# A model for structured document retrieval: empirical investigations

**Mounia Lalmas** **Ian Ruthven**

University of Glasgow

**Abstract**

Documents often display a structure, e.g., several sections, each with several sub-sections and so on. Taking into account the structure of a document allows the retrieval process to focus on those parts of the document that are most relevant to an information need. In previous work, we developed a model for the representation and the retrieval of structured documents. This paper reports the first experimental study of the effectiveness and applicability of the model.

## 1 Introduction

In traditional *information retrieval* (IR), a document is considered as an atomic entity that is represented 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*, for example determined by the author. For instance, a document may have several sections, each with several sub-sections and so on. With a *structured document*, the representable, 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, and hence should be returned to the user. Moreover, the representation of a structured document must allow for the retrieval process to return *aggregated* components, e.g., a section, a set of sections, or all sections of the document that are relevant to a query, instead of delivering the whole document.

The requirements for representing and retrieving structured documents have been studied in [CMF96], and were the basis of a model proposed in [Lal97]. The theory used to build the model was *Dempster-Shafer's* (*D-S*) *Theory of Evidence* [Sha76] which provides a sound, formal framework for considering the intrinsic uncertainty of the representation of documents and the retrieval process. Furthermore, the theory provides an aggregation operator, *Dempster's combination rule*, that allows the expression of the representation and the uncertainty of aggregated components.

In [Lal97], we demonstrated the relationships between the requirements of a model for the representation and the retrieval of structured documents and some of the functions offered by D-S theory. We described the model based on the theory, and we showed that the model appropriately provides for: (i) representing individual and aggregated document components, and the uncertainty of their representation; (ii) calculating the relevance of a document or

document component to a query; and (iii) retrieving document components that are most relevant to a particular information need.

The next step is to study the effectiveness and applicability of our model. In particular we are interested in: (i) how well does the model capture the relevance of individual components? (ii) does Dempster's combination rule model aggregation appropriately with respect to the representation of document components? and (iii) what are the experimental behaviours of the criteria used to select the relevant document components?

To investigate these issues, an implementation and an evaluation of the model were performed. The aim of this paper is to report our initial investigations in implementing and evaluating our model. At this early stage, we concentrated on two specific tasks:

(i) **Implementation of basic features of the model:** We discuss some of the problems in adapting D-S theory of evidence to IR, in particular the difficulties in defining the notion of uncommitted belief. This leads to a basic implementation that allows us to investigate a primitive version of our model but does give a good understanding of how the model could behave in practice.

(ii) **Design of an evaluation method for structured document retrieval:** Most methods used to evaluate the effectiveness of IR models are based on non-structured document retrieval. It was therefore necessary to develop a method to carry out our investigations. We present an evaluation method for structured document retrieval that allows the examination of the behaviour of our model under varying conditions.

The paper is organised as follows: We outline the main concepts of D-S theory of evidence in section 2. We give an overview of the model in section 3. We describe the implementation of the basic features of our model and discuss the problems encountered in section 4. We explain the design of our evaluation method in section 5 and we report our experiments in section 6. The results and their analysis are discussed in section 7. We finish with some thought for future work in section 8, and conclude in section 9.

# 2 Dempster-Shafer's Theory of Evidence

The following explanation of Dempster-Shafer's Theory of Evidence is based on IR. A more general view can be found in [Sha76]. D-S theory is based on the view that propositions are represented as subsets of a given set. Assume that we have a set $T$ of indexing elements (e.g., terms, phrases, etc.). This set is referred to as a *frame of discernment*. The powerset of $T$, $\wp(T)$, defines a set of potentially overlapping sets, $A_1, \cdots, A_n$. Example of propositions are: "$A_i$ is relevant to a query", or "$A_i$ indexes the document".

Evidence can be associated with subsets of $\wp(T)$. This evidence is assigned to each subset (proposition) to express the evidence that it is observed or *discerned*. The evidence is computed based on a density function $m : \wp(T) \rightarrow [0, 1]$ called a *basic probability assignment*

(*bpa*):

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

Any $A$ such that $m(A) > 0$ is called a *focal element*. Each source of evidence (e.g., traditional weighting scheme or user evidence) may define a different set of focal elements, referred to as a *body of evidence*. Given a body of evidence, the total belief in any set $A$ is found from the sum of all its subsets. This defines a *belief function* $Bel : \wp(T) \rightarrow [0, 1]$:

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

$Bel(A)$ is the total belief committed to A, that is, the total positive effect the body of evidence has on the truth being in $A$. For sets composed of a single element, $Bel(A)$ will equal $m(A)$.

D-S theory has an operation, *Dempster's rule of combination*, for the combination of evidence from a variety of sources. This rule aggregates two bodies of evidence defined within the same frame of discernment into one body of evidence. More precisely, it computes a measure of agreement between two bodies of evidence concerning various propositions discerned from a common frame of discernment. Let $m_1$ and $m_2$ be two bpas associated to two bodies of evidence defined in $T$. The new body of evidence is defined by the bpa $m = m_1 \oplus m_2$ as follows:

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

# 3 The model for structured document retrieval

In this work, the structure of a document corresponds to a tree whose nodes are the components of the document (e.g., chapters, sections, etc.) and whose edges represent the composition relationship (e.g., a chapter contains several sections). The *root* node of the tree embodies the whole document, and the *leaf* nodes comprise the raw content of the document (e.g., a text, an image, etc.). A leaf node can be at any granularity, e.g., single term, phrase, sentence, etc. Any non-leaf node is referred to as an *aggregated* component (the root node included), and its information content is defined by the aggregation of its components. The model does not place any restrictions on the location of relevant components within a document; that is, the relevance of a document component is only dependent on the number and relevance of its sub-components, not the order in which they appear.

Next, we describe the model developed for the representation and retrieval of structured documents: the representation (section 3.1) and retrieval (section 3.2) of structured documents[1].

[1] See [CMF96, Lal97] for the complete description. Also, the reader should refer to these papers to obtain explanations for any of the modelling decisions.

## 3.1 Representing structured documents

In the model, the information content of a component $c$ is represented by a body of evidence defined on a frame of discernment $T$. This frame is defined as a set of indexing terms[2]. The focal elements define the propositions (a set of terms) describing the information content of the component $c$. A bpa $m_c$ expresses the uncertainty attached to these propositions. For any focal element $A$, $m_c(A)$ is the belief that the set of terms in $A$ is a good representation of the information content of the component $c$.

The computation of the focal elements and their associated bpa depends on whether the component is a leaf or aggregated. For a leaf component, the two entities are defined from the indexing process, whereas for an aggregated component, they are determined from the application of Dempster's combination rule. More precisely, let $c$ be an aggregated component, and let $c_1, \ldots, c_k$ be its sub-components, where each $c_i$ is defined by a body of evidence with bpa $m_{c_i}$, for $i = 1, \ldots, k$. The body of evidence associated to the component $c$ is defined with the bpa $m_c = m_{c_1} \oplus \cdots \oplus m_{c_k}$.

In [Lal97], we show that it was necessary to assign a belief to the frame of discernment, and this to each leaf component. This is so that any proposition discerned by a leaf component remains discerned when the component is combined with other components to form aggregated components[3]. That is, the frame of discernment constitutes itself a focal element. By having the frame itself as a focal element, the case where some beliefs remain uncommitted can be captured. In the context of IR, uncommitted beliefs may be used to represent the uncertainty (overall ignorance) associated to the indexing of a component.

## 3.2 Retrieval of structured documents

With structured documents, retrieving a leaf component means that only the component is relevant to the information need, whereas retrieving an aggregated component means that all its sub-components are relevant to the information need[4]. Furthermore, several related components may be retrieved in response to a query. In this case, the adopted approach is to retrieve the component deeper in the structure. This choice corresponds to the most *specific* component of the document that satisfies the information need, but which remains *exhaustive* to the information need[5]. Therefore, the retrieval of structured document needs to cater for the following: to express the relevance of a document component to an information need, and to focus on those parts of the document that are most relevant to an information need. The model uses the criteria of specificity and exhaustivity for this effect.

[2]This is not specific to the model; the frame may be defined as a set of phrases, logical formulae or any other indexing features.

[3]This behaviour is not specific to the way Dempster's combination rule is used, and has been well acknowledged.

[4]In the latter case, only the aggregated component could be displayed to the user and then constitutes an access point from where the user can decide to browse the structure if needed.

[5]We recall that a 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 information required by the query.

We represent an information need by a set of terms $Q$. Let $c$ be a document component with bpa $m_c$. The relevance of this component to the query is expressed by:

$$Bel_c(Q) = \sum_{A \subseteq Q} m_c(A)$$

$Bel_c(Q)$ captures relevance because it is based on all terms sets defining the information content of the component $c$ that are included in the query set. It also takes into account the beliefs associated to these sets.

Documents are first *fetched* based on the exhaustivity criterion. These are any document whose root component $c$ is such that $Bel_c(Q) > 0$ since there exists at least a set of terms that is part of the query set. It was shown in [Lal97] that any root component $c$ such that $Bel_c(Q) = 0$ is not exhaustive to the query, and hence neither of its sub-components.

A fetched root component $c$ may not be the most specific to the query $Q$. All the components that constitute the document with root $c$ are *browsed* based on the specificity criteria. In each *branch*[6], the most specific component to the query $Q$ is the component $c'$ with the highest $Bel_{c'}(Q)$. The reason is that if all the terms discerned are included in the query set, then all the component's information content concerns the query, and vice versa.

With the above strategy, two components may be related (i.e., belong to the same branch). In this case, the component with the higher relevance is retrieved since it corresponds to the most specific to the query. If the components have the same relevance, the component deeper in the structure is retrieved.

# 4 Implementation of the basic features of the model

In the previous section, we presented a model for the representation and retrieval of structured documents. The next step is to provide an implementation of the model to investigate its effectiveness and applicability. In particular, we want to provide answers to the following three questions: (i) does the belief function model appropriately the relevance of a document component to a query? (ii) does the Dempster combination rule model aggregation appropriately with respect to the representation of documents? (iii) is our expression of exhaustivity and specificity adequate; in other words does it allow the retrieval of those document components that are most specific to the query?

To answer those questions, we must implement the basic features of our model. Four steps are required: choosing the frame of discernment and focal elements (section 4.1), selecting how to assign evidence for each focal element (section 4.2), computing the uncommitted belief for each document component (section 4.3), and deciding how to combine evidence for each proposition (section 4.4).

[6]A branch is used to refer to the related components starting from the root component, and ending with a leaf component, all organised along one "line" of the document structure.

## 4.1 Frame of discernment and focal elements

Other major investigations in the use of D-S to model IR, e.g. [SH93, dSM93], have relied on external knowledge representation such as thesauri, to define how elements interact, and none have been applied to structured documents and very few have been implemented. Therefore it is difficult to know how D-S theory behaves in practice. This work starts at a very basic level to provide an experimental underpinning to more complex representational techniques. Therefore, the frame of discernment is taken to be the set of indexing terms. The focal elements, then, correspond to the terms used to index document components, and the frame itself (the set of all indexing terms).

## 4.2 Assigning evidence

Let $c$ be a document component with bpa $m_c$. The computation of $m_c$ associated to each document component presented major problems. None of the formulations we proposed, [RL97], led to a belief function that models appropriately the relevance of a document component to a query. This is due to the fact that a bpa must be normalised: the sum of the bpa over all subsets of $T$ should equal 1; i.e. all the documents should have an equal sum. For example, we carried out some experiments, reported separately in [RL97], on non-structured documents normalising each term according to the following formula[7]:

$$m_c(t) = \begin{cases} 0 & \text{if } t \notin c, \\ \frac{w(t)}{\sum_{t \in c} w(t)} & \text{if } t \in c. \end{cases}$$

where $w(t)$ is $idf(t)$ or $tf(t) \times idf(t)$ weighting functions [vR79], and $t \in c$ means that the term $t$ occurs in the document component $c$. The results showed very low effectiveness. This is because the normalisation tends to give a higher weight to a term appearing in a short document than a long document. The issue of *normalising* is an important and controversial issue in D-S theory. Similar counter-intuitive results have also been found in other D-S applications (e.g., [Zad96]), and other IR theories (e.g., [CRSvR95]).

As a consequence, we did not normalise the weights. The assignment of bpa $m_c$ to each focal element of a component $c$ was calculated as follows: $m_c(t) = w(t)$ where $w(t)$ is defined as above.

## 4.3 Uncommitted beliefs

We also investigated the estimation of the uncommitted beliefs. For example, we tried the following formulation:

$$\sum_{t \in T} m_c(t) - \sum_{t \in c} m_c(t)$$

[7]For simplicity of notation, we write $m_c(t)$ instead of $m_c(\{t\})$.

The results, also reported in [RL97], showed very low effectiveness. This is because, as for normalisation, a short document tends to have larger uncommitted belief than a long document. As a consequence, we did not consider at this stage uncommitted beliefs. This has some consequences which are discussed in the next section.

## 4.4 Representing aggregated component

The representation of the aggregated components is a direct application of Dempster's combination rule. Let us suppose that we have two components (sub-documents) $c_1$ and $c_2$ with respective bpas $m_1$ and $m_2$. Suppose that the uncommitted belief is for each bpa respectively $u_1$ and $u_2$. Then the bpa associated to the aggregated component is:

$$m(t) = \frac{1}{K} \times \begin{cases} 0 & \text{if } t \notin c_1 \text{ and } t \notin c_2, \\ m_1(t) \times m_2(t) + m_1(t) \times u_2 + m_2(t) \times u_1 & \text{if } t \in c_1 \text{ and } t \in c_2, \\ m_1(t) \times u_2 & \text{if } t \in c_1 \text{ and } t \notin c_2, \\ m_2(t) \times u_1 & \text{if } t \notin c_1 \text{ and } t \in c_2, \\ u_1 \times u_2 & \text{the uncommitted belief.} \end{cases}$$

where $K$ is defined as follows:

$$K = \sum_{t \notin c_1 \text{ and } t \notin c_2} m_1(t) \times m_2(t)$$

However, since we did not take into account the uncommitted beliefs (see previous section), the formula that we used instead is[8]:

$$m(t) = \begin{cases} 0 & \text{if } t \notin c_1 \text{ and } t \notin c_2, \\ m_1(t) + m_2(t) & \text{if } t \in c_1 \text{ and } t \in c_2, \\ m_1(t) & \text{if } t \in c_1 \text{ and } t \notin c_2, \\ m_2(t) & \text{if } t \notin c_1 \text{ and } t \in c_2. \end{cases}$$

Two consequences arise with the use of the above formula as the aggregation function. First, it is always the case that an aggregated component has a higher relevance than its individual sub-components. Therefore, we cannot determine whether our specificity criterion is valid or not. Second, the aggregation function is different to that defined by Dempster's combination rule, and hence the conclusions drawn from the experimental results may not be applicable to Dempster's combination rule. However, as reported in section 7, we can derive some preliminary conclusions on the effectiveness of our model.

# 5 Design of evaluation method

In this section we detail our evaluation method for assessing the performance of our model. As mentioned in the previous section our implementation uses a primitive version of Demp-

[8]It is was not possible to simply have $u_1 = u_2 = 0$ because very few terms will have non-null weights in the aggregated component (see section 3.1).

ster's combination rule but it does allow us to perform a first stage in the evaluation of our model. This section is structured as follows: first we describe the need for a new evaluation method for structured document retrieval, then we describe our approach to this problem.

In IR, the effectiveness of a model is evaluated by testing the model implemented on standard test collections. These consist of queries, documents, and relevance judgements. To evaluate a structured document retrieval model, one would require a number of test collections, each with relevance assessments for whole documents and each document component. There are two current problems with this approach. First only one collection of this type exists, and this collection is not freely available. Secondly, any experimental validation would require an evaluation of a model's performance over a number of collections.

Our alternative approach is to generate artificial test collections based on existing test collections. There are two main advantages in using this methodology: first, we can create any number of test collections, and secondly we can investigate the performance of our model on retrieval effectiveness by varying the relative similarity of the document components. We do not want simply to test how well our technique targets relevant components but we also want to investigate how retrieval effectiveness changes depending on the relative relevance of different components. This "relativeness" must be made explicit somehow.

We constructed a set of test collections, each based on the standard Cranfield and Cacm collections. In each collection, each document $d_i$ was combined with another document $d_j$ to form a *pseudo-structured document* $d_{ij}$. The combination criterion was different for each test collection, and reflected the similarity between $d_i$ and $d_j$:

- **collection one:** each document in the collection was combined with the document most similar to it. The document-document similarity was calculated using the cosine similarity measure. This simulates the case where each pseudo-structured document is composed of two sub-components about the same topic.

- **collection two:** each document in the collection was combined with the document that is the most different to it with at least one term in common. The dissimilarity was also calculated using the cosine measure. This simulates the case where the two sub-components of a document are on different topics (e.g., two articles in a newspaper).

  Collections three and four are more specialised:

- **collection three:** for each query:

  - for each relevant document $d_i$ (as given by the relevance judgements accompanying the test collection), $d_j$ is the relevant document most similar to $d_i$; and
  - for each retrieved, non-relevant document $d_i$, $d_j$ is the retrieved, non-relevant document most similar to $d_i$.

  In this case, each pseudo-structured document is composed by adding more relevant

information to an already relevant component[9]. **Collection one** combines documents according to similarity, whereas this case combines documents according to similarity *and* relevance. The sub-components of each relevant document is relevant (see section 7 for more explanation on how this is useful). Note that this gives a different collection for each query.

- **collection four:** for each query, for each retrieved document $d_i$, $d_j$ is the retrieved document most similar to $d_i$. This also gives a different collection for each query. **Collection one** matched documents over the whole collection, and **collection three** only matched documents over the relevant set. This case is to ensure that any results obtained from **collection three** are not due to the document-document similarity measure being over the relevant documents rather than the whole collection.

The four collection types described here differ in the level of similarity between $d_i$ and $d_j$, from a minimal overlap (collection two), through similar (collection one) to similar *and* relevant (collection three). Each of the four collection types contain the same number of documents as the original collections. The constructed collections are intended as a means of comparing the performance of the aggregation function against standard retrieval. There still remains the question of how we decide whether a structured document is relevant or not. In order to directly compare the retrieval effectiveness, we regard a structured document $d_{ij}$ as relevant if $d_i$ is relevant. This means that the structured document collections have the same number of relevance assessments for each query as the original collections.

At present the set-up is intended to provide an investigation into the performance of a basic implementation of the model on extreme examples of structured document collections. For example, we would expect the collection two results from a collection of structured documents whose components were very different in content. It is also intended as a means of testing whether describing a document as the composition of its components is appropriate for information retrieval.

# 6 Experiments

Our experiments are designed to investigate the effectiveness of the model in retrieving either sub-components or whole documents. We have based our experiments on three cases:

- **Case one:** Using the $idf$ weighting scheme on the Cacm collection.
- **Case two:** Using the $idf$ weighting scheme on the Cranfield collection.
- **Case three:** Using the $tf \times idf$ weighting scheme on the Cacm collection. The $idf$ weighting scheme does not take into account document length, only the weights

[9]We are not suggesting that the material in this pseudo-structured document is coherent, or that it would be assessed relevant if presented to the user, only that it is combined of two assessed relevant documents.

attached to the terms. The $tf \times idf$ scheme takes into account the number of times a term appears relative to the number of terms in the document. This case, then, is to test whether the aggregation operator is affected by the different weighting scheme used.

For each query and for each collection we scored the three components of each pseudo-structured document (the document $d_{ij}$, and each of its sub-components $d_i$ and $d_j$). The score of sub-components $d_i$ and $d_j$ were calculated using the belief function (section 3.2) based on the $idf$ or $tf \times idf$ weighting schemes (section 4.2). The score of the structured document $d_{ij}$ was calculated using the belief function on the weights of the terms in the aggregated document (section 4.4).

# 7 Results and analysis

The experiments described in the previous section allow us to construct recall_precision (RP) graphs for each of our three cases and for each collection type described in section 5. As mentioned in section 5, the structured document collections contain the same number of documents as the collections from which they were derived.

Figures 1, and 2 summarise our results. For each case, we are comparing the retrieval effectiveness of retrieving a sub-component against retrieving the whole document. In Figures 1 and 2, $idf/tf \times idf$ represents the results from retrieving the sub-component $d_i$. This is equivalent to standard $idf$ or $tf \times idf$ retrieval on each collection. The other four lines represent the results from the four collections described in section 5. As we have not considered the effect of uncommitted belief so far, this is equivalent to the retrieval score of the sub-component $d_i$ added to the retrieval score of the sub-component $d_j$. The labels for each line reflect the collections described in section 5.

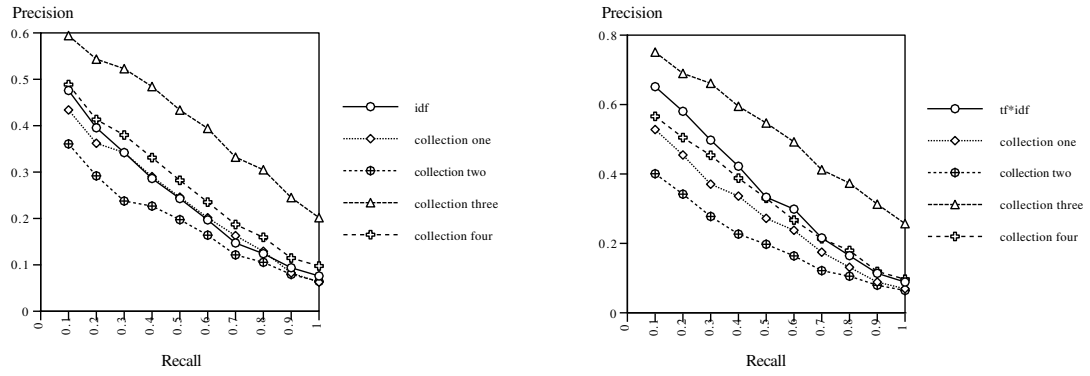


Figure 1: RP figures for $idf$ (lhs) and $tf \times idf$ (rhs) weighting scheme on the Cacm collection

The results using $idf$ scheme show that for both the Cacm and the Cranfield collections, the poorest retrieval comes from retrieving a structured document whose components are

on different topics. The best retrieval performance comes from retrieving a structured document whose components are both relevant. This increase in effectiveness is particularly noticeable when compared against the effectiveness of retrieving a document whose components are similar (**collection one**) or even similar in query terms (**collection four**) but are not necessarily relevant.

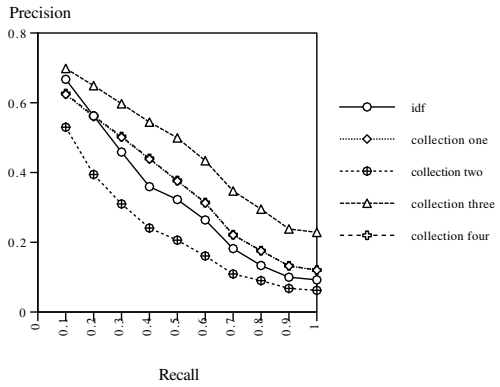


Figure 2: RP figures for $idf$ weighting scheme on the Cranfield collection

As discussed in section 2, the D-S theory computes a measure of agreement between two sources of evidence. In our work this is taken to be a measure of coherence between two document components. As seen in our results, the similarity or coherence between the documents is a factor in predicting whether to retrieve a document component or the whole document.

The results from using the $tf \times idf$ weighting scheme suggest that using the aggregation function at present, results are not affected by the particular weighting scheme used.

Although these results need further investigation on collections of varying size and document length, they can be summarised as follows:

- in all cases, retrieving a document containing two relevant components is significantly better than retrieving only one of the relevant components;
- retrieving a structured document whose components are topically similar is roughly as effective as retrieving a structured document whose components are topically similar as regards to a particular query;
- retrieving a structured document whose components are topically similar is about as effective in terms of recall_precision figures as retrieving a sub-component;
- retrieving a structured document whose components are topically dissimilar is never as effective as retrieving the most relevant sub-component.

# 8 Discussion and Future work

In previous work, we proposed a model for the representation and retrieval of structured documents [Lal97]. This model was expressed using Dempster-Shafer's Theory of Evidence. The aim of this paper was to investigate the effectiveness of our model. In particular, we were interested in the following three issues: **Issue 1:** the relevance of individual components as determined by the belief function; **Issue 2:** the aggregation of the representation of components as defined by the D-S combination rule; and **Issue 3:** the experimental behaviour of the criteria used to select the relevant document components.

Due to the normalisation problem (see section 4.2 and [RL97]), the relevance of document components as computed by the belief function (Issue 1) to queries was not adequately captured. This must be overcome otherwise, as it is the case in the work reported in this paper, it becomes difficult, if not impossible, to investigate the other issues 2 and 3 listed above. Representing a document component by a body of evidence is very expressive (e.g., it captures incompleteness, impreciseness, uncertainty), but there are major theoretical and implementational issues in deciding how these aspects should be captured. For example, it is not obvious how to implement focal elements, and in particular, how to compute their associated uncertainty. We are currently carrying two types of experiments to study how to implement bodies of evidence as an appropriate way to represent documents:

- **Type 1:** We are investigating different weighting mechanisms on which to base the bpas. This type of experiments has two aims. The first one is to ensure that the bpas do indeed capture the beliefs associated to term sets in representing document components. The second one is that it is necessary to study more carefully the normalisation process as imposed by the D-S framework, so that to overcome any counter-intuitive behaviour as those we observed while carrying out some of our initial experiments.

- **Type 2:** We are studying different methods that can be used to estimate the uncommitted beliefs, [RL97]. So far, the different formulations that we have tried lead to very low effectiveness. Therefore, we have ignored them in the implementation of our model, and as a result, we cannot affirmatively conclude on the effectiveness our model when implemented.

Regarding Issue 2, as discussed in section 4.3 and above, we did not take into account uncommitted beliefs in the implementation of our model. As a result, the aggregation operation did not perform as well as expected. At this stage, we should not yet conclude that the aggregation operation as defined in the D-S framework is not appropriate. Further experiments (of Type 2 above) are necessary to establish this. We have however a better understanding of the behaviour of the uncommitted belief in the context of IR modelling in the sense that document length must be better encapsulated when computing uncommitted beliefs.

We do not have at this stage any conclusion regarding Issue 3. This is due to the fact that as discussed in section 4.4, with the way we implemented the aggregation operation, it is

never the case that sub-components are retrieved. Therefore, we still do not know whether our criterion of specificity is valid or not.

# 9 Conclusion

Dempster-Shafer Theory of Evidence is a theory of uncertainty that captures many features of IR (e.g., the uncertainty of indexing, the uncertainty of retrieval, and the combination of evidence). It also allows us to assign different levels of evidence to individual elements (e.g., indexing terms) and to their conjunction (e.g., a document, or document component). This theory encapsulates probability theory as a special case and is hence more expressive than probability theory.

This paper describes some initial attempts to investigate the requirements of implementing an IR system for the representation and retrieval of structured documents based on D-S theory. This investigation has highlighted some of the difficulties in successfully implementing a D-S based IR system but has also shown that even such a basic implementation does capture some intuitive aspects of structured document retrieval. In particular, we have demonstrated that the retrieval of documents whose components are dissimilar will significantly reduce retrieval effectiveness and that retrieval of documents whose components are relevant will significantly increase retrieval performance.

Although it is not possible at this stage to draw clear conclusions on the effectiveness and the applicability of our model, two major achievements arose from these initial experiments. First, we now have a clearer understanding of the behaviour of the D-S functions in the context of IR modelling. We can therefore accomplish in future work a more appropriate implementation of the functions, thus allowing us to examine the effectiveness of our model. Second, we developed a method that allowed us to investigate how retrieval effectiveness changes depending on the relative relevance of different components (e.g., a component with two relevant sub-components, a component with only one relevant component). Standard IR evaluation methods would only allow us to to test how well our technique targets relevant components.

# 10 Acknowledgement

We would like to thank Norbert Fuhr for hosting one of the authors, Mounia Lalmas, while carrying out this work. This work has been conducted under the framework of the Esprit project FERMI, Basic Research Action 8134.

# References

[CMF96] Y Chiaramella, P Mulhem, and F Fourel. A model for multimedia information retrieval. Technical report, Basic Research Action FERMI 8134, 1996.

[CRSvR95] F Crestani, I Ruthven, M Sanderson, and C J van Rijsbergen. The troubles with using a logical model of ir on a large collection of documents. Experimenting retrieval by logical imaging on TREC. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, Washington D.C., USA, November 1995.

[dSM93] W Teixeira de Silva and R L Milidiu. Belief function model for information retrieval. *Journal of the American Society for Information Science*, 44(1):10 – 18, 1993.

[Lal97] M Lalmas. Dempster-shafer's theory of evidence applied to structured documents: capturing uncertainty. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, PA, USA, July 1997.

[RL97] I Ruthven and M Lalmas. Experimenting on dempster-shafer's theory of evidence in information retrieval. Technical report, Dept of Computing Science, University of Glasgow, 1997. forthcoming.

[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.

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

[vR79] C J van Rijsbergen. *Information retrieval*. Butterworths, 1979. 2nd edition.

[Zad96] L A Zadeh. A simple view of the dempster-shafer theory of evidence. *The AI Magazine*, pages 85 – 90, 1996.