ISSN 1470-5559

Best Entry Pages for the Topic Distillation Task

Theodora Tsikrika and Mounia Lalmas

Queen Mary
University of London

RR-05-02

April 2005

Department of Computer Science

Best Entry Pages for the Topic Distillation Task

Theodora Tsikrika
Queen Mary University of London
London, E1 4NS
UK
theodora@dcs.qmul.ac.uk

Mounia Lalmas
Queen Mary University of London
London, E1 4NS
UK
mounia@dcs.qmul.ac.uk

ABSTRACT

In a typical web search, users consider entry pages to relevant sites as more valuable than isolated pieces of relevant text. The Topic Distillation Task aims at identifying the page at the right level of site hierarchy considered to provide optimal access, by browsing, to relevant pages within the site, i.e. its *Best Entry Page*. Our aim is to estimate a measure of how good a page is as an entry page to the site it belongs, by aggregating the page's system-assessed relevance with that of its structurally related, Web pages belonging to the same site. To model this aggregation, we propose a framework which is expressed within *Dempster-Shafer Theory of Evidence.* Furthermore,we generalise our model by taking into account other system-assessed properties of Web pages. Apart from their relevance, the authority and hub properties of Web pages are considered in the aggregation. We evaluate our approach by performing experiments using the .GOV test collection. The results of these experiments are promising.

Keywords

Web IR, Dempster-Shafer theory of evidence

1. INTRODUCTION

In the World Wide Web, a document covering a broad topic may be distributed over a number of interlinked pages, which belong to the same site. Users, however, consider it redundant for a Web Information Retrieval (IR) system to return many relevant pages from the same site, since, in practice, they are able to easily reach all these pages by browsing, when given an appropriate entry page to the site. Therefore, Web IR systems should quantify not only how relevant Web pages are, but also how good they are as entry pages to the site they belong. A *Good Entry Page* (GEP) measure should reflect how well a page enables a user to obtain access, by browsing, to the relevant pages within the site. Web IR systems could then employ this measure in order to *focus* retrieval [13], by presenting to the user, not all the relevant pages from a site, but only the page considered to provide **optimal** access, by browsing, to relevant pages within the site, i.e. its *Best Entry Page* (BEP).

TREC Web Track's Topic Distillation Task was introduced in 2002 and one of its objectives was to capture a typical web search, where users consider entry pages to relevant sites as "more valuable than isolated pieces of relevant text" [11]. To be more specific, the aim of the task was to identify *key resources* on a broad topic. In TREC-11 [4], key resources were defined as the type of resources that a human editor might compile. This broad definition encompassed, among others, the notion of the BEP to a site, when multiple pages from the same site are retrieved. In TREC-12 [5], key resources were redefined and constrained to correspond to pages at the right level of site hierarchy acting as the BEPs to the retrieved sites. To simplify and clarify the task, these key resources were biased towards the sites' entry pages (often referred to as their *homepages*[1]).

Various Web IR approaches have been applied in the context of this task [4, 5]. Their aim is to identify a site's BEP, by ranking all the pages in that site with the respect to an estimated GEP measure. To estimate this measure for a particular page, most of these approaches consider that the measure should reflect not only the relevance of that page, but also that of the pages within the site that are accessible from it by browsing. Therefore, by exploiting the structural relations between pages belonging to the same site (i.e. the *site structure*), a page's GEP measure is estimated based on the **aggregation** of its own system-assessed relevance score and that of the pages within the site that are linked by it.

In this work, we aim at estimating how good a page is as an entry page to the site it belongs, by also employing an aggregation-based approach. To model the aggregation of the system-assessed relevance scores of structurally related Web pages belonging to the same site, we propose a framework, which is formally expressed within *Dempster-Shafer (D-S) theory of evidence* [21]. D-S is a theory of uncertainty that supports the explicit representation of combination of evidence, expressed by Dempster's combination rule. This makes the use of D-S theory particularly attractive in this work, as it allows us to model the aggregation in a straightforward manner. Furthermore, we consider that, apart from their property of relevance, other system-assessed properties of Web pages, such as their authority and hub [12], could be taken into account in this aggregation. This allows us

[1] For instance, TREC's homepage is trec.nist.gov, whereas the homepage of TREC's publications site is trec.nist.gov/pubs/ [11].

to consider a generalised view of a GEP measure defined in reference not only to relevance, but to any other property.

The remainder of the paper is organised as follows. Section 2 contains an overview of related work. Section 3 gives an introduction to D-S theory and discusses what makes its use attractive in this work. Our model is described in Section 4, by first considering in the aggregation only the system-assessed relevance of Web pages, and then by incorporating other of their system-assessed properties, such as their authority and hub. The description of the setting for performing experiments in order to evaluate our approach using the .GOV test collection, is provided in Section 5. The results of these experiments are reported and analysed in Section 6. Section 7 provides some concluding remarks and outlines future research.

2. RELATED WORK

Documents displaying **logical structure** can be considered not as atomic entities, but as aggregates of interrelated objects that can be retrieved separately [3]. Examples of such documents include structured documents (e.g. XML), hypertext/ hypermedia and Web documents. Given a query, one may retrieve objects that may be related to each other, e.g sub-components of the same document, linked hypertext nodes or linked Web pages. These related objects may be displayed at distant locations in the result, and this can waste user time and lead to user disorientation [3]. This motivated the introduction of the concept of *best entry points* (BEPs), which correspond to document components from which users can browse to access further relevant document components. The *best entry pages* introduced in the previous section are their Web-specific equivalent.

Most of the approaches employed in identifying BEPs are aggregation-based and exploit the content and the structural knowledge associated with the documents. These approaches can be viewed from two different perspectives depending on how the aggregation is performed.

In the first case, the aggregation is performed at indexing time through the propagation of index term weights. This results in the representation of a document component to be defined as the aggregation of the representation of its own content and the representation of its structurally related components [3]. Given a query, document components of varying granularity are ranked based on their aggregated representation and the top ranking ones are selected as BEPs. In structured document retrieval, proposed models have been based on Dempster-Shafer theory of evidence [13] and Bayesian inference networks [7]. The former has also been applied and evaluated on a small Web test collection, constructed from a single museum Web site [14]. Although the results of this latter study indicated effective focussed retrieval of hierarchically structured Web documents, one of the limitations of this aggregation-based approach is that it does not scale well with the size of standard Web test collections or the real Web.

In the second case, the aggregation is performed at query time through the propagation of system-assessed relevance scores. A ranking is produced by estimating a score for each document component, defined as the aggregation of its own system-assessed relevance score and that of its structurally related components. This approach has been applied in hypertext environments [9], the Web [15] and in the Topic Distillation Task [4, 5]. Most of the approaches employed this task are based on various spreading activation mechanisms [6], where a fraction of the relevance score of each page propagates to the pages linked by it, assuming that pages linked by other relevant pages are possibly relevant as well.

Our aim is to employ the aggregation-based approach based on the propagation of system-assessed relevance scores, and model this aggregation using Dempster-Shafer theory of evidence. Furthermore, apart from system-assessed relevance scores, other system-assessed properties of Web pages, such as their authority and hub [12], are also considered in this aggregation.

3. DEMPSTER-SHAFER THEORY OF EVIDENCE

In this section, we describe the main concepts of Dempster-Shafer (D-S) Theory of Evidence, a mathematical theory of evidence and plausible reasoning, which has been developed by Shafer [21] based on earlier work by Dempster [8].

Frame of discernment. Suppose that we are concerned with the value of some quantity u and that the (non-empty) set of its possible values is Θ. In the D-S framework, this set Θ of mutually exhaustive and exclusive events is called a *frame of discernment*. Propositions are represented as subsets of this set. An example of a proposition is “the value of u is in A” for some $A \subseteq \Theta$. For $A = \{a\}$, $a \in \Theta$, “the value of u is a” constitutes a *basic proposition*. *Non-basic propositions* are defined as the union of basic propositions. Therefore, propositions are in a one-to-one correspondence with the subsets of Θ.

Basic probability assignment. Beliefs can be assigned to propositions to express their certainty. The beliefs are usually computed based on a density function $m : \wp(\Theta) \rightarrow [0, 1]$ called a basic probability assignment (bpa):

$$m(\emptyset) = 0 \text{ and } \sum_{A \subseteq \Theta} m(A) = 1$$

The quantity $m(A)$ represents the belief assigned to *exactly* the set A (and not to any proper subset of A), that is the exact evidence that the value of u is in A. If there is positive evidence for the value of u being in A, then $m(A) > 0$ and A is called a *focal element*. The proposition A is said to be discerned. No belief can ever be assigned to the false proposition (represented as $\emptyset$). The sum of all non-null bpas must equate 1. The focal elements and the associated bpas define a *body of evidence*.

A δ-*discounted bpa* $m^{\delta}(.)$ (with $0 \leq \delta \leq 1$) can be obtained from the original bpa m as follows:

$$\begin{array}{rcl} m^{\delta}(A) & = & \delta m(A) \; \forall A \subseteq \Theta, A \neq \Theta \\ m^{\delta}(\Theta) & = & \delta m(\Theta) + 1 - \delta \end{array}$$

The discounting factor δ represents some form of meta-knowledge regarding the reliability of the body of evidence, which could not be encoded in m.

Belief function. Given a body of evidence with bpa m, one can compute the *total* belief provided by the body of evidence for a proposition. This is done with a belief function $Bel : \wp(\Theta) \mapsto [0, 1]$ defined upon m, so that it takes into account the measures of belief assigned to more specific propositions, i.e. to subsets of A:

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

$Bel(A)$ is the total belief committed to A, that is the total positive effect the body of evidence has on the value of u being in A. Complete ignorance with respect to the frame of discernment Θ is represented by the *vacuous belief* function over Θ, induced by the mass function m defined by $m(\Theta) = 1$ and for all $A \subset \Theta$, $m(A) = 0$.

Dempster's combination rule. This rule aggregates two independent bodies of evidence with bpas m_1 and m_2 defined with the same frame of discernment Θ, into one body of evidence defined by a bpa m on the same frame Θ:

$$m(A) = m_1 \oplus m_2(A) = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B)m_2(C)}$$

Dempster's combination rule, then, computes a measure of agreement between two bodies of evidence concerning various propositions discerned from a common frame of discernment. The rule focuses only on those propositions that both bodies of evidence support. The denominator of the equation is a normalisation factor that ensures that m is a bpa.

The use of D-S theory is particularly attractive in this work, as it provides a rule to combine the effect of different bodies of evidence (i.e. Dempster's combination rule), which allows us to explicitly model the aggregation. Furthermore, D-S theory allows the representation of evidence of different levels of abstraction, which can be used to express the system-assessed properties of the Web pages, which are defined at different levels of abstraction (as it will be explained in Section 4.2.1). Finally, D-S theory supports the possibility of discriminating between uncertainty associated with a source of evidence and any ignorance regarding that source.

4. DESCRIPTION OF THE MODEL

Being a Good Entry Page (GEP) to a site reflects how well the page enables the user to obtain access, by browsing, to pages within this site. Our aim is to estimate a GEP measure of each page as the aggregation of its own system-assessed properties and those of its structurally related Web pages belonging to the same site. This aggregation is modelled within Dempster-Shafer (D-S) theory of evidence.

First, we describe our model by considering in the aggregation only the system-assessed relevance of Web pages (Section 4.1) and then we generalise our model by incorporating other of their system-assessed properties, such as their authority and hub (Section 4.2).

4.1 The model

This section describes the proposed framework for estimating a GEP measure of a page as the aggregation of its own system-assessed relevance and that of its structurally related Web pages belonging to the same site. To model this aggregation within D-S theory, we do the following. First, we define a frame of discernment based on the property of relevance and then describe how Web pages are represented as bodies of evidence within the defined frame of discernment (Section 4.1.1). The aggregation of the bodies of evidence, corresponding to pages which are structurally related to a particular page, allows us to estimate a GEP measure of a page (Section 4.1.2).

4.1.1 Representation of objects

In our framework, the definition of the frame of discernment Θ is based on the system-assessed properties to be considered in the aggregation. When the only property to be taken into account is the system-assessed relevance of Web pages, the elements of Θ are defined as the mutually exclusive propositions $\theta_0 = \{\neg R\}$ and $\theta_1 = \{R\}$. The proposition corresponding to $\{R\}$ reflects that the page has been assessed as relevant by the system (with respect to the submitted query). On the other hand, the proposition corresponding to $\{\neg R\}$ reflects that the page has been assessed as non-relevant by the system.

Each Web page is referred to as an *object* and is represented by a body of evidence defined in Θ, through a set a focal elements for which there is positive evidence. Since a bpa m represents the uncertainty associated to a proposition, $m(p)$ corresponds to the degree to which the system has assessed an object as p. For instance, $m(R)^2$ corresponds to the degree to which the system has assessed an object as relevant with respect to a query. The value of $m(p)$ is estimated by employing an appropriate IR approach, such as a probabilistic or a vector space model. For instance, if we suppose that an object o has been retrieved with relevance score 0.6, $m(R) = 0.6$. In this case, since the propositions correspond to singleton sets, the overall belief $Bel(R) = m(R)$.

From the definition of the bpa, each body of evidence must assign the same total amount of belief to the entire set of properties exhibited by the objects and which define the frame of discernment. One approach in ensuring that this condition holds is to treat it as an *uncommitted belief*, which can be used to represent the uncertainty (overall ignorance) associated with the available evidence regarding these properties. It is defined as $1 - \sum_{p_k \in \Theta} m(p_k)$ and it is assigned as the bpa value of the proposition corresponding to the frame of discernment. If not null, this proposition constitutes a focal element. For instance, for the above example where $m(R) = 0.6$, the uncommitted belief is $m(\Theta) = 0.4$.

4.1.2 Object aggregation

To estimate a GEP measure of a page, we aggregate its own system-assessed relevance and that of its structurally related Web pages belonging to the same site. However, users tend, intuitively, to browse down from starting points [14]. Therefore, they may consider a BEP as one that enables them to access pages that are deeper, or at the same level, in the hierarchy of the site. Initially, we concentrate on the Web pages which are structurally related by *hierarchical down* links and then also consider *same directory* links[3].

A page containing hierarchical down links is represented as an *aggregate object.* This object is derived from the aggregation of the bodies of evidence of its *component objects* (i.e. the objects linked by it with hierarchical down links) and the object corresponding to the page itself. For instance, consider a site consisting of five pages connected by hierarchical down links (Figure 1). Page 3 is then represented as aggregate object a_3 derived from the aggregation of o_1, o_2 and o_3, and page 5, represented as a_5, is derived from the aggregation of a_3 and o_4.

In our example, consider that pages 1 and 2 are retrieved, whereas pages 3, 4 and 5 are not. Suppose that objects o_1 and o_2 corresponding to the retrieved pages, have been assessed as relevant $\{R\}$, with belief 0.8 and 0.6 respectively.

[2] For simplicity, we use $m(R)$ instead of $m(\{R\})$.

[3] Hierarchical down links are intra-domain Web links whose source is higher in the directory path than their target, while for same directory links, their source is in the same directory path as their target.

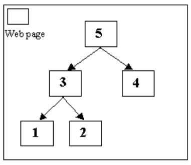

Figure 1: Web Site

Therefore $m_1(R) = 0.8$, $Bel_1(R) = 0.8$ and $m_2(R) = 0.6$, $Bel_2(R) = 0.6$. The uncommitted belief is: $m_1(\Theta) = 0.2$, $m_2(\Theta) = 0.4$, and $m_3(\Theta) = m_4(\Theta) = m_5(\Theta) = 1$.

The aggregation process is applied to the whole site starting with the retrieved pages deepest in the hierarchy, where no aggregation is performed. At the first step of the aggregation, the component objects of o_3 are aggregated into an intermediate aggregate c_3. Its body of evidence is computed using Dempster's combination rule: $m_{c_3} = m_1 \oplus m_2$. For the propositions supported by both bodies of evidence, we have $m_{c_3}(R) = 0.92$ and $m_{c_3}(\Theta) = 0.08$. At the next step, the body of evidence of a_3 is computed: $m_{a_3} = m_{c_3} \oplus m_3$, with $m_{a_3}(R) = 0.92$, $m_{a_3}(\Theta) = 0.08$ and $Bel_{a_3}(R) = m_{a_3}(R)$. Similarly, for a_5: $m_{a_5}(R) = 0.92$, $m_{a_5}(\Theta) = 0.08$ and $Bel_{a_5}(R) = m_{a_5}(R)$.

We consider the belief in the property of relevance $Bel(R)$, as this is computed through the aggregation process, to reflect a GEP measure, since it captures, for each page, its own system-assessed relevance and that of its linked pages. For our example, the ranking with respect to this GEP measure is: $Bel_{a_3}(R) = Bel_{a_5}(R) > Bel_{a_1}(R) > Bel_{a_2}(R)$. However, it would be more intuitive to consider page 3 as the BEP, i.e. to have $Bel_{a_3}(R) > Bel_{a_5}(R)$. To model this, we can employ the following approaches.

First of all, since component objects reflect information deeper in the hierarchy, the contribution of this information should diminish as we move further up. This can be modelled by a *discounted bpa*, with a discounting factor reflecting a *propagation* [9] (or *fading* [15]) factor. This can be applied to intermediate aggregate c_i to reflect the uncertainty associated with the propagation process. Discounted bpas can also be applied to model the contribution of each of the component objects forming an aggregate object. The extent of each contribution, referred to as *accessibility* [19], captures the uncertainty related to the structure of the site. For instance, if an object o has n hierarchical down links to objects o_i, their accessibility could be set to $\frac{1}{n}$, and that of object o itself could be set to 1.

A second approach would be to reflect the contribution of the non-retrieved pages employed in the aggregation, by setting the value of their $m(\neg R)$. In the example described above, when a page is not retrieved by the system, it is represented by $m(\Theta) = 1$. This expresses complete ignorance with respect to Θ. However, the evidence that the page has not been retrieved by the system, can be used to express our belief in the page being non-relevant. This can be expressed by setting $m(\neg R) \neq 0$ for the non-retrieved pages.

Finally, additional types of links, such as same directory links can be taken into account in the aggregation. Consider, for example that in the site depicted in Figure 1, there exists a same directory link between pages 3 and 4. In this case, the component objects of the new aggregate a'_3 are: o_1, o_2, o_3 and o_4. $m'_{a_3} = m_{a_3} \oplus m_4 = m_1 \oplus m_2 \oplus m_3 \oplus m4$ and the GEP measure is estimated by $Bel'_{a_3}(R)$. To estimate m_{a_5}, we could propagate m'_{a_3} and m_4. However, this would result in m_4 being considered twice in the aggregation. Therefore, m_{a_3} is propagated instead.

4.2 Generalisation of the model

This section provides a generalisation of our model by incorporating other system-assessed properties in the aggregation. First, we describe these properties (Section 4.2.1) and define a frame of discernment based on them (Section 4.2.2). We discuss how Web pages are represented as bodies of evidence within the defined frame of discernment (Section 4.2.3). The aggregation of these bodies of evidence is analogous to the process detailed in Section (Section 4.1.2), and therefore it will not be further discussed.

4.2.1 Properties of objects

So far, we have considered in the aggregation only the system-assessed relevance of Web pages. In essence, we view relevance as a static and objective concept, which is assessed by the IR system without a user's intellectual involvement. This corresponds to the *objective class* of relevance [1, 10, 20]. Within this class, there exists a specific *type* [1] (or manifestation [20]) of relevance called *algorithmic relevance*, which describes the **kind of the relation** between the query and the retrieved Web pages, and constitutes one of the system-assessed *properties* of the Web pages.

By referring to the system-assessed property of relevance of Web pages, we have actually been implying the most prominent kind of algorithmic relevance commonly known as "topicality". This is defined in terms of " ... how well the topic of the retrieved information matches the topic of the request" [10]. We label this kind of algorithmic relevance as **topical relevance**, and consider that *topic* refers to the contents of the Web pages and the query. Therefore, this kind of system-oriented algorithmic relevance, as expressed by the retrieved Web pages, is assessed by employing content-based evidence. A specification of topical relevance could be derived by considering the IR model employed by the system and, for instance, we could have probabilistic topical relevance or vector space topical relevance.

In the Web, however, other kinds of algorithmic relevance (or properties) could be assessed, by employing further, Web-specific, evidence. Here, we consider the source of evidence most commonly exploited by Web IR approaches, i.e. the connectivity of a page within the Web graph. This allows us to define Web-specific algorithmic relevance in terms of not only "how well the topic of the retrieved information matches the topic of the request", but also in terms of "how well the retrieved page is connected within the Web graph". The assumption underlying the use of this evidence by Web IR systems, which usually employ *link analysis ranking* algorithms [12], is that the Web's link structure can be viewed as a network of recommendations[4] between pages [22]. When a page is pointed by other pages, it is considered to be recommended by them and vice versa.

The connectivity of a Web page can be determined in terms of its incoming links or its outgoing links. This leads us in considering two kinds of algorithmic relevance.

The first one is determined in terms of "how well the topic of the retrieved page matches the topic of the request and

[4] This network takes into account only inter domain links, since the underlying assumption is that they are the ones conveying endorsement [12].

additionally how well the retrieved page is linked by other pages within the Web graph". This is usually determined in a recursive manner, where "how well the retrieved page is linked by other pages" is assessed by considering how well these pages are linked and so on. When a page is pointed by other pages, it is considered to be recommended by them and regarded as an authority [12]. Therefore, we refer to this kind of algorithmic relevance as **authority relevance**. A page assessed as authoritatively relevant is typically defined as a page that is not only topically relevant, but it is also a "trusted source of correct information" [22].

Similarly, the second kind of algorithmic relevance is determined in terms of "how well the topic of the retrieved page matches the topic of the request and how well the retrieved page links to other pages within the Web graph". We refer to this property as **hub relevance**. A hub relevant page [12], provides a comprehensive list of links to authority relevant pages on the topic of the query.

So far we have considered system-assessed *properties* of Web pages determined with respect to a specific query. However, we can assess "how well a page is connected within the Web graph", irrespective of a query. Therefore, by taking into account the incoming links of a page and employing an appropriate link analysis ranking algorithm (such as PageRank [16]), the *query-independent authority* of a page can be determined. This assessment could be combined with a topical relevance measure, in order to estimate the page's authority relevance. The same applies when considering the incoming links of a page. In this case, the *query-independent hub* of a page (or its *utility* [18]) can be assessed.

In summary, the system-assessed **relevance** of a Web page can be refined into its topical relevance, authority relevance and hub relevance depending on the type of evidence considered. Similarly, the system-assessed **authority** of a Web page can be refined into its authority relevance and query-independent authority and its system-assessed **hub** into its hub relevance and query-independent hub. We introduce the notion of *composite* properties to refer to the ones that are more specific than the *elementary* properties of a page. In our case, the topical, authority and hub relevance, and query-independent authority and hub of a page constitute its composite properties, which are specified with respect to the elementary properties of relevance $\{\mathbf{R}\}$, authority $\{\mathbf{A}\}$ and hub $\{\mathbf{H}\}$.

4.2.2 *Frame of discernment*

To define a frame of discernment based on the above properties, we consider $\mathbb{E} = \{e_1, \cdots, e_E\}$ to be the set of elementary properties and $\mathbb{C} = \{c_1, \cdots, c_C\}$ the set of composite properties, with $c_i \subseteq \mathbb{E}$. The frame of discernment Θ is constructed based on the set $\mathbb{E}$. The elements of the frame are defined as the mutually exclusive propositions, derived by considering all the possible boolean conjunctions of all the elements $e_i \in \mathbb{E}$, containing either e_i or its negation $\neg e_i$. There are 2^E elements in Θ and each is denoted as $\theta_{b_1 b_2 \cdots b_n}$, where $b_1 b_2 \cdots b_n$ is 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 $b_i = 0$.

Since we consider that the set of elementary properties of a Web page consists of the relevance $\mathbf{R}$ of the page, its authority $\mathbf{A}$ and hub $\mathbf{H}$, we define $\mathbb{E} = \{R, A, H\}$. The propositions then forming the frame of discernment Θ are listed in Table 1.

Table 1: Propositions forming Θ

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

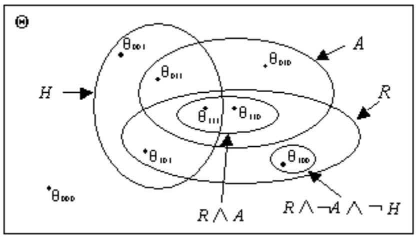

Figure 2: Example of an object in Θ

Each element $\theta_{b_1 b_2 \cdots b_n} \in \Theta$ corresponds to the property $\theta_{b_1 b_2 \cdots b_n}$ assessed by the system for a Web page. For instance, θ_{100} corresponds to $\{R \wedge \neg A \wedge \neg H\}$, reflecting that the page has been assessed as being topically relevant. Therefore, θ_{100} provides a more refined representation of the notion of relevance compared to that provided by the proposition $\{R\}$. The latter corresponds to $\theta_{100} \vee \theta_{101} \vee \theta_{110} \vee \theta_{111}$, and reflects the overall relevance, without specifying what evidence have been considered. Consequently, θ_{100} corresponds to the topical relevance as this is defined in classical IR, where the connectivity of a Web page into the Web graph is not considered.

4.2.3 *Representation of objects*

Each Web page, referred to as an *object*, is represented by a body of evidence defined in Θ. Every elementary property $e_i \in \mathbb{E}$ for which there is positive evidence supporting it, defines a focal element, the proposition p_i. Every composite property c_k also defines a focal element, the proposition $p_k = \bigwedge_l p_l$, where each p_l is the proposition associated to the elementary property e_l for $e_l \in c_k$.

If we consider an object o assessed as: $p_1 = \{R\}$, $p_2 = \{A\}$, $p_3 = \{H\}$ and $p_4 = \{R \wedge A\}$, then these properties are defined in terms of the propositions in Θ as: $p_1 = \theta_{100} \vee \theta_{101} \vee \theta_{110} \vee \theta_{111}$, $p_2 = \theta_{010} \vee \theta_{011} \vee \theta_{110} \vee \theta_{111}$, $p_3 = \theta_{001} \vee \theta_{011} \vee \theta_{101} \vee \theta_{111}$ and $p_4 = \theta_{110} \vee \theta_{111}$. If we further assess property $p_5 = \{R \wedge \neg A \wedge \neg H\}$, then $p_5 = \theta_{100}$ (Figure 2).

The value of $m(p)$ is estimated by employing an appropriate Web IR approach. For instance, if p_i corresponds to the authority relevance of an object $\{R \wedge A \wedge \neg H\}$, $m(p_i)$ could be estimated using HITS algorithm [12].

For the object o defined above, if we suppose that $m(p_1) = 0.2$, $m(p_2) = 0.1$, $m(p_3) = 0.05$, $m(p_4) = 0.15$ and $m(p_5) = 0.1$, then the uncommitted belief $m(\Theta) = 1 - (0.2 + 0.1 + 0.05 + 0.15 + 0.1) = 0.4$. The belief $Bel(R) = m(p_1) + m(p_4) + m(p_5) = 0.2 + 0.15 + 0.1 = 0.45$ can be considered to reflect the object's overall relevance.

The aggregation of these bodies of evidence is analogous to the process detailed in Section 4.1.2. In this case, we can estimate a GEP measure with respect to a property at any level of abstraction. For instance, $Bel(R \wedge \neg A \wedge \neg H)$ can be considered the GEP measure with respect to topical relevance and $Bel(R \wedge A \wedge \neg H)$ with respect to authority relevance. $Bel(R)$, on the other hand, is the measure with respect to the overall relevance.

5. DESCRIPTION OF THE EXPERIMENTS

To evaluate the proposed framework, we perform experiments using the .GOV test collection, employed in the Topic Distillation Task in TREC-11 [4] and TREC-12 [5]. .GOV is a 1.25 million pages crawl of the .gov Internet domain, collected in early 2002. There are 100 available topics: 50 were employed in TREC-11 [4] and 50 in TREC-12 [5].

Our system consists of a retrieval component and a post-retrieval processing component for the identification of BEPs. The retrieval component is the InQuery retrieval system [2], used to index the collection (by applying stopword removal and stemming) using only the content (**C**) of the pages. The top X pages retrieved by submitting the titles of the topics constitute the baseline of our experiments, **C**(top X pages). For this set of experiments, X was set to 100 and 1000.

Processing the **C** results, so that only the top ranking page from each domain was kept and replaced by its domain name, produced a ranking of the top retrieved domains, **C**(top X domains). Processing the relevance assessments in the same manner, produced the set of domains containing key resources. The retrieved domains were evaluated against the processed relevance assessments, in order to measure the effectiveness of the retrieval component in identifying pages from the domains containing the key resources.

The post-retrieval processing component for BEP identification used **C**(top X domains) and replaced the domain name with its BEP. The BEP was identified using the pages retrieved by **C**(top X pages). Two simplistic BEP identification approaches were evaluated: the selection of the top ranking retrieved page from each domain (*TopRanking*) and of the shallowest retrieved one (*Shallowest*). Two aggregation-based approaches were evaluated: a linear combination (**LC**) and the proposed Dempster-Shafer framework (**DS**). The **LC** was used to represent a baseline with respect to the aggregation process. The **DS** was used with frame of discernment $\Theta = \{\neg R, R\}$. $m(R)$ corresponds to the belief estimated by the retrieval component and $Bel(R)$ is the GEP measure.

Initially, we consider only hierarchical down (*Down*) links. The accessibility of the component pages is represented by *acc* and the propagation factor by *prop*. The accessibility of the parent pages is set to 1 for all experiments. The aggregation-based methods are denoted as **LC**(*acc*, *prop*) and **DS**(*acc*, *prop*). For instance, **DS**(1, 0.5) represents the application of the framework with *acc* = 1 and *prop* = 0.5. *prop* and *acc* were experimentally set to 0.25, 0.5, 0.75, 1 and $\frac{1}{n}$, where n is the number of component objects. We also set $m(\neg R)$ for the non-retrieved pages to values from 0.3 to 0.7, at step 0.1. We also consider the case where a second type of links, same-directory (*SameDir*) links, are taken into account in the aggregation.

Finally, we evaluate **DS**, with the frame of discernment formed from $\mathbb{E} = \{R, A\}$. $m(R \wedge \neg A)$ corresponds to the belief estimated by the retrieval component, $m(R \wedge A)$ corresponds to the authority values estimated using HITS [12] on **C**(top 100 pages) and $Bel(R)$ is the GEP measure. In this instance, only *Down* links were considered and the approach is denoted as **DS** $\{R, A\}$.

In TREC-11, the evaluation measure was precision at 10. In TREC-12, the redefinition of the task resulted in a lower number of key resources. This affected the stability of precision at 10 and R-precision (precision at R, where R is the number of relevant documents for a query). We use precision at 5,10, mean average precision (MAP) and R-precision.

6. ANALYSIS OF THE RESULTS

This section discusses the results of our experiments. The baseline content-based (**C**) results for TREC-11 are presented in Table 2. Due to the confusion about the definition of the task, the relevance assessments were carried out in a manner appropriate for the Topic Relevance and not the Topic Distillation task [11]. This implied that content-based approaches performed better than ones biased towards retrieving BEPs. Precision at 10 for our baseline **C** is 0.2204 (Table 2), when the best submitted run in TREC-11 achieved 0.2510 precision at 10. While the content-based approach performed well, any further processing aiming at the identification of BEPs decreased the effectiveness (Table 3). Since this set of relevance assessements does not allow us to investigate the effectiveness of our framework, no further discussion about them will be included.

The baseline content-based (**C**) results for TREC-12 are presented in Table 2. R-precision is 0.0774 and precision at 10 is 0.0760. The highest R-precision score achieved by the runs submitted to TREC-12 was 0.1636, while precision at 10 was 0.1280 [5]. These figures correspond to two different runs and were achieved after the processing for identification of the BEP. One reason explaining the lower effectiveness of our **C** results is that only the content was used when indexing the Web pages. Additional Web evidence, such the referring anchor text, which has shown to improve the effectiveness [5, 11, 17], was not included. For instance, it was reported in [17], that by using the PL2 weighting scheme for TREC-12 topics, the R-precision achieved was equal to 0.0730 when using content only, and equal to 0.1325 when using both content and referring anchor text. In this work, we concentrate on investigating the effect of the application of our framework on the results produced by the baseline. It is an objective of future research to investigate whether improvements in the baseline can lead to improvements in our BEP identification method.

Although, we use precision at 10 and R-precision ($R < 100$) as evaluation measures, the retrieval of **C**(top 1000 pages) is employed in order to increase the number of retrieved pages from each domain and investigate their effect when identifying the BEP. The evaluation of the top ranking domains (Table 2) is a first indication that **C** retrieves pages from domains containing pages assessed as key resources.

By comparing the two simplistic BEP identification approaches, *Shallowest* performs better when using the top 100 retrieved pages (Table 4), while *TopRanking* performs better when using the top 1000 (Table 5). These results are consistent with those produced when these methods were applied in the Topic Distillation Task [4, 5] and indicate that these techniques do not work particularly well. Therefore, a more sophisticated approach is required.

The **LC** aggregation approach is applied by employing only the system-assessed relevance scores. It achieves R-precision of 0.1207 when the top 100 retrieved pages are used (Table 4), and 0.1226 for the top 1000 (Table 5). These figures constitute a significant improvement over **C**. When *acc* = $\frac{1}{n}$, **LC** initially estimates the average of the relevance

Table 2: C results for TREC-11 & TREC-12

TREC-11	MAP	Pr. at 5	Pr. at 10	R-Pr.
C (top 100 pages)	0.1460	0.2531	0.2204	0.1713
C (top 1000 pages)	0.1642	0.2531	0.2204	0.1733
C (top 100 domains)	0.4088	0.4898	0.3959	0.4294
C (top 1000 domains)	0.4583	0.4898	0.3959	0.4388
TREC-12	MAP	Pr. at 5	Pr. at 10	R-Pr.
C (top 100 pages)	0.0899	0.1000	0.0760	0.0774
C (top 1000 pages)	0.0985	0.1000	0.0760	0.0774
C (top 100 domains)	0.3762	0.3760	0.2920	0.3394
C (top 1000 domains)	0.4097	0.3760	0.2920	0.3423

Table 3: TREC-11 top 100 pages

TREC-11	MAP	Pr. at 5	Pr. at 10	R-Pr.
C	0.1460	0.2531	0.2204	0.1713
TopRanking	0.0723	0.2082	0.1816	0.1215
Shallowest	0.0589	0.1837	0.1714	0.1122
LC ($\frac{1}{n}, 0.25$)	0.0710	0.2041	0.1796	0.1215
LC ($\frac{1}{n}, 0.5$)	0.0678	0.1959	0.1755	0.1191
DS ($\frac{1}{n}, 0.5$)	0.0713	0.2041	0.1796	0.1211
DS ($m(\neg R) = 0.5$)	0.0670	0.2000	0.1776	0.1203

scores of the component pages. This average is weighted by *prop* before being added to the score of the parent page. Tables 4, 5 indicate that the lower the value of *prop* in **LC**, meaning the lower the impact of the component pages in the aggregation, the better the results. The best results are achieved when $prop = 0.25$, while an increase in the value of *prop*, decreases R-precision.

The results of the **DS** aggregation approach employing only the system-assessed relevance scores are listed in Tables 4, 5. For $acc = 1$, the best results are achieved when $prop = 0.5$, with R-precision 0.1113 when the top 100 retrieved pages are used (Table 4), and 0.1061 for the top 1000 (Table 5). These results are an improvement over **C**. However, the results of the **LC** aggregation are slightly better. **DS** achieves its best results for $prop = 0.5 > 0.25$ used by **LC**. This could indicate that for **DS** to achieve better results, the component pages should contribute more in the aggregation than they do in **LC** . The worst results for **DS** with $acc = 1$, are observed when $prop = 1$. This indicates that the contribution of the component pages in the aggregation should be lower than that of the parent page.

For **DS** with $acc = \frac{1}{n}$, the contribution of the component pages is considered within the computation of their intermediate aggregate, which is propagated and aggregated with the parent. This means that since the impact of the component pages is already diminished within their aggregation, an additional propagation factor is probably not needed. This is verified in the **DS** for $acc = \frac{1}{n}$, where the best results are achieved for $prop = 1$ when the top 100 retrieved pages are used (Table 4), and for $prop = 0.75$ for the top 1000 pages (Table 5). These results are an improvement over **C**, with the **LC** aggregation still performing slightly better.

So far, we have investigated the effect on the aggregation of the system-assessed relevant pages. The contribution of the non-retrieved pages participating in the aggregation can be captured in the **DS** approach, by setting $m(\neg R) \neq 0$. The values of *acc* and *prop* are set to 1, so that changes in the effectiveness can be attributed to the influence of $m(\neg R)$. When the aggregation is performed using the top 100 retrieved pages (Table 4), the best results are observed for $m(\neg R) = 0.4$, with R-precision equal to 0.1229. This constitutes a significant improvement over **C**, equivalent to that achieved by **LC**. Similar results are obtained when the aggregation is performed using the top 1000 retrieved pages (Table 5), This indicates that considering the contribution of the non-retrieved pages in the aggregation, could be an important source of evidence and should be further investigated.

Table 4: TREC-12 top 100 pages

TREC-12	MAP	Pr. at 5	Pr. at 10	R-Pr.
C	0.0899	0.1000	0.0760	0.0774
TopRanking	0.0699	0.0920	0.0760	0.0787
Shallowest	0.0862	0.1080	0.0960	0.0965
LC ($acc, prop$)				
LC ($\frac{1}{n}, 0.25$)	0.1036	0.1120	0.0880	0.1207
LC ($\frac{1}{n}, 0.5$)	0.0939	0.1080	0.0880	0.1107
LC ($\frac{1}{n}, 0.75$)	0.0881	0.1040	0.0880	0.1107
LC ($\frac{1}{n}, 1$)	0.0819	0.1000	0.0860	0.1057
DS ($acc, prop$)				
DS $(1, 0.25)$	0.0851	0.1040	0.0820	0.0893
DS $(1, 0.5)$	0.1006	0.0960	0.0840	0.1113
DS $(1, 0.75)$	0.0842	0.0840	0.0800	0.0995
DS $(1, 1)$	0.0445	0.0560	0.0620	0.0709
DS ($\frac{1}{n}, 0.5$)	0.0716	0.1000	0.0800	0.0843
DS ($\frac{1}{n}, 0.75$)	0.0885	0.1160	0.0860	0.1007
DS ($\frac{1}{n}, 1$)	0.0955	0.1160	0.0880	0.1107
DS ($m(\neg R)$)				
DS (0.3)	0.0881	0.1000	0.0880	0.1129
DS (0.4)	0.1030	0.1160	0.0900	0.1229
DS (0.5)	0.1096	0.1160	0.0900	0.1207
DS (0.6)	0.1068	0.1200	0.0880	0.1140
DS (0.7)	0.0953	0.1120	0.0860	0.1074
DS {R, A} ($acc, prop$)				
DS {R, A} $(1, 0.5)$	0.0763	0.1120	0.0800	0.0920
DS {R, A} $(1, 0.75)$	0.0640	0.0920	0.0680	0.0847
DS {R, A} ($\frac{1}{n}, 0.5$)	0.0746	0.1120	0.0820	0.0854
DS {R, A} ($\frac{1}{n}, 0.75$)	0.0733	0.1120	0.0820	0.0854

The results of the **DS** which takes into account both *Down* and *SameDir* links are listed in Table 6. R-precision increases with respect to **C**, but it does not improve over the **DS** approach when only *Down* links are employed. However, this method was evaluated using only a limited number of combinations of values for the *acc* and *prop* parameters and $m(\neg R)$ was set to 0. Further experiments are needed to conclude whether it is worth including additional types of links and what their contribution should be.

Finally, the results of the **DS** $\{R, A\}$ approach, where both the topical relevance and the authority relevance scores are taken into account, are listed in the lower part of Table 4. The results indicate that the incorporation of the HITS authority relevance scores does not improve on the effectiveness. Further investigation is needed in what Web evidence associated with the hyperlink structure could be considered.

The results of the experiments are promising. They indicate that our **DS** approach improves the effectiveness over the baseline and is at least as effective as a simple **LC** aggregation. In addition, the framework's flexibility allows us to incorporate and combine various sources of evidence, ranging from the system-assessed properties of the aggregated Web pages to their contribution in the aggregation process.

7. CONCLUSIONS AND FUTURE WORK

We proposed a framework for estimating a measure of how good a Web page is as an entry page to the Web site it belongs. This measure is estimated by aggregating the page's system-assessed properties with those of its structurally related Web pages. The framework is expressed within Dempster-Shafer theory of evidence, with the properties of the Web pages, such as their relevance, authority and

Table 5: TREC-12 top 1000 pages

TREC-12	MAP	Pr. at 5	Pr. at 10	R-Pr.
C	0.0985	0.1000	0.0760	0.0774
TopRanking	0.0725	0.0920	0.0760	0.0794
Shallowest	0.0448	0.0800	0.0680	0.0680
LC $(acc, prop)$				
LC $(\frac{1}{n}, 0.25)$	0.1148	0.1320	0.1000	0.1226
LC $(\frac{1}{n}, 0.5)$	0.1026	0.1360	0.0940	0.1135
LC $(\frac{1}{n}, 0.75)$	0.0836	0.1160	0.0860	0.0989
LC $(\frac{1}{n}, 1)$	0.0741	0.1040	0.0780	0.0998
DS $(acc, prop)$				
DS $(1, 0.25)$	0.0869	0.1000	0.0860	0.0909
DS $(1, 0.5)$	0.1046	0.1080	0.0800	0.1061
DS $(1, 0.75)$	0.0635	0.0920	0.0720	0.0820
DS $(\frac{1}{n}, 0.25)$	0.0742	0.1000	0.0820	0.0859
DS $(\frac{1}{n}, 0.5)$	0.0751	0.0920	0.0800	0.0836
DS $(\frac{1}{n}, 0.75)$	0.1088	0.1160	0.0940	0.1139
DS $(\frac{1}{n}, 1)$	0.0897	0.1160	0.0940	0.1010
DS $(m(\neg R))$				
DS (0.3)	0.0728	0.0840	0.0680	0.0829
DS (0.4)	0.0871	0.0840	0.0680	0.0991
DS (0.5)	0.1188	0.1034	0.0862	0.1208
DS (0.6)	0.1022	0.1000	0.0800	0.1115
DS (0.7)	0.1315	0.1241	0.0966	0.1337

Table 6: TREC-12 with Down and SameDir links

	TREC-12 top 100	MAP	Pr. at 5	Pr. at 10	R-Pr.
DS	Down $(1, 0.5)$ SameDir $(1, 0.25)$	0.1054	0.1120	0.0920	0.1130
DS	Down $(1, 0.5)$ SameDir $(1, 0.5)$	0.0899	0.1160	0.0880	0.1091
	TREC-12 top 1000	MAP	Pr. at 5	Pr. at 10	R-Pr.
DS	Down $(1, 0.5)$ SameDir $(1, 0.25)$	0.1085	0.1400	0.0980	0.1113
DS	Down $(1, 0.5)$ SameDir $(1, 0.5)$	0.0872	0.1320	0.0900	0.1165

hub represented at various levels of abstraction and various aggregation methods expressed by modelling the contribution of the components in the aggregation. The results of our experiments using the .GOV collection are promising.

We are currently performing experiments in order to further evaluate our framework. Our aim is to employ additional Web evidence, such as the referring anchor text in the indexing and the URL length and counts of incoming and outgoing links in the BEP identification. We would also like to investigate ways in which the uncommitted belief, representing the uncertainty associated with the available evidence, can be exploited. Finally, since the use of BEPs is intended to support users' information seeking behaviour, we are interested in conducting interactive experiments using the .GOV document collection. These would aim at eliciting from the users criteria of what constitutes a BEP, which could be subsequently incorporated in our framework.

8. ACKNOWLEDGEMENTS

9. REFERENCES

[1] P. Borlund. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.

[2] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of DEXA-92, 3rd International Conference on Database and Expert Systems Applications*, pages 78–83, 1992.

[3] Y. Chiaramella, P. Mulhem, and F. Fourel. A model for multimedia information retrieval. Technical Report Technical Report Fermi, ESPRIT BRA 8134, University of Glasgow, 1996.

[4] N. Craswell and D. Hawking. Overview of the trec-2002 web track. In *Proceedings of 11th Text Retrieval Conference (TREC-2002), NIST Special Publication 500-251. Gaitensburg, MD*, 2002.

[5] N. Craswell and D. Hawking. Overview of the trec-2003 web track. In *Proceedings of 12th Text Retrieval Conference (TREC-2003), NIST Special Publication 500-255. Gaitensburg, MD*, 2003.

[6] F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482, 1997.

[7] F. Crestani, L. de Campos, J. F. Luna, and J. Huete. Ranking structured documents using utility theory in the bayesian network retrieval model. In *Proceedings of SPIRE*, pages 168–182, Manaus, Brasil, October 2003.

[8] A. P. Dempster. A generalization of bayesian inference. *Journal of Royal Statistical Society*, 30:205–447, 1968.

[9] H. P. Frei and D. Stieger. Making use of hypertext links when retrieving information. In *Proceedings of the ACM conference on Hypertext*, pages 102–111. ACM Press, 1992.

[10] S. Harter. Physchological relevance and information science. *Journal of the American Society for Information Science and Technology*, 43:602–615, 1992.

[11] D. Hawking and N. Craswell. Very large scale retrieval and web search (preprint version). In E. Voorhees and D. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

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

[13] M. Lalmas. Dempster-shafer's theory of evidence applied to structured documents: Modelling uncertainty. In *Proceedings of SIGIR 1997*, pages 110–118, Philadelphia, PA, USA, 1997.

[14] M. Lalmas and E. Moutogianni. A dempster-shafer indexing for the focussed retrieval of hierarchically structured documents: Implementation and experiments on a web museum collection. In *Proceedings of RIAO*, pages 442–456, Paris, France, 2000.

[15] M. Marchiori. The quest for correct information on the web: hyper search engines. In *Proceedings of the sixth international conference on World Wide Web*, pages 1225–1235. Elsevier Science Publishers Ltd., 1997.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[17] V. Plachouras and I. Ounis. Usefulness of hyperlink structure for query-biased topic distillation. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 448–455. ACM Press, 2004.

[18] V. Plachouras, I. Ounis, and G. Amati. A utility-oriented hyperlink analysis model for the web. In *Proceedings of the First Latin American Web*

Congress (LA-WEB 2003), 2003.

[19] T. Roelleke, M. Lalmas, G. Kazai, I. Ruthven, and S. Quicker. The accessibility dimension for structured document retrieval. In *Proceedings of the Twenty-Fourth European Colloquium on Information Retrieval Research (ECIR02).*, pages 382–302, 2002.

[20] T. Saracevic. Relevance reconsidered. In *Proceedings of the Second international conference on conceptions of library and information science:Integration in perspective*, pages 201–218, Copenhagen:Royal School of Librarianship, 1996.

[21] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.

[22] P. Tsaparas. *Link Analysis Ranking*. PhD thesis, Dept of Computer Science, University of Toronto, 2004.