. For specific queries, or queries on topics not adequately represented in the collection, $scope_q \rightsquigarrow 0$, while for generic queries, or queries on topics well represented in the collection $scope_q \rightsquigarrow 1$. Therefore, we propose that the uncertainty assigned to the content-related body of evidence m_c be $m_c(\Theta) = scope_q$, while the uncertainty assigned to the link-related body of evidence m_l be $m_l(\Theta) = 1 - scope_q$. The explanation is that for specific queries, i.e. when $m_c(\Theta) = scope_q \rightsquigarrow 0$, 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 $m_l(\Theta) = 1 - scope_q$ uncertainty value associated to the body of evidence m_l, and vice-versa.

To summarize 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:

$$m_{c,l}(d) \propto m_c(d) * m_l(d) + (1 - scope_q) * m_c(d) + scope_q * m_l(d) \quad (8)$$

4. EXPERIMENTS AND RESULTS

To evaluate our combination of evidence proposed mechanism, we conduct experiments involving content and link analyses combined as described in Section 3 (Equation 8). For indexing our testing collection, the WT10g collection used for TREC10 [3], a classical stop word list is used and no stemming is applied [21]. The queries and the relevance assessments used are the ones supplied for TREC10. More specifically, we use only the provided query titles, since the number of terms in the query titles is closer to the average number of terms in queries submitted by users in the Web, as reported in [20]. For the mapping of terms to concepts, we use the hierarchical structure of concepts provided by WordNet. We use only the hierarchy of concepts associated to nouns following [6]. We evaluate results in terms of the average precision and precision at recall of $1, 5$ and 10 retrieved documents.

Our system consists of a module for content analysis and a module for link analysis. For the content-related module we use the probabilistic framework described in [1]. More specifically, we employ the weighting scheme, which returned the best results for the content analysis on the WT10g collection. Our link analysis module is based on a novel probabilistic approach[3] where each document is assigned a precomputed score corresponding to its popularity, similarly to PageRank [7]. For creating the combined result list, the set of the 1000 top ranked documents is formed, taking into account only the results from the content analysis. Then the relevance of each document in that set is re-computed according to the combination mechanism we defined in Section 3 (Equation 8).

Based on these preliminary results, we consider the content analysis module alone as our baseline system. We just combine the results obtained from content and link analyses using constant values of uncertainty in Dempster's combination rule (Equation 8). The levels of uncertainty we assigned to the content analysis are 0.5, 0.25, 0.05, and the corresponding levels of uncertainty we assigned to link analysis are 0.5, 0.75, 0.95 respectively (Table 3). The introduction of the combination of evidence mechanism does not improve the performance, compared to the baseline system. However, as the level of uncertainty decreases for the content analysis and increases respectively for the link analysis, the average precision, as well as the precision at recall of $1, 5$ and 10 documents increases.

[3]This probabilistic approach to link analysis is currently under development in our group.

Table - 3. Results for the baseline and combination with constant measure of uncertainty.

Experiment	Av. Precision	Prec. at 1	Prec. at 5	Prec. at 10
Baseline	0.1907	0.4800	0.3760	0.3240
DSCL(0.50, 0.50)	0.0847	0.4000	0.2520	0.2200
DSCL(0.25, 0.75)	0.0980	0.4200	0.2480	0.2260
DSCL(0.05, 0.95)	0.1134	0.4200	0.2640	0.2260

Table - 4. Results for baseline and combination with query-biased measure of uncertainty.

Experiment	Av. Precision	Prec. at 1	Prec. at 5	Prec. at 10
Baseline	0.1907	0.4800	0.3760	0.3240
DSLA-SUMT	0.1100	0.4200	0.2640	0.2280
DSLA-PRDT	0.1156	0.4200	0.2680	0.2300
DSIA-SUMT	0.1149	0.4200	0.2680	0.2300
DSIA-PRDT	0.1161	0.4200	0.2600	0.2240

Following, we replace the constant value of uncertainty assigned to each source of evidence with a measure of uncertainty based on the query scope, as defined in Section 2. LA denotes the approach based on the interpretation of the structure of concepts as a lattice, while the approach based on the interpretation of the structure of concepts as a set of independent concepts is denoted by IA. Moreover, for combining the probabilities of concepts into a value corresponding to the term scope, we used the maximum function for LA, and a weighted sum of the probabilities of concepts for IA. For each of the two approaches, we tested both the sum (SUMT) and the product (PRDT) of the term scopes of single terms to calculate the query scope (Table 4). The results obtained are still lower than those obtained by using the baseline system. However, the introduction of the query-biased measure of uncertainty results in a small improvement over the combination of evidence based on constant-levels of uncertainty.

Following, we investigate the causes of the decrease in precision resulting from the introduction of the combination of evidence mechanism. Starting form Equation 8 in Section 3, we test separately two cases. In the first case, denoted by PD, Equation 8 is replaced by the product of the probability mass functions of the content-related and the link-related bodies of evidence:

$$m_{c,l} = m_c(d) * m_l(d)$$

In the second case, denoted by WS, Equation 8 is replaced by the weighted sum of the probability mass functions of the content-related and the link-related bodies of evidence, where the weights in the summation are the uncer-

Table - 5. Results for baseline and parts of Dempster-Shafer with query scope.

Experiment	Av. Precision	Prec. at 1	Prec. at 5	Prec. at 10
Baseline	0.1907	0.4800	0.3760	0.3240
PD	0.0866	0.4000	0.2520	0.2200
WSCL(0.05, 0.95)	0.1827	0.5200	0.3600	0.3160
WSLA-SUMT	0.1763	0.5200	0.3720	0.3120
WSLA-PRDT	0.1828	0.5000	0.3600	0.3120
WSIA-SUMT	0.1864	0.5600	0.3600	0.3160
WSIA-PRDT	0.1889	0.5400	0.3800	0.3220

tainty of the link-related body of evidence and the uncertainty of the content-related body of evidence respectively:

$$m_{c,l} = m_c(d) * m_l(\Theta) + m_c(\Theta) * m_l(d)$$

For the latter case, we test a constant measure of uncertainty and a query-biased measure of uncertainty. We employ both approaches defined in Section 2 for calculating the query scope (LA, IA), using both the product and the sum of term scopes (SUMT, PRDT). The results are shown in Table 5.

If we compare the performance of the system using the product of scores returned from content and link analyses (PD) to our baseline system, we obtain a similar decrease in performance as in the cases where the Dempster's combination rule (Equation 8) was used. This decrease of performance can be attributed to the fact that a relevant document might not be highly connected in the Web graph, resulting into a low score from the link analysis. Therefore, the product of scores in the Dempster's combination rule (Equation 8) introduces a level of noise when relevant documents are not also popular. However, by using only the weighted sum of scores from content and link analyses, we obtain a similar level of precision to that of the baseline system. Moreover, in all cases, an improvement in the precision of the top retrieved document is achieved, which varies from 4% to 8%, and for the case of WSIA-PRDT we get approximately the same level of precision at 5 documents.

5. CONCLUSIONS

In this paper we present a query-biased combination of evidence mechanism for the Web. We propose two methods for estimating probabilistically the scope of a query, based on 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.

The experiments conducted show that our proposed methods do not work well when Dempster's combination rule is used for the combination of evidence. The results are more promising when the query scope is used as a weighting factor in a linear combination of the sources of evidence, showing that potentially, we can increase the precision among the top ranked documents. Also, TREC10 experiments reported in [22, 10] show that a combination of content and link analyses does not outperform systems using only content analysis and this is attributed to the structure of the WT10g collection. We also believe that the link structure of the WT10g collection does not adequately represent the actual link structure of the Web, and therefore this collection does not favour the application of link analysis.

In future work we will test the proposed methods using other document collections, which represent better the Web in its current state. We will seek ways to incorporate a hierarchical structure based only on the document collection. We also intent to examine the results in a per query basis in order to specify the characteristics of queries that benefit from the query-biased combination of evidence mechanism.

ACKNOWLEDGMENTS

We would like to thank Gianni Amati (Fondazione Ugo Bordoni, Italy) for providing the content-related module of the described experiments. This work is funded by a UK Engineering and Physical Sciences Research Council (EPSRC) grant, number GR/R90543/01.

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 500-250: The Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[2] G. Amati and I. Ounis. Conceptual Graphs and First Order Logic. *The Computer Journal*, 38(1), 2000.

[3] P. Bailey, N. Craswell, and D. Hawking. Engineering a Multi-Purpose Test Collection for Web Retrieval Experiments. Accepted by the Information Processing and Management journal.

[4] J.A. Barnett. Computational methods for a mathematical theory of evidence. In *the 7th International Joint Conference on Artificial Intelligence*, pages 868–875, 1981.

[5] K. Bharat and M.R. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Research and Development in Information Retrieval*, pages 104–111, 1998.

[6] D. Brezeale. The Organization Of Internet Web Pages Using Wordnet, Master Thesis, University of Texas at Arlington, 1999.

[7] 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.

[8] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, 1998.

[9] C. Fellbaum(ed.). *WordNet An Electronic Lexical Database*. MIT Press, 1998.

[10] C. Gurrin and A.F. Smeaton. Dublin City University experiments in connectivity analysis for TREC-9. In *E.M. Voorhes and D.K. Harman (Eds.), The Nineth Text Retrieval Conference (TREC-9), Department of Commerce, National Institute of Standards and Technology*, 2001.

[11] T.H. Haveliwala. Topic-Sensitive PageRank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

[12] J.M. Jose. An Integrated Approach for Multimedia Information Retrieval, PhD Thesis, The Robert Gordon University, Scotland Aberdeen, 1998.

[13] J.M. Jose and D.J. Harper. A Retrieval Mechanism for Semi-Structured Photographic Collections. In *Proceedings of DEXA'97, LNCS 1308*, pages 276–292, 1997.

[14] S.J. Kim and S.H. Lee. Improved computation of the pagerank algorithm. In F. Crestani, M. Girolami, and C.J. Van Rijsbergen, editors, *Proceedings of the 24th ECIR*, pages 73–85, 2002.

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

[16] G. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38:39–41, 1995.

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

[18] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*, 2002.

[19] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[20] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a Very Large AltaVista Query Log, SRC Technical note 1998-014, 1998.

[21] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworth-Heinemann, 1979.

[22] K. Yang. Combining text- and link-based retrieval methods for Web IR. In *E.M. Voorhes and D.K. Harman (Eds.), The Nineth Text Retrieval Conference (TREC-9), Department of Commerce, National Institute of Standards and Technology*, 2001.