Using Dempster-Shafer's Theory of Evidence to combine aspects of information use

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

Mounia Lalmas
Department of Computing Science
Queen Mary
University of London.
Email: *mounia@dcs.qmw.ac.uk*

Abstract

In this paper we propose a model for relevance feedback. Our model combines 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 combination of evidence. This model also incorporates the uncertain nature of information retrieval and relevance feedback. We discuss the sources of uncertainty in combining evidence in information retrievel and the importance of combining evidence in relevance feedback. We also present results from a series of experiments that highlight various aspects of our approach and discuss our findings.

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

1. Introduction and background

Information seeking 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 an Information Retrieval (IR) system, and they will generally not have a 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* information when they saw it. This lead to the notion of *relevance feedback* (Rocchio, 1971): users marking information objects as relevant to their needs. The system can then use this information quantitatively - retrieving more documents similar to the relevant documents - and qualitatively - retrieving documents that are more similar to the relevant documents before those documents that are less similar.

Relevance feedback (RF) techniques aim to improve retrieval 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. (Robertson and Sparck Jones, 1976), or to add or remove query terms. In practice, most IR systems will use a combination of reweighting and query modification techniques, e.g. (Rocchio, 1971).

The majority of RF techniques are based on the presence or absence of keywords in 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 (Barry and Schamber, 1998). As indicated by Denos et al (Denos et al., 1997), 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 IR systems to extend the range of aspects that are used in RF. This means increasing the power of the IR system in detecting those criteria a user may be employing in making relevance assessments.

In (Ruthven and Lalmas, 1999) 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 showed, experimentally, that the best performance came from *selecting* which set of characteristics, for each query term, best indicated relevant material. *In other words, different combinations of characteristics are better at detecting relevance for different queries.*

In (Ruthven and Lalmas, 1999) we 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, (Ruthven et al., 1999), 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 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. In particular we propose a model for combination of evidence based on Dempster-Shafer's Theory of Evidence.

Section 1.1 examines related work in evidence combination, and section 1.2 summarises our previous work on evidence combination.

1.1. Combination of evidence in information retrieval

There are two main reasons for why evidence combination can improve retrieval:

i. empirical evidence that different retrieval functions retrieve different documents, e.g. (Lee, 1998).

ii. theoretical: different query representations can provide different interpretations of a user's underlying information need. This has a strong connection to Ingwersen's work on polyrepresentation. This theory states that multiple representations of the same object can provide better insight into what constitutes relevance than a single representation, (Ingwersen, 1994).

A number of approaches for evidence combination techniques have been proposed. These either combine query representations (different versions of the same query), or rankings (combining the results of different retrieval techniques).

Most of these methods have been based on empirical research: they have not been based on an underlying formal model of evidence combination.

Belkin et al (Belkin et al., 1995), for example, examined the role of multiple query representations in ad-hoc IR. A number of experiments were carried out with the conclusion that, although the combination of evidence had the potential to improve retrieval effectiveness, the actual performance gain is variable. One reason for this is that it is difficult to predict what combination of evidence will be effective for individual queries. An important conclusion from this work was that choosing which set of evidence best suits an individual query is an important stage in evidence combination. In this paper, section 5.5, we demonstrate how this can be achieved.

In (Lee, 1998), Lee proposed a RF technique based on multiple relevance feedback algorithms. Different RF techniques may produce different modified queries, and different queries will retrieve different documents. Therefore using a combination of RF techniques to modify queries will retrieve more relevant documents. An initial experiment validated this hypothesis. Combination of the rankings given by individual RF techniques *can* provide significant improvements in effectiveness over single RF methods, although combining techniques that retrieve different relevant documents does not *necessarily* improve retrieval effectiveness. In our approach we can incorporate information on the uncertainty attached to the combination process, and information on the quality of individual algorithms.

The second line of combination research is based on more formal models of evidence combination. The major example of this is the inference network model by described by Haines and Croft (Haines and Croft,1983). Inference networks are composed of nodes - representing documents, terms, phrases, etc. - and arcs representing the dependencies between the nodes. Each node contains a ‘link matrix’ that calculates the belief for a node given the belief on its parent nodes. Inference networks allow the combination of different representations of the same query terms, e.g. individual query terms, or phrases composed of the query terms. Silva et al (Silva et al., 2000), also proposed evidence combination techniques based on inference network models.

Our approach to evidence combination is also based on a formal model, namely Dempster-Shafer’s Theory of Evidence. The attraction for this theory over other formal techniques such as inference networks is that it allows us to represent the uncertainty attached to the evidence combination process. As we will describe later, this is a powerful and coherent way of representing aspects of combination such as the quality of evidential sources, the user's assessment's of evidence, and the reliability of evidence.

1.2 Previous research

In (Ruthven and Lalmas, 1999), 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* (term frequency) characteristic, (Harman, 1992), measures the frequency of a term within a document. *idf* (inverse document frequency) (Sparck Jones, 1972), based on the number of documents containing a term, measures the frequency of a term within a collection. Both these measures, usually in combination, have been demonstrated to give good retrieval results on a wide range of collections.

The *theme* characteristic 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[1]. 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* characteristic 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 of term. For example, a value of 50 for the *theme* characteristic of a term 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 in 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.

In (Ruthven and Lalmas, 1999) we demonstrated that for each query we could select which set of characteristics of each query term - *theme*, *context*, *tf* and *idf* - to use to improve RF performance. We also demonstrated that optimal

[1] Here we are talking about content-bearing words, and so do not include prepositions or other terms that occur frequently in the document collection. These terms tend to be poor at discriminating between relevant and non-relevant documents, and are usually not considered during retrieval.

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 was an important variable. The method of combination used had as big an impact on the quality of the RF effectiveness as which characteristics were combined (Ruthven and Lalmas, 1999). **Formally specifying a model of combination that can be used to understand how the combination process should operate is then necessary to fully exploit this approach.**

1.3. Aims and outline of paper

This paper describes such a formal model. We use Dempster-Shafer's Theory of Evidence as the basis of our modelling approach. We also carried out a number of experiments to investigate the effectiveness of our approach.

The paper is structured as follow. In section 1.4, 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 and we also motivate the suitability of this theory in modelling relevance feedback. In section 3 we discuss the combination of evidence without relevance information - ranking the documents after the user has submitted a query but not yet assessed any documents. This models the situation in which the user has submitted a new query to the system. In sections 4 and 5 we deal with combination of evidence when the user has assessed some documents as relevant. Throughout sections 3, 4 and 5 we present experimental results and discuss our findings. We conclude in section 5.

We should note here that our approach does not depend on a particular definition of relevance. A user may assess a document as relevant for many reasons, the assessment of relevance may change over time (section 4.1), and some documents may be considered to be more relevant than others (section 4.1). What we do claim for the relevance assessments is that by a user assessing a document as relevant, s/he is indicating that the document contains information of the kind s/he is looking for at the current point in the search. The actual mechanisms by which the user makes a relevance assessment (the details of the IR system interface) are not important to this paper.

1.4. 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 two 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: Example document representations

2. Dempster-Shafer's Theory of Evidence

Our interest is in investigating the effect of combining evidence from different characteristics of term use in documents. There are a variety of formal theories we could use for this purpose. We have chosen *Dempster-Shafer's (DS) Theory of Evidence* as it is a well-understood, formal framework for combining sources of evidence. The mathematical connection between IR and DS Theory was suggested by Van Rijsbergen (Van Rijsbergen, 1992), although this work concentrated on retrieval functions in general rather than specifically on relevance feedback. A continuing stream of research has investigated how theories based on DS can be used to model various aspects of the IR process, e.g. da Silva and Milidiu (1993), Schocken and Hummel (1993) and Lalmas and Ruthven (1998).

DS is a theory of uncertainty (Saffioti, 1987) that was first developed by Dempster (Dempster, 1968) 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. This explicit representation of ignorance, or the imprecision of evidence, makes the use of the DS theory particularly attractive for modelling complex systems. The combination of evidence is expressed by *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 is the combination of evidence explicitly captured as a fundamental property.

In this section we describe the main concepts of DS theory, based on the description given by Shafer (1976), presented within the context of RF.

2.1. Frame of discernment

The DS 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:

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

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 ∅). The focal elements and the associated bpa define a *body of evidence*.

In our work term characteristics, which assign mass only to singleton sets, act as a body of evidence assigning mass values to individual documents[2]. Each term characteristic acts as *bpa*. Our approach is slightly different from most DS 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 1.3). The focal elements are then the documents that have a positive mass value assigned to them, i.e. display the term characteristic.

[2] 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.

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*, we can compute the total belief provided by the body of evidence for a proposition. This is done with a *belief function* $Bel\colon \wp(U) \to [0,1]$ defined upon *m* as follows:

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

$Bel(A)$ is the total belief committed to *A*, that is, the mass of *A* itself plus the mass attached to all subsets of *A*. $Bel(A)$ is then 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 DS framework (one which makes it different from probability theory) is that if *Bel(A)*<1, then the remaining evidence 1-*Bel(A)* needs not necessarily refute *A* (i.e., support 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 the light of new information. This is formally represented by a plausibility function $Pl\colon \wp(U) \to [0,1]$ defined upon a bpa m as follows:

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

Pl(*A*) is the mass of *A* and the mass of all sets that intersect with *A*, i.e those that could transfer their mass to *A* or a subset of *A*. *Pl*(*A*) is the extent to which the available evidence fails to refute *A*.

2.5 Dempster's combination rule

DS 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*:

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

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 *U*. 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 1.3). However, the maximum mass value will not be assigned to all documents, as each term does not appear in every document. 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 with 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 the 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.

The second approach, taken by probability theory, assumes that any evidence that does not support a proposition is evidence against that proposition, i.e. $P(A) = 1 - P(\overline{A})$. DS 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 DS approach.

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)

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,

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

Equation 5 calculates the uncommitted belief for a term characteristic *bpa*, where *n*= number of documents in a collection, d_i is the *i*th document in the collection, and $m(d_\text{i})$ is the mass assigned to document 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 the 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,

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

Equation 6 defines 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 document’s 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 characteristic. 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[3], 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 that has better computational properties.

iii. the *quality* of the characteristic. Some characteristics may be better at indicating relevant material than others. The focus of our work is to select which characteristics best indicate 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 characteristic may become better/worse at discriminating relevant material.

For example the *theme* characteristic may be very good at indicating 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 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. Evidence supporting changes in users' criteria of this kind has been shown by, for example, Vakkari (2000) and Ellis (1989), and other studies that show that relevance, and the process of making relevance assessments, are dynamic processes.

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 (Ruthven and Lalmas, 1999) 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*) (Sparck Jones, 1972). 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 each 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 all these aspects but in a practical implementation this will rely on experimentation.

[3] See Harman (Harman, 1992) for an overview of term frequency measures.

2.7 Conclusion

DS 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 4. The combination of evidence in DS is not only conceptually simple but it is easily implemented. DS 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. DS provides this with the notion of uncommitted belief.

iii. functions to score documents: As will be discussed in sections 3.2. and 4.2 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 retrieval situations.

Our main interest is in providing a model for RF. This is accomplished in two stages. The first stage is to develop a method of retrieving documents when we have no relevance information from the user. This provides an initial set of documents that the user can assess for relevance. In the next section we describe how we use DS in combining evidence from term characteristics to provide such a retrieval function.

The second stage is to combine the retrieval function for retrieving documents with information from the users' relevance assessments, the feedback situation. This is described in section 4. The feedback model is, then, an extension of our initial retrieval model.

3. Initial document retrieval

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 and 3.2. With relevance information we can use both sources; this is described in sections 4.1 and 4.2.

3.1 Combining term characteristic information

The evidence given by the term characteristics is assigned to individual documents (singleton sets) with each characteristic of a term describing a mass function. This mass function will assign zero mass to each non-singleton set[4] and a non-zero score to each document that contains a positive score for a 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 7, 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 7: 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 7, 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.

[4] With the exception of the frame of discernment itself.

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.

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 8. 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 8: Mass function gained by combining three characteristics of terms t_3 and t_4

3.2 Ranking and retrieval

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 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 to the user in the following order: d_3, d_1, d_5, d_4, and finally d_2. d_3, the only document that contains both query terms (t_3 and t_4) is retrieved first, all the other documents only contains one query term each.

In the next section, we describe an experiment to test the effectiveness of the DS retrieval model for ranking documents.

3.3 Experiment

In this section we shall first describe the data we used for this experiment, section 3.3.1, a baseline experiment using no combination of evidence, section 3.3.2, and finally results of combining evidence from the term characteristics, section 3.3.3.

3.3.1 Experimental setup

In these experiments we used the Wall Street Journal (1990-92) (**WSJ**) and the Associated Press (1988) (**AP**) test collections from the TREC-5 set of collections (Voorhees and Harman, 1996). The details of these collections are summarised in Table 9. We applied common IR indexing steps such as the removal of highly frequent terms and the reduction of terms to their root variant (Van Rijsbergen, 1979).

Collection	**AP**	**WSJ**
Number of documents	79 919	74 580
Number of queries used[5]	48	45
Average words per query	3	3
Number of unique terms in collection	129 240	123 852

Table 9: Details of collections used

Each test collection comes with a set of fifty topics, each describing an information need and which criteria relevant documents should fulfil to be assessed relevant. A TREC topic has a number of sections (see Figure 1 for an example topic). In this experiment we only used the short **Title** section as a query, as using any more of the topic description may be an unrealistic user query, which are typically fairly short.

Number: 301

Title: International Organized Crime

Description:

Identify organisations that participate in international criminal activity, the activity, and, if possible, collaborating organisations and the countries involved.

Narrative:

A relevant document must as a minimum identify the organisation and the type of illegal activity (e.g., Columbian cartel exporting cocaine). Vague references to international drug trade without identification of the organisation(s) involved would not be relevant.

Figure 1: Example of a TREC topic

[5] Although each collection has 50 topics, not all topics have relevance assessments. Therefore we omitted these topics/queries in our experiments.

Associated to each topic is a list of documents that have been independently assessed as being relevant to the topic. These relevance assessments are used in measuring IR system effectiveness. Two standard evaluation measures are commonly used with IR: *precision* and *recall*. The recall of a system for a query is measured as the ratio of relevant documents *retrieved* to the total number of relevant documents for the query. Precision is the ratio of relevant documents retrieved to the total number of documents retrieved, (Van Rijsbergen, 1979).

IR systems typically rank documents in decreasing order of estimated likely relevance to a query. Recall and precision figures can be calculated at various points in this document ranking to give an indication of effectiveness at different levels of retrieval, (for example at 10% recall, i.e. 10% of relevant documents retrieved, 20% recall, 30% recall, etc. to give a set of 10 recall-precision figures), Figure 2.

Recall	Precision
10	67.3
20	65.9
30	59.2
40	45.3
50	36.7
60	33.3
70	21.9
80	19.7
90	15.3
100	12.1
Avg	37.7

Figure 2: Example recall and precision figures

Recall-precision (RP) figures are averaged over the set of queries or topics, to give a single set of RP figures for a collection. A common summary for a set of RP figures is to take the *average precision* value, the average of the precision measures at all relevance levels. We will use this measure to describe the majority of the results in this paper.

3.3.2 Retrieval by single characteristic

In the first experiment we carried out a retrieval using each characteristic as a single retrieval function (retrieval only by *idf* score of each query term in a document, retrieval only by *tf* score, etc.) The overall performance of each characteristic is measured by the average precision of the characteristic as a retrieval function for a set of queries, shown in Table 11.

Recall	**AP**				**WSJ**			
	idf	***tf***	***theme***	***context***	***idf***	***tf***	***theme***	***context***
10	22.5	9.1	5.0	9.2	16.5	12.7	4.7	3.2
20	19.6	4.5	1.9	5.7	16.1	11.8	2.1	3.1
30	18.9	3.0	1.4	1.6	15.8	8.1	0.7	2.7
40	16.6	2.4	1.3	1.5	14.2	7.1	0.4	2.2
50	12.4	2.1	1.2	0.8	14.0	6.9	0.4	2.1
60	7.7	1.0	1.2	0.8	10.7	6.3	0.3	2.0
70	6.7	0.7	1.1	0.6	10.4	6.2	0.3	1.4
80	6.1	0.6	1.1	0.5	8.7	5.1	0.3	1.3
90	5.4	0.4	1.1	0.2	8.1	5.0	0.3	1.3
100	4.9	0.4	1.1	0.1	7.5	4.9	0.3	1.3
Avg	**12.1**	**2.4**	**1.6**	**2.1**	**12.2**	**7.4**	**1.0**	**2.1**

Table 11: Retrieval by single characteristic. **Avg** is the average precision.

In both collections the *idf* function performed best, followed by *tf*, *context* and finally *theme*. These figures act as baseline figures for our next experiments on combination of evidence, described in the next section.

3.3.3 Retrieval by combination of evidence

In this experiment we compared the performance of using each combination of characteristics as a retrieval function. We compared two methods of combination; Dempster's combination rule and a simple summation method that consisted of summing the characteristic scores for each query term in a document. Table 12 (columns 2 and 3) shows the average precision for this experiment (full tables are in the Appendix, Tables A.1 - A.8)[6].

As indicated in sections 1.1 and 2.6 it may not be appropriate to treat each characteristic as equally important in retrieving relevant documents. Consequently we also tried weighting each characteristic with different values to investigate the effect of different uncommitted beliefs on the combination. The results from this experiment are shown in Table 12 (columns 4 and 5).

[6] As yet we lack a formal theory to decide how we should select good values to alter the uncommitted belief for characteristics. Consequently, we weighted each characteristics in an ad-hoc manner with the following values: *idf* -1, *tf* - 0.75, *theme* - 0.15, *context* - 0.5. Different weights give different results, as indicated in Table A.17, for the CISI collection. The appendix is available electronically at http://www.dcs.gla.ac.uk/ir/papers/Postscript/igr_ml_jiis_appendix.ps.gz and http://www.dcs.gla.ac.uk/ir/papers/Pdf/igr_ml_jiis_appendix.pdf

AP				
Combination	**simple,** **no weighting**	**DS,** **no weighting**	**simple,** **weighting**	**DS,** **weighting**
all	6.7	5.1	8.4	**10.4**
idf **+** ***context***	10.1	**12.2**	10	**12.2**
idf **+** ***tf***	10	5.1	**10.5**	10.4
idf **+** ***tf*** **+** ***context***	8.3	1.6	**8.6**	1.4
idf **+** ***tf*** **+** ***theme***	5.4	5.1	9.2	**10.4**
idf **+** ***theme***	5.1	**12.2**	10.4	**12.2**
idf **+** ***theme*** **+** ***context***	7.2	**12.1**	9.3	10.4
tf **+** ***context***	**6.9**	1.7	**6.9**	1.6
tf **+** ***theme***	1.9	1.6	**2.1**	1.6
tf **+** ***theme*** **+** ***context***	4.9	1.6	**6.8**	1.6
theme **+** ***context***	5.3	0.1	**7.7**	0.1

WSJ				
Combination	**simple,** **no weighting**	**DS,** **no weighting**	**simple,** **weighting**	**DS,** **weighting**
all	9.6	5.4	**11.1**	10.4
idf **+** ***context***	11	**12.3**	11	**12.3**
idf **+** ***tf***	10.2	5.4	**10.5**	10.4
idf **+** ***tf*** **+** ***context***	11.4	5.4	**11.5**	10.4
idf **+** ***tf*** **+** ***theme***	6	5.4	9.8	**10.4**
idf **+** ***theme***	5.3	**12.3**	10.2	**12.3**
idf **+** ***theme*** **+** ***context***	9.4	**12.3**	11	**12.3**
tf **+** ***context***	**10.4**	0.9	10.2	0.7
tf **+** ***theme***	3.5	0.9	**3.7**	0.7
tf **+** ***theme*** **+** ***context***	6.9	0.9	**10.1**	0.7
theme **+** ***context***	6.3	0	**8.3**	0

Table 12: Summarised results of combining characteristics, using Dempster's combination rule (**DS**), summing characteristic scores (**simple**), either weighting the characteristic scores (**weighting**) or treating characteristics as equally important (**no weighting**). ***all*** is the combination of all characteristics. Highest value for each combination is shown in bold.

Table 13 summarises how often each strategy obtained the highest average precision for a given combination, excluding single characteristics.

AP				**WSJ**			
	No weighting	**Weighting**	**Total**		**No weighting**	**Weighting**	**Total**
simple	1	6	**7**	**simple**	1	6	**7**
DS	2	4	**6**	**DS**	3	4	**7**
Total	**3**	**10**		**Total**	**4**	**10**	

Table 13: Number of times each strategy gave highest average precision for a combination of characteristics, using Dempster's combination rule (**DS**), summing characteristic scores (**simple**), either weighting the characteristic scores (**weighting**) or treating characteristics as equally important (**no weighting**). This count omits the single characteristic combinations as these are unaffected by the combination strategy or weighting.

We can compare the results under two conditions: the different combination methods and the effect of weighting the importance of the characteristics relative to each other.

i. Method of combination. From Tables 12 and 13 we can see that the method of combining the characteristic information does not have a big effect on how successful the strategies were overall. That is, using Dempster's combination rule instead of simply summing the characteristic scores did not significantly increase the number of combinations that gave higher average precision. This is not surprising as the way we have used the DS theory so far is basically also a summation method.

However, from Tables A.9 - A.10 in the Appendix, it is clear that the combination rule is having an effect. In particular, the different combination methods change the relative ordering of which combination of characteristics give better results, i.e. some combinations perform better using Dempster's combination rule and some perform better using the simple addition method. The combinations that involve a combination of *tf* and another characteristic tend to perform worse with the DS method than the simple method, whereas methods that combine *idf* do better with the DS method. One possible reason for this the way we assign the mass function, which we shall discuss below.

A further difference between the two methods is that using Dempster's rule tends to even out differences in the recall-precision values of the combinations (see Tables A.9 and A.10). Although the two methods did not give vastly differing *ranges* of recall-precision values under the two conditions, in the results from the Dempster combination case there were sets of results that were identical. For example, in Table A.9 (**simple** method) the results from combining *idf* with *theme* or *context* were different from the results from *idf* alone. In Table A.10 (**DS** method) these three results were identical. In the simple case the recall-precision values tended to be more distinct.

One possible cause of this effect is due to the way we assign the mass function. Although we manipulate the amount of mass assigned to each document by varying the uncommitted belief function, each characteristic will assign mass to a different *number* of focal elements. For example, the *idf* characteristic of a term will assign evidence to every document that contains the term; the other characteristics will only assign evidence to documents for which the characteristic has a non-zero value. As the values of *theme*, *context*, and *tf* may be zero for a number of documents in each case, it is likely that each of these characteristics will not only assign different values to each document, but also assign values to a variable number of documents.

In the DS method this will have the effect of increasing the uncommitted belief for the characteristics which assign a mass value to fewer focal elements. Thus the characteristics that assign mass to the fewest number of focal elements will have the least effect on scoring the documents. The DS method, then, biases retrieval in favour of characteristics that assign evidence to more characteristics. In our case this is *idf* so the results of a combination of *idf* will be closer to the results given by *idf* alone. As *idf* is the best single retrieval function, DS generally gives better results for combinations with *idf*. The different characteristics also assign values to different numbers of characteristics using the simple method. However as the combination in the simple method is not affected by the total mass assigned to the documents (as is the case in the DS method, through the uncommitted belief) this bias does not occur.

ii. Weighting of characteristics. Although the method of combination did not produce any significant effects, treating different characteristics with varying importance to other characteristics did produce better overall results than treating all characteristics as equally important. Weighting of characteristics not only increases the average precision of most combinations of characteristics, it also modifies which combinations give better results in both methods of combination. For example, in Table 12 (AP), the combination of all characteristics performs better than the combination of *idf*, *theme* and *context* information, if we use weighting and poorer if we do not.

In both collections the combination of DS and weighting can improve retrieval effectiveness although only slightly. Although we have not shown a clear advantage in using the DS combination rule in combining evidence from characteristics, we believe that the flexibility of the uncommitted belief in representing the various forms of uncertainty discussed above hold the potential for improved results. This is the subject of ongoing research.

3.4 Summary

Sections 3.1 and 3.2 described how to score and rank documents using term characteristics. We have demonstrated, in section 3.3, that combining characteristics of information use under two methods (DS and simple) can increase, although modestly, average precision. We have also shown that Dempster's combination rule performs in the same range as a standard method of scoring documents and that characteristics should be treated as of varying importance.

We now turn to relevance feedback. Our approach is to treat the relevance information from the user - the list of documents they regard as containing relevant information - as an additional source of evidence to be combined. Our RF model is an extension of the model outline in the previous section but extended to incorporate relevance feedback information.

4. Relevance feedback

In a relevance feedback situation we want to extrapolate from the information in the relevant documents to facilitate the retrieval of more relevant documents. That is we want to use the information in the documents the user has marked relevant to help retrieve documents that the user may also consider relevant. In this section we suggest how this might be achieved in our model, section 4.1, and how documents should be ranked when we have relevance feedback information. In section 5 we describe a set of experiments designed to test the effectiveness of our approach.

4.1 Combination of characteristics with relevance information

When we have relevance information from the user, we have two sources of evidence to rank documents: the term characteristic information and the relevant documents. We have described how we use the term characteristic information to rank documents in section 3. The question now is how to use the term characteristic information in relevant and non-relevant documents? That is, how do we integrate evidence from the user with our DS model to define 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 relation.

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 contains one term, $t_{4,}$ 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 14.

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

Table 14: Contingency table based on the presence/absence of the *theme* characteristic of t_4 in the relevant and non-relevant documents

The first set of documents contain those that 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 the 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[7] 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 15. The uncommitted belief is 205 (250-(25+20)).

The other two cells (right hand column of Table 14) contain the sets that 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[8].

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.

[7] Calculated from the values given in Table 1.

[8] D-S expressly forbids the use of negative evidence (something that does not happen) being used to assign evidence. In this situation we differ from the F_4 weighting scheme (Robertson and Sparck Jones, 1976) which uses statistical information and a similar contingency table to derive weights that incorporate information on the absence of a term in a relevant/non-relevant document.

Example three:

The simplest case is to consider 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[9]. For each query term in document d_3 we have some indication of how useful the term may be in detecting relevance[10].

The current query is composed of the terms t_4, and t_5. In Table 15 we show the various sets that are assigned a mass value based on this document selection. Also we have filled in 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 15: 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 16(**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 16(**b**), to form a single mass function, Table 16(**c**). In none of the mass functions in Table 16 do we assign all the possible evidence - there is uncommitted belief at each stage.

[9] It is customary in IR to assume that the documents that 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.

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

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 16: **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

The results of the final combination, Table 16(**c**), is represented diagramatically in figure 3.

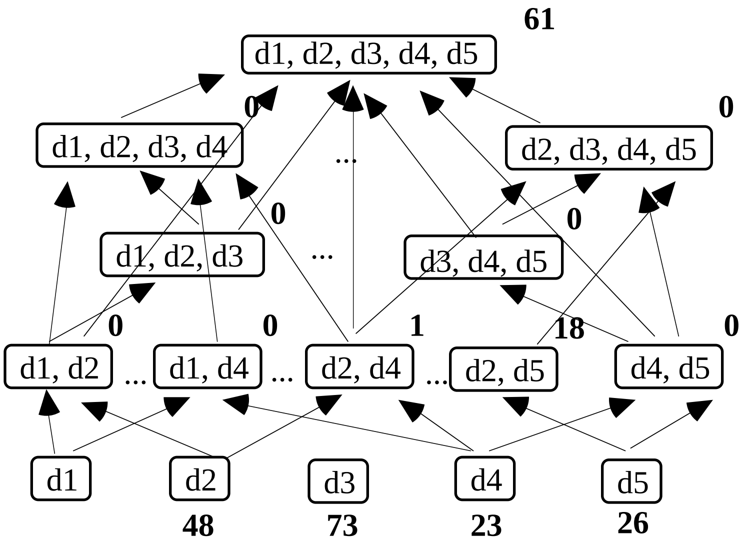

Figure 3: Diagrammatic representation of the combination of characteristics in a relevance feedback situation. → represents subset relation. Figures indicate mass values

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 three 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. However, researchers such as Borlund and Ingwersen (Borlund and Ingwersen, 1997) have investigated the use of partial relevance assessments: asking users to give a numerical value describing the relevance of a document. 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. *source of evidence* - biasing evidence between relevance assessments and query. Evidence from research such as Salton and Buckley (Salton and Buckley, 1990) indicates that relevance information and query information should not always be treated as being equally important. Furthermore, Haines and Croft (Haines and Croft,1983) showed that this is collection dependent; in some collections, better retrieval effectiveness is achieved by treating query terms as more important, and in other collections we should treat user relevance as being more important. The uncommitted belief, then, may be used to bias retrieval in favour of term characteristics appearing in the original query or those added from the user-selected relevant documents. If we extend our approach to include query term expansion, e.g. (Rocchio, 1971), we could also bias evidence between the original query characteristics of terms and characteristics of new query terms suggested by the system.

iii. *time of evidence*. In section 2.6 **iii.**, we argued the characteristics of a term that best indicate relevance can change over time. One reason for this is that a user may change her criteria for assessing relevance in the light of the relevant material. Typically RF algorithms do not consider time in deciding how to modify queries: each relevant document is considered to be an equal contributor to RF regardless of when in the search a document was assessed relevant. New relevance assessments can gradually change the system's view of which characteristics indicate relevance but a better way of handling the order in which assessments are made is by the use of ostensive weighting, suggested by Campbell and Van Rijsbergen (Campbell and Van Risjsbergen, 1996). Ostensive weighting of evidence, in a RF context, means treating the most recent relevance assessments as the best source of evidence regarding what the user regards as relevant material. Relevance assessments made early in the search, on the other hand, should be regarded as poorer indications of relevance. We can use the uncommitted belief to reflect this. If a term only appears in documents assessed early in the search, we should increase our uncertainty (uncommitted belief) regarding the term's utility for RF; if a term appears in the most recent relevant documents, they should be regarded as better evidence for RF and have a lower uncommitted belief.

4.2 Ranking and retrieval with relevance information

To re-rank documents after RF we need to obtain a score for each document; the characteristics give us a score for each document (section 3.1) and the relevance assessments can be used to give us a score for sets that represent the useful characteristics (section 4.1). We have three ways to score a document: mass, belief and plausibility functions, which we 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 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). This method will score all sets (the singleton document sets and those sets containing more than one document). However when ranking the documents we need consider the singleton document sets as the user will only be presented with a list of ranked documents.

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

Table 17: Documents scored by plausibility function

If we score the documents from Example 3, Table 16(**c**), according to the plausibility function, we arrive at the scores in Table 17 for the singleton document sets. 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.

5. Experiments on relevance feedback

We now describe our experiments on relevance feedback. In this experiment we investigate the use of term characteristics and DS in the context of relevance feedback. We introduce the data we used in these experiments in section 5.1, our baseline comparison measures in section 5.2, our methodology in section 5.3 and the results of our experiments in sections 5.4 - 5.6. We summarise the results of our findings in section 5.7.

5.1 Data

In this experiment we used a different collection from the experiments in section 3, as our particular implementation of the model is computationally expensive. The collection we used is the CISI collection, details of which are given in Table 18. This collection contains fewer and shorter documents than either the AP or WSJ collection making it an easier collection upon which to experiment. This collection has much higher number of query terms per query, although the average query term count is skewed somewhat by some very long queries.

Collection	**CISI**
Number of documents	1 460
Number of queries used	76
Average words per query	27.3
Number of unique terms in the collection	7 156

Table 18: Details of CISI collection

We carried out identical combination experiments to those described in section 3.3 for the CISI collection. These are reported in Appendix, Tables A.11 - A.16. The results we have previously obtained hold: combining information can improve retrieval effectiveness, weighting characteristics often improves retrieval effectiveness and DS and the simple combination method perform approximately as well as each other. The main differences between the two collections used previously and the CISI collection is that *tf* is a better single retrieval function than *idf*, and *theme* and *context* give higher average precision when used as a single retrieval function. These differences may arise from different features of the document collections, such as document length.

5.2 Baseline measures

In sections 5.2.1 - 5.2.3 we introduce the three baseline measures we used to compare our RF method.

5.2.1 No feedback

Our first baseline is the retrieval results obtained from doing no relevance feedback. For the CISI collection this is the combination of all characteristics combined using Dempster's combination rule. The characteristics were weighted as follows: *idf* - 1, *tf* - 0.75, *theme* - 0.5, *context* - 0.25.

5.2.2 Best combination

It may be that a better retrieval result could be obtained by using a good combination of characteristics rather than using RF. That is, we want to test whether the quality of the retrieval function is more important than the quality of the query: is developing a good query (through RF) more important than developing a good retrieval function (selecting the best overall combination of characteristics)? To test this, our second baseline is the best combination of characteristics from the experiments on combination of evidence. This is a combination of *tf* and *idf* for the CISI collection, Table A.12.

5.2.3 $F_{4.5}$

We should compare our technique for relevance feedback against another relevance feedback algorithm. For this we have chosen the $F_{4.5}$ weighting algorithm (Robertson and Sparck Jones, 1976), equation 7, which uses relevance feedback information to assign a new weight to a term. This technique for reweighting query terms was chosen partly because it has been shown to give good results but also because it does not add any new terms to the query. The experiments in this paper are concerned with selecting aspects of term use for RF, not expanding the query with new concepts. Our approach, can however, be extended to cover query expansion.

$$w_q(t) = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(n-r+0.5)(R-r+0.5)} \qquad \textbf{(7)}$$

Equation 7 show the $F_{4.5}$ function assigns a weight to term *t* for a given query. *r* = the number of relevant documents containing the term *t*, *n* = the number of documents containing *t*, *R* = the number of relevant documents for query *q*, and *N* = number of documents in the collection

5.3 Methodology

We carried out three experiments to test the performance of three aspects of our approach; weighting of characteristics, selecting characteristics of terms and method of combination of characteristic information. We isolate these three stages to allow us to investigate what aspects of our general approach are successful.

The first experiment was a version of the previous experiment on combination of evidence, section 3.3. The previous result indicated that characteristics should be treated as of varying importance. In this new experiment we used RF information to derive values to weight characteristics to examine whether relevance feedback information could lead to better weighting values for the different characteristics.

In the second experiment we used the relevant documents to select characteristics of query terms on a query-query basis. In this experiment we investigate whether relevance feedback information can lead to better combinations of characteristics for individual queries.

The final experiment is an implementation of the model of RF developed in section 4. This experiment looks at whether our proposed method of combining relevance feedback information with term characteristic information can improve retrieval effectiveness.

In each of the three experiments we used the following methodology:

i. documents were ranked using the combination of all characteristics, combined using Dempster's combination rule. This is the same ranking function as the first baseline.

ii. a cut-off was applied at rank position 30. Documents at or above this rank position were used to modify the query.

iii. documents in positions 30 - *N* (where *N* is the number of documents in the collection) were rescored by one of the methods described in sections 5.4 - 5.6. Each method corresponds to one of the experiments outlined above.

iv. recall-precision figures were calculated over the whole document ranking. The documents in rank positions 1 - 30 were fixed, no document in the collection could be ranked above them. This processing of reranking documents not used for query modification is known as freezing (Chang et al., 1971) and one common method of evaluating RF algorithms.

These steps were applied for 4 iterations, or cycles, of relevance feedback (steps i. - iv. were followed for a cut-off at 30 documents, then steps ii. - iv. were followed for a cut-off at 60 documents, a cut-off at 90 documents, etc). This resulted in five document rankings. Results will be presented as the average precision of each ranking.

5.4 Experiment one - relevance feedback using derived weighting factors

In section 3.3, we demonstrated that treating characteristics as of varying importance to a query was important: some characteristics should be treated as more important than others. We varied the uncommitted belief attached to a characteristic of a term by weighting the characteristics. In this experiment, we attempted to automatically *derive* good weights for characteristics. These weights were calculated for each characteristic of each query term. The weight corresponded to the ratio of the average characteristic value of a term in the relevant documents to the average characteristic value of a term in the non-relevant documents. We are then considering how good a characteristic of a term is at discriminating relevance. This method will be referred to as the *ratio* method.

We tried four versions of this approach. The first version ranked documents by the first baseline method (section 5.2.1) to provide an initial ranking, then weighted each characteristic of each query terms by derived weights by the method described above. In Table 19 (Ratio 1, column 5) we see that this method performed better than no feedback, Best Combination and $F_{4.5}$ after relevance feedback (Iterations 1 - 4).

	CISI						
Iteration	**No feedback**	**Best combination**	**$F_{4.5}$**	**Ratio 1**	**Ratio 2**	**Ratio 3**	**Ratio 4**
0	11.7	**12.9**	11.7	11.7	10.3	11.5	11.7
1	11.7	12.9	8.3	14.4	12.6	14.0	**14.6**
2	11.7	12.9	8.3	14.4	13.0	14.4	**14.6**
3	11.7	12.9	8.5	14.8	13.2	14.3	**14.9**
4	11.7	12.9	8.5	14.9	13.3	14.5	**15.0**

Table 19: Results of ratio methods. Highest value for each iteration is shown in bold.

A second approach used the same technique as before but with an initial ranking given by combination of all characteristics with no weighting. From Table 19 (Ratio 2, column 6) this also gave better results than $F_{4.5}$ and no feedback and from iterations 2 - 4, it was better than the Best Combination baseline.

A third approach used the same technique as the two previous attempts but with an initial ranking given by *idf*, rather than the baseline ranking. From Table 19 (Ratio 3, column 7) this also performs better than the baseline measures.

The only difference between these three approaches is how the initial ranking was created. The difference between these results demonstrates that better initial rankings can give better overall results after feedback.

In section 2.6 we proposed several reasons for varying the uncommitted belief of a term characteristic. In section 3 we weighted each characteristics to reflect its *strength*, in this section so far we weight each characteristic according to its *quality*. We can combine these two sources of uncertainty by weighting each characteristic by the product of these two weights; the ones derived from feedback and the ones reflecting the relative effectiveness of the characteristic. Table 19 (Ratio 4, column 8) shows that this method performs better than the three baselines and better than the other three ratio methods. The difference over the initial ratio method (Column 5) was slight but consistent.

Weighting characteristics by how well they discriminate can, then, improve feedback without any other query modification.

5.5 Experiment two - relevance feedback using selective combination of evidence

In previous work, (Ruthven and Lalmas, 1999), we demonstrated that an important aspect of query modification was *selecting* which characteristics of terms were important in detecting relevance. The previous experiment weighted characteristics of terms according to their discriminatory power in detecting relevance. In this experiment we ignore characteristics of terms that are poor discriminators of relevant material.

We calculate for each term the average score for each characteristic in the relevant and non-relevant set, e.g. the average *tf* for term t_1 in relevant documents, the average *tf* for term t_1 in non-relevant documents. If the average score in the relevant documents is greater then the characteristic will contribute to the document score. If the average score in the relevant documents is less than in the non-relevant documents, then we drop this characteristic of the query term from the query. In this way, we only combine evidence from good discriminators of relevance.

In this experiments we explored several cases,

i. $F_{4.5}$ baseline, with initial ranking given the combination of all characteristics using the default weights, (baseline 3, section 5.2.3), results shown in Table 20, column 2

ii. $F_{4.5}$ baseline, with initial ranking given the combination of all characteristics using no weighting of characteristics, results shown in Table 20, column 3.

$F_{4.5}$ will produce weights for the query terms once we have relevance feedback information, i.e. after an initial ranking. Cases **i** .and **ii.** are designed to investigate the effect of different initial rankings on the performance of $F_{4.5}$.

iii. The combination of all characteristics of all query terms using no weighting of characteristics, results shown in Table 20, column 4

iv. The combination of the characteristics of terms selected by the method outlined above, using no weighting of characteristics, results shown in Table 20, column 5

v. The combination of all characteristics of all query terms using the default weights, results shown in Table 20, column 6

vi. The combination of the characteristics of terms selected by the method outlined above, using the default weights, results shown in Table 20, column 7

Comparing the two versions of $F_{4.5}$ (Columns 2 and 3) and the two versions using no selection of characteristics (Columns 4 and 6), we see again that better initial rankings gives better results at each iteration. The results of selecting term characteristics demonstrate the value of this approach (Columns 5 and 7). All figures for the first selection test (with no weighting of characteristics) are better than either $F_{4.5}$ or no selection. Combining weighting and selection gives the best results overall (Column 7).

CISI						
Iteration	**$F_{4.5}$ Weighting**	**$F_{4.5}$ No Weighting**	**No Weighting No Selection**	**No Weighting Selection**	**Weighting No Selection**	**Weighting Selection**
0	**11.7**	10.3	10.3	10.3	**11.7**	**11.7**
1	8.3	7.2	10.3	12.0	11.7	**13.1**
2	8.3	7.2	10.3	12.3	11.7	**13.3**
3	8.5	7.3	10.3	12.3	11.7	**13.4**
4	8.5	7.3	10.3	12.3	11.7	**13.5**

Table 20: Average precision figures for selection experiments. Highest values at each iteration shown in bold.

We should note that, although both weighting and selection of characteristics gives the highest performance overall, the increase in effectiveness by using selection as a percentage of the original ranking is greater than that gained by weighting alone. Table 21 column 2 shows the percentage increase, over the original ranking, at each iteration when we use selection of characteristics and no weighting over no weighting and no selection (Table 20, column 4). Table 21, column 3 shows the percentage increase at each iteration when we use selection and weighting over weighting alone (Table 20, Column 5).

	CISI	
Iteration	**Selection and no weighting**	**Selection and weighting**
0	0.00	0.00
1	16.50	11.97
2	19.42	13.68
3	19.42	14.53
4	19.42	15.38

Table 21: Percentage increase in effectiveness over the original ranking when incorporating selection

Selecting term characteristics on a query-query basis, then, can improve retrieval effectiveness over what we can achieve from weighting alone, and over the best individual combination of characteristics.

5.6 Experiment three - relevance feedback based on full model

Our final experiment explores the method of combination of evidence; either only using values of characteristics derived from indexing (as in section 3) or combining these values according to the model outlined in section 4.

In this experiment we again compared four combinations of weighting and selection (no weighting and no selection, Table 22 column 5; selection and no weighting, column 6; no selection and weighting, column 7; selection and weighting, column 8). The baselines are shown in Table 22, columns 2 - 4.

CISI							
Iteration	**No feedback**	**Best combination**	**$F_{4.5}$**	**DS no selection no weighting**	**DS selection no weighting**	**DS no selection weighting**	**DS selection weighting**
0	11.7	**12.9**	11.7	10.3	10.3	11.7	11.7
1	11.7	**12.9**	8.3	10.2	11.9	11.5	**12.9**
2	11.7	12.9	8.3	10.2	12.2	11.5	**13.2**
3	11.7	12.9	8.5	10.2	12.2	11.5	**13.2**
4	11.7	12.9	8.5	10.2	12.3	11.6	**13.3**

Table 22: Results of using full DS model. Highest average precision figures are shown in bold.

Comparing the performance of the baselines in Table 22, we can see that the Best Combination method performed most effectively overall, with the no feedback baseline outperforming the $F_{4.5}$ measure.

The results of our model of RF again show the merits of weighting and selecting characteristics of terms, with the biggest increase in average precision given by the combination of weighting and selection. Comparing these results against those obtained in sections 5.4 and 5.5 we see that this model slightly decreases performance in most cases, and only one of the four versions (column 8) outperforms the Best Combination baseline. when we use selection of characteristics, the model performs fairly well as a RF technique.

5.7 Summary

In this section we summarise the results of these experiments under three conditions:

i. *weighting of characteristics.* Incorporating evidence on the relative importance of terms is important for two reasons. Firstly, it will generally improve initial rankings, bringing more relevant documents higher up the ranking. This means that more relevant documents are likely to come into the documents we use for query modification and so increase the evidence we have to differentiate relevant documents from irrelevant ones. Secondly, as shown in section 5.4 we can use the discriminatory power of a term in discriminating relevant and non-relevant documents to weight characteristics to give improved retrieval of relevant documents. Combining more than one source of uncertainty of term characteristics can improve retrieval effectiveness even more than when only using one source.

This latter finding is significant as it demonstrates that incorporating information on the various sources of uncertainty in the retrieval process can improve retrieval effectiveness. This combination of uncertainty is an important aspect of our DS model, and the use of a formal model, such as DS, means that we can start isolating exactly how the different sources of uncertainty affect retrieval effectiveness.

ii. *selection of characteristics.* Selecting good characteristics of terms - those that are more likely to retrieve relevant documents than irrelevant ones also improves retrieval effectiveness, section 5.5. Combining this information with weighting can improve retrieval effectiveness even more than either technique alone. The weighting of characteristics incorporates the uncertainty regarding the evidence we use in combination, the selection procedure dictates to what evidence the combination is applied. This reflects back to the work described in section 1.1 by Belkin et al, who suggest evidence combination should be tailored to individual queries. This is one aspect of such a tailoring process.

iii. *method of combining evidence.* Our final experiment compared the effect of treating relevance information from the user as an additional source of evidence, as outlined in section 4, against query modification alone. The results from this experiment were not as effective as we hoped, in that incorporating relevance feedback information in the way we implemented it, tended to decrease performance. This may be because our model is not yet sophisticated enough in the manner in which it handles user relevance information. However the particular model we outlined in section 4 is only one method of exploiting relevance feedback information, and the general approach to RF is still valid. The use of such a formal model allows us, however, to analyse where and in what way individual interpretations of this model are successful. This is the subject of ongoing research.

6. Characteristics used in feedback

One important factor in the success of our RF approach was the weighting of characteristics. We were interested in whether there were any differences in which characteristics were chosen to represent query terms when used or did not use weighting. In this section we compare which characteristics were selected when we weight characteristics (Table 22, column 8) and when we do not weight characteristics (Table 22, Column 6) in the final experiment.

Table 23 indicates how many times each characteristic was selected. In both cases the percentage of use of each characteristic was relatively similar and the relative ordering of use was also similar; *theme* most often, followed by *context* and then *tf*.

CISI			
No weighting of characteristics			
Total	***tf***	***theme***	***context***
33440	3259 (39%)	3603 (43.1%)	3373 (40.3%)
Weighting of characteristics			
Total	***tf***	***theme***	***context***
33440	3246 (38.8%)	3544 (42.4%)	3357 (40.2%)

Table 23: Number of times a characteristic was used to describe a query term in relevance feedback

Table 24 gives the number of times a particular combination of characteristics was used to describe a query term. Again, there are strong similarities between the two cases; each combination was used a similar proportion of times when using weighting or no weighting, with combinations of all characteristics, *idf* and *theme*, or *idf*, *tf* and *context* being most used in either case.

CISI						
No weighting						
idf + tf	***idf + theme***	***idf + context***	***idf + tf + theme***	***idf + tf + context***	***idf + theme + context***	***idf + tf + theme + context***
139 (1.4%)	1440 (14.1%)	214 (2.1%)	88 (0.9%)	1084 (10.6%)	127 (1.2%)	1948 (19.0%)
Weighting						
idf + tf	***idf + theme***	***idf + context***	***idf + tf + theme***	***idf + tf + context***	***idf + theme + context***	***idf + tf + theme + context***
137 (1.4%)	1393 (13.9%)	209 (2.1%)	77 (0.8%)	1076 (10.7%)	115 (1.15%)	1930 (19.2%)

Table 24: Number of times each combination of characteristic was used to describe a query term in relevance feedback

These two tables show that, although different documents may being retrieved when characteristics are weighted similar query modification is taking place: similar term characteristics are being selected under both conditions. A slightly higher number of term characteristics are being chosen when we do not weight the characteristics which may reflect differences in the performance of the two conditions.

7. Conclusion

In IR, we may have a potentially large numbers of users who are unfamiliar with electronic searching. It is therefore important that systems are as effective as possible in targeting 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. The core of this approach is the combination of evidence from algorithms describing the information use of terms and relevance information from users. This model is based on Dempster-Shafer's Theory of Evidence which 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 feedback 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).

We also showed how the notion of uncommitted belief can be used to represent and combine various sources of uncertainty in the RF process. These aspects are described in sections 2.6 and 4.1, and are summarised in Table 25.

Characteristic	**Term**	**Document**
uncertainty	importance	partial relevance assessment
imprecision	source	
quality		time of assessment
strength		

Table 25: Sources of uncertainty that can be incorporated via the uncommitted belief of a mass function

These sources of uncertainty arise from different parts of the retrieval process: indexing the documents, retrieval of documents, relevance feedback and how the user assesses documents. In our model we can incorporate them into a unified framework.

Our approach of including information on how terms are used within documents can increase the flexibility of IR 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, (Ruthven et al., 1999). However we believe that we have demonstrated at a theoretical and experimental level the suitability and flexibility of using characteristics of information use and their combination. We have also shown that the Dempster-Shafer approach can capture many important aspects of this combination, in particular the representation and manipulation of the uncertainty involved in relevance feedback.

Acknowledgements

We would like to thank Keith van Rijsbergen for his helpful comments on this work. We are also grateful the anonymous reviewers for their comments and advice. This research is supported in part by the British Library and Information Commission funded project '*Retrieval through explanation*' (http://www.dcs.gla.ac.uk/ir/projects/explanation/).

References

(Belkin et al., 1995) N. J. Belkin, P. Kantor, E. A. Fox and J. A. Shaw. *Combining the evidence of multiple query representations for information retrieval*. Information Processing and Management. **31**. 3. pp 431-448. 1995.

(Barry and Schamber, 1998) C.L. Barry and L Schamber. *Users' criteria for relevance evaluation: a cross-situational comparison.* Information Processing and Management. **34**. 2/3. pp 219 - 236. 1998.

(Borlund and Ingwersen, 1997) P. Borlund and P. Ingerwersen. *The development of a method for the evaluation of interactive information retrieval systems*. Journal of Documentation. **53**. 5. pp 225 - 250. 1997.

(Campbell and Van Risjsbergen, 1996) I. Campbell and C. J. van Rijsbergen. *Ostensive model of information needs.* Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective (CoLIS 2). Copenhagen. pp 251-268. 1996.

(Chang et al., 1971) Y K Chang, C Cirillo and J Razon. *Evaluation of feedback retrieval using modified freezing, residual collection & test and control groups*. The SMART retrieval system - experiments in automatic document processing. G. Salton (ed). Chapter 17. pp 355-370. 1971.

(Dempster, 1968) A. P. Dempster. *A generalization of the Bayesian inference.* Journal of Royal Statistical Society. **30**. pp 205 - 447. 1968.

(Denos et al., 1997) 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. Toulouse. pp 214 - 224. 1997.

(Ellis, 1989) D. Ellis. *A behavioural approach to information system design.* Journal of Documentation. **45**. 3. pp 171-212. 1989.

(Haines and Croft,1983) 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. Pittsburgh. pp 2 - 11. 1993.

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

(Ingwersen, 1994) P. Ingwersen. *Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction.* Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 101-110. Dublin. 1994.

(Lalmas and Ruthven, 1998) M. Lalmas and I. Ruthven. *Representing and Retrieving Structured Documents using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation.* Journal of Documentation. **54.** 5. pp 529 - 565. 1998

(Lee, 1998) J. H. Lee. *Combining the evidence of different relevance feedback methods for information retrieval.* Information Processing and Management. **34**. 6. pp 681-691. 1998.

(Robertson and Sparck Jones, 1976) S. E. Robertson and K. Sparck Jones. *Relevance weighting of search terms.* Journal of the American Society of Information Science. **27**. pp 129 - 146. 1976.

(Rocchio, 1971) J. J. Rocchio. *Relevance feedback in information retrieval.*The SMART retrieval system: experiments in automatic document processing. (G. Salton, ed.). Ch. 14. pp 313 - 323. Prentice-Hall.

(Ruthven and Lalmas, 1999) I. Ruthven and M. Lalmas. *Selective relevance feedback using term characteristics.* Proceedings of the Third International Conference on Conceptions of Library and Information science. CoLIS 3. Dubrovnik. 1999.

(Ruthven et al., 1999) I. Ruthven, M. Lalmas and C. J. van Rijsbergen. *Retrieval through explanation: an abductive inference approach to relevance feedback.* 10th Irish Conference on Artificial Intelligence & Cognitive Science. Cork. 1999.

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

(Salton and Buckley, 1990) G. Salton and C. Buckley. *Improving retrieval performance by relevance feedback.* Journal of the American Society for Information Science. **41**.4**.** pp 288 - 297. 1990.

(Shafer, 1976) G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press. 1976.

(Schocken and Hummel, 1993) S. Schocken and R. A. Hummel*. On the use of the Dempster Shafer model in information indexing and retrieval applications*. International Journal of Man-Machine Studies. **39**. pp 1 - 17. 1993.

(Silva et al., 2000) I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura and N. Ziviani. *Link-based and content-based evidential information in a belief network model.* Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. Athens. pp 96 – 103. 2000.

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

(da Silva and Milidiu, 1993) W. Teixeira da Silva and R. Luiz Milidiu. *Belief function model for information retrieval.* Journal of the American Society for Information Science. **44**. 1. pp 10 - 18. 1993.

(Vakkari, 2000) P. Vakkari. *Relevance and contributing information types of searched documents in performance.* Proceedings of the 23rd ACM Sigir Conference on Research and Development in Information Retrieval. Athens. pp 2 – 9. 2000.

(Van Rijsbergen, 1979) C. J. van Rijsbergen. *Information retrieval.* Butterworths. 2nd edition. 1979.

(Van Rijsbergen, 1992) C. J. van Rijsbergen. *Probabilistic retrieval revisited.* The Computer Journal. **35**. 3. pp 291 - 298. 1992.

(Voorhees and Harman, 1996) E. M. Voorhees and D. Harman. *Overview of the Fifth Text REtrieval Conference (TREC-5).* Proceedings of the 6th Text Retrieval Conference. Gaitherburg. pp 1-28. Nist Special Publication 500-238. 1996.