

Modelling Good Entry Pages on the Web

Theodora Tsirikika and Mounia Lalmas

Department of Computer Science, Queen Mary University of London,
London E1 4NS, UK

{theodora, mounia}@dcs.qmul.ac.uk

<http://qmir.dcs.qmul.ac.uk>

Abstract. Being a *good entry page* to a Web site reflects how well the page enables a user to obtain optimal access, by browsing, to relevant and quality pages within the site. Our aim is to model a measure of how good an entry page is, as a combination of evidence of the properties exhibited by the Web pages, which belong to the same site and are structurally related to it. The proposed model is formally expressed within *Dempster-Shafer's Theory of Evidence* and can be applied in the context of Web Information Retrieval tasks.

Keywords: Web Information Retrieval, best entry pages, formal model, Dempster-Shafer theory of evidence

1 Introduction

The aim of Information Retrieval (IR) systems is to assist users in satisfying their information needs by providing them with a ranked list of documents relevant to their queries. In the context of the Web, where hypermedia document authoring is encouraged, a document on a single topic may be distributed over a number of pages, which belong to a single site and are linked to each other. Users, however, consider it redundant for a Web IR system to return many relevant pages from the same site [3], since, in practice, they are able to easily reach all these pages by browsing, when given an appropriate entry page to the site [1].

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 optimal access, by browsing, to the relevant pages within the site. Web IR systems could then employ this measure in order to *focus* retrieval [5], by presenting to the user, not all the relevant pages from a site, but only the one considered to be its *Best Entry Page* (BEP).

We could also consider a more generalised view of a GEP measure, where being a GEP is determined not only in reference to the relevance of the pages it provides access to, but also with respect to other properties of these pages. In the Web, for instance, where there can exist thousands or even millions of pages relevant to a topic, other properties of Web pages, such as their *quality*, are taken into account when retrieving them. However, quality is a subjective notion

and can be interpreted in different ways. Here, we consider that the quality of a Web page can be captured by its *authority* [6] or its *utility* [10] on the topic. An authoritative page can be defined as a page that is not only topically relevant, but it is also a "trusted source of correct information" [13]. A *topical utility* (or hub [6]) page, on the other hand, provides a comprehensive list of links to authoritative pages on the topic.

In this paper, we adopt this generalised view and our aim is to quantify how good a page is as an entry page to a site, in reference to a particular property. This reflects how well a page enables the user to obtain access to other pages, belonging to the same site and exhibiting this property with respect to the topic. In Web IR, various approaches, such as spreading activation mechanisms [11] and site compression techniques [1] have been used in quantifying GEPs, by exploiting the structural relations among pages belonging to the same site. The underlying assumption is that a GEP measure should reflect not only the relevance or quality of a page, but also that of the pages within the site, that are accessible from it, by browsing.

To achieve our aim, we model Web pages and the properties they exhibit in ways that enable us to model a GEP measure of a page as a combination of evidence of the properties of the pages, which are structurally related to it and belong to the same site. This model is formally expressed within *Dempster-Shafer's (D-S) Theory of Evidence* [12]. D-S is a theory of uncertainty, that allows the explicit representation of combination of evidence, which allows us to model the aggregation of the properties of pages.

The paper is organised as follows. Section 2 contains a brief introduction to D-S theory. The model is described in Section 3 and some concluding remarks are provided in Section 4.

2 Dempster-Shafer's Theory of Evidence

In this section, we describe the main concepts of Dempster-Shafer's (D-S) Theory of Evidence [4, 12]. The combination of evidence, expressed by Dempster's combination rule, makes the use of D-S theory particularly attractive in this work, since it allows the expression of the aggregation of the properties of structurally related pages, which belong to the same site.

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): $m(\emptyset) = 0$ and $\sum_{A \subseteq \Theta} m(A) = 1$. $m(A)$ represents the belief exactly committed to 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 $m^\delta(\cdot)$ can be obtained from the original bpa m as follows: $m^\delta(A) = \delta m(A), \forall A \subseteq \Theta, A \neq \Theta$ and $m^\delta(\Theta) = \delta m(\Theta) + 1 - \delta$, with $0 \leq \delta \leq 1$. 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 as follows: $Bel(A) = \sum_{B \subseteq A} m(B)$. $Bel(A)$ 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 .

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 above equation is a normalisation factor that ensures that m is a bpa.

3 Description of the Model

Being a good entry page to a site with respect to a specific topic and in reference to a particular property, reflects how well the page enables the user to obtain access to pages within this site, which are exhibiting this property with respect to the topic. Our aim is to model a measure of *topical GEP* of a page, as a combination of evidence of the properties of the pages which are structurally related to it and belong to the same site.

First, we describe the properties exhibited by Web pages (Section 3.1) and define a frame of discernment based on these properties (Section 3.2). Web pages are represented as bodies of evidence within the defined frame of discernment (Section 3.3). The aggregation of the bodies of evidence, corresponding to pages which are structurally related to a particular page, allows us to model a measure of the topical GEP of a page (Section 3.4). Finally we extend the model, by presenting a refinement of the frame of discernment (Section 3.5).

3.1 Properties of Objects

We consider that each Web page exhibits a number of properties. These are either determined with respect to a particular topic (*topic-dependent properties*) or are considered to be intrinsic attributes of the pages (*topic-independent properties*). In this work, we concentrate on the topic-dependent properties of topical relevance, topical authority and topical utility, on the one hand, and the topic-independent properties of authority and utility, on the other, which reflect the quality of a Web page irrespective of any specific topic.

The authority of a Web page is usually determined by *link analysis ranking* algorithms, which view the Web's link structure as a network of recommendations¹ between pages [13]. When a page is pointed by other quality pages, it is considered to be recommended by them and therefore regarded as an authority. When a link analysis ranking algorithm takes into account the whole of the Web's link structure [9], a topic-independent measure of a page's authority is determined. This can then be combined with a topical relevance measure, in order to estimate the page's topical authority. In this case, the topical authority property of a page is directly related both to its topical relevance property, and to its topic-independent authority property.

Link analysis ranking algorithms can also be applied to a sub-network of the Web's link structure, generated by the links between the top ranked topical relevant pages, already retrieved by a Web IR system, and their immediate neighbours (forming the *base set* of pages). In this case, the measure of the page's authority, as this is determined by the algorithm, becomes topic-dependent and the relation among the properties indirect. HITS algorithm [6] and its extensions [2] adopt this approach in quantifying the topical authority of a page.

The same applies to the topical utility of a Web page, which can be quantified, either as a direct combination of its topical relevance with a topic-independent measure of its utility [10], or by employing an appropriate link analysis ranking algorithm [6] on the network of links connecting the pages in the base set.

To capture these relations and dependencies among the properties of a Web page, we introduce the notion of *composite* properties to refer to the ones that are related to the more *elementary* properties of a page. In our case, the topical authority and the topical utility of a page constitute its composite properties. These are related to the elementary properties of topical relevance and topic-independent authority of the page and to topical relevance and topic-independent utility of the page, respectively.

3.2 Frame of Discernment

To define a frame of discernment based on the above properties, we consider $\mathbb{E} = \{e_1, \dots, e_E\}$ to be the set of elementary properties and $\mathbb{C} = \{c_1, \dots, c_C\}$ 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

¹ This network takes into account only inter-domain links, since the underlying assumption is that they are the ones conveying endorsement [6].

exclusive propositions, which are 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 $\theta_{b_1 b_2 \dots b_n}$, where $b_1 b_2 \dots b_n$ is an n -bit binary number, such that $\theta_{b_1 b_2 \dots b_n}$ corresponds to the proposition “ $x_1 \wedge x_2 \wedge \dots \wedge x_n$ ”, 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 topical relevance R exhibited by the page, its topic-independent authority A and topic-independent utility U , we define $\mathbb{E} = \{R, A, U\}$. Each element $\theta_{b_1 b_2 \dots b_n} \in \Theta$ corresponds to the property $\theta_{b_1 b_2 \dots b_n}$ exhibited by a Web page. For instance, θ_{100} corresponds to $\{R \wedge \neg A \wedge \neg U\}$, reflecting that the page is exhibiting topical relevance, but not authority or utility. 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 topical relevance as this is exhibited by a page, without specifying whether the authority or utility of the page have been considered. Consequently, θ_{100} corresponds to the topical relevance as this is defined in classical IR, where authority or utility are not considered. In essence, we consider that a page can exhibit any of the properties defined in Θ or any of the elementary or composite properties, which were described in the previous section, and can be expressed in terms of the propositions in Θ .

3.3 Representation of Objects

In our framework, each Web page is referred to as an *object* and is represented by a body of evidence defined in Θ , through a set of focal elements for which there is positive evidence. Therefore, the focal elements can be used as propositions modelling the properties exhibited by the objects. Every elementary property $e_i \in \mathbb{E}$ exhibited by an object, meaning that 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 to exhibit the following properties: $p_1 = \{R\}$, $p_2 = \{A\}$, $p_3 = \{U\}$ 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 assume that object o exhibits property $p_5 = \{R \wedge \neg A \wedge \neg U\}$, then $p_5 = \theta_{100}$ (Figure 1).

A bpa m represents the uncertainty associated to a property and $m(p)$ corresponds to the degree to which an object exhibits property p . The value of $m(p)$ is estimated by employing an appropriate Web IR approach. For instance, if p corresponds to the topical authority of an object, $m(p)$ could be estimated using HITS algorithm [6] or one of its modifications [2]. However, we are not interested in which approach has been actually employed, just in that $m(p)$ has been estimated. The higher the $m(p)$, the more the object exhibits this property, whereas $m(p) = 0$ means that there is no evidence that the object exhibits property p .

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

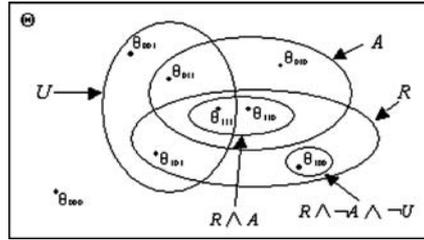


Fig. 1. Example of an object in the frame of discernment

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 the properties exhibited by an object. 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 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 overall relevance exhibited by the object.

3.4 Object Aggregation

To model a measure of the topical GEP of a page, we consider the combination of evidence of the properties of the pages which belong to the same site and are accessible from it. However, there is a tendency of users and it is more intuitive for them to browse down from starting points [7]. Therefore, they consider a GEP as one that enables them to access pages that are deeper in the hierarchy of the site. In this work, we concentrate on users following hierarchical down links² and consider only this kind of structural relation between Web pages.

In our framework, 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*, which are 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 2a). Page 3 is then represented as aggregate object a_3 derived from the aggregation of o_1 , o_2 and o_3 (Figure 2b). The aggregate object a_3 corresponds to the same Web page as object o_3 . We use Dempster's combination rule to compute the body of evidence of the aggregate object a_3 : $m_{a_3} = m_1 \oplus m_2 \oplus m_3$. Uncommitted belief is also assigned to the aggregate object.

² Hierarchical down links are intra-domain Web links whose source is higher in the directory path than their target.

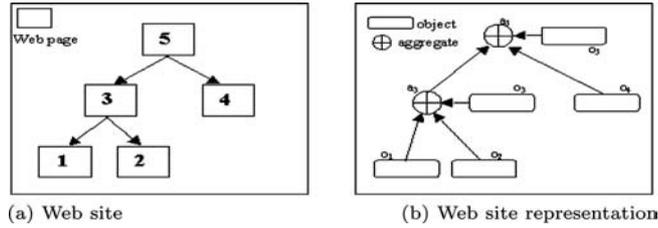


Fig. 2. Web Site

Table 1. Aggregation process

Step 1 of the aggregation							
object o_1	m_{o_1}	object o_2	m_{o_2}	object o_3	m_{o_3}	aggregate a_3	m_{a_3}
R	0.8	R	0.6	Θ	1	R	0.92
Θ	0.2	Θ	0.4			Θ	0.08
Fading factor - Step 2 of the aggregation							
aggregate a_3	$m_{a_3}^{\beta_3}$	object o_4	m_{o_4}	object o_5	m_{o_5}	aggregate a_5	m_{a_5}
R	0.46	Θ	1	Θ	1	R	0.46
Θ	0.54					Θ	0.54

The aggregation process is applied to the whole site starting with the pages deepest in the hierarchy, where no aggregation is performed. At the first step, we move up one level in the hierarchy to the page which has links to these pages and compute the body of evidence of the aggregate object (a_3 in our example).

In our example, consider that pages 1 and 2 are retrieved, whereas pages 3, 4 and 5 are not. We consider the objects o_1 and o_2 , corresponding to the retrieved pages, to exhibit the property $\{R\}$, with belief 0.8 and 0.6 respectively, whereas there is no evidence regarding any other properties for these or for the rest of the objects. The uncommitted belief is then equal to: $m_1(\Theta) = 0.2$, $m_2(\Theta) = 0.4$, and $m_3(\Theta) = m_4(\Theta) = m_5(\Theta) = 1$.

In the first step of the aggregation process (Table 1), Dempster’s combination rule yields the aggregate object a_3 , exhibiting property R with belief 0.92. Despite page 3 not being initially retrieved, it still can be considered to exhibit topical relevance, by reflecting that of its descendants. We consider the overall belief $Bel_a(R)$ (which is equal to $m_a(R)$ in this case) to reflect a measure of the topical GEP of a page in reference to property R . One can see that the belief in a property increases, when there is positive evidence for that property in more than one of the bodies of evidence being aggregated. Within this step of the aggregation, we could to model the contribution of each of the component objects forming an aggregate object, by using discounted bpas. The extent of each contribution could capture 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 contribution could be set to $\frac{1}{n}$, whereas that of object o itself could be set to 1.

Before moving one level up in the hierarchy to the next step of the aggregation, a *fading factor* [8] is applied to the aggregate objects already formed. As aggregate objects reflect information deeper in the hierarchy, the contribution of

this information should diminish as we move further up. This is modelled by a β_i -discounted bpa m^{β_i} , where β_i reflects the fading factor applied to aggregate object a_i . If we set $\beta_3 = 0.5$, then $m_{a_5} = 0.46$. By comparing the values of the topical GEP measure of the Web pages in reference to property R (depicted in bold in Table 1), we see that page 3 is considered the best entry page to the site.

3.5 Refinement of the Frame of Discernment

The refinement of a frame of discernment Θ (*coarse frame*) into a frame of discernment V (*refined frame*) is defined by splitting each element of Θ into a set of elements, that can be viewed as the latter representing more precise items of information than the former.

For instance, element $\theta_{100} \in \Theta$, reflects the proposition “the Web page exhibits topical relevance, but not authority or utility”. This could be split into k elements θ_{100-r_i} , each corresponding to “the Web page exhibits topical relevance, as this is determined by retrieval algorithm r_i , but not authority or utility”, where r_i with $i = 1, \dots, k$ corresponds to a ranking approach that can be applied in order to determine the topical relevance of an object. For instance, consider that θ_{100} is split into 2 elements of V , each more precise than θ_{100} : θ_{100-p} denoting “the Web page exhibits topical relevance as this is determined by the probabilistic model, but not authority or utility” and θ_{100-v} denoting “the Web page exhibits topical relevance as this is determined by the vector space model, but not authority or utility”.

The refinement is formally defined by $\omega : 2^\Theta \rightarrow 2^V$: (i) $\omega(p) \neq 0$ for all $p \in \Theta$, meaning that every element in Θ is split into elements in V ; (ii) $\omega(p) \cap \omega(p') = 0$ if $p \neq p'$ for all $p, p' \in \Theta$, meaning that two elements cannot be split into the same element, and (iii) $\bigcup_{p \in \Theta} \omega(p) = V$, meaning that the result of a refinement is a frame of discernment. For instance, refinement for element θ_{100} is expressed as: $\omega(\theta_{100}) = \{\theta_{100-p}, \theta_{100-v}\}$. The refinement function is also extended to a set $A \subseteq \Theta$: $\omega(A) = \bigcup_{p \in A} \omega(p)$, where $\omega(A)$ consists of all the elements in V that are obtained by splitting all the elements in A . Therefore, this refinement function links two frames of discernment, such that one is defined in terms of the other. If the properties exhibited by a Web page are modelled by the coarse frame, then the refinement function can give us a more detailed representation in terms of how these properties are actually determined.

Let Bel_Θ and Bel_V be the belief functions defined on Θ and V , respectively. These belief functions must satisfy the criteria that these two frames are *compatible*, which means that the two frames must agree on the information defined in them. Therefore, although refining a set means that more precise items of information are obtained, the union of these items carries the same information as the original set. The belief functions are compatible, if for given set A on the frame Θ , the following holds: $Bel_\Theta(A) = Bel_V(\omega(A))$.

4 Conclusions

We presented a model expressed within the theoretical framework of Dempster-Shafer’s theory of evidence, which quantifies a measure of how good a Web

page is as an entry page to a Web site. Web pages are represented in terms of topic-dependent and topic-independent properties. Structurally related pages belonging to the same site are combined using Dempster's combination rule, to produce an aggregate, which exhibits properties reflecting those of its structurally related components. We consider that the belief assigned to a particular property of an aggregate object corresponds to a measure of its topical GEP, in reference to that property.

In this work, we considered only intra-domain hierarchical down links to reflect the structural relations between Web pages. The next step is to consider other kinds of intra- and inter-domain structural relations. Currently, we are performing large-scale experiments to evaluate our approach using test collections of Web documents.

References

1. E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. Topic distillation with knowledge agents. In *Proceedings of 11th Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251. Gaitensburg, MD, 2002.
2. K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of ACM SIGIR*, pages 104–111, 1998.
3. 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.
4. A. P. Dempster. A generalization of bayesian inference. *Journal of Royal Statistical Society*, 30:205–447, 1968.
5. G. Kazai, M. Lalmas, and T. Roelleke. Focussed structured document retrieval. In *Proceedings of String Processing and Information Retrieval*, 2002.
6. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
7. 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, 2000.
8. 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.
9. 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.
10. 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.
11. V. Plachouras, I. Ounis, G. Amati, and C. J. van Rijsbergen. University of glasgow at the web track of trec 2002. In *Proceedings of 11th Text Retrieval Conference (TREC-2002)*, NIST Special Publication 500-251. Gaitensburg, MD, 2002.
12. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, NJ, 1976.
13. P. Tsaparas. *Link Analysis Ranking*. PhD thesis, Department of Computer Science, University of Toronto, 2004.