

Dempster-Shafer's Theory of Evidence applied to Structured Documents: modelling Uncertainty

Mounia Lalmas

Department of Computing Science
University of Glasgow
G12 8QQ Scotland
mounia@dcs.gla.ac.uk

Abstract

Documents often display a structure determined by the author, e.g., several chapters, each with several sub-chapters and so on. Taking into account the structure of a document allows the retrieval process to focus on those parts of the documents that are most relevant to an information need. Chiaramella et al advanced a model for indexing and retrieving structured documents. Their aim was to express the model within a framework based on formal logics with associated theories. They developed the logical formalism of the model. This paper adds to this model a theory of uncertainty, the Dempster-Shafer theory of evidence. It is shown that the theory provides a rule, the Dempster's combination rule, that allows the expression of the uncertainty with respect to parts of a document, and that is compatible with the logical model developed by Chiaramella et al.

1 Introduction

In traditional information retrieval (IR), a document is considered as an atomic entity that is indexed and retrieved as a whole by the system, and is presented to the user as a query result. Documents, then, constitute the basic information units on which IR systems are based. However, documents often display a structure determined by the author. For instance, a document may have several chapters, each with several sub-chapters and so on. With a structured document, the indexable, and consequently retrievable units, should be the document components instead of the document because, often, only parts of the document are relevant to an information need. Moreover, the indexing of a structured document must allow for the retrieval process to return aggregated components, e.g., a chapter, a set of chapters, or all chapters of the document, relevant to a query, instead of delivering only a reference to the whole document.

Chiaramella et al [CMF96] proposed a model for indexing and retrieving structured documents. Their aim was to express the

model within a framework based (i) on formal logics to capture the semantics of information and (ii) with associated theories of uncertainty capable of handling uncertainty inherent in the IR process and intrinsic to information. They developed the logical formalism of the model. This work adds to this model a *theory of uncertainty*.

A well-known theory of uncertainty used in IR is probability theory (see [Fuh92] for a survey of probabilistic IR models). Here, an alternative theory of uncertainty is used, the *Dempster-Shafer theory of evidence* [Sha76], because it is more flexible in modelling the IR process, but still theoretically sound. This has been discussed by several authors (see [Lal96a, SH93, TdSM93]). In particular, the theory provides a rule, *the Dempster's combination rule*, that allows the expression of the uncertainty with respect to aggregated components, and that is compatible with the logical model developed by Chiaramella et al.

The paper is organised as follows. Section 2 describes the logical model for structured documents developed by Chiaramella et al. Section 3 presents the Dempster-Shafer theory of evidence. The theory is then used to construct a model that takes into account the uncertainty that arises in IR to index (section 4) and retrieve (section 5) structured documents. The paper finishes with an example (section 6) and some thoughts for future work (section 7).

2 A logical model for structured documents

The requirements to index and retrieve structured documents have been studied in [KC95, Mec95], and were the bases of the model proposed by Chiaramella et al. The model is described in this section. This is a simplified version of the model, but sufficient for the purpose of this work. See [CMF96] for the complete description¹.

2.1 Structure

In [CMF96], the structure of a document corresponds to a tree whose nodes (*objects*) are the components of the document (e.g., chapters, sections, etc.) and whose edges represent the *composition* relationship (e.g., a chapter contains several

¹The reader should refer to that paper and also to [CK96] to obtain explanations for any of the modelling decisions made by the authors.

Permission to make digital/hard copies of all or part of this material for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copyright is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires specific permission and/or fee

SIGIR 97 Philadelphia PA, USA
Copyright 1997 ACM 0-89791-836-3/97/7..\$3.50

sections). The document's semantic content becomes defined by the *aggregation* of its components. The *root* object of the tree, which is unique for each document, embodies the whole document, and the *leaf* objects comprise the raw content of the document (i.e., a piece of text, an image, etc.). Any non-leaf object is referred to as a *composite* object (the root object included).

More formally, the structure of documents is defined by a set of objects O and a composition function formalised by the relation $\pi \subseteq O \times O$. The set O corresponds to the components constituting documents, and the composition function defines how the objects are organised. For any two objects $o_i, o_j \in O$, $(o_i, o_j) \in \pi$ means that o_j is a direct component of o_i , or that o_i is a parent of o_j . $\pi(o_i)$ denotes the set of direct component

objects of o_i , and is formally defined as $\pi(o_i) = \{o_j \in O \mid (o_i, o_j) \in \pi\}$. Chiamarella et al consider only hierarchical structure; that is, an object has at most one parent. A root object has no parent.

Consider the example of a structured document of Figure 1, which will be used throughout this paper. This document contains seven objects $o_1, o_2, o_3, o_4, o_5, o_6$ and o_7 . The object o_5 has two components, the objects o_1 and o_2 , hence $\pi(o_5) = \{o_1, o_2\}$. The objects o_1, o_2, o_3 , and o_4 are leaf objects. They correspond to pieces of texts, images, speeches. The objects o_5, o_6 and o_7 are composite objects. Composite objects do not contain information that is independent of their components; only the leaf objects contain information. The object o_7 is the root object.

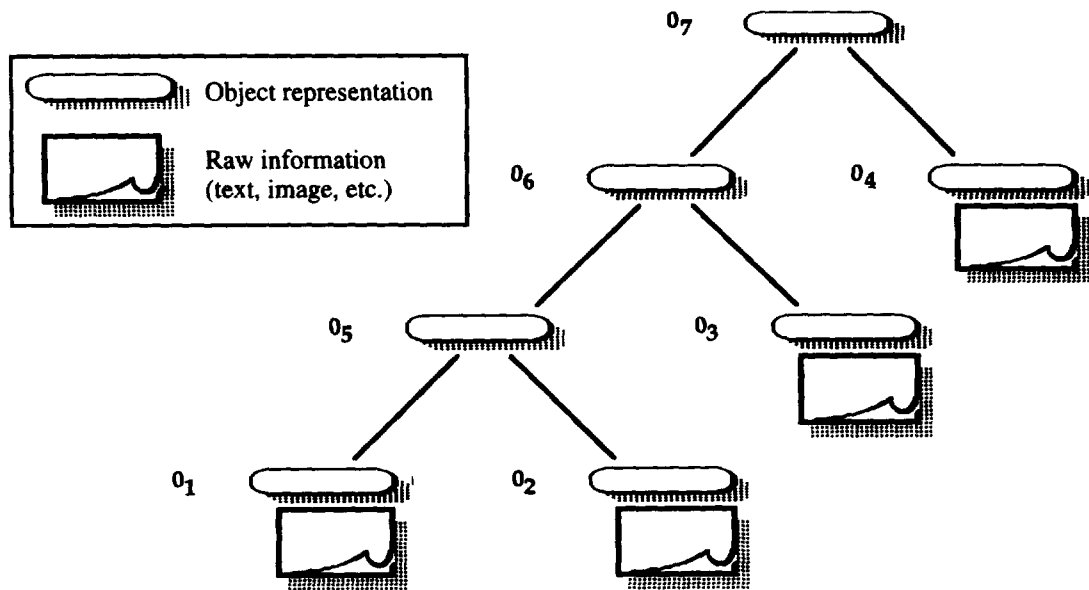


Figure 1: Example of a structured document

2.2 Semantic content

Each object has an associated semantic content defined within a given logic. This is represented by the function *semantic*: $O \rightarrow LI$, where LI is a language (a set of formulae) defined by the logic. The semantic content of an object is determined by whether the object is a leaf or composite. The semantic content of a leaf object is obtained from the indexing process, which is not discussed in this paper. It is assumed determined. The semantic content of a composite object is determined from its component objects. More precisely, for any composite object o , *semantic*(o) depends on *semantic*(o_i) for all o_i in $\pi(o)$. This is referred to as an *aggregation*, and is formally expressed as follows:

$$semantic(o) = \bigoplus_{o_i \in \pi(o)} semantic(o_i)$$

where \bigoplus is a semantic aggregation operator. Chiamarella et al do not specify which aggregation operator to use, but they impose on it the *dependency constraint*:

$$semantic(o) \rightarrow semantic(o_i)$$

meaning that the formula representing the semantic content of an object *semantically implies* that of each of its component objects. The implication \rightarrow is a logical semantic operator² defined on LI . For example, assume that LI is propositional logic and o_1 and o_2 are two objects such that *semantic*(o_1) = *wine* and *semantic*(o_2) = *grape*, where the propositions *wine* and *grape* represent indexing concepts. If $\pi(o) = \{o_1, o_2\}$, then *semantic*(o) should be defined such that it semantically implies both *wine* and *grape*. One possible aggregation would be to define *semantic*(o) as *wine* \wedge *grape*, where \wedge is the classical logical conjunction, and \rightarrow is the material implication.

2.3 The retrieval strategy: fetch and browse

Taking into account the structure of a document allows the retrieval process to focus on those parts of the documents that are most relevant to an information need. Take for instance the document pictured in Figure 1: retrieving o_1 , o_3 , or o_7 means, respectively, that only the leaf object (e.g., one chapter of a book), the objects o_1 and o_2 (e.g., two chapters of a book), or the whole document (e.g., all chapters of a

²The implication is not necessarily part of the logic for LI .

book) is/are relevant to the information need. Only the retrieved objects (e.g., o_1 , o_2 , or o_7) are displayed to the user, and then constitute access points from where the user can decide to browse the structure if needed.

Several related objects may be retrieved as answers to a query; for example, the objects o_1 and o_7 . Chiaramella et al decided that, in this case, the response should contain the deepest component in the structure, which is then displayed to the user; in this case, the object o_1 . This choice corresponds to the most *specific* component of the document that satisfies the information need, but which remains *exhaustive* to the information need³.

To implement this retrieval strategy, Chiaramella et al use part of the logical approach to IR developed by Nie [Nie90], which is re-expressed in terms of objects vs query. An information need is represented by a query expressed as a formula q of the language LI . The relevance of an object o to this information need is evaluated as follows:

$$R(o,q)=F[P(\text{semantic}(o) \rightarrow q), P'(q \rightarrow \text{semantic}(o))]$$

The matching between o and q is determined by the evaluation of two uncertain implications $P(\text{semantic}(o) \rightarrow q)$ and $P'(q \rightarrow \text{semantic}(o))$, which represent measures, respectively, of how exhaustive and how specific is the object o to the information need. P and P' are two measures of uncertainty (e.g., probability functions). F is a function that combines these measures. Chiaramella et al developed only the logical formalism of the model, and hence ignore the uncertainty issue (P and P'). Their retrieval strategy uses only the implications $\text{semantic}(o) \rightarrow q$ and $q \rightarrow \text{semantic}(o)$, and the function F . The two implications express whether the object o is, respectively, exhaustive and specific, to the information need. The function F is defined as a fetch and browse strategy that allows the retrieval of the most specific objects. With fetch, a pre-selection is done made on the criteria of exhaustivity ($\text{semantic}(o) \rightarrow q$). This is the classical retrieval process on unstructured documents. With browse, the structure of the pre-selected documents is investigated. This is done by browsing within the structure, using both implications as described next.

2.3.1 Fetch Phase

All root objects o such that $\text{semantic}(o) \rightarrow q$ are fetched. These objects are exhaustive to the query. All other objects are not exhaustive to the query, and neither are their components (if any). The latter is due to the dependency constraint.

2.3.2 Browse Phase

Given a fetched object o , two cases arise:

1 - $q \rightarrow \text{semantic}(o)$: the formula representing the semantic content of the object o and the query are logically equivalent. This is the ideal case, and the object o is an optimal response, being both

exhaustive and specific, to the query. The object o is retrieved.

2 - not ($q \rightarrow \text{semantic}(o)$): the object o is relevant to the query considering the criteria of exhaustivity, but not specificity. One of its components may be a more specific response to the query. Let o' in $\pi(o)$:

- a- $\text{semantic}(o') \rightarrow q$: the object o' is exhaustive to the query. The object is therefore fetched. It may still be possible to find a more specific object a level deeper in the structure, so the same browse phase is applied to the components of o' .
- b - not ($\text{semantic}(o') \rightarrow q$): the object o' does not satisfy anymore the exhaustivity criteria; we have been too deep in the structure of the document. The optimal response is either o , or another component of o .

The above strategy works because of the dependency constraint. Without it, it would be impossible to assert the possible relevance of related components as above.

2.4 Uncertainty in IR

Due to the complex nature of information, determining the semantic content of a document is a highly uncertain task, because it often relies on incomplete evidence (e.g., terms that appear in a text document, pixels of an image, etc.). The notions of evidence and uncertainty are not specific to IR, and frameworks have been developed to formally express them. These frameworks are referred to as theories of uncertainty [Saf87]. The one adopted in this work is the *Dempster-Shafer theory of evidence* (D-S) [Sha76]. It cannot yet be claimed that this is the best theory to use. Further investigations are necessary to either prove or refute this claim. It is however my belief that the D-S theory of evidence is both promising and sufficient for the modelling of uncertainty inherent to structured documents. The reasons are manifolds: (i) expressive IR models based on the D-S framework have been proposed (see [SH93, TdSM93]), which can be used to represent leaf objects; (ii) the theory is particularly appropriate in capturing the uncertainty associated with composite objects because it provides an aggregation operator, the *Dempster's combination rule*, that allows the expression of the uncertainty with respect to aggregated components; and (iii) the properties of the aggregation operator are compatible with those defined by the logical model developed by Chiaramella et al. This paper concentrates on points (ii) and (iii).

3 Dempster-Shafer's Theory of Evidence

The D-S framework is based on the view whereby propositions are represented as subsets of a given set W , referred to as a *frame of discernment*. The propositions of interest are in a one-to-one correspondence with the subsets of W . The correspondence between propositions and subsets is useful because it translates the set-theoretic notions of intersection, union and complementation into the logical notion of conjunction, disjunction and negation; thus yielding the semantics of classical logic: the frame of

³A document is specific to a query if all its information content concerns the query. A document is exhaustive to the query if the document contains all the required information.

discernment W is a set of *possible worlds* and a proposition p is equivalent to the set of possible worlds where p is true⁴.

Evidence can be associated to each proposition (subset) to express the uncertainty (*belief*) that it has been observed or *discerned*. The evidence is usually computed based on a density function m called a *basic probability assignment* (bpa):

$$m(\perp) = 0 \text{ and } \sum m(p) = 1$$

$m(p)$ represents the belief exactly committed to the proposition p . If $m(p) > 0$, then p is said to be discerned by the bpa m , and is called a *focal element*.

A bpa m defines a *body of evidence*, from where a *belief function* Bel is defined:

$$Bel(q) = \sum_{p \rightarrow q} m(p)$$

$Bel(q)$ is the total belief committed to q , that is, the total positive effect the body of evidence has on q being true, and its value comes from the bpa of all discerned propositions of the frame that imply q .

The D-S theory has an operation, the Dempster's rule of combination that aggregates two (or more) bodies of evidence defined within the same frame of discernment into one body of evidence. Let m_1 and m_2 be two bpas defined in W . The new body of evidence is defined by the bpa m :

$$m(r) = m_1 \oplus m_2(r) = \frac{\sum_{p \wedge q = r} m_1(p) * m_2(q)}{\sum_{p \wedge q \neq \perp} m_1(p) * m_2(q)}$$

In words, the Dempster's combination rule 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 bpa m takes into account the bpas associated to the propositions discerned by m_1 and m_2 that yield the propositions discerned by m . The denominator is a normalisation factor that ensures that m is a bpa (without it, a belief may be associated to \perp which goes against the definition of a bpa⁵).

4 Indexing structured documents

⁴The set of possible worlds where p is true is denoted $w(p)$ where $w(p) \subseteq W$. The following properties hold between propositions p and q and their respective $w(p)$ and $w(q)$:

(i) $p \rightarrow q$ is equivalent to $w(p) \subseteq w(q)$.

where \rightarrow is the material implication as defined in classical logic. The above means that $p \rightarrow q$ is true if q is true in all possible worlds where p is true.

(ii) $p \wedge q$ is equivalent to $w(p) \cap w(q)$.

(iii) $p \vee q$ is equivalent to $w(p) \cup w(q)$.

⁵The aggregation is not defined when no proposition is discerned by the two bodies of evidence (see [SHA76]).

The logic defining the language LI was not specified by Chiaramella et al. In this paper, classical logic is adopted as the logic for LI . That is, formulae of LI are propositions. Furthermore, the semantic implication used in the dependency constraint is defined as the material implication of classical logic. I am aware that classical logic is not the best logic to capture semantics (see [vRL96]) in IR. However, a correspondence exists between classical logic and the D-S framework, which can be exploited as a first step in adding the representation of uncertainty to the logical model for structured documents.

The semantic content of each object o is represented by a body of evidence: the focal elements define the propositions representing the semantic content⁶ of the object and a bpa m expresses the uncertainty attached to these propositions. For any discerned p , $m(p)$ is the belief in p being a good description of the semantic content of the object o . This generalises the model initially proposed in [CMF96], where for each object o , one focal element would be defined, let us say p , and $m(p) = 1$, meaning that the proposition p was certain.

The computation of bodies of evidence depends on whether objects are leaf or composite objects. It is assumed that these have been defined for leaf objects of any type (textual, image, etc.) from the indexing process. How to obtain the bodies of evidence of composite objects is discussed next.

4.1 Composite objects

The Dempster's combination rule is used to compute the body of evidence of composite objects. Let $o \in O$, and let $o_1, \dots, o_k \in \pi(o)$, where each direct component o_i has an associated bpa m_i , for $i = 1, \dots, k$. The bpa m associated to the object o is defined as $m = m_1 \oplus \dots \oplus m_k$.

The aggregation operator in Chiaramella et al's model has been replaced by that defined by the Dempster's combination rule. The aggregation operator used in Chiaramella et al's model possessed properties, which have their counterpart with the aggregation as defined by the Dempster's combination rule. The properties are that the aggregation operator is commutative, associative, has a neutral element, and is idempotent. It is shown in [LC97] that the aggregation operator as defined by the Dempster's combination rule satisfies the first three properties, but not the last one. Idempotence means that it should be that $m \oplus m = m$. Suppose that the proposition $a = \text{information retrieval}$ is based on the evidence that the two terms *information* and *retrieval* appear in the document. Suppose that the proposition $b = \text{information management}$ is based on the evidence that *information* and *management* appear in the document. Having this evidence twice, it seems intuitive that some belief should be transferred to some proposition $c = \text{information}$, meaning that the composite object is about information related topics. Therefore, the fact that the Dempster's combination rule does not satisfy the idempotence property is in fact more intuitive than if it did

⁶In contrast to the model developed by Chiaramella et al, the semantic content of an object is represented by a set of propositions, and not a single proposition as given by the function *semantic*.

accept the property. This discussed in details in [LC97]. This paper only concentrates on the dependency constraint.

4.2 Dependency constraint

The dependency constraint states that the proposition representing the semantic content of an object semantically implies that of each of its direct components. The constraint is also satisfied but in a slightly different way because a set of propositions is used to represent the semantic content of an object. By definition of the Dempster's combination rule, for any proposition r discerned by m , for each $o_i \in \pi(o)$, there exists p_i discerned by m_i such that $p_1 \wedge \dots \wedge p_k = r$. Therefore, since classical logic is used, for any such p_i , $r \rightarrow p_i$. Thus, the propositions representing the semantic content of a composite object imply those of its components.

The relation between $m(r)$, and $m_i(p_i)$ for $i=1, \dots, k$ is however not as straightforward. No relation exists between

m_1		m_2		m_3	
$a=\{1,2\}$	0.3	$d=\{1\}$	0.6	$d=\{1\}$	0.39
$b=\{3\}$	0.3	$e=\{2,4\}$	0.4	$f=\{2\}$	0.26
$c=\{3,4\}$	0.4			$g=\{4\}$	0.34

Table 1: Bodies of evidence of objects o_1 , o_2 and o_3

For example, suppose that the aggregation is defined as \wedge the classical logical conjunction⁷, and \rightarrow as the material implication. Let $a = \text{wine}$, $b = \text{colour}$, $c = \text{white}$, $d = \text{Chardonnay}$ and $e = \text{grape}$, so $f = \text{grape} \wedge \text{wine}$ and $g = \text{white} \wedge \text{grape}$. It can easily be proved that the above implications hold.

It can also be observed that no systematic relationship exists between the bpa of the propositions discerned by m_3 and those discerned by m_1 and m_2 . For example, $m_3(g)$ is smaller than $m_1(c)$ and $m_2(e)$, whereas $m_3(d)$ is higher than $m_1(a)$ and smaller than $m_2(d)$.

4.4 Discussion

In the previous example, no proposition discerned by m_3 implies the proposition b discerned by m_1 . This arises because b is incompatible with the two propositions discerned by m_2 ⁸: m_2 does not support the proposition b , nor any proposition that implies it. This behaviour is not specific to the way the Dempster's combination rule is used, and has been well acknowledged⁹. This behaviour is,

⁷This is in fact what (partly) happens with the use of the Dempster's combination rule as the aggregation operator. The rule also combines beliefs.

⁸Both $w(b) \cap w(d) = \emptyset$ and $w(b) \cap w(e) = \emptyset$ hold.

⁹Assume that two medical experts believe that a patient has, respectively, disease A and disease B, where A and B are incompatible with each other. Each expert's belief is modelled by a body of evidence. When combining the two bodies of evidence, any information about diseases A and B is discarded. However, some argue that this information should

the different values because $m(r)$ does not depend only on $m_i(p_i)$ for $i=1, \dots, k$. This is illustrated in the next section.

4.3 Example

Take the document represented by Figure 1. Only the objects o_1 , o_2 and o_3 are considered. Suppose that the set of possible worlds is $W = \{1, 2, 3, 4\}$ which constitutes the frame of discernment, and that the bpas of the three leaf objects are defined in Table 1 (the propositions associated to each set of possible worlds are explicitly represented). m_3 is defined in terms of m_1 and m_2 , and is also shown in Table 1. a, b, c, d, e, f, g, h correspond to propositions of LI , and f and g are the aggregations of, respectively, a and e , and c and e .

It can be shown that the followings hold: $d \rightarrow d, a$, $f \rightarrow a, e$ and $g \rightarrow c, e$, where d, f and g are the propositions that constitute the core of o_3 . Any proposition discerned by m_3 implies two propositions discerned, respectively, by m_1 and m_2 ; thus satisfying the dependency constraint.

however, a disadvantage if the fetch and browse approach is used. With the representation of composite objects, given that a query is represented by a proposition q (more about this in section 5.1), it may be the case that no proposition discerned in m_3 implies q , but a proposition deeper in the structure implies q (this is the case for $q = b$).

This behaviour can be overcome by assigning a belief to the frame of discernment W , and this for each leaf object. That is, the frame of discernment constitutes a focal element. The proposition that corresponds to the frame is the true proposition T. Let m be bpa associated to a leaf object. This means that $m(T) > 0$. As a result, any proposition discerned by a leaf object remains discerned when the object is combined with other objects to form a composite object¹⁰. Moreover, the frame of discernment remains a focal element of the composite object¹¹. Consequently, any proposition discerned by a leaf object remains as such at each level of the structure. The fetch and browse strategy can then be used (see section 5.3).

By having the frame itself as a focal element, the case where some beliefs remain uncommitted can be captured. In the

still be retained, for example, by assigning small belief values to each disease.

¹⁰This is because for any p , we have $w(p) \cap W = w(p)$.

¹¹For example, let o_1 and o_2 be objects with respective bpas m_1 and m_2 . Suppose that these are the direct components of the object o with body of evidence m . If $m_1(W) > 0$ and $m_2(W) > 0$, then $m(W) > 0$ because W is a focal element with respect to m_1 and m_2 , and $W \cap W = W$, so W is also a focal element with respect to m .

context of IR, uncommitted beliefs may be used to represent the uncertainty associated to the indexing of an object. Total ignorance, i.e., where nothing is known about how to index the object, can also be represented: setting $m(T)=1$.

A different way to overcome the behaviour of the Dempster's combination is to base the logic of the model on a non-classical logic. The behaviour of the rule may be different in this case. For example, Mirlog [MS96] has been advanced as an appropriate logic to capture semantics and relevance in the context of IR, and when combined with the D-S framework may lead to more intuitive behaviours. This will be the object of future research.

5 Retrieving structured documents

The retrieval strategy that takes into account uncertainty is described. First, the relevance of an object to a query is defined where objects are represented as bodies of evidence. Second, the criteria of exhaustivity and specificity are expressed, upon which the fetch and browse strategy is based. Third, these are used to determine the most specific components of a document.

5.1 Relevance of an object to an information need

An information need is represented by a query formula q . Let o_i be an object with bpa m_i . The relevance of this object to the query is expressed by:

$$Bel_i(q) = \sum_{p \rightarrow q} m_i(p)$$

$Bel_i(q)$ captures relevance because it is based on all propositions defining the semantic content of the object o_i that imply the query formula. It also takes into account the beliefs associated to these propositions; the higher their beliefs, the higher the relevance. Also, the greater their number, the higher the relevance.

5.2 Exhaustivity and specificity

In logical IR models, the implication $d \rightarrow q$ (where d is the document formula) has been acknowledged to capture the document's exhaustivity to the query [Nie90, CC92]. The quantity $Bel_i(q)$ if not null indicates that the object o_i is exhaustive to the query. This is because $Bel_i(q) > 0$ means that there exists a proposition p discerned by m_i that implies the query ($p \rightarrow q$). Furthermore, $m_i(p)$ can be viewed as a degree of exhaustivity¹², which depends on the uncertainty associated with p indexing the object.

The higher $Bel_i(q)$, the more specific is the object o_i to the query. The reason is that if all the propositions discerned by m_i imply the query, then all the object's semantic content concerns the query, and vice versa. In the former case, $Bel_i(q)$ will tend to be high, whereas in the latter case, it will

¹²In [Lal96a], a degree of exhaustivity was proposed to capture partial relevance: we may not have the document implying the query, but we still have part of the document that concerns the query. This could be captured by using a plausibility function instead of a belief function (see [Sha76]).

tend to be low. Of course, this also depends on the uncertainty associated with each proposition. For example, two objects may have the same proposition implying the query, but the belief associated to the proposition is different in the two objects. The object with greater belief associated with the proposition can be viewed as more specific to the query than the other object. Note that it will rarely obtain $Bel_i(q)=1$ because a belief is assigned to the whole frame, which in many cases will not imply the query.

In Chiamarella et al's model, the specificity of an object o_i was expressed by evaluating the inverse implication $q \rightarrow semantic(o_i)$ (see section 2). This cannot be used anymore since an object is represented by a set of propositions¹³. Some of these may be implied by the query formula, but nothing can be concluded from this since the object representation will usually involve other propositions.

To conclude, any object o_i such that $Bel_i(q) > 0$ is exhaustive to the query, and the higher $Bel_i(q)$, the most specific is this object to the information need.

5.3 Retrieval strategy: Fetch and Browse

The fetch and browse strategy is discussed where objects are represented by bodies of evidences, some of which are computed in terms of others via the Dempster's combination rule.

5.3.1 Fetch Phase

Documents are first retrieved based on the exhaustivity criteria: any document whose root object o is such that $Bel(o) > 0$. Let $Fetch[q]$ be the set of such root objects. Any root object not in $Fetch[q]$ is not exhaustive to the query, and hence neither of its components. This is due to the dependency constraint discussed in section 4.2. This strategy can only be used if a belief is associated to the whole frame of discernment, and this for each leaf object (see section 4.4).

5.3.2 Browsing Phase

The objects obtained in the set $Fetch[q]$ may not be those most specific to the query q . Let o be in $Fetch[q]$. All the objects that constitute the document with root o must be examined.

The term *branch* is used to refer to the related objects, starting from the root object, and ending with a leaf object, all organised along one "line" of the document structure¹⁴. For example, the document represented in Figure 1 has four branches: $\{o_1, o_5, o_6, o_7\}$, $\{o_2, o_5, o_6, o_7\}$, $\{o_3, o_6, o_7\}$ and $\{o_4, o_7\}$.

¹³Using the reverse implication, for an object to be specific to the query, all the object propositions must be implied by the query formula. The commonality number function (see [Sha76]) may be used to express this.

¹⁴Formally, a branch is a set of objects $b \subseteq O$ that are related as follows: (i) for any $o, o' \in b$, either $o \in \pi(o')$ or $o' \in \pi(o)$, (ii) exactly one object in b is a root, and (iii) exactly one object in b is a leaf.

In the browse phase, all branches must be examined to define the object, in each, with the higher relevance (as computed by the belief functions). Furthermore, the relevance of all objects in a branch must be computed (except for the case described below) because no systematic relationship exists between the beliefs associated to a composite object and its direct components (see section 4.3). Let $o_i \in \pi(o)$. If $Bel_i(q) \leq Bel(q)$, the object o_i is less specific to the query than the object o is, and hence must not be retrieved. If $Bel_i(q) > Bel(q)$, the object o_i is more specific to the query than the object o is, and hence may be retrieved instead of o . However, in both cases, a component of o_i may be more specific to the information need than are the objects o or o_i . Consequently, a decrease or an increase of belief (i.e., relevance) along a branch is not a stopping condition for browsing. Some more specific components may still be found deeper in the structure.

There is however a stopping condition for browsing. As soon as an object o_i is reached such that $Bel_i(q) = 0$, then there is no need to continue along this branch because the object o_i is not exhaustive to the query, and neither are its components (if any). Due to the dependency constraint, all the components will have a null relevance to the query.

5.3.3 Discussion

With the above strategy, two objects may be related (i.e., belong to the same branch). For example, in the branches $\{o_1, o_3, o_6, o_7\}$ and $\{o_3, o_6, o_7\}$, respectively, the object o_1 and o_6 may have the higher relevance. The approach adopted by Chiaramella et al is that the object deeper in the structure is retrieved and displayed to the user. In this paper, the object with the higher relevance is retrieved since it corresponds to the most specific to the information need. If the objects o_1 and o_6 have the same relevance, the object deeper in the structure is retrieved. I would like to add that it is not yet clear that this is the best approach which may mislead the user who then may not browse up the structure to consult other parts of the document. Some user studies are necessary at this stage to determine whether this method is best with respect to the user.

The browse phase is different from that proposed by Chiaramella et al. In this paper, a degree of specificity is used, and not only whether an object is specific or not to browse along the structure of a document. In Chiaramella et al's browse phase, as soon as an object is reached that is exhaustive and specific to an information need, the browsing ceases, which is as soon as an object o is reached such that $semantic(o) \rightarrow q$ and $q \rightarrow semantic(o)$ hold. However, for any object $o_i \in \pi(o)$ that remains exhaustive to the query, due to the dependency constraint (i.e., $semantic(o) \rightarrow semantic(o_i)$), then $q \rightarrow semantic(o_i)$ holds. This means that the object o_i is itself specific to the information need. The browse phase stops in Chiaramella et al's case because the two objects o and o_i are both specific to the query, and there is no notion of one being more specific than the other. The browse phase in this paper takes this into account.

6 Example

Suppose that with the structured document of Figure 1, the leaf objects o_1, o_2, o_3 , and o_4 are defined by the bodies of evidence in Table 2 (beliefs are assigned to the entire frame in each case). By using the Dempster's combination rule, the bodies of evidence shown in Table 3 are obtained for the composite objects. The relevance of different queries with respect to each object is evaluated. The values are represented in the Table 4. With the strategy described in the previous section, the retrieved objects are shown in Table 5.

Some of the results are discussed. Take the query a , for which the three leaf objects o_1, o_2 , and o_3 are highly relevant to the query (high beliefs). It is appropriate that the composite of these three leaf objects, the object o_6 , is retrieved. Consider now the query c . The only exhaustive object to this query is o_2 , which is effectively retrieved. Consider the query k in Tables 1 and 2. Correctly, the object o_3 is retrieved, being the most relevant since both the objects o_1 , and o_2 are relevant to the information need. Take query e . Two objects are retrieved, o_1 and o_4 , which are the only objects that are exhaustive to the query.

m_1		m_2		m_3		m_4	
$a=\{1,2\}$	0.3	$c=\{1\}$	0.6	$a=\{1,2\}$	0.4	$a=\{1,2\}$	0.2
$b=\{3\}$	0.3	$T=\{1,2,3,4\}$	0.4	$d=\{2,3\}$	0.4	$e=\{3,4\}$	0.5
$T=\{1,2,3,4\}$	0.4			$T=\{1,2,3,4\}$	0.2	$T=\{1,2,3,4\}$	0.3

Table 2: Body of evidence associated to the leaf objects

m_5		m_6		m_7	
$c=\{1\}$	0.51	$c=\{1\}$	0.32	$c=\{1\}$	0.27
$b=\{3\}$	0.14	$b=\{3\}$	0.1	$f=\{2\}$	0.26
$a=\{1,2\}$	0.14	$f=\{2\}$	0.28	$b=\{3\}$	0.2
$T=\{1,2,3,4\}$	0.19	$a=\{1,2\}$	0.17	$a=\{1,2\}$	0.16
		$d=\{2,3\}$	0.08	$d=\{2,3\}$	0.04
		$T=\{1,2,3,4\}$	0.05	$e=\{3,4\}$	0.04
				$T=\{1,2,3,4\}$	0.03

Table 3: Body of evidence associated to the composite objects

Query	Bel_1	Bel_2	Bel_3	Bel_4	Bel_5	Bel_6	Bel_7
$a=\{1,2\}$	0.3	0.6	0.4	0.2	0.65	0.77	0.69
$g=\{1,2,3\}$	0.6	0.6	0.8	0.2	0.81	0.95	0.93
$h=\{1,2,4\}$	0.3	0.6	0.4	0.2	0.65	0.77	0.69
$i=\{1,4\}$	0	0.6	0	0	0.51	0.32	0.27
$c=\{1\}$	0	0.6	0	0	0.51	0.32	0.27
$b=\{3\}$	0.3	0	0	0	0.14	0.1	0.2
$j=\{2,3,4\}$	0.3	0	0.4	0.5	0.14	0.46	0.5
$T=\{1,2,3,4\}$	1	1	1	1	1	1	1
$e=\{3,4\}$	0.3	0	0	0.5	0.14	0.1	0.24
$d=\{2,3\}$	0.3	0	0.4	0	0.14	0.46	0.5
$k=\{1,3,4\}$	0.3	0.6	0	0.5	0.65	0.42	0.51

Table 4: Relevance of all objects to queries

Query	Retrieved objects
$a=\{1,2\}$	o_6
$g=\{1,2,3\}$	o_6
$h=\{1,2,4\}$	o_6
$i=\{1,4\}$	o_2
$c=\{1\}$	o_2
$b=\{3\}$	o_1
$j=\{2,3,4\}$	o_7
$T=\{1,2,3,4\}$	o_7
$e=\{3,4\}$	o_1, o_4
$d=\{2,3\}$	o_7
$k=\{1,3,4\}$	o_5

Table 5: The retrieved objects

8 Conclusion and future work

Chiaromella et al [CFM96] have developed a logical model for structured documents. The model captures well the challenge where information cannot be considered as a flat structure any longer by allowing the retrieval of parts of documents that are exactly relevant to an information need. The model was specified within a logic. In the model, the uncertainty issue that often accompanies IR was ignored. This paper is a *first step* to extend the logical model so that uncertainty is formally represented. The D-S theory of evidence is used because this framework provides many features that are important in the context of IR (this was extensively discussed in [Lal96a, SH93]). It is assumed that the logic used in the model was classical logic. In that case, the D-S formalism has a straightforward counterpart. The paper shows that the theory allows the expression of the uncertainty with respect to components of a document, and that is compatible with the logical model developed by Chiaromella et al. This looks promising for future IR development for the modelling of structured documents within a formal and theoretically sound framework.

The next step is to perform some experimentation to study the effectiveness and the applicability of the model. It is important to know whether the belief functions capture well

the relevance of objects, or whether others measures, offered by the D-S framework, would be better. Also, the criteria of exhaustivity and specificity need to be further investigated experimentally. Finally, it is essential to know how the Dempster's combination rule performs experimentally, meaning whether it models aggregation appropriately with respect to semantic contents. Various implementations of the model are currently being explored and will be discussed in future papers.

Other models for structured documents have been proposed. In [RF96], a model is investigated based on probability theory, and a prototype for investigating representations of structured documents and retrieval functions has also been developed. The different characteristics of this work and the model proposed in this paper must be compared and contrasted. One way to achieve this would be to apply the model using the developed prototype as an implementation environment. This will allow the comparison of the results of the two models for IR and the evaluation of the relationship between the theories

There has been other work combining retrieval and browsing techniques in the context of hypermedia IR (see [AS96] for an overview). Whether the model developed in this paper can be applied to this work should be investigated. Although the

model has been developed for a certain kind of structured documents, I believe it may be applied to hypermedia documents, so that the indexing of an object takes into account not only the semantic content of the object, but also that of related objects. Since (i) the model is formally expressed within a logic, (ii) the D-S theory is more general than several other theories of uncertainty (e.g., probability theory, possibility theory), and (iii) an IR model based on D-S was shown to cover other IR models (e.g., the vector space model [TdSM93]), the outcome could be that the model proposed in this paper can be reasoned about, thus eventually leading to a framework for evaluating hypermedia systems, which is still an open problem.

The view that classical logic is not the best logic to model IR is not new. This has been established by several authors already [CC92, vRL96, Lal96a]. The D-S framework has already been used in conjunction with situation theory [Dev90] to develop a logical IR model in [Lal96a], although that work was not concerned with structured documents as studied in this paper. Also, a non-classical logic, Mirlog, has been proposed as a logic for IR [CS96]. It would be interesting to re-express the model in the context of Mirlog or situation theory.

In summary, the work described in this paper shows that the D-S theory of evidence is both promising and sufficient for the modelling of uncertainty inherent to structured documents.

Acknowledgement

I would like to thank the Department of Computing Science at the University of Glasgow for hosting my research this year. I would also like to thank Yves Chiamella, Jean-Pierre Chevallet and Iadh Ounis for spending the time describing their logical model to me. And thanks to Ian Ruthven, Thomas Rölleke, Iain Campbell, Mark Dunlop, Martin Gardner and Keith van Rijsbergen for their insightful comments. This work has been carried out in the framework of the Esprit project FERMI, Basic Research Action 8314.

9 Reference

- [AF96] M. Agosti and A. Smeaton, eds., *Information Retrieval and Hypertext*, Kluwer Academic Publishers, 1996.
- [CC92] Y. Chiamella, and J.P. Chevallet, About Retrieval Models and Logic. *The Computer Journal*, 35 (3): 233-242, 1992.
- [CK96] Y. Chiamella and A. Kheirbek, An integrated model for hypermedia and information retrieval, In M. Agosti and A. Smeaton, eds., *Information Retrieval and Hypertext*, Kluwer Academic Publishers, 1996.
- [CMF96] Y. Chiamella, P. Mulhem and F. Fourel, A model for multimedia information retrieval, Technical Report, Basic Research Action FERMI 8134, 1996.
- [Dev91] K.J. Devlin, *Logic and Information*, Cambridge University Press, 1991.
- [Fhu92] N. Fuhr, Probabilistic Models in Information Retrieval, *The Computer Journal*, 35 (3): 243-255, 1992.

[KC95] A. Kheirbeck and Y. Chiamella, Integrating hypermedia and information retrieval with conceptual graphs formalism, *Proceedings of Hypertext - Information Retrieval - Multimedia*, Konstanz, pp 47-60, 1995.

[Lal96a] M. Lalmas, *Theories of Information and Uncertainty for the modelling of Information Retrieval: an application of Situation Theory and Dempster-Shafer's Theory of Evidence*, Ph.D. Thesis, University of Glasgow, 1996.

[Lal96b] M. Lalmas, Modelling information retrieval with Dempster-Shafer's theory of evidence: a case study. *Proceedings of ECAI Workshop on Uncertainty in Information Systems: Questions of Viability*, Budapest, pp 29-36, 1996.

[LC97] M. Lalmas and Y. Chiamella, Dempster-Shafer's Theory of Evidence applied to Structured Documents: modelling Uncertainty. Fermi Technical Report, Esprit BRA 8134, Department of Computing Science, University of Glasgow, 1997. (Extended paper).

[Mec95] M. Mechkour, Emir 2: An extended model for image representation and retrieval. Technical Report, Basic Research Action FERMI, n. 8134, 1995.

[MS96] C. Meghini and U. Straccia, A relevance terminological logic for information retrieval, *Proceedings of ACM-SIGIR Annual International Conference on Research and Developments in Information Retrieval*, Zurich, pp 197-205, 1996.

[Nie90] J. Nie, Un modele de logique general pour les systemes de recherche d'informations. Application au prototype RIME. Ph.D. Thesis, University Joseph Fourier, Grenoble, 1990.

[RF96] T. Rölleke and N. Fuhr, Retrieval of complex objects using a four-valued logic, *Proceedings of ACM-SIGIR Annual International Conference on Research and Developments in Information Retrieval*, Zurich, pp 206-214, 1996.

[Saf87] A. Saffioti, An AI view of the treatment of uncertainty, *The Knowledge Engineering Review*, 2 (2): 75-97, 1987.

[Sha76] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[SH93] S.S. Schoken and R.A. Hummel, On the use of the Dempster Shafer model in information indexing and retrieval applications, *Int. J. Man-Machine Studies*, 39: 1-37, 1993.

[TdSM93] W. Teixeira de Silva and R.L. Milidui, Belief Function Model for Information Retrieval, *Journal of the American Society for Information Science*, 44 (1): 10-18, 1993.

[vRij86] C.J. van Rijsbergen, A non-classical logic for Information Retrieval, *The Computer Journal*, 29: 481-485, 1986.

[vRL96] C.J. van Rijsbergen and M. Lalmas, An Information Calculus for Information Retrieval, *Journal of the American Society of Information Science*, 47 (5): 385-398, 1996.