ISSN 1369-1961

Department of
Computer Science

Technical Report No. 766

Integrating information use into relevance feedback

Ian Ruthven and
Mounia Lalmas

QUEEN MARY
AND WESTFIELD COLLEGE
UNIVERSITY OF LONDON

March 1999

Integrating information use into relevance feedback

Ian Ruthven
Department of Computing Science,
University of Glasgow,
Email:igr@dcs.gla.ac.uk

Mounia Lalmas
Department of Computing Science,
Queen Mary and Westfield College,
University of London.
Email: mounia@dcs.qmu.ac.uk

March 1999

Abstract

In this paper we propose a model of relevance feedback based on Dempster-Shafer's Theory of Evidence. This model is founded on combining evidence from user's relevance assessments with algorithms describing how words are used within documents. We motivate the use of the Dempster-Shafer framework as an appropriate theory for modelling the relevance feedback process. This model also incorporates the uncertain nature of information retrieval and relevance feedback.

Keywords: Information retrieval, relevance feedback, uncertain evidence, digital libraries, Dempster-Shafer's Theory of Evidence

1. Introduction and background

In Digital Libraries, seeking information is an inherently uncertain activity: searchers may not have a developed idea of what information they are searching for, they may not be able to transfer their conceptual idea of what information they want into a suitable query to present to the system, and they will generally not a have good idea of what information is available for retrieval. However, early in the field, researchers recognised that although users had difficulty expressing exactly the information they want, they could recognise *relevant* and *non-relevant* information when they saw it. This lead to the notion of *relevance feedback* [Roc71]: users marking information objects as relevant or non-relevant to their needs. The system can then use this information quantitatively - retrieving more relevant documents - and qualitatively - retrieving relevant documents before non-relevant ones.

Relevance feedback (RF) techniques aim to improve retrieval effectiveness by preferring retrieval of relevant documents and suppressing retrieval of non-relevant documents, based on relevance information given by the user. This relevance information, or *relevance assessments*, can be used in two ways: to alter the weights attached to query terms, e.g. [RSJ76], or to add or remove query terms, e.g. [Roc71]. In practice, most Information Retrieval (IR) systems, used in Digital Libraries will use a combination of reweighting and query modification techniques.

The majority of RF techniques are based on the presence or absence of keywords in relevant and non-relevant documents. However the reasons *why* a user may select a document as relevant can depend on many more aspects than simply which terms appear in the document [BS98]. As indicated by Denos et al [DBM97], although users can give explicit reasons for assessing a document as relevant, IR systems cannot use this information to improve a search because they lack the flexibility to detect *why* a user has marked a document as relevant. It is imperative, then, for access methods to digital libraries to extend the range of aspects that are used in relevance feedback. This means increasing the power of the IR system in detecting those criteria a user may be employing in making relevance assessments.

In [RL99] we demonstrated that incorporating information on how words are *used* within documents - *term characteristics*, in a RF situation, can lead to significant improvements in retrieval effectiveness across collections. We also demonstrated, experimentally, that different combinations of characteristics are more suitable for different queries. **In other**

words, different combinations of characteristics are better at detecting relevance for different queries.

This information was used in [RL99] to motivate biasing retrieval in favour of how a term was used in relevant, as opposed to irrelevant documents. In our experiments on the TREC[1] collections [VH96] we demonstrated that this technique of selective relevance feedback - selecting which characteristics to use for which query - not only performed well but outperformed standard relevance feedback algorithms such as the $F_{4.5}$ measure, [RSJ76].

However we also highlighted the need for a formal model to reason about how information was used within documents, what aspects of the use of information is likely to retrieve relevant documents and how this information should be used in relevance feedback. The development of this model is in two stages: the first stage is to define a reasoning module to select which characteristics of term usage are important in a search (which terms are important, which characteristics of those terms are important, how important are the term characteristics in the current search); the second stage is to construct a method of *combining* the information from the reasoning module in order to display the optimal ranking of documents to the user.

In this paper we concentrate on the second stage of our model: combining evidence about which terms and term characteristics are good at retrieving relevant documents. This approach is based on Dempster-Shafer's Theory of Evidence. The next section summarises our previous work.

1.1. Previous work

In [RL99], we investigated how information on how words, or terms, are used within documents can be used to improve RF. This was based on two standard IR measures, *tf*, and *idf*, and two novel measures, *theme* and *context*.

The *tf* measure, [Har92], measures the frequency of a term within a document. *idf*, based on the number of documents containing a term, measures the frequency of a term within a collection.

The *theme* measure is based on the distribution of a word's occurrences in a document. If the occurrences of a word are spread evenly throughout the document then the word is likely to be related to the main topic of the document[2]. If, on the other hand, the occurrences of a word only occur in one section of the document then the word is more likely to be related to a sub-topic. This assumption is reflected in the *theme* relation: the higher the *theme* value of a word, the more evenly distributed the term is throughout the document.

The *context* relation measures how closely associated a query term is with other query terms occurring in the same document. So if two query terms occur very closely together in the same document, e.g. in the same sentence, then there is a higher likelihood that they are contextually related. The *context* relation gives a higher value to a query term if it occurs in close proximity to another query term.

Each function was designed to assign a value between 0-50 to each document, according to how well a document displays the characteristic. For example, a value of 50 for the *theme* characteristic means that the term is exactly distributed throughout the document (likely to be the main topic or related to the main topic) whereas a low value means that the term is only used locally (or as a sub-topic). Obviously these measures interact somewhat, for example a low *theme* value and high *tf* value means that the term is used often but only in one part of the document.

---

[1] These document collections are comprised of full-length newspaper articles. The size of the collections which we used ranged from around 70,000 to 200,000 documents.

[2] Here we are talking about content-bearing words, and so do not include prepositions or other stop terms.

In [RL99] we demonstrated that for each query we could select which set of characteristics - *theme*, *context*, *tf* and *idf* - to use to improve RF performance. We also demonstrated that optimal performance is achieved when we vary the *amount* of evidence coming from each characteristic. For example, for some queries we should score documents by the *context* and *theme* characteristics, but we can improve performance by varying how much of the individual term characteristic value contributes to the document score, e.g. by counting the *context* characteristic as only half as important as the *theme* characteristic. Over a series of experiments we concluded that *how* the evidence coming from each characteristic was combined had as big an impact on the quality of the RF effectiveness as which characteristics were combined. Formally specifying a model of combination which can be used to understand how the combination process should operate is then necessary to fully exploit this approach.

1.2. Outline of paper

This paper describes such a formal model and the paper is structured as follows: in section 1.3, we give a working example which we use to illustrate our approach and highlight the salient modelling issues. In section 2 we give a brief introduction to Dempster-Shafer's Theory of Evidence as we have applied it; and we also motivate the suitability of this theory in modelling relevance feedback. In section 3 we discuss the combination of evidence in detail and we conclude in section 4.

1.3. Working example

The discussion in the rest of the paper will be illustrated by examples based on a simple document representation.

Consider five documents each containing three terms: $d_1\{t_1, t_2, t_3\}$, $d_2\{t_4, t_5, t_6\}$, $d_3\{t_3, t_4, t_5\}$, $d_4\{t_1, t_3, t_5\}$, and $d_5\{t_2, t_4, t_6\}$. Table 1 shows the values for three characteristics of the terms used in the documents. All characteristics scores for terms that do not occur in a document are taken to be zero. Note that the context relation, as defined at present is query dependent as well as document dependent as it is measured by the proximity of two query terms. Values for this characteristic will be defined further in the examples.

Document	Term	*theme*	*tf*
d_1	t_1	50	30
	t_2	25	15
	t_3	45	20
d_2	t_4	30	10
	t_5	10	10
	t_6	30	15
d_3	t_3	15	50
	t_4	25	30
	t_5	0	30
d_4	t_1	10	45
	t_3	0	30
	t_5	0	30
d_5	t_2	10	10
	t_4	50	20
	t_6	0	0

Table 1: Sample document representations

2. Dempster-Shafer's Theory of Evidence

Dempster-Shafer's (D-S) Theory of Evidence is a theory of uncertainty [Saf87] that was first developed by [Dem68] and extended by Shafer [Sha76]. Its main difference to probability

theory, which is treated as a special case, is that it allows the explicit representation of ignorance and combination of evidence. The explicit representation of imprecision of evidence makes the use of the D-S theory particularly attractive for modelling complex systems. The combination of evidence is expressed by the so-called *Dempster's combination rule*, which allows the expression of aggregation necessary in a model using multiple sources of evidence. In no other theory of uncertainty, the combination of evidence is as explicitly captured as a fundamental property.

In this section we describe the main concepts of D-S theory, based on the description given in [Sha76].

2.1. Frame of discernment

The D-S framework is based on the view whereby propositions are represented as subsets of a given set. Suppose that we are concerned with the value of some quantity u, and the set of its possible values is U. The set U is called a *frame of discernment.* An example of a proposition is "the value of u is in A" for some $A \subseteq U$. Thus, the propositions of interest are in a one-to-one correspondence with the subsets of U. The proposition $A = \{a\}$ for $a \in U$ constitutes a basic proposition "the value of u is a". In our approach the frame of discernment is taken to be the set of available documents, which in our example is the set $\{d_1, .., d_5\}$.

2.2 Basic probability assignment

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

$$m(\varnothing) = 0 \text{ and } \sum_{A \subseteq U} m(A) = 1 \qquad (1)$$

$m(A)$ 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 $\varnothing$). The focal elements and the associated bpa define a *body of evidence.*

In our work individual term characteristics, which assign mass only to singleton sets, act as a body of evidence assigning mass values to individual documents[3]. Each term characteristic acts as *bpa*. Our approach is slightly different from most D-S applications as we have, *a priori*, fixed the maximum mass value that can be assigned to a set. The maximum value that can be attached to a document is 50, which is the maximum value that can be attached to a term characteristic (section 1.2). The focal elements are then the documents which have a positive mass value assigned to them, i.e. display the term characteristic.

From the definition of the bpa, in equation 1, the sum of the non-null bpas must equate to 1, i.e. each body of evidence must assign the same amount of evidence to the frame of discernment. In our example, each term characteristic assigns a total evidence of 250 (5 documents * maximum characteristic value of 50). The total evidence can be scaled to fall between 0 and 1.

2.3 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* $Bel: \wp(U) \rightarrow [0,1]$ defined upon m as follows:

---

[3]The user's relevance assessments, which can assign mass values to singleton sets or sets with multiple elements also act as a bpa. This will be discussed separately in section 3.

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

$Bel(A)$ 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.

2.4 Plausibility function

A particular characteristic of the D-S framework (one which makes it different from probability theory) is that if $Bel(A)<1$, then the remaining evidence $1\text{-}Bel(A)$ needs not necessarily refute A (i.e., supports its negation $\overline{A}$). That is we do not have the so-called *additivity rule* $Bel(A) + Bel(\overline{A}) = 1$. Some of the remaining evidence may be assigned to propositions which are not disjoint from A, and hence could be plausibly transferable to A in light of new information. This is formally represented by a plausibility function $Pl: \wp(U) \rightarrow [0,1]$ defined upon a bpa m as follows:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) \tag{3}$$

$Pl(A)$ is the extent to which the available evidence fails to refute A.

2.5 Dempster's combination rule

D-S theory has an operation, *Dempster's rule of combination*, for the pooling of evidence from a variety of sources. This rule aggregates two *independent* bodies of evidence defined within the same frame of discernment into one body of evidence. Let m_1 and m_2 be the bpas associated to two independent bodies of evidence defined in a frame of discernment U. The new body of evidence is defined by a bpa m on the same frame U:

$$m(A) = m_1 \otimes m_2(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \tag{4}$$

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 new bpa takes into account the bpa associated to the propositions in both bodies that yield the propositions of the combined body. The denominator of the above equation is a normalisation factor that ensures that m is a bpa. In our approach, we use the combination rule to combine the *bpa*s from the term characteristics. This combination produces a single *bpa* over the documents in the collection derived from the combination of the individual term characteristic information.

2.6 Uncommitted belief

From the definition of the *bpa*, each body of evidence must assign the same total amount of belief to the frame of discernment, T. The total amount of evidence that can be assigned to the documents is N*50 (where N is the number of documents in the collection and 50 is the maximum mass value that can be assigned to each document, see section 1.2). However, the maximum mass value will not be assigned to all documents. Consequently there will be evidence which is unassigned, violating the definition of the *bpa*. There are three possible ways to avoid this violation: (1) normalise the bpa values assigned to the focal elements such that each *bpa* sums to the same value, (2) assign the remainder of the belief equally to the documents in the collection that do not display the characteristic, or (3) treat it as *uncommitted belief*.

In the first approach - normalisation - we scale the bpas for each body of evidence such that the sum of the evidence attached to the focal elements sum to the same amount. Let us consider the example of two bodies of evidence, the *theme* values for terms t_1 and t_5, shown in Table 2. The total amount of evidence to be assigned is 250. The mass values for each term are then scaled so that they sum to 250 (column 4, Table 2). However as only evidence assigned by t_5 is to document d_2, then all the evidence is assigned to this document, irrespective of how well the document reflects the theme characteristic. Worse, the mass value assigned to d_1 by term t_1 is lower than that assigned to document d_2 by t_5 after normalisation, even though before normalisation it had a higher value. Normalisation, then, can give counter-intuitive results, changing the relative amount of evidence assigned to documents without justification.

Term	Document	Mass	Normalised mass
t_1	d_1	50	208.3
	d_2	0	0
	d_3	0	0
	d_4	10	41.7
	d_5	0	0
t_5	d_1	0	0
	d_2	10	250
	d_3	0	0
	d_4	0	0
	d_5	0	0

Table 2: Normalising mass values for theme characteristics (terms t_1 and t_5)

The second approach, taken by probability theory, assumes that any evidence that does not support a proposition is evidence against that proposition, i.e. $P(A) = 1 - P(\overline{A})$. D-S theory views this as untenable, as any evidence that is not assigned to a proposition could turn out to support the proposition. It is merely evidence that has not been assigned. This leads to the notion of uncommitted belief, which is specific to the D-S approach.

In our approach the uncommitted belief is the evidence not directly assigned by a term characteristic to a focal element (a document or a set of documents), and is given by,

$$ub = N*50 - \sum_{i=1}^{N} m(d_i) \qquad (5)$$

Equation 5: uncommitted belief for a term characteristic *bpa*, where N = number of documents in a collection, d_i is the ith document in the collection, and $m(\mathrm{d_i})$ is the mass assigned to document $\mathrm{d_i}$, for that term.

This equation will give us a direct calculation of the uncommitted belief, based on the mass values assigned to the focal elements. However, we can further utilise the uncommitted belief by treating is as a measure of the *quality* of the evidence supplied by the term characteristic. This means using the uncommitted belief as a regulating device, controlling how much of the value of a characteristics are converted into the mass function. We take the example of the *tf* values for term t_5 (shown in Table 3, column 3). If the *tf* measure is unreliable, or is less accurate at measuring the term frequency than another algorithm, we could increase the measure of uncommitted belief and rescale the mass values accordingly (Table 3, column 4). The rescaling is based on a constant factor given by,

$$m'(d_i) = \frac{m(d_i)}{\sum_{i=1}^{n} m(d_i)} \times ((n \times 50) - ub') \qquad (6)$$

Equation 6: rescaling the mass for a term characteristic, where $m(d_i)$ is the original mass assigned to document d_i, $m'(d_i)$ is the new mass value. n is the number of documents in the collection, ub' is the value of the uncommitted belief in the new *bpa* .

$\sum_{i=1}^{n} m(d_i)$ is the amount of evidence assigned to the focal elements of the original *bpa.*

This differs from the normalisation approach in two ways: firstly, the mass values for each focal element are still within the same range, 0-50, as we only ever decrease the mass values. Secondly *all* the bpas for each characteristic are scaled so the values are not affected by how many focal elements (documents displaying the characteristic) are present for each *bpa.* We are only recalculating the mass values for a term characteristic - asserting that a characteristic as a whole is better or worse than another characteristic.

Document	Term	Mass m	Mass m'
d_1	t_5	0	0
d_2	t_5	10	7.14
d_3	t_5	30	21.43
d_4	t_5	30	21.43
d_5	t_5	0	0
$\sum_{i=1}^{5} m(d_i)$		70	50
uncommitted belief		180	200

Table 3: Using uncommitted belief to reflect the quality of a term characteristic

Using the uncommitted belief in this fashion we can reflect a number of aspects of a term characteristics:

i. the *uncertainty* of the characteristic. Some characteristics may reflect aspects of the documents information content that are more easily measurable. For instance the term frequency, *tf*, is an easier characteristic to provide an algorithm for, as it is more objective in nature than measuring the topical nature of the document, which is dependent on the interpretation of what constitutes the topical nature of the document.

ii. the *imprecision* of the characteristics. One algorithm may be more accurate at describing a characteristic than another. For example, there are several ways to calculate the term frequency (*tf*) in a document[4], some of which are more effective on different collections or for different types of documents but which require more or less computation. So we may choose a less precise (less effective) algorithm which has better computational properties.

iii. the *quality* of the characteristics. Some characteristics may be better at detecting relevance than others. The focus of our work is to select which characteristics best describe relevance at a particular point in a search. As this may change over time, as the user refines what they are looking for, or as the information need changes, the characteristics may become better/worse at detecting relevance. For example the *theme* characteristic may be very good at detecting relevance at the start of the search (looking for documents about a particular topic) but later in the search the *context* may become more important (looking for documents in

[4]See Harman [Har92] for an overview of term frequency measures.

which a term appears only in a particular context). The uncommitted belief can then be used to reflect the changing importance of each term characteristic at different points in the search.

iv. the *strength* of the characteristic. Some characteristics should be considered to be more important than others independent of any other information. For example in [RL99] we showed that certain characteristics worked better on different collections independent of any other evidence. This may be due to the idiosyncrasies of individual collections but means that some characteristics may need to be treated as more/less important than others, regardless of the user's relevance assessments. The *strength* of the characteristic reflects the difference in quality of term characteristics reflecting different aspects of information use (*tf* as opposed to *theme*) rather than different implementation of the same characteristic (given by the *imprecision* of the characteristic).

v. the *importance* of the term. The uncommitted belief can also be used to represent information that is not document or query dependent, for example information on the frequency of the term in the collection, or the inverse document frequency (*idf*) [SJ72]. Also, some terms may be better at retrieving relevant documents than others, or we may be more certain of their utility, e.g. query terms. So we may want to treat the evidence regarding these terms as more certain.

The first four uses of uncommitted belief, **i.-iv.**, describe various aspects of term characteristics as a whole. These four values may be combined to a single value of the overall uncommitted belief for the term characteristic. The fifth use can be used to modify the evidence supplied by any characteristic of a term. In this paper we do not discuss how we obtain values for these aspects but in a practical implementation this will rely on experimentation.

2.7 Conclusion

D-S is a suitable framework for integrating term characteristic information into the relevance feedback process for three reasons:

i. combination of evidence: Evidence in a relevance feedback situation comes from two sets of evidence - evidence derived from algorithms describing how words are used within documents, section 1.2, and evidence from the user in the form of relevance assessments, see section 3.2.1. The combination of evidence in D-S is not only conceptually simple but it is easily implemented. D-S then provides a formal but manageable method of combining evidence from a variety of sources.

ii. representation of imprecision: All evidence is not equal, especially in relevance feedback, where the reasons for relevance may change over a search. So we need to be able to represent the quality of evidence. D-S provides this with the notion of uncommitted belief.

iii. functions to score documents: As discussed in sections 3.1.2. and 3.2.2 we show that we do not always want to score documents based on the same evidence at every stage in the search. The three functions - mass, belief and plausibility functions - provide alternative methods for different circumstances.

In the next section we describe how we use D-S in combining evidence from term characteristics and users' relevance assessments in a RF situation.

3. Combining characteristics of use

IR systems normally present a ranking of documents to the user: the documents are ranked in decreasing order of retrieval score. There are two sources of evidence we can employ to decide on the score of a document: - the evidence given by the term characteristics and the evidence given by the user's relevance assessments. For initial retrievals we have no evidence from the user (no relevance assessments) and can only use term characteristic information,

sections 3.1.1. and 3.1.2. With relevance information we can use both sources; this is described in sections 3.2.1 and 3.2.2.

3.1.1 Combination of evidence with no relevance information

The evidence given by the term characteristics is assigned to individual documents (singleton sets) with each term's characteristics describing a mass function. This mass function will assign zero mass to each non-singleton set[5] and a non-zero score to each document which contains a positive score for each term characteristic. We use the combination rule to calculate the score of each document, thus taking into account all the term characteristics of a term.

Example one:
Suppose we only consider the single word query t_3. The combination of two characteristics - *theme* and *tf* - for this term allow us to score the documents in order of estimated relevance based on how this term is used in the documents, as shown in Table 4, Column 4.

Documents	Theme	Tf	Combined score initial *ub*	Combined score altered *ub*
d_1	45	20	55	35
d_2	0	0	0	0
d_3	15	50	60	28
d_4	0	30	27	11
d_5	0	0	0	0
ub	190	150	108	176

Table 4: Mass function gained by combining two characteristics of term t_3 where *ub* = uncommitted belief

In this example we have calculated the uncommitted belief according to equation 5. If the uncommitted belief for the *theme* characteristic is increased from 190 to 210 and for *tf* is increased from 150 to 210, then we get the scores in Table 4, Column 5.

The mass function is then altered by the uncommitted belief. The combination with unaltered uncommitted belief assigns most evidence to d_3, followed by d_1, d_4, and none to d_2 or d_5. Treating the *tf* characteristic as less reliable than *theme*, by assigning a greater degree of uncommitted belief, changes the mass function to assigning most evidence to d_1, then d_3, d_4 and none to d_2 or d_5. Thus the use of the uncommitted belief can shift the emphasis of the combined mass function in the direction of one or other sources of evidence.

As noted in section 2.2, the maximum mass that can be assigned to a document by a term characteristic is 50, but a term can receive a higher mass as the result of combination. This is not a problem as the total evidence (total mass function) still sums to 250, i.e. the combination does not alter the total evidence over the frame of discernment.

[5]With the exception of the frame of discerment itself.

Example two:
As Dempster's rule is associative and commutative we can combine multiple characteristics of multiple terms. If we consider a two term query, say t_3 and t_4 we obtain Table 5. We then obtain a ranking that takes into account how the terms are used in the different documents.

Documents	t_3			t_4			Combined score
	theme	*tf*	*context*	*theme*	*tf*	*context*	
d_1	45	20	0	0	0	0	48
d_2	0	0	0	30	10	0	17
d_3	15	50	25	25	30	25	128
d_4	0	30	0	0	0	0	19
d_5	0	0	0	50	20	0	32

Table 5: Mass function gained by combining three characteristics of terms t_3 and t_4

3.1.2 Ranking and retrieval with no relevance information

Given a mass function over the documents in the collection, how should we rank the documents for presentation to the user? DS provides three functions for scoring documents: mass, belief and plausibility functions. In this case, as all the evidence is divided between the frame of discernment (the uncommitted belief) and the singleton sets so the belief function equates to the mass function. So our choice is then between the mass/belief functions and the plausibility function. In this situation the plausibility is equal to the sum of the mass assigned to the document and the uncommitted belief. As the uncommitted belief is the same for each document, i.e. not document dependent, then the plausibility and mass functions will give identical rankings although different scores. As we are only interested in ranking the documents we choose the mass function, as the simplest of the three available functions, to rank documents. In example two the documents would then be presented: d_3, d_1, d_5, d_4, and finally d_2. d_3, the only document which contains both query terms (t_3 and t_4) is retrieved first, all the other documents only contains one query term each.

Sections 3.1.1 and 3.1.2 described how to score and rank documents in the absence for relevance information from the user. Sections 3.2.1 and 3.2.2 extend this approach to include relevance information.

3.2.1 Combination of characteristics with relevance information

So far in this section we have assumed a query which has allowed us to directly score the documents according to the value of the characteristics of the query terms which it contains. In a relevance feedback situation we want, in addition, to extrapolate from the information in the relevant documents to facilitate the retrieval of more relevant documents. The question is how to use the term characteristic information in relevant and non-relevant documents, that is how to integrate evidence from the user, defining a *bpa* over the frame of discernment? We have a number of options:

i. we can treat the *value* of a term characteristic as important. In our example the *theme* value of term t_3 in document d_1 is 45, if d_1 is relevant then we could say that a value of 45 for this characteristic of this term is a good indicator of relevance. However we cannot with any credibility say that individual values of a term characteristic leads to relevance, we can only say that a thematic relation for a term indicates relevance better than no thematic relationship.

ii. we can treat the values for individual documents as a range, e.g. the *theme* value of term t_3 in document d_1 is 45 and in document d_3 it is 15. If both these documents, and no others, are relevant then we could assume that only documents which have t_3 *theme* values in the range 15-45 should be considered. However the users may make few relevant judgements and we cannot assert for certain that one particular characteristic is the one that

defines relevance. Also we cannot guarantee that users will have seen or assessed documents with *theme* values outside this range so we have no certainty that this range is significant.

iii. we can treat the evidence more generally by asserting that the *value* of particular term characteristics do not define which values are important, as in **i.** and **ii.**, but instead define how well the characteristic predicts relevance based on its appearance in the relevant and non-relevant documents. Let us assume that the query is t_4 and documents d_2 and d_5 have been marked relevant. For each term characteristics there are four cases to consider, based on the presence/absence of the term t_4 in the relevant and non-relevant documents. These are outlined in Table 6.

	t_4 theme characteristic	
Relevance	**Present**	**Absent**
Relevant	$\{d_2, d_5\}$	$\{\}$
Non-relevant	$\{d_3\}$	$\{d_1, d_4\}$

Table 6: Contingency table based on the presence/absence of the *theme* characteristic of t_4 in the relevant and non-relevant documents based

The first set of documents contain those which are relevant and display the term characteristic ($\{d_2, d_5\}$), the second contain the non-relevant documents that display the term characteristic ($\{d_3\}$).We can derive values for each of these cells that display the term characteristic by simply averaging the characteristic value of the term in each document in the cell. In our example the average *theme* score for query term t_4 is 20 in the relevant set displaying the characteristic and 25 in the non-relevant set displaying the characteristic so we assign a mass of 20 to the set $\{d_2, d_5\}$ and 25 to the set $\{d_3\}$, shown in Table 7. The uncommitted belief is 205 (250-(25+20)).

The other two cells (right hand column of Table 6) contain the sets which do not display the term characteristic and are either relevant or not-relevant. As the term characteristic of a term that does not appear in a document is automatically 0, the mass assigned to these sets is 0. In this way, we only consider the cells that indicate presence of a term[6].

Repeating this for the *tf* characteristic would give us a mass of 15 to the set $\{d_2, d_5\}$ and 30 to the set $\{d_3\}$ with an uncommitted belief of 210. These two mass functions can be combined using Dempster's combination rule to provide a single mass function based on the two term characteristics as demonstrated in example two.

We demonstrate the full model of relevance feedback, incorporating user's relevance assessments and term characteristics in Example three.

Example three:
The simplest case is to consider is relevance feedback with one relevant document. Assume that the user has issued a query, has marked document d_3 as relevant and has made no relevance decision on the other four documents[7]. For each query term in document d_3 we have some indication of how useful the term may be in detecting relevance[8].

---

[6]D-S expressly forbids the use of negative evidence (something which does not happen) being used to assign evidence. In this situation we differ from the F4 weighting scheme [RSJ76] which uses statistical infromation and a similar contingency table to derive weights which incorporate information on the absence of a term in a relevant/non-relevant document.

[7]It is customary in IR to assume that the documents which have not been marked explicitly as relevant or non-relevant can be assumed non-relevant, although they in all likelihood will contain a number of relevant documents that have either not been retrieved by the system or not been assessed by the user.

[8]Of course, it may be that a characteristic only appears by chance, and relevance is better described by another characteristic. By taking into account the characterisitics of terms in non-relevant documents we can limit this to

The current query is composed of the terms t_4, and t_5. In Table 7 we show the various sets which are assigned a mass value based on this document selection. Also we have filled in suitable values for the context characteristic.

	t_4		t_5	
	set	**mass**	**set**	**mass**
theme				
relevant	$\{d_3\}$	25	$\{d_3\}$	0
non-relevant	$\{d_2, d_5\}$	20	$\{d_2, d_4\}$	5
context				
relevant	$\{d_3\}$	15	$\{d_3\}$	40
non-relevant	$\{d_2, d_5\}$	20	$\{d_5\}$	20
tf				
relevant	$\{d_3\}$	30	$\{d_3\}$	30
non-relevant	$\{d_2, d_5\}$	15	$\{d_2, d_4\}$	20

Table 7: Mass functions based on relevance assessments

Dempster's combination rule can then be used to obtain a single mass function based on the mass functions from t_4, and t_5, Table 8(**a**). All other subsets of the frame of discernment are assumed to have zero mass. The evidence from the relevance assessments can be combined with the evidence from term characteristics for t_4, and t_5, Table 8(**b**), to form a single mass function, Table 8(**c**).

Set	mass
$\{d_1\}$	0
$\{d_2\}$	7
$\{d_2, d_4\}$	11
$\{d_2, d_5\}$	40
$\{d_3\}$	86
$\{d_4\}$	0
$\{d_5\}$	10

a

Set	mass
$\{d_1\}$	0
$\{d_2\}$	70
$\{d_2, d_4\}$	0
$\{d_2, d_5\}$	0
$\{d_3\}$	43
$\{d_4\}$	37
$\{d_5\}$	32

b

Set	mass
$\{d_1\}$	0
$\{d_2\}$	48
$\{d_2, d_4\}$	1
$\{d_2, d_5\}$	18
$\{d_3\}$	73
$\{d_4\}$	23
$\{d_5\}$	26

c

Table 8: **a.** mass function from combing relevance information only
b. mass function from combining term characteristic information only
c. mass function from combining relevance information and term characteristic information

a certain extent - by only considering characteristics that better describe relevant documents that non-relevant documents.

The results of the final combination, Table 8(**c**), is represented diagramatically in figure 1.

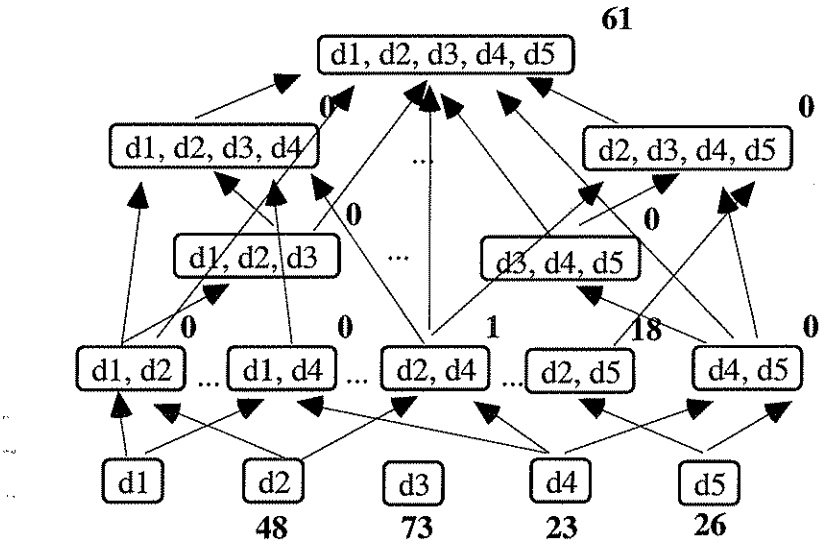

Figure 1: Diagrammatic representation of the combination of characteristics in a relevance feedback situation. -> represents subset relation. Figures indicate mass values

In section 2.6 we enumerated a number of uses for the uncommitted belief (four of which reflected the quality of term characteristics, one which reflected the quality of individual terms). There are two further uses for the uncommitted belief when we have relevance information:

i. *partial* relevance assessments. Most IR systems only allow users to mark a document as relevant or not-relevant. Recently, however, researchers such as [BI97] have investigated the use of partial relevance assessments: asking users to give a numerical value describing how relevant a document is. We can use this information to modify the uncommitted belief of a term according to whether it appears in a highly-relevant or slightly-relevant document.

ii. *biasing evidence between relevance assessments and query.* Evidence from research such as [SB90] indicates that relevance information and query information should not always be treated as being equally important. Furthermore, Haines and Croft [HC93] showed that this is collection dependent, i.e. in some collections, better retrieval effectiveness is achieved by treating query terms as more important, and in other collection we should treat user relevance as being more important. The uncommitted belief, then, may be used to bias retrieval in favour of terms appearing in the original query or those added from the user-selected relevant documents.

3.2.2 Ranking and retrieval with relevance assessments

The aim is to get a score for each document, the characteristics give us a score for each document (section 3.1.1) and the relevance assessments can be used to give us a score for sets that represent the useful characteristics (section 3.2.1). We have three ways to score a document: mass, belief and plausibility functions, which we will discuss in turn below.

i. mass function. The mass function considers the score for each set, and only that score. Intuitively this is not what we want as the characteristic evidence only gives a score to singleton sets and the relevance feedback evidence will only tend to give evidence to non-singleton sets. We want a method that will score the documents on all the evidence available.

ii. belief function. The belief function measures the total evidence supporting a set, based on the mass assigned to itself and its subsets. If we were working on a model for calculating the score of a set of documents, e.g. in a clustering model, then this is exactly what we would

want because it would calculate the score of all the sets including the non-singleton sets. However we are at the moment only interested in ranking the singleton sets (individual documents) so the belief function is the exact opposite of what we require because it uses the evidence of the singleton sets to score the non-singleton sets, rather than the other way round.

iii. plausibility function. The plausibility function considers the total plausible evidence for a set. This is the mass for a set and all the sets with which it intersects. This is then what we want - a function that combines the evidence from the characteristics (attached mainly to the singleton sets) and for the usefulness of the characteristics (attached to the non-singleton sets). Although this method will potentially score all sets, we only need to consider the singleton document sets when ranking the documents for retrieval.

Document d_i	$Pl(d_i)$
$\{d_1\}$	61
$\{d_2\}$	128
$\{d_3\}$	134
$\{d_4\}$	85
$\{d_5\}$	105

Table 9: Documents scored by plausibility function

If we score the documents from Example 3, Table 8(**c**), according to the plausibility function, we arrive a the scores in Table 9. In this case we would retrieve the documents in the order d_3 then d_2, d_5, d_4 and finally d_1. As d_3 is the only document marked relevant by the user, we should expect this to come at the top of the retrieved documents. d_2 is retrieved second as it contains both query terms and both query terms display the term characteristics. Documents d_5 and d_4 which both contain one query term appear next. d_5 is retrieved ahead of d_4 as the one query term it contains better displays the *theme* and *tf* characteristics than the query term contained within d_4. d_1 correctly appears at the bottom of the ranking as it does not contain either query term.

4. Conclusion

For digital libraries, with potentially large numbers of users who are unfamiliar with electronic searching, it is important that systems are as effective as possible in targetting relevant information. One of the strengths of relevance feedback is that it only requires a user to indicate relevant material, as opposed to *describing* an information need.

In this paper we have proposed a model for relevance feedback that allows the integration of how terms are used within documents into the relevance feedback process in a Digital Library system. This model is based on Dempster-Shafer's Theory of Evidence. The core of this approach is the combination of evidence from algorithms describing the information use of terms and relevance information from users. Dempster-Shafer's Theory of Evidence allows flexibility in how we combine this evidence: it allows us to include the quality of evidence (via the uncommitted belief), whilst providing a uniform framework for combining evidence. It also allows us to use information in different ways to retrieve documents, so we retrieve documents using different scoring functions in the presence/absence of relevance information (when we have no relevance information we use the mass function, and when we have relevance information from the user we use the plausibility function).

Our approach of including information on how words are used within documents can increase the flexibility of digital library systems in detecting relevant information without increasing the complexity of the users' role in the process.

In this paper we have only concentrated on *how* evidence is combined, not how we select which evidence to combine. Which terms to choose in relevance feedback and which characteristics of those terms to use in relevance feedback is outside of this process, and is

being developed separately. However we believe that we have demonstrated at a theoretical level the suitability and flexibility of the Dempster-Shafer approach in relevance feedback.

References

[BS98] C.L. Barry and L Schamber. Users' criteria for relevance evaluation: a cross-situational comparison. *Information, Processing and Management.* 34 (2/3). 1998.

[BI97] P. Borlund and P. Ingerwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation.* 53. **5.** 225 -250. 1997.

[Dem68] A. P. Dempster. A generalization of the Bayesian inference. *Journal of Royal Statistical Society*, 30, 1968, 205-447.

[DBM97] N. Denos, C Berrut and M Mechkour. An image system based on the visualization of system relevance via documents. *Database and Expert Systems Applications (DEXA '97)*, 8th International Conference, September, Toulouse, France. 1997.

[HC93] D. Haines and W. B. Croft. Relevance Feedback and Inference Networks. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2-11. 1993.

[Har92] D. Harman. Ranking algorithms. In: *Information retrieval : data structures & algorithms* . (W. B. Frakes and R. Baeza-Yates, ed.). Ch. 14. pp 363 - 392. 1992

[RSJ76] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society of Information Science*. **27**. p129-146.1976.

[Roc71] J. J. Rocchio. Relevance feedback in information retrieval. In: *The SMART retrieval system: experiments in automatic document processing.*(G. Salton, ed.). Ch. 14, pp 313-323. Prentice-Hall.

[RL99] I. Ruthven and M. Lalmas. *Selective relevance feedback using term characteristics.* CoLIS 3. 1999. *to appear*

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

[SB90] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science.* 41.**4.** 288-97. 1990.

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

[SJ72] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* **28**. (1). 11-20. 1972

[VH96] E. M. Voorhees and D. Harman. Overview of the Fifth Text REtrieval Conference (TREC-5). in *Proceedings of the 5th Text Retrieval Conference*. Gaitherburg, MD. 1996.