Combining Evidence for Relevance Criteria: a Framework and Experiments in Web Retrieval

Theodora Tsikrika and Mounia Lalmas

Department of Computer Science, Queen Mary, University of London, UK
{theodora, mounia}@dcs.qmul.ac.uk, http://qmir.dcs.qmul.ac.uk

Abstract. We present a framework that assesses relevance with respect to several relevance criteria, by combining the query-dependent and query-independent evidence indicating these criteria. This combination of evidence is modelled in a uniform way, irrespective of whether the evidence is associated with a single document or related documents. The framework is formally expressed within Dempster-Shafer theory. It is evaluated for web retrieval in the context of TREC's Topic Distillation task. Our results indicate that aggregating content-based evidence from the linked pages of a page is beneficial, and that the additional incorporation of their homepage evidence further improves the effectiveness.

Key words: Dempster-Shafer theory, topic distillation, best entry point

1 Motivation, Background, and Aim

In ad hoc Information Retrieval (IR), multiple *criteria* are applied when assessing the relevance of documents. The relevance criterion at the heart of IR, and the one usually employed by IR systems, is the *topical* relevance (or *topicality*) of documents [1]. From a user's perspective, though, empirical studies have reached a consensus that users are influenced by beyond topical factors when assessing retrieved documents [1]. Therefore, IR systems need to consider beyond topical relevance criteria. For instance, on the Web, due to its size and unregulated nature, users desire authoritative information, without explicitly stating so.

An IR system assesses relevance by using *evidence of relevance* in its retrieval function. In essence, each source of evidence indicates relevance with respect to a specific criterion. For instance, content-based evidence is used for capturing a document's *topicality*. In web environments, link-based query-independent evidence, such as a page's PageRank [2], indicate a page's *authority*. Algorithms such as HITS [11], on the other hand, express a query-dependent view of a page's authority, i.e. its *topical authority*. In addition, URL-based query-independent evidence (e.g. URL length [14] or URL types [12]) is used for assessing a page's "*homepageness*" [5] (i.e. how likely it is for a page to be a site's homepage).

To assess relevance that reflects various criteria, IR systems combine evidence indicating the criteria of interest. The predominant *combination of evidence* approaches that incorporate, in a principled manner, evidence indicating beyond topical criteria are probabilistic frameworks. These estimate the belief in

relevance given query-dependent and query-independent features. For instance, in language modelling frameworks (e.g. [12]), prior probabilities of relevance, estimated using query-independent features, are embedded in the framework, and combined with the (content-based) language modelling probability. Other frameworks (e.g. [7]) transform each feature's value into a feature-based relevance score, and subsequently linearly combine all available relevance scores.

Our aim is similar: to estimate, in a principled manner, the belief in relevance, when various criteria are of interest, by combining the (query-independent and query-dependent) features indicating these (or any combination of these) criteria. However, unlike others, our aim also is to estimate the belief in each of the relevance criteria of interest by decomposing relevance into the criteria involved. For instance, for web retrieval, by decomposing relevance into topicality, authority and homepageness, the *combination of evidence for relevance criteria* allows us to estimate the belief, for each page, in being each of the following: on the topic, a topical authority, a topical homepage, a topical authoritative homepage, and any other possible combination of these criteria.

In addition, since on the Web, and other hyperlinked environments, users browse, they assess a web page in terms of the information it contains, and the information it provides access to [4, 11], i.e. as an entry point to the Web's structure. In particular, when many interlinked pages from the same site are retrieved, users would rather not be presented with all of them, but with only a *Best Entry Point* (BEP) [4] to the site, i.e. a page at a suitable level in the site's hierarchy providing access, by browsing, to the relevant information in the site.

For instance, BEPs could correspond to homepages, viewed as good entry points for users to follow the flow of information in the site, or to be presented with an overview of its content [9]. To identify BEPs as *topical homepages*, web IR systems could combine content-based and homepage evidence. Alternatively, they could assess each page in terms of its own features and those of the pages it provides access to. Such approaches aggregate the features of each page with those of its linked pages, by propagating them through the site's structure [4].

The aggregation can be performed by propagating: (i) *term weights* [10, 15] or (ii) *relevance scores* [13, 15]. The former identifies BEPs with respect to topicality, whereas the latter is able to capture multiple relevance criteria, depending on the relevance scores incorporated in the aggregation. For instance, by aggregating relevance scores indicating the topicality and authority of pages, we can model BEPs that provide access to pages containing authoritative information on a topic. This flexibility has not been fully exploited in the context of the Web, where relevance scores reflecting only a single criterion, e.g. content-based relevance scores reflecting topicality [13, 15, 5, 6], are usually aggregated.

Therefore, our aim is twofold: (i) to assess relevance with respect to relevance criteria, by combining the evidence indicating these criteria, and (ii) to model this combination of evidence in a uniform way, irrespective of whether the evidence is associated with a single information item (e.g. a single web page) or with related information items (e.g. linked web pages). Although our aim is to provide a framework applicable to various environments, we focus on web

retrieval, where we assess each web page as an entry point with respect to any relevance criterion, given either its own features, or also those of its linked pages.

To estimate the belief in relevance by combining the available evidence, various formalisms for reasoning with uncertainty can be employed. We explore the possibility of modelling our framework using Dempster-Shafer theory of evidence [17], an alternative formalism to probability theory. We consider this theory to be useful at the conceptual design level, and for providing guidance in expressing and performing the combination of evidence. We apply our framework to a web retrieval task, TREC's Topic Distillation [5,6], an informational task concerned with retrieving for a broad topic key resources, interpreted as BEPs (that correspond to *homepages*) of sites providing *credible* information on the *topic*.

Section 2 introduces Dempster-Shafer theory. The framework, expressed within this theory, is described in Section 3. It is evaluated for TREC's Topic Distillation task. Section 4 describes the experimental setting, and presents and discusses the results of the experiments. Section 5 provides some concluding remarks.

2   Dempster-Shafer Theory of Evidence

Dempster-Shafer (DS) theory of evidence is a formalism for representing, manipulating and revising *degrees of belief* rendered by multiple sources of evidence to a common set of propositions. It concerns the same concepts as those considered by Bayesian probability theory. It does not rely, though, on the probabilistic quantification of degrees of belief, but on a more general system based on *belief functions*. This theory was developed by Shafer [17], based on Dempster's earlier work [8]. We summarise the necessary background of the theory, by adopting Shafer's [17] initial terminology, notation and interpretation of the formalism.

Frame of discernment. Suppose we are concerned with the value of some quantity θ and the set of its possible values is Θ. In DS theory, this set Θ of exhaustive and mutually exclusive events is called *frame of discernment*. There is an one-to-one correspondence between subsets of Θ and propositions. The propositions of interest could be: "the value of θ is in A", $A \subseteq \Theta$. If $A = \{a\}$, $a \in \Theta$, the proposition is expressed as "the value of θ is a" and constitutes an *elementary proposition. Non-elementary propositions* are disjunctions of elementary ones.

Basic probability assignment. The belief committed to a proposition given some evidence is quantified by a function $m : 2^{\Theta} \rightarrow [0,1]$ called a *basic probability assignment* (bpa). Bpas can assign belief to any proposition in the frame and not only to the elementary ones. No belief can ever be assigned to the false proposition ($m(\emptyset) = 0$) and the sum of all bpas must equate 1: $\sum_{A \subseteq \Theta} m(A) = 1$. The quantity $m(A)$ represents the belief committed *exactly* to A, which due to lack of evidence (ignorance) cannot be committed to any proper subset of A.

Belief assignments are carried out only for propositions for which there is evidence. Consequently, committing belief to a proposition A does not necessarily imply that the remaining belief is committed to its negation $\neg A$. Therefore, if $m(A) = 0.6$, and there is no further evidence for or against A or any other proposition in Θ, then, the remaining $1 - 0.6 = 0.4$ is assigned to the frame:

$m(\Theta)$=0.4. This represents a state of ignorance and implies that this remaining belief could be assigned to any proposition in Θ, when new evidence becomes available. Complete ignorance with respect to the frame Θ is represented by the *vacuous* bpa: $m(\Theta)$=1 and $m(A)$=0, $\forall A \subset \Theta$. In any case, if $m(A) > 0$, A is called a *focal element*. The focal elements and associated bpa define a *body of evidence*.

We can also obtain a δ-*discounted bpa* m^{δ} $(0 \leq \delta \leq 1)$ from the original bpa m as follows: $m^{\delta}(A) = \delta * m(A)$, $\forall A \subset \Theta$ and $m^{\delta}(\Theta) = \delta * m(\Theta) + 1 - \delta$. The discounting factor δ represents a form of knowledge on the reliability of the body of evidence.

Belief function. Given a body of evidence with bpa m, one can compute the *total* belief committed to a proposition $A \subseteq \Theta$. This is done with a *belief function* $Bel : 2^{\Theta} \mapsto [0, 1]$ defined upon m, so that it considers the belief assigned to the more specific propositions (i.e. to the subsets) of A: $Bel(A) = \sum_{B \subseteq A} m(B)$.

Dempster's combination rule. This rule aggregates two distinct bodies of evidence, with bpas m_1 and m_2, defined within the same frame Θ, into one body of evidence defined by a bpa m on the same frame: $m(A) = m_1 \oplus m_2(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)}$. The rule is commutative and associative. It computes a measure of agreement between two bodies of evidence concerning propositions discerned from a common frame. It focuses only on propositions that both bodies of evidence support. The numerator is the sum over all conjunctions that support a proposition. The denominator is a normalisation factor ensuring m is a bpa. Combining bpa m_1 with a vacuous bpa m_v, has no effect on m_1: $m_1 \oplus m_v = m_1$.

3 Combining Evidence for Relevance Criteria

This section presents our framework, expressed within DS theory, for modelling the combination (and aggregation) of evidence (features) for relevance criteria. Our presentation focuses on web retrieval. In Section 3.1, we assess the relevance of each page, given either its own features, or also those of its linked pages, without considering what the underlying criteria are. In Section 3.2, we extend our framework and explicitly consider the criteria of interest. In both cases, we consider that the features' values have been transformed to relevance scores.

3.1 The Basic Framework: Combining Evidence for Relevance

We define the frame of discernment Θ in terms of the relevance criteria of interest. When we only consider the relevance of web pages without explicitly specifying the underlying criteria, the elements of Θ are defined as the mutually exclusive propositions $\theta_0 = \{\neg R\}$ and $\theta_1 = \{R\}$. Proposition $\{R\}$ reflects "a good point to enter for accessing R information", R being relevant. Each page x, referred to as *object* o_x, is represented by a body of evidence defined in Θ. Its associated bpa $m_x(A)$ quantifies the belief in $A \subseteq \Theta$, given all available evidence for x.

Representation. When a single source of evidence of relevance is available, $m_x(\{R\})$ (denoted $m_x(R)$ for simplicity) quantifies the degree to which this evidence indicates that this is a good point to enter to access relevant information. Suppose page x has a relevance score 0.6 given evidence e, then $m_x(R) = 0.6$. The

remaining belief is assigned to Θ, $m_x(\Theta) = 0.4$, representing that, at this stage, we have no further evidence for any other proposition in Θ. The total belief is $Bel_x(R) = m_x(R)$. Suppose page y has a zero relevance score given evidence e. A first approach is to associate o_y with a *vacuous bpa* $m_y(\Theta) = 1$. This expresses complete ignorance with respect to Θ, i.e. we consider that our evidence does not allow us to assign any belief in $\{R\}$ or $\{\neg R\}$. However, we do know that, given evidence e, page y was assessed as non relevant. This can be used to express our belief in $\{\neg R\}$. Therefore, a second approach is to set $0 < m_y(\neg R) < 1$.

When more than one source of evidence is available for each page, a separate bpa is defined in terms of each source of evidence taken into account. Suppose that page z is assigned relevance score 0.6 given evidence e_1, and relevance score 0.7 given evidence e_2. Then, page z is represented by 2 separate bpas: $m_{z:e_1}(R) = 0.6$ ($m_{z:e_1}(\Theta) = 0.4$) and $m_{z:e_2}(R) = 0.7$ ($m_{z:e_2}(\Theta) = 0.3$).

Combination. To combine the available evidence associated with a page, we combine the bodies of evidence using Dempster's combination rule. For instance, given page z as above, the combination yields $m_z = m_{z:e_1} \oplus m_{z:e_2}$, with $m_z(R) = (m_{z:e_1}(R) * m_{z:e_2}(R) + m_{z:e_1}(R) * m_{z:e_2}(\Theta) + m_{z:e_1}(\Theta) * m_{z:e_2}(R))/1 = 0.6 * 0.7 + 0.6 * 0.3 + 0.4 * 0.7 = 0.88$ and $m_z(\Theta) = (m_{z:e_1}(\Theta) * m_{z:e_2}(\Theta))/1 = 0.4 * 0.3 = 0.12$.

Aggregation. To assess each page in terms of its own features and those of the pages it provides access to, we aggregate the bodies of evidence of linked pages belonging to the same site using Dempster's combination rule.

Consider the web sites in Figure 1. Each page i, referred to as object o_i, is represented by a body of evidence in Θ and its associated bpa is m_i. Given the evidence from page p (site A) and its linked children pages c_k, $k = 1, ..., 5$, the aggregation is expressed as: $m_{p,c_{1-5}} = m_p \oplus m_{c_1} \oplus ... \oplus m_{c_5}$.

As the user enters site A at page p, the actual information accessed is the one contained in p. The information contained in its children should be considered as "potential" [13], since the user needs to traverse the links in order to fully access it. Hence, the contribution of evidence from the children as a whole should be weighed appropriately, to reflect the uncertainty associated with their propagation to the parent page. This is expressed with a *propagation* (or *fading* [13]) factor, and is modelled by a *discounted bpa*. For instance, the bpa associated with the aggregate $o_{c_{1-5}}$, formed from the children of page p, is $m_{c_{1-5}}^{prop}$, where *prop* is the propagation factor. This is expressed as: $m_{p,c_{1-5}} = m_p \oplus m_{c_{1-5}}^{prop}$.

We can also express the contribution of each child o_{c_k} in forming the aggregate. The extent of this contribution, referred to as *accessibility* (*acc*) [16], is modelled by a discounted bpa $m_{c_k}^{acc_k}$. The bpa for c_{1-5} is: $m_{c_{1-5}} = m_{c_1}^{acc_1} \oplus ... \oplus m_{c_5}^{acc_5}$. Thus, the belief in $\{R\}$ is (see also the definition of discounted bpas):

$$m_{c_{1-5}}(R) = (acc_1 * m_{c_1}(R)) \oplus ... \oplus (acc_5 * m_{c_5}(R)) \tag{1}$$

$$m_{p,c_{1-5}}(R) = m_p(R) \oplus (prop * m_{c_{1-5}}(R)) \tag{2}$$

To determine the BEP in a site with respect to relevance, we rank the pages in the site by their total belief in $\{R\}$: $Bel(R) = m(R)$.

Aggregation methods. By appropriately setting the accessibility and propagation factors, we can express various aggregation methods.
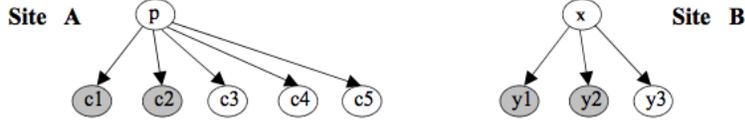
Site A   p
c1 c2 c3 c4 c5

x   Site B
y1 y2 y3

Fig. 1. Examples of linked pages in web sites

Table 1. Examples of aggregation methods applied to site A

Site A	Aggregation method **acc1**				Aggregation method **accn**				Aggregation method **notR**			
	$m_i(\cdot)$			$Bel_i(\cdot)$	$m_i(\cdot)$			$Bel_i(\cdot)$	$m_i(\cdot)$			$Bel_i(\cdot)$
o_i	R	¬R	Θ	R	R	¬R	Θ	R	R	¬R	Θ	R
o_{c_1}	0.8	0	0.2	0.8	0.16	0	0.84	0.16	0.8	0	0.2	0.8
o_{c_2}	0.6	0	0.4	0.6	0.12	0	0.88	0.12	0.6	0	0.4	0.6
$o_{c_{1-2}}$	0.92	0	0.08	0.92	0.261	0	0.739	0.261	0.92	0	0.08	0.92
o_{c_3}	0	0	1	0	0	0	1	0	0	0.1	0.9	0
$o_{c_{1-3}}$	0.92	0	0.08	0.92	0.261	0	0.739	0.261	0.91	0.01	0.08	0.91
o_{c_4}	0	0	1	0	0	0	1	0	0	0.1	0.9	0
$o_{c_{1-4}}$	0.92	0	0.08	0.92	0.261	0	0.739	0.261	0.90	0.02	0.08	0.90
o_{c_5}	0	0	1	0	0	0	1	0	0	0.1	0.9	0
$o_{c_{1-5}}$	0.92	0	0.08	0.92	0.261	0	0.739	0.261	0.89	0.03	0.08	0.89
o_p	0	0	1	0	0	0	1	0	0	0.1	0.9	0
$o_{p,c_{1-5}}$	0.92	0	0.08	0.92	0.261	0	0.739	0.261	0.88	0.04	0.08	0.88

Suppose that only pages c_1, c_2 (site A) and pages y_1, y_2 (site B) are assigned non-zero relevance scores given all available evidence e, i.e. are retrieved given evidence e. Suppose also that the bpas for these retrieved pages (given evidence e) are: $m_{c_1}(R) = m_{y_1}(R) = 0.8$ and $m_{c_2}(R) = m_{y_2}(R) = 0.6$. and that we associate the non-retrieved pages with vacuous bpas: $m_j(\Theta) = 1$, $j = \{p, c_3, c_4, c_5, x, y_3\}$.

Method **acc1** sets the accessibility of each child, i.e. its individual contribution to the aggregation, equal to 1. The aggregation of objects o_{c_1}, o_{c_2} (site A) yields object $o_{c_{1-2}}$ (Table 1). The belief of the aggregate object in $\{R\}$, $m_{c_{1-2}}(R) = 0.92$, is greater than that of either of its component objects. Since, the non-retrieved children, o_{c_3}, o_{c_4} and o_{c_5}, are associated with vacuous bpas, their aggregations with $o_{c_{1-2}}$, for forming $o_{c_{1-5}}$, leave $m_{c_{1-2}}$ unaffected, i.e. $m_{c_{1-5}}(R) = m_{c_{1-2}}(R) = 0.92$. Similarly for site B, $m_{y_{1-3}}(R) = m_{y_{1-2}}(R) = 0.92$. If the propagation factor is uniformly set across sites (e.g. $prop = 1$), the belief in pages p and x is the same, despite having different numbers of non-retrieved children. Method **acc1** considers only the contribution of the retrieved children.

To model page x as a better BEP than page p, since it provides access to less non-relevant information, we need to consider the non-retrieved pages. One way is to set the propagation factor $prop = \frac{1}{n}$ (n is the number of children). Another is to set the accessibility $acc = \frac{1}{n}$ (method **accn** in Table 1). With a propagation factor uniformly set across sites (e.g. $prop = 1$), page x is now considered a better BEP than page p ($m_{x,y_{1-3}}(R) = 0.416 > 0.261 = m_{p,c_{1-5}}(R)$). Method **accn** greatly discounts the contribution of the children in the aggregation.

Method **notR** explicitly takes into account the non-retrieved pages, by modelling them not with a vacuous bpa, but with a bpa that assigns belief to $\{\neg R\}$.

Suppose $m_j(\neg R) = 0.1$, $j = \{p,c_3,c_4,c_5\}$. We set $acc = 1$ and form object $o_{c_{1-2}}$ as before. Objects $o_{c_{1-2}}$, o_{c_3} support conflicting propositions. Their aggregation erodes the beliefs in them, and $m_{c_{1-2}}(R) = 0.92$ becomes $m_{c_{1-3}}(R) = (m_{c_{1-2}}(R) * m_{c_3}(R) + m_{c_{1-2}}(R) * m_{c_3}(\Theta) + m_{c_{1-2}}(\Theta) * m_{c_3}(R)) / (1 - m_{c_{1-2}}(R) * m_{c_3}(\neg R) - m_{c_{1-2}}(\neg R) * m_{c_3}(R)) = (0.92 * 0 + 0.92 * 0.9 + 0.08 * 0) / (1 - 0.92 * 0.1 - 0 * 0) = 0.91$ (Table 1). Greater values of $m_i(\neg R)$ for non-retrieved pages lead to even greater erosion. Also, the more non-retrieved children are included in the aggregation, the more the belief in $\{R\}$ is reduced. By setting $prop = 1$, $m_{p,c_{1-5}}(R) = 0.88 < 0.90 = m_{x,y_{1-3}}(R)$. The values of $m(\neg R)$ can be determined experimentally or by evidence reflecting, for instance, the system's reliability or the query's difficulty.

This aggregation of linked pages belonging to the same site can be performed in an ascending manner (**bottom-up** propagation), starting from the pages deepest in the site's hierarchy. To remove the cycles from the site's structure, we construct a sitemap tree (similarly to [15]), using only *Down* type links (i.e. those linking pages with those below in the site's directory path [9]). Alternatively, we can perform an **1step** propagation, by considering for each page only its immediate neighbours (not necessarily just those connected with *Down* links).

3.2 The Extended Framework: Combining Evidence for Relevance Criteria

In this section, we explicitly consider the criteria underlying relevance.

Frame of discernment. The frame Θ is constructed based on the set of criteria of interest: $\mathbb{E} = \{e_1, \cdots, e_E\}$. The mutually exclusive elementary propositions of Θ are all the possible Boolean conjunctions of all the elements $e_i \in \mathbb{E}$, containing either e_i or $\neg e_i$. There are 2^E elements in Θ, each denoted as $\theta_{b_1 b_2 \cdots b_n}$, with $b_1 b_2 \cdots b_n$ an n-bit binary number, such that $\theta_{b_1 b_2 \cdots b_n}$ corresponds to the proposition "$x_1 \wedge x_2 \wedge \cdots \wedge x_n$", where $x_i = e_i$ if $b_i = 1$, and $x_i = \neg e_i$ if $b_i = 0$.

Suppose the criteria of interest are topicality (T), authority (A), and homepageness (HP): $\mathbb{E} = \{T, HP, A\}$. Then, the propositions forming the frame Θ are listed in Table 2. For instance, θ_{111} corresponds to $\{T \wedge HP \wedge A\}$, reflecting that a page is "a good point to enter to access *homepages* containing *authoritative* information *on the topic*". Analogously, $\{T \wedge A\}$ reflects that a page is "a good point to enter to access *topical* and *authoritative* information". Therefore, θ_{111} provides a more refined representation of the notion of topical relevance compared to $\{T \wedge A\}$, $\{T \wedge HP\}$ or $\{T\}$. In this work, we focus on $\mathbb{E} = \{T, HP\}$. (Due to space limitations, we do not report on criterion $\{A\}$, that we also considered.)

Representation. Consider we have two sources of evidence for each page: one capturing topicality $\{T\}$, and the other homepageness $\{HP\}$. Then, each page is represented by two separate bpas. Suppose the content-based (**C**) score, for page c_1 (site A), reflecting its topicality, is 0.8 and its URL-based (**U**) one, reflecting its homepageness, is 0.6. Then, we have $m_{c_1:C}$ and $m_{c_1:U}$ (Table 3).

Combination. The combination $m_{c_1} = m_{c_1:C} \oplus m_{c_1:U}$ (Table 3) assigns belief to propositions $\{T\}$, $\{HP\}$, and their conjunction $\{T \wedge HP\}$. Given the initial belief $m_{c_1:C}(T) = 0.8$, we were unable to draw any finer distinction about the type of topicality supported, i.e $\{T \wedge HP\}$ or $\{T \wedge \neg HP\}$. Once evidence for $\{HP\}$

Table 2. Propositions forming the frame of discernment Θ in the extended framework

θ_{000}	$\neg T \wedge \neg HP \wedge \neg A$	θ_{010}	$\neg T \wedge HP \wedge \neg A$	θ_{100}	$T \wedge \neg HP \wedge \neg A$	θ_{110}	$T \wedge HP \wedge \neg A$
θ_{001}	$\neg T \wedge \neg HP \wedge A$	θ_{011}	$\neg T \wedge HP \wedge A$	θ_{101}	$T \wedge \neg HP \wedge A$	θ_{111}	$T \wedge HP \wedge A$

Table 3. Combination in the extended frame

Site A	$m_i(\cdot)$				$Bel_i(\cdot)$		
o_i	T	HP	$T \wedge HP$	Θ	T	HP	$T \wedge HP$
$o_{c_1:C}$	0.8	0	0	0.2	0.8	0	0
$o_{c_1:U}$	0	0.6	0	0.4	0	0.6	0
o_{c_1}	0.32	0.12	0.48	0.08	0.8	0.6	0.48
$o_{c_2:C}$	0.6	0	0	0.4	0.6	0	0
$o_{c_2:U}$	0	0.7	0	0.3	0	0.7	0
o_{c_2}	0.18	0.28	0.42	0.12	0.6	0.7	0.42

Table 4. Aggregation in the extended frame

Site A	$m_i(\cdot)$				$Bel_i(\cdot)$		
o_i	T	HP	$T \wedge HP$	Θ	T	HP	$T \wedge HP$
o_{c_1}	0.32	0.12	0.48	0.08	0.8	0.6	0.48
o_{c_2}	0.18	0.28	0.42	0.12	0.6	0.7	0.42
$o_{c_{1-2}}$	0.11	0.07	0.81	0.01	0.92	0.88	0.81
o_{c_3}	0	0	0	1	0	0	0
...							
$o_{p,c_{1-5}}$	0.11	0.07	0.81	0.01	0.92	0.88	0.81

became available, some of this initial belief was assigned to subset $\{T \wedge HP\}$, but the total belief in $\{T\}$, $Bel_{c_1}(T) = m_{c_1}(T) + m_{c_1}(T \wedge HP) = 0.8$, remained the same. We also combine the evidence for o_{c_2}: $m_{c_2} = m_{c_2:C} \oplus m_{c_2:U}$ (Table 3).

Aggregation. Suppose we use aggregation method acc1. The aggregation $m_{c_{1-2}} = m_{c_1} \oplus m_{c_2}$ (Table 4) further redistributes the belief among non-disjoint propositions. The aggregation in terms of Bel is not affected by this distribution of belief, since it is only concerned with the total belief assigned to propositions. For instance, $Bel_{c_{1-2}}(T) = Bel_{c_1}(T) \oplus Bel_{c_2}(T) = 0.8 \oplus 0.6 = 0.92$. Furthermore, this, in essence, corresponds to $Bel_{c_1:C}(T) \oplus Bel_{c_2:C}(T) = 0.8 \oplus 0.6 = 0.92$, i.e. the aggregation in terms of Bel, irrespective of the additional evidence incorporated, produces the same results as if a single source is considered. Aggregating with o_{c_3}, o_{c_4}, o_{c_5}, and o_p, while setting $prop=1$, leads to $m_{p,c_{1-5}}$ (Table 4).

When multiple evidence are aggregated, we can produce many rankings using the belief Bel in different propositions, and determine BEPs with respect to different criteria. For instance, $Bel(T)$ identifies BEPs for accessing topically relevant pages, while $Bel(T \wedge HP)$ BEPs for accessing topical homepages.

Advantages of using DS theory include assigning belief to criteria for which there is evidence, rather than, of necessity, to every criterion. Also, we can assign belief to a set of propositions, without having to distribute belief among its individual propositions. Finally, the relaxation of the law of additivity ($m(A) + m(\neg A) \leq 1$) allows us to flexibly represent web pages given the available evidence.

4 Experiments

We perform evaluation experiments using the .GOV corpus and the topics and relevance assessments from TREC's Topic Distillation (TD) task (50 topics from TD2003 [5] and 75 topics from TD2004 [6]). We index the pages in the collection by combining their content and incoming anchor text. We apply stopword removal and stemming and use the weighting scheme and retrieval component employed in InQuery [3]. This content-based retrieval approach (**C**) is our baseline. To select the BEP from each site, we group, by their domain name, the top 500 pages retrieved by C, and apply aggregation approaches to each group.

We perform the aggregation in our extended DS framework with criteria of interest *topicality* (T) and *homepageness* (HP), i.e. we form Θ based on $\mathbb{E} = \{T, HP\}$. In Section 4.1, we focus on topicality and rank the pages by their $Bel(T)$. In Section 4.2, we also consider their homepageness and rank them by their $Bel(T \wedge HP)$. The belief $m\{T\}$ is quantified by the *content-based relevance score* (C), whereas $m\{HP\}$ by a query-independent *URL-based relevance score* (**U**), computed using each page's URL path length: $\frac{1}{log_2(urlpathlen+1)}$ [14].

We apply the following *aggregation methods*: **DS acc1**(*prop*), **DS accn**(*prop*), and **DS notR**(*prop*, not*T*), where *prop* is the propagation factor, and not*T* the belief experimentally assigned to proposition $\{\neg T\}$ for pages not in the top 500 retrieved by C: $m(\neg T)$=not*T*. We compare these DS aggregations to linear combination (**LC**) aggregations, which can be considered to derive from equations (1) and (2) (Section 3.1), by replacing DS combination ($\oplus$) with addition ($+$). These aggregation methods are **LC acc1**(*prop*) and **LC accn**(*prop*).

For each of these aggregation methods (DS acc1,LC acc1,DS accn,LC accn,DS notR), we apply the *propagation strategies*: *bottom-up* and *1step Down*. These two strategies consider only the *Down* type links and our results indicate that they perform similarly. Therefore, we only present the more efficient *1step Down* propagation. We also apply *1step* propagation by aggregating linked pages connected with all, not only *Down*, types of intra-site links (*1step All*).

To tune parameters *prop* and not*T*, we use TD2003 as our training set, with TD2004 becoming our test set. Tuning *prop* involved an exploration from 0.1 to 1 at step 0.1, and tuning not*T* an exploration from 0.1 to 0.9 at step 0.1. These tunings aimed at maximising P@10. We select P@10 because, in TD2003, mean average precision (MAP) and R-precision (precision at R, R = number of relevant documents for a query) are more sensitive than P@10 [18]. We also set $prop=\frac{1}{n}$ (*n*=number of children) which led to poor results and is not reported.

In all the presented tables, the effectiveness values improving over the baseline are depicted in **bold**. Statistically significant results, indicated by a $*$, are determined by applying a Wilcoxon matched-pairs signed ranks test (α=0.05).

4.1 Experiments in Aggregating Evidence for Topicality

First, we select the BEP from each site with respect to topical relevance criteria, i.e. we select pages that provide access to topically relevant information. To this end, we aggregate, in a DS or linear manner, the content-based relevance scores of linked pages. Previous research has already indicated that the within-site linear aggregation of content-based relevance scores is effective for Topic Distillation [15]. Our objectives are: (i) to examine the effectiveness of this aggregation when modelled within our DS framework (and also compare it to a linear aggregation) and (ii) to gain an insight into the workings of the aggregation, by studying the effect of the various aggregation methods and propagation strategies.

In the training set, for most propagation strategies (except for DS accn and LC accn for *1step Down*), the lower the contribution of the children as a whole (determined by the propagation factor), the better the results. The best results

were achieved for $prop=0.1$. Also, the more links were considered (*1step All* vs. *1step Down*), the more the effectiveness improved, suggesting that evidence from pages connected with all types of links is beneficial. We applied each propagation strategy, with its most effective parameter(s) for each aggregation method, to our test set (Table 5). Our training set observation, that considering low contributing evidence from all children is beneficial, is confirmed by our test set results.

While the contribution of the children pages as a whole is determined by *prop*, the contribution of each individual child is determined by the aggregation method. Method acc1 considers only the children retrieved by C. Method accn greatly discounts the contribution of each child and thus is *indirectly* affected by the non-retrieved children. Method DS notR is *directly* affected by non-retrieved children, with their contribution expressed through not$T$.

In the training set, acc1 was the most effective method, followed by notR, whereas accn did not perform particularly well. These observations are confirmed by the test set results, indicating that although the contribution of the retrieved children should be low (expressed through low *prop* values), it should not be too greatly discounted (as achieved by accn). This is further supported by DS notR being most effective for low not$T$ values, i.e. not$T=0.1$, which discount the contribution of retrieved pages more gradually than accn (see Table 1). The most effective methods, DS acc1, LC acc1, and DS notR, for *1step All*, improve P@10 significantly over the baseline, with DS acc1 *1step All* also improving MAP.

Overall, our results confirm previous findings that aggregating content-based evidence from the retrieved children of a web page is beneficial for Topic Distillation [15, 5, 6]. Our framework allowed us to study these aggregations further, indicating that the contribution of the retrieved children should be low, but not too greatly discounted. In addition, considering only the immediate neighbours is sufficient, with the most effective and robust strategy (*1step All*) taking into account all linked (immediate) children. These findings apply for both DS and linear aggregations, with the DS aggregation performing comparatively better. Our DS framework also provides the expressiveness and flexibility to incorporate evidence for additional relevance criteria, e.g. homepageness, discussed next.

4.2 Experiments in Combining and Aggregating Evidence for Topicality and Homepageness

These experiments aim at assessing relevance with respect to topicality (T) and homepageness (HP) relevance criteria, by considering the available evidence capturing each criterion, i.e. the content-based (C) and URL-based (U) scores of web pages. First, we combine, for each page, its two scores, producing, in essence, a reranking of the C baseline. Next, we express within our DS framework the aggregation of these two scores of linked pages belonging to the same site. In that way, we identify each site's BEP as the page that provides access to homepages containing topically relevant information. We denote this aggregation as T⊕HP.

The combination of the C and U scores is performed in our DS framework and compared to a linear combination. The DS combination is expressed as $m(T)\oplus m(HP)$ resulting in belief also assigned to $m(T\wedge HP)$ (see Table 3). We

Table 5. Aggregating content-based evidence (top 500 pages retrieved by C)

TD2004		MAP	P@5	P@10	R-Prec.
C	(baseline)	0.1237	0.2187	0.1893	0.1622
BEP DS acc1	1step Down $prop = 0.1$	0.1064	**0.2480***	**0.2013**	**0.1628**
	1step All $prop = 0.1$	**0.1347**	**0.2827***	**0.2187***	**0.1974***
BEP LC acc1	1step Down $prop = 0.1$	0.0998	**0.2213**	**0.1947**	0.1537
	1step All $prop = 0.1$	0.1136	**0.2453***	**0.2120***	**0.1873***
BEP DS accn	1step Down $prop = 0.9$	0.0977	**0.2213**	0.1880	0.1541
	1step All $prop = 0.1$	0.1121	**0.2373***	**0.2027**	**0.1638**
BEP LC accn	1step Down $prop = 0.8$	0.0981	**0.2213**	0.1880	0.1541
	1step All $prop = 0.1$	0.1121	**0.2453***	**0.2013**	0.1579
BEP DS notR	1step Down $prop = 0.2$ $notT = 0.1$	0.1069	**0.2373**	**0.2000**	0.1566
	1step All $prop = 0.1$ $notT = 0.1$	0.1116	**0.2533***	**0.2067***	**0.1669**

Table 6. Combining/Aggregating content- and URL-based evidence (top 500 pages retrieved by C)

TD2004		MAP	P@5	P@10	R-Prec.
C	(baseline for CU and C+0.2U)	0.1237	0.2187	0.1893	0.1622
CU	(baseline for BEP T⊕HP approaches)	**0.1478***	**0.2827***	**0.2227***	**0.1881***
C+0.2U	(baseline for BEP T⊕HP approaches)	**0.1504***	**0.2827***	**0.2213***	**0.1752**
BEP T⊕HP DS acc1	1step Down $prop = 0.1$	0.0971	0.1947	0.1800	0.1623
	1step All $prop = 0.1$	0.1014	0.2267	0.1920	0.1663
BEP T⊕HP DS accn	1step Down $prop = 0.1$	0.1312	0.2720	**0.2413***	**0.1985**
	1step All $prop = 0.1$	0.1287	0.2773	**0.2373**	**0.2027**
BEP T⊕HP DS notR	1step Down $prop = 0.1$ $notT = 0.9$	0.1238	0.2560	**0.2387**	**0.1934**
	1step All $prop = 0.1$ $notT = 0.1$	0.1245	0.2773	0.2200	0.1876

rerank the top 500 pages retrieved by C in terms of $Bel(T \wedge HP) = m(T \wedge HP)$. Since this corresponds, in essence, to a multiplication of the C and U scores, we denote it as CU. The linear combination is expressed as $C + w*U$. We tune w in TD2003 for values 0.1 to 1 at step 0.1, and achieve the best results for $w = 0.2$.

Both combinations improve significantly over C in TD2004 (Table 6), confirming the usefulness of homepage evidence for this task [5,6]. They are also more effective than the aggregations of content-based evidence (see Table 5). Next, we perform the T⊕HP aggregation, using CU and C+0.2U as baselines.

Our training set results for the T⊕HP aggregation indicate that, when also considering homepage evidence, the contribution of the retrieved children is still beneficial, but should be greatly discounted. In fact, accn was the most effective followed by notR and then acc1, with all achieving their best results for $prop = 0.1$. Also considering only few of the children might be sufficient, since *1step Down* performed comparably to *1step All*. We apply the most effective approaches for the T⊕HP aggregation to the test set (Table 6). Our training set observations are confirmed by our test set results. The most effective method is accn, with *1step Down* improving P@10 significantly over all baselines. Method notR is slightly less effective, but still improves, though not significantly, over the baselines, whereas acc1 only improves over the content-based baseline (C).

Previous research has examined either the aggregation of content-based evidence from linked pages, or the combination of content-based and homepage evidence for a single page. We integrate these approaches, and indicate that by incorporating beyond content-based evidence when aggregating linked pages, as modelled by our DS framework, we can further improve the effectiveness.

5 Conclusions

We proposed a framework that assesses relevance with respect to any of the relevance criteria of interest, by combining the evidence indicating these criteria, derived both from a web page and its linked web pages. We estimate the belief in relevance and perform this combination using Dempster-Shafer (DS) theory of evidence. The expressiveness and flexibility of the framework is demonstrated by the ease with which the combination with respect to any relevance criterion is expressed, the aggregation of evidence from linked pages is incorporated, and negated evidence can be considered. We evaluated the framework in the context of TREC's Topic Distillation task, and in terms of the topicality and homepageness relevance criteria. Our experiments indicated the effectiveness of aggregating content-based evidence on their own, or together with homepage evidence, and allowed us to study the workings of aggregation methods.

References

1. C. L. Barry. User-defined relevance criteria: An exploratory study. *JASIS*, 45(3):149–159, 1994.
2. 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.
3. J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *DEXA'92, pp. 78-83*.
4. Y. Chiaramella. Information retrieval and structured documents. In *European Summer School in IR*, volume 1980 of *LNCS*, pages 286–309, 2001.
5. N. Craswell and D. Hawking. Overview of the trec-2003 web track. In *TREC-2003*.
6. N. Craswell and D. Hawking. Overview of the trec-2004 web track. In *TREC-2004*.
7. N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR'05*, pages 416–423, 2005.
8. A. Dempster. A generalization of bayesian inference. *Journal of Royal Statistical Society*, 30:205–247, 1968.
9. N. Eiron and K. S. McCurley. Untangling compound documents on the web. In *ACM Hypertext and Hypermedia conference*, pages 85–94, 2003.
10. N. Fuhr, M. Lalmas, S. Malik, and Z. Szlavik, editors. *Proceedings of INEX 2004*.
11. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
12. W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR'02*, pages 27–34, 2002.
13. M. Marchiori. The quest for correct information on the web: hyper search engines. In *Proceedings of the 6th WWW conference*, pages 1225–1235, 1997.
14. V. Plachouras and I. Ounis. Usefulness of hyperlink structure for query-biased topic distillation. In *SIGIR'04*, pages 448–455, 2004.
15. T. Qin, T.-Y. Liu, Z. X.-D., Z. Chen, and W.-Y. Ma. A study of relevance propagation for web search. In *SIGIR'05*, pages 408–415, 2005.
16. T. Roelleke, M. Lalmas, G. Kazai, I. Ruthven, and S. Quicker. The accessibility dimension for structured document retrieval. In *ECIR'02*, pages 382–402, 2002.
17. G. Shafer. *A mathematical theory of evidence.* Princeton University Press, 1976.
18. I. Soboroff. On evaluating web search with very few relevant documents. In *SIGIR '04*, pages 530–531, 2004.