ON THE USE OF THE DEMPSTER SHAFER MODEL

IN INFORMATION INDEXING AND RETRIEVAL APPLICATIONS

Shimon Schocken
Department of Information Systems
Stern School of Business
New York University

Robert A. Hummel
Courant Institute of Mathematical Sciences
New York University

Forthcoming in the:
"INTERNATIONAL JOURNAL OF MAN-MACHINE STUDIES"

Working Paper Series

STERN IS-93-16

*Replaces: IS-92-27

On the Use of the Dempster Shafer Model in Information Indexing and Retrieval Applications

Shimon Schocken

Stern School of Business, New York University

Robert A. Hummel

Courant Institute of Mathematical Sciences, New York University

May 27, 1993

On the Use of the Dempster Shafer Model in Information Indexing and Retrieval Applications

The Dempster Shafer theory of evidence concerns the elicitation and manipulation of degrees of belief rendered by multiple sources of evidence to a common set of propositions. Information indexing and retrieval applications use a variety of quantitative means – both probabilistic and quasi-probabilistic – to represent and manipulate relevance numbers and index vectors. Recently, several proposals were made to use the Dempster Shafer model as a relevance calculus in such applications. The paper provides a critical review of these proposals, pointing at several theoretical caveats and suggesting ways to resolve them. The methodology is based on expounding a canonical indexing model whose relevance measures and combination mechanisms are shown to be isomorphic to Shafer's belief functions and to Dempster's rule, respectively. Hence, the paper has two objectives: (i) to describe and resolve some caveats in the way the Dempster Shafer theory is applied to information indexing and retrieval, and (ii) to provide an intuitive interpretation of the Dempster Shafer theory, as it unfolds in the simple context of a canonical indexing model.

Keywords: Theory of evidence, Dempster Shafer model, relevance measures, information indexing and retrieval.

Running Head: *On the Use of the Dempster Shafer Model*

1 Introduction

Consider a finite and exhaustive set of mutually-exclusive propositions and a body of evidence that supports some subsets of propositions and discounts others. Many theories were put forward to describe how one should represent and update one's degrees of belief in such propositions when new or additional evidence is brought to bear. The classical approach is to cast degrees of belief as probabilities - a set of numbers between 0 and 1 that obeys the axioms of subjective probability - and use Bayesian inference rules to revise them in light of new evidence. One problem with this approach is that it does not offer a clear way to model the various degrees of 'uncommitted beliefs,' or 'second order uncertainties,' that characterize most realistic inference problems. For example, consider the extreme case of 'insufficient reason,' in which one knows absolutely nothing about a given set of n propositions. The common solution, which goes back to LaPlace, is to assign a degree of belief of $1/n$ to each of the propositions under consideration. Incidently, this is also the solution that emerges from maximizing the unconstrained entropy function associated with the n unknown probabilities.

Over the years, many students of belief revision theories have objected to this crude quantification of insufficient reason. Why, the argument goes, should ignorance be translated to the strong statement that every proposition (or state of nature) is equally likely? This criticism has led to several alternative models that attempt to capture the elusive notion of uncommitted belief by modifying the axiomatic framework of probability theory. Perhaps the best known model in this category is the 'theory of evidence,' originated by Dempster's (1967, 1967a) work on upper and lower probabilities. Dempster's ideas, which were based on a frequentist view of inference, were refined and extended by Shafer (1976), who also gave them a subjective interpretation. This led to the Dempster Shafer model – an elaborate formalism for representing and revising degrees of support rendered by multiple sources of evidence to a common set of propositions[1].

When the work of Dempster and Shafer was 'discovered' by the artificial intelligence community, it immediately stirred a considerable interest in two application areas in which normative models of belief formation play a key role: expert systems, and information indexing and retrieval systems. For expert systems, the Dempster Shafer (DS) model provides a mathematically-sound model for representing and manipulating rule-based degrees of belief, an area that was traditionally dominated by ad-hoc belief revision calculi whose relationship to probability theory was at best murky. For information indexing and retrieval systems, the DS model can be used as a *relevance calculus*, designed to quantify and revise the degrees of association between documents, keywords, and user-supplied queries.

This line of thought has led to the development of several DS-based information indexing and retrieval applications. For example, Biswas, Bezdek, Marques, and Subramanian (1987) built a document retrieval system in which the relevance of documents to taxonomical classes was measured and manipulated, respectively, by belief functions and Dempster's rule: *"We choose to define similarity functions based on the Dempster Shafer theory of evidence ... one of the advantages of this approach is that it reflects the process of belief revision and updating just as in human reasoning processes."* (Biswas et al, 1987). Coming from a different direction, Turtle and Croft (1991) describe a canonical representation in which relevance is handled through inference networks that are structured as directed

[1] In this paper, the terms the *Dempster Shafer theory of evidence* and the *Dempster Shafer model* are used interchangeably.

acyclic graphs. The nodes in the networks correspond to keywords, documents, and queries, and *"the arcs joining the nodes are interpreted as assertions that the parent node provides support for the child node."* Turtle and Croft proposed to operationalize these degrees of support through either subjective probabilities, or DS belief functions. A similar approach was undertaken in RUBRIC, a full-text information retrieval system described by Tong and Shapiro (1985). RUBRIC can be instantiated to operate with several alternative relevance calculi, the DS model being a prime example.

The importance of such applications is obvious, as they attempt to take the DS model 'out of the lab' and implement it in realistic settings. In doing so, however, many adopters of the DS model have taken the model's validity for granted, either explicitly or implicitly. With that in mind, it is important to point out that both the cognitive and the normative roots of the DS model are still a matter of intense controversy: whereas Shafer (1987) argues that the theory of evidence is a natural extension of probability theory, many critics, e.g. Lindley (1987), view it as a reformulated version of a specialized, albeit interesting, case of classical probabilistic inference. The debate is not helped by the somewhat forbidding notation of the DS model, which prevents an intuitive understanding of its underlying structure and philosophy.

In fact, the gap between the theory and practice of the DS model seems to be two-directional. On the one hand, many practitioners believe that the normative correctness of the DS model is a 'closed case,' proceeding to implement it without questioning its underlying rationale. On the other hand, many researchers try to defend the DS model on abstract philosophical or mathematical grounds, without realizing that simpler justifications can be found *in the field*, i.e. in the way the model is actually used in certain canonical settings.

The latter point is worth emphasizing: a close examination of certain applications of the DS model can provide not only a better understanding of the model, but, furthermore, a compelling normative justification.

The research reported here builds on previous work by Schocken and Pyun (1990) and by Hummel and Landy (1988). Schocken and Pyun presented a simple DS-based relevance calculus which operated over a hierarchical space of keywords. Although the approach involved a frequentist interpretation of DS mass and belief functions, no attempt was made to interpret or justify the theory's algebra – Dempster's rule – using a domain semantics. One such interpretation was given by Hummel and Landy, who viewed the rule as analogous to a certain mechanism for pooling expert opinions. However, the Hummel and Landy work was an abstract mathematical analysis, detached from specific domains of application. The present research integrates and extends these two papers, resulting in a complete (although not unique) interpretation of the DS theory, as it unfolds in the context of an information indexing and retrieval (IR) application.

The plan of the paper is depicted in figure 1 and described as follows. §2 presents the notion of index vectors and the challenge of eliciting and measuring relevance in a normative, rather than ad-hoc, fashion. §3 gives an overview of the DS model and illustrates how it is commonly used in the context of IR systems. This sets the stage to four critical questions regarding the theoretical fit between the general features of the DS model and the specific requirements of IR applications. In order to answer these questions, §4 presents a canonical indexing model in which the notions of keywords, taxonomies, and relevance, are treated formally and unambiguously. It is then shown that the canonical model expounded in §4 is isomorphic to the DS model, leading to a new intuitive understanding of the latter. §5 offers concluding remarks about the implications of the research on the DS model and on IR applications. The paper ends with two appendices. Appendix A presents an extension of the DS model that enables the handling of conjunctive indexing opinions (the need for this extension is discussed at the end of §4.1). Appendix B gives the proofs of the propositions presented in the body of the paper.

Put figure 1 around here

2 The Problem

Models of bibliographical indexing concern the construction of data structures that enable rapid keyword-based access to vast collections of documents. Given a document, on the one hand, and a *keyword-list*, on the other, the goal of the indexing model is to select a subset of keywords that 'best' describes the document to its potential users. This is done either by human catalogers or by automatic keyword extraction algorithms. Since some keywords are more relevant to the document than others, a numeric scale is often used to express the strength of association between the document and the selected keywords. The result is an *index vector*, consisting of pairs of keywords and their respective 'relevance numbers.' Several models exist for representing and manipulating such indexing vectors, and the reader is referred to Salton and McGill (1983) and to Salton and Buckley (1988) for comprehensive treatments of the general approach to the subject.

Relevance is an elusive concept that defies simplistic definitions. Generally, the term is taken to refer to a relation between a document and an information need, and as such it is strictly definable only by a library patron (a searcher), in the context of a particular search process. In this paper, however, when we say 'relevance' we refer to a static association relation between documents and keywords. That is, we say that document d is relevant to keyword k if we believe that k is a good descriptor of d; Similarly, the term 'relevance number' is used to represent the intensity of that relation. When there will be a potential for confusion between this type of relevance and the relevance feedbacks obtained from library patrons, we will make the distinction explicit.

Formally, let D be a set of documents about a certain domain of interest, and let $\mathcal{K} = \{k_1, \ldots, k_m\}$ be a list of domain keywords. The index of each document $d \in D$ is a set of pairs of the form:

$$S_d = \{(K_1, r_1), \ldots, (K_n, r_n)\} \tag{1}$$

where $K_i \subseteq \mathcal{K}$ and $0 \leq r_i \leq 1, i = 1, \ldots, n$. The K_i's are *lexical subsets*, representing different groupings of keywords, and the r_i's are *relevance numbers.* Taken together, the pair $(K_i, r_i) \in S_d$ says that the degree of relevance between the document d and the lexical subset K_i is r_i. Had we restricted the K_i's to be singletons only, (1) would become the 'term-weight vectors' that are normally used in information indexing and retrieval applications. Further, had we required that all the r_i be 0 or 1 only, (1) would be reduced to the familiar index terms (also called 'subject headings') that are normally used to classify articles in professional journals. Given the obvious simplicity of a *Boolean* indexing scheme, why bother about developing formalisms for *weighted* indexing?

The answer is that relevance is a subjective and composite relation which is often based on aggregating several indexing opinions that come from different sources. Specifically, each document has many *classifiers*, or discerning characteristics, that determine its relevance to different keywords. For example, the *title* of a document can suggest one index, whereas the *abstract* can suggest another. Other aspects of the document, obtained through lexical, linguistic, and citations, analyses will yield additional indexing opinions that must be taken into consideration. Hence, even if the individual opinions were forced to be binary, their aggregation would probably induce a continuous index. In addition, many indexing opinions are not cast automatically; rather, they are elicited from human catalogers who inject yet another level of uncertainty and subjectivity to the indexing process. That is, when two or more catalogers are asked to index the same document, they may well supply different (but hopefully overlapping) indexing opinions. Indeed, empirical research indicates that a great deal of indexing inconsistency characterizes novice as well as well-trained catalogers (Jacoby & Slamecka, 1962, Stevens, 1965).

Different IR applications use different models to handle this pluralism, and the validity of these models can be studied on empirical as well as on normative grounds. From an *empirical* perspective, an IR model must perform well in terms of functional criteria such as recall and precision, and whether or not the model makes sense on normative grounds is of a lesser importance. From a *normative* perspective (to which this paper belongs), the credibility of an IR model hinges on its capacity to elicit, represent, and synthesize, relevance opinions in a *formal*, rather than *ad-hoc*, fashion. In order to do so, the relevance numbers and the rules that combine them must be given a compelling interpretation. So far, the leading normative interpretation of relevance has been probabilistic. Beginning with the seminal work of Maron and Kuhns (1960), probabilistic IR models were developed by Bookstein (1983), Cooper & Huizinga (1982), Cooper and Maron (1978), Croft & Harper (1979), Harter (1975), Radecki (1988), Robertson & Sparc Jones (1976), Thompson (1990), and Yu & Salton (1976), among others. Recently, however, several attempts were made to handle relevance in IR applications using the Dempster Shafer model, which is widely considered to be a less restricted extension of probability theory. The strengths and weaknesses of the latter approach are discussed in the next section.

3 A Dempster Shafer Indexing Model

The DS theory of evidence concerns the representation and manipulation of degrees of support rendered by different sources of evidence to a common set of propositions, denoted θ and called the *frame of discernment.* In contrast to a standard Bayesian design, in which degrees of support are normally assigned to elements of θ directly, the DS model assigns degrees of support to *subsets* of propositions, i.e. to members of the power-set 2^{θ}, also called 'possibilities.' The DS model offers several complementary ways to express evidential support in possibilities. In particular, the model defines three mappings from 2^{θ} to $[0,1]$ termed *mass*, *belief*, and *plausibility*, functions. The three mappings are mathematically equivalent in the sense that knowledge of any one of them (for every possibility) can be used to compute the other two. Therefore, we view them as alternative means to keep score of the same primitive set of degrees of support. In the standard model, when several sources of evidence support a common set of possibilities (the support can be cast in terms of either mass, belief, or plausibility functions), the overall support in the possibilities is computed through Dempster's rule of combination.

What is the nexus of the DS model and information indexing and retrieval applications? In one way or another, all DS-based IR applications are based on the following premises: (i) The DS notion of *degrees of support* can be used to operationalize the IR notion of *relevance numbers*; and (ii) When two or more classification criteria supply different sets of relevance numbers concerning the same document, Dempster's rule provides a plausible mechanism to combine them into a composite index (said otherwise: revise the relevance of the document to certain keywords in light of new evidence). The goal of this section is to motivate a critical analysis of these premises. Specifically, we intend to:

- Provide a rigorous but accessible overview of the DS model, as it unfolds in the familiar context of an IR application;
- Present a series of questions regarding the *theoretical fit* between the general features of the DS model and the special requirements of IR applications.

The Frame of Discernment: The frame of discernment θ is an exhaustive set of mutually exclusive elements that can be interpreted as hypotheses, propositions, or simply 'labels.' The power-set that contains all the subsets of θ (including θ itself and the empty set) is denoted 2^θ. In general, the semantics of the labels depends on the context in which the DS model is applied. In information indexing and retrieval applications, the frame of discernment is normally taken to be a keyword-list $\mathcal{K} = \{k_1, \ldots, k_n\}$. If we fix d on a particular document, each keyword k_i can be interpreted as the proposition "*k_i is relevant to d.*" As we will see later, a DS-based indexing model seeks to compute the degrees of support in such propositions in light of different bodies of evidence. To illustrate the lexical interpretation of a frame of discernment, a keyword-list that supports a collection of documents about modern art might be $\mathcal{K} = \{\texttt{Arp}, \texttt{Braque}, \texttt{Cezanne}, \ldots,$ $\texttt{Zorn}\}$, enumerating all the major artists of the Twentieth Century. The power-set in this case is $2^{\mathcal{K}} = \{\{\texttt{Arp}\}, \{\texttt{Braque}\}, \{\texttt{Cezanne}\}, \ldots, \{\texttt{Arp}, \texttt{Braque}\}, \{\texttt{Arp}, \texttt{Cezanne}\}, \{\texttt{Braque},$ $\texttt{Cezanne}\}, , \ldots, \{\texttt{Arp}, \texttt{Braque}, \texttt{Cezanne}\}, \ldots, \emptyset, \mathcal{K}\}$, the last two elements being the empty set and $\mathcal{K}$ itself. Each element in $2^{\mathcal{K}}$ represents a disjunction of keywords, denoted hereafter a *lexical subset.* The act of indexing a document using $\mathcal{K}$ amounts to choosing, among all the indexing possibilities in $2^{\mathcal{K}}$, the one or more lexical subsets that best describe the document to its potential users.

For example, suppose that an art scholar is asked to index the document *"The Influence of Cezanne on early Cubism"* using $\mathcal{K}$, based on partial information such as the document's title or abstract. Without loss of generality, assume that (i) the main focus of the document is Cezanne; and (ii) the only Cubist artists in the current keyword-list are Braque and Picasso. Under these assumptions, the scholar will probably supply an index of the form $S = \{(\{\texttt{Cezanne}\}, r_1), (\{\texttt{Braque}, \texttt{Picasso}\}, r_2)\}$, with $r_1 > r_2$. This would entail the following information: (i) the document is relevant to Cezanne; (ii) it is also relevant, to a lesser extent, to either Braque or to Picasso. This is quite different from the indexing opinion $S' = \{(\{\texttt{Cezanne}\}, r_1), (\{\texttt{Braque}\}, r_2), (\{\texttt{Picasso}\}, r_2)\}$, which would be more appropriate if the document's title were, say, *"The Influence of Cezanne on the early work of Braque and Picasso"*.

We arrive at our first question:

> **Question Q1:** When the DS model is applied to information indexing and retrieval applications, the keyword-list $\mathcal{K}$ is taken to be the *frame of discernment*, and indexing possibilities are taken to be elements of the lexical

power-set $2^{\mathcal{K}}$. What are the taxonomical implications and limitations of this representation?

To motivate this question, consider again the document *"The Influence of Cezanne on early Cubism"*. Note that the most reasonable index of this document would be $S'' = \{(\{\texttt{Cezanne}\}, r_1), (\{\texttt{Cubism}\}, r_2)\}$, especially if the document's abstract makes no references to specific artists other than Cezanne. However, `Cubism` is not an element of the original keyword-list $\mathcal{K}$, so it does not entail an indexing possibility. To solve the problem, we may want to extend the original frame of discernment, creating a new keyword-list of the form $\mathcal{K}' = \mathcal{K} \cup \{\texttt{Cubism}\}$. However, the keywords `Braque`, `Picasso` and `Cubism`, have a great deal in common from a bibliographical standpoint. Therefore, $\mathcal{K}'$ is not a valid frame of discernment, because some of its elements are no longer mutually exclusive. Before we present a solution to this problem, we have to be very specific about the proper relationship among *frames of discernment*, *keyword-lists*, and *taxonomies of classes.* We will return to this issue in section 3, where an answer to `Q1` is given.

<u>**Mass Functions:**</u> A mapping $m : 2^\theta \rightarrow [0,1]$ with the properties:

$$m(\emptyset) = 0 \tag{2}$$

$$\sum_{X \in 2^\theta} m(X) = 1 \tag{3}$$

is called a *mass function*[2]. In the DS model, the mass $m(X)$ represents the degree to which a certain source of evidence supports the possibility X, where $X \subseteq \theta$. As a convention, the mass which is 'left over' after all the *proper* subsets of θ have been assigned masses

[2] Throughout the paper, upper-case variables refer to sets and lower-case variables refer to scalars.

is allocated to θ itself and denoted the *uncommitted belief* displayed by m, or $m(\theta)$. In DS-based IR applications, where θ is taken to be a keyword-list $\mathcal{K}$, the mass $m(X)$ is taken to represent (to a first approximation that will be discussed shortly) a degree of relevance, or, more accurately, the degree of belief that the document is relevant to the lexical subset $X \subseteq \mathcal{K}$, according to a certain classifier. Hence, if a classifier (say, classifier number 1) supplies the indexing opinion $S_1 = \{(\{\texttt{Cezanne}\}, 0.6), (\{\texttt{Braque}, \texttt{Picasso}\}, 0.3)\}$, then the mass function that is induced by this opinion is defined as follows:

$$\begin{array}{rcl} m_1(\{\texttt{Cezanne}\}) & = & 0.6 \\ m_1(\{\texttt{Braque}, \texttt{Picasso}\}) & = & 0.3 \\ m_1(\mathcal{K}) & = & 0.1 \\ m_1(X) & = & 0 \ \ for\ all\ other\ proper\ subsets\ of\ \mathcal{K} \end{array} \tag{4}$$

Note that the uncommitted belief induced by the opinion is assigned by default to the frame of discernment by means of $m_1(\mathcal{K}) = 1 - 0.6 - 0.3 = 0.1$. The rationale for this assignment is as follows. If a certain classifier provides no information whatsoever about indexing possibilities, the classifier's 'ignorance' can be represented by the index $S = \{(\mathcal{K}, 1)\}$. This implies the mass function $m(\{\texttt{Arp}, \texttt{Braque}, \texttt{Cezanne}, \ldots, \texttt{Zorn}\}) = 1$ and $m(X) = 0$ elsewhere, reflecting the (not very useful) opinion that the document is relevant to Arp, or to Braque, or to Cezanne, or to any other artist in the keyword-list. Other classifiers can provide more focused relevance opinions, resulting with lower levels of $m(\mathcal{K})$. Hence, unlike a standard probabilistic design, where the notion of uncommitted belief is not well-defined, the DS model provides explicit means to quantify and manipulate it via $m(\mathcal{K})$. Although uncommitted beliefs, or 'second-order uncertainties,' can and have been treated in the standard framework of subjective probability, (e.g. Baron, 1987), there is no *simple* way to do it. The theory of evidence is unique in that it treats the notion of uncommitted belief explicitly, at the axiomatic level.

It is important to observe that mass functions represent indivisible, or atomic, degrees of belief. For example, the magnitudes of $m(\{\texttt{Braque},\texttt{Picasso}\})$, $m(\{\texttt{Braque}\})$, and $m(\{\texttt{Picasso}\})$ are unrelated, and a mass function like $m(\{\texttt{Braque},\texttt{Picasso}\}) = 0.9$, $m(\{\texttt{Braque}\}) = 0$, and $m(\{\texttt{Picasso}\}) = 0$ is not inconsistent with the theory. This particular function represents a cataloger who strongly believes that the document is relevant to either Braque or to Picasso, although he is not willing to say anything more specific beyond this assessment.

But what does this notion of relevance *mean*? We arrive at our next question:

> **Question Q2:** A mass function is a formal, domain-independent, component of the DS model. Relevance is an informal, but highly intuitive, concept that plays a key role in information indexing and retrieval applications. If a mass function is taken to represent relevance, then what is the semantics of this representation? Said otherwise, what *type* of relevance do mass functions represent?

Question Q2 suggests the premise that mass functions are not necessarily a natural representation of the intuitive notion of relevance, as it is typically construed in information indexing and retrieval applications. To illustrate this reservation, consider the following example. If mass functions are used to represent relevance, then the relevance numbers in each index must sum up to 1. That is, the set of allowable indexing opinions $\{(K_1, r_1), \ldots, (K_n, r_n)\}$ is constrained by $\sum_1^n r_i = 1$. Many would argue that this constraint does not make sense, and that an indexing opinion like, say, $\{(\{\texttt{Albers}\},0.8), (\{\texttt{Kandisnki}\},0.4), (\{\texttt{Klee}\},0.4)\}$ is perfectly reasonable. The only 'wrong' thing about this opinion is that it is inconsistent with the DS notion of a mass function, but this seems to be a limitation of the model's application, not of the opinion.

One pragmatic solution is to treat the relevance numbers not as *absolute*, but rather as *relative*, measures of subjective relevance. According to this position, the two indexes $S = \{(A, 0.8), (B, 0.4,)(C, 0.4)\}$ and $S' = \{(A, 0.4), (B, 0.2), (C, 0.2)\}$ are equally informative, as both imply exactly the same relative information: the document is twice as relevant to A as it is to B, and it is as relevant to B as it is to C. However, this immediately leads to another snag: according to the same principle, the index is also equivalent to $S'' = \{(A, 0.2), (B, 0.1), (C, 0.1)\}$. Yet S' and S'' reflect two different states of uncommitted belief (0.2 and 0.6, respectively), and thus they do not induce the same mass function.

To get around the problem, we can elicit uncommitted beliefs directly from the catalogers[3]. For example, having specified an indexing opinion, say $\{(A, 0.8), (B, 0.4), (C, 0.4)\}$, the cataloger can be asked to rate his confidence in the opinion on a scale of 0 to 1. If the confidence level is 1, the index is normalized to $\{(A, 0.5), (B, 0.25), (C, 0.25)\}$, reflecting an uncommitted belief of 0. If the confidence level is 0.8, the index is normalized to $\{(A, 0.4), (B, 0.2), (C, 0.2)\}$, reflecting an uncommitted belief of 0.2. In general, for any unconstrained indexing opinion $\{(K_1, r_1), \ldots, (K_n, r_n)\}$ and a confidence level $0 \leq c \leq 1$, we can find a unique mass function $\{m(K_1), \ldots, m(K_n), m(\mathcal{K})\}$ such that (i) the $m(K_i)$'s preserve the relative properties of the unconstrained r_i's; and (ii) $m(\mathcal{K}) = 1 - c$.

The shift from an absolute to a relative scale of relevance has several justifications. First, a significant body of psychological evidence indicates that relevance is indeed a relative

[3] In this section, the terms *classifier* and *cataloger* are used interchangeably. The distinction between the two terms is made explicit in the next section.

property (Saracevic, 1975). Second, we must remember that ultimately, an IR application must satisfy the information needs of library patrons, and that relevance numbers should be used pragmatically to that end. For example, according to Maron (1982)'s 'Ranking Principle,' the chief objective of relevance numbers is to present to the patron a set of documents, sorted by decreasing order of perceived relevance to his or her query. A similar principle is used in diagnostic expert systems, where ordinal, rather than cardinal, degrees of beliefs are often used to guide the inference engine to promising directions and to explain the system's reasoning to the people who consult it. If we accept Maron's Ranking Principle as a working assumption, then normalization is not an issue, since rankings are invariant under normalization. However, when multiple indexing opinions are aggregated into a pooled index (something that we have not done yet), normalization becomes a tricky manipulation. Specifically, let S_1 and S_2 be two indexing opinions, $\oplus$ an aggregation operator, and N a normalization operator. In many cases (depending on the specific definitions of $\oplus$ and N), it can be shown that $N(S_1 \oplus S_2) \neq N(S_1) \oplus N(S_2)$, i.e. that N is not homomorphic.

In conclusion, we see that even though relevance numbers can be represented by mass functions, the representation has some theoretical caveats. Clearly, these limitations are related to the fact that we are still lacking explicit domain semantics. That is, we do not know yet what is the exact *meaning* of relevance numbers. This analysis is taken up in section 3, where an answer to question Q2 is presented.

__The Core:__ The *core* of a mass function $m : 2^\theta \rightarrow [0,1]$ is the union of all the possibilities $X \in 2^\theta$ for which $m(X) > 0$. When the frame θ is taken to be a keyword-list $\mathcal{K}$, the core becomes a list of indexing possibilities, in the view of one particular classifier. For example, the core of the mass function induced by classifier 1 (Eqn. 4) is $C_1 = \{\{\texttt{Cezanne}\}, \{\texttt{Braque}, \texttt{Picasso}\}, \mathcal{K}\}$. Suppose now that the *same* document is indexed by another classifier (classifier no. 2), whose indexing opinion is captured by the following mass function:

$$\begin{aligned} m_2(\{\texttt{Picasso}\}) &= 0.8 \\ m_2(\mathcal{K}) &= 0.2 \\ m_2(X) &= 0 \ \ for\ all\ other\ proper\ subsets\ of\ \mathcal{K} \end{aligned} \tag{5}$$

The core of this mass function is $C_2 = \{\{\texttt{Picasso}\}, \mathcal{K}\}$. Is there a credible way to combine the two indexing opinions (4-5) into an aggregate index? As a first approximation, one can focus on all the lexical subsets that both classifiers agree are relevant to the document. In particular, if classifier 1 thinks that X is relevant and classifier 2 thinks that Y is relevant, then *both* classifiers agree that $X \cap Y$ is relevant (recall that both X and Y are interpreted as disjunctions of keywords). This leads to the following definition of a *pooled core*: Let $m_1, m_2 : 2^{\theta} \rightarrow [0, 1]$ be two mass functions with cores C_1 and C_2. The pooled core $C = C_1 \oplus C_2$ will be:

$$C_1 \oplus C_2 = \{X \cap Y \mid X \in C_1 . Y \in C_2, X \cap Y \neq \emptyset\}. \tag{6}$$

For example, the pooled core of $C_1 = \{\{\texttt{Cezanne}\}, \{\texttt{Braque}, \texttt{Picasso}\}, \mathcal{K}\}$ and $C_2 = \{\{\texttt{Picasso}\}, \mathcal{K}\}$ is $C_1 \oplus C_2 = \{\{\texttt{Cezanne}\}, \{\texttt{Picasso}\}, \{\texttt{Braque}, \texttt{Picasso}\}, \mathcal{K}\}$[4]. In general, then, the pooled core can be viewed as a first approximation of the degree of consensus or disagreement displayed by two independent indexing opinions. If $C_1 \oplus C_2 = C_1 = C_2$, we have a consensus regarding which possibilities are likely. If $C_1 \oplus C_2 = \emptyset$, the classifiers agree on nothing. If $C_1 \oplus C_2$ is not empty, we have an overlap of some opinions. Of course the problem of (6) is that it merely *identifies* areas of mutual agreement (or lack thereof)

[4] Note that $\mathcal{K}$ acts as an attractor, in that $A \cap \mathcal{K} = A$ for all $A \subseteq \mathcal{K}$.

between two classifiers. In order to compute the *intensity* of such agreements, a more sensitive pooling mechanism is required. Dempster's rule provides one such mechanism.

Dempster's Rule: The most fundamental (and debateable) pillar of the DS model is the convention that once degrees of support are cast in terms of mass functions, Dempster's rule provides a proper mechanism to combine them. Let m_1 and m_2 be two mass functions defined over the same frame of discernment: $m_1, m_2 : 2^{\theta} \rightarrow [0,1]$, with cores $C_1 = \{A_1, \ldots, A_{n_1}\}$ and $C_2 = \{B_1, \ldots, B_{n_2}\}$, respectively. Dempster's rule computes the pooled mass function $m = m_1 \oplus m_2 : 2^{\theta} \rightarrow [0,1]$ as follows:

$$m'(X) = \sum_{A_i \cap B_j = X} m_1(A_i) \cdot m_2(B_j), \tag{7}$$

$$m(X) = \begin{cases} \frac{1}{1-m'(\emptyset)} \cdot m'(X) & X \neq \emptyset \\ 0 & X = \emptyset \end{cases} \tag{8}$$

The rationale behind (7-8) can be explicated through an 'intersection table.' In our two-classifiers scenario (4-5), the table has the following form:

	$m_1(\texttt{Cezanne}) = 0.6$	$m_1(\texttt{Picasso},\texttt{Braque}) = 0.3$	$m_1(\mathcal{K}) = 0.1$
$m_2(\texttt{Picasso}) = 0.8$	$m'(\emptyset) = 0.48$	$m'(\texttt{Picasso}) = 0.24$	$m'(\texttt{Picasso}) = 0.08$
$m_2(\mathcal{K}) = 0.2$	$m'(\texttt{Cezanne}) = 0.12$	$m'(\texttt{Picasso},\texttt{Braque}) = 0.06$	$m'(\mathcal{K}) = 0.02$

The top row of the table records the mass function of the first classifier excluding its zero elements, i.e. the set of values $m_1(A_1), \ldots, m_1(A_{n_1})$ for elements A_i in the core C_1. The left column of the table records the mass values of the second classifier for its core elements, i.e. the set of values $m_2(B_1), \ldots, m_2(B_{n_2})$ (The curly brackets are dropped for the sake of brevity, e.g. $m(\texttt{Picasso},\texttt{Braque})$ stands for $m(\{\texttt{Picasso},\texttt{Braque}\})$, etc.). Inside the table, the (i,j)'th cell records the pooled mass contributed to $A_i \cap B_j$ by the pair A_i and B_j, which is taken to be the product $m_1(A_i) \cdot m_2(B_j)$. Using these entries and combining cells with equivalent intersections following (7-8), one obtains:

$$\begin{aligned} m'(\texttt{Cezanne}) &= 0.12, \\ m'(\texttt{Picasso}) &= 0.24 + 0.08 = 0.32, \\ m'(\texttt{Picasso},\texttt{Braque}) &= 0.06, \\ m'(\mathcal{K}) &= 0.02, \\ m'(\emptyset) &= 0.48, \end{aligned} \tag{9}$$

After multiplying by $\frac{1}{1-m'(\emptyset)} = 1.923$ one obtains:

$$\begin{aligned} m(\texttt{Cezanne}) &= 0.23 \\ m(\texttt{Picasso}) &= 0.62 \\ m(\texttt{Picasso,Braque}) &= 0.11 \\ m(\mathcal{K}) &= 0.04 \\ m(\emptyset) &\overset{\text{def}}{=} 0 \end{aligned} \tag{10}$$

Since the $m(\cdot)$'s sum up to 1 and $m(\emptyset) = 0$, the mapping $m = m_1 \oplus m_2$ that emerges from Dempster's rule is also a mass function, consistent with (3).

In words, Dempster's rule computes a measure of agreement between two sources of evidence concerning various possibilities drawn from a common frame of discernment. The rule is conservative in that it focuses only on those possibilities that *both* sources support. The magnitude of the pooled support that a possibility X collects is computed by summing the products of the two masses $m_1(X)$ and $m_2(X)$, which explains the product operator in (7). Because the sources of evidence express their opinions over 2^θ rather than over θ, a joint agreement on a possibility can occur in more than one way, i.e. whenever the two sources support possibilities whose intersection gives X. This explains the summation operator in (7). Finally, when a pairing of two opinions results in a null possibility (the empty set), the multiplication of their individual masses may still be positive. This is an anomaly, since the definition of a mass function (3) requires that the mass of the null possibility be zero. This explains the role of (8), in which $m'(\emptyset)$ is deducted from the total mass and the remaining mass is divided by $(1 - m'(\emptyset))$ to ensure that the pooled mass will sum up to 1.

Dempster's rule is often compared to and contrasted with Bayes rule, because both rules concern the combination of probabilistic opinions into an aggregate (posterior) opinion. It is crucial to observe however that unlike Bayes rule, which is a trivial consequence of the axioms of probability theory, Dempster's rule is a *prescriptive* pooling mechanism which is neither right nor wrong, and thus it is less of a 'rule,' and more of a 'recipe.' Therefore, we take the position that the ultimate justification of Dempster's rule should be sought in the field, i.e., in the various applications in which the rule is supposed to have a certain sense of domain validity. This leads to the following question:

> **Question Q3:** What is the intuitive justification of Dempster's rule in the context of information indexing and retrieval applications? If one wishes to aggregate indexing opinions via a certain pooling mechanism, then why use (7-8) and not another set of formulae?

A typical way to avoid this question is to invoke the argument: "If one uses mass functions to represent relevance numbers, then one should combine them using Dempster's rule, because that is how mass functions are combined in the DS model." This argument could have been valid if Dempster's rule had a normative, domain-independent, and non-controversial justification. But this is not the case. In fact, many researchers have struggled to make sense of Dempster's rule, and the debate is still going strong: *"Shafer's theory has been strongly criticized for its failure to give a meaning to the measures of belief and plausibility, or to show how someone might arrive at a particular numerical assessment. In the absence of a definite interpretation, it is difficult to see how the rules of the theory, and in particular Dempster's rule, can be justified "* (Buxton, 1989). Given this controversy, the importance of question Q3 is obvious. Hence, our goal is to interpret, and to a certain extent defend, the *meaning* of Dempster's rule in the specific context of an information indexing and retrieval application. This analysis is taken up in section 3, where we return to question Q3.

Belief Functions: Building on the elementary notion of a mass function $m : 2^{\theta} \rightarrow [0, 1]$, the function $\text{Bel} : 2^{\theta} \rightarrow [0, 1]$, denoted a *belief* function, can be defined as follows:

$$\text{Bel}(X) = \sum_{A \subseteq X} m(A) \tag{11}$$

Whereas $m(X)$ measures the support rendered to X (a subset of propositions) directly, Bel(X) measures the total support rendered to X and to all its subsets (each being a more specific proposition than X). This relationship is depicted in figure 2, which illustrates how a belief function can be derived from the mass function given by (10). Note that (3) and (11) imply that Bel($\emptyset$) = 0 and Bel(θ) = 1 always. Also, (11) implies that the Bel function is completely determined by the m function, and, likewise, that m can be recovered from Bel's definition (Shafer, 1976, p. 39).

Put figure 2 around here

Plausibility Functions: Whereas Bel(X) measures the *total* support rendered to a possibility X, the plausibility of X, denoted Pl(X), measures the *maximal* support that X can possibly attain under a given mass function m. Specifically:

$$\text{Pl}(X) = \sum_{A \cap X \neq \emptyset} m(A) \tag{12}$$

In words, Pl(X) records the total mass allocated to all the possibilities with which X intersects. For a pictorial description of this relationship, refer again to figure 2.

The intuitive relationship between the three functions $m(\cdot)$, Bel($\cdot$), and Pl($\cdot$) can be described as follows. Beginning with Bel's definition, consider the two possibilities $X, A \subseteq \theta$.

Since both X and A are disjunctions of propositions, the set-theoretic statement $A \subseteq X$ is equivalent to the logical rule $A \rightarrow X$, which we will interpret as: 'If the truth lies in A, it must also lie in X.' Therefore, the sum of all the masses associated with premises A that imply X can be viewed as a measure of the total support rendered to X. As regards Pl's definition, suppose now that $A \cap X \neq \emptyset$ (but A is not necessarily a subset of X). Since the possibility A is a disjunction of propositions, the mass $m(A)$ rendered to it can 'float' freely to any one of its subsets, including those that intersect X. In the extreme case, the intersection $A \cap X$ may inherit the *entire* mass of A. It follows that $\text{Pl}(X)$ is the upper bound of $\text{Bel}(X)$.

To do justice to the theory of evidence, it should be noted that the construction of Bel and Pl using m is only one way to define these functions. Shafer provided direct definitions of mass, belief and plausibility functions in terms of each other. He has also emphasized the key role that *subadditivity* plays in the theory of evidence, a point which we now turn to discuss in the specific context of information indexing and retrieval.

Sub Additivity: The *complement* of a set $X \subseteq \theta$, i.e. the set of all propositions that are in θ and not in X, is denoted hereafter $\overline{X}$. Definitions (11) and (12) imply the following important relationships:

$$\text{Pl}(X) = 1 - \text{Bel}(\overline{X}) \tag{13}$$

$$0 \leq \text{Bel}(X) \leq \text{Pl}(X) \leq 1 \tag{14}$$

If a certain Bel_b were a *Bayesian* representation of degrees of belief, the additivity axiom of probability theory ($X \cap Y = \emptyset$ implies $\text{Bel}_b(X \cup Y) = \text{Bel}_b(X) + \text{Bel}_b(Y)$) would mean that

$$\text{Bel}_b(X) = 1 - \text{Bel}_b(\overline{X}), \tag{15}$$

yet (13) and (14) imply that in the general case $\mathrm{Bel}(X) \leq 1 - \mathrm{Bel}(\overline{X})$, leading to the famous subadditivity property of the theory of evidence:

$$\mathrm{Bel}(X) + \mathrm{Bel}(\overline{X}) \leq 1 \tag{16}$$

In other words, the belief that one holds in a possibility does not automatically imply one's disbelief in the negation of that possibility. In information indexing and retrieval applications, where θ is taken to be a keyword-list $\mathcal{K}$, this tenent has important implications. For example, if the admittance of new evidence causes a cataloger to increase his belief in the document's relevance to a lexical subset X, the same evidence should not necessarily decrease his belief in the document's relevance to lexical subsets in $\overline{X}$, especially if the cataloger is not confident in his indexing opinion. In particular, the difference $1 - \mathrm{Bel}(X) - \mathrm{Bel}(\overline{X})$ is called the uncommitted belief with respect to X. If Bel were a Bayesian representation of degrees of belief, the uncommitted belief would be zero by definition. This is best illustrated in the 'state of insufficient reason,' in which one knows absolutely nothing about a set of propositions $\theta = \{q_1, \ldots, q_n\}$. Whereas the common solution is to set $\mathrm{Bel}(q) = 1/n$ for all $q_i \in \theta$, the theory of evidence would set $\mathrm{Bel}(\theta) = 1$ and $\mathrm{Bel}(X) = 0$ for all the other proper subsets of θ. This is the case when the uncommitted belief is at maximum.

The interpretation of $\mathrm{Bel}(\cdot)$ and $\mathrm{Pl}(\cdot)$ as lower and upper-probabilities has led many to view the theory of evidence as a novel calculus for eliciting and manipulating interval-valued, rather than point-valued, degrees of beliefs. Indeed, the theory allows one to express the belief in every hypothesis X by means of the interval $[\mathrm{Bel}(X), \mathrm{Pl}(X)]$, which may be updated as new evidence about X is admitted. Further, the width of the interval, $\mathrm{Pl}(X) - \mathrm{Bel}(X)$, which by definition equals $1 - \mathrm{Bel}(X) - \mathrm{Bel}(\overline{X})$, represents the uncommitted belief with respect to X. If the uncommitted beliefs induced by a certain mass function

m were zero for all the hypotheses under consideration, the intervals would degenerate to zero widths and Bel would be a standard probability function. Yet in the more general case in which the mass reflects some 'second-order uncertainty,' or 'ambiguity,' the degree of belief in possibilities X drawn from θ is allowed to 'float' between Bel(X) and Pl(X). One benefit of such a model is that it is more robust and less prone to human errors in assessing subjective degrees of support.

We arrive at our last question:

> **Question Q4:** The designer of a DS-based IR application can choose to elicit and represent relevance through three alternative languages: mass functions, belief functions, and belief intervals. What is the relationship among these three representation in the specific context of information indexing and retrieval applications?

Recall that the three functions m, Bel, and Pl, are mathematically equivalent, in the sense that knowledge of any one of them (for every possibility) can be used to compute the other two. This equivalence might lead one to concur that the question of whether to use m, Bel, or [Bel, Pl] to elicit and manipulate degrees of support depends on cognitive and efficiency considerations. As it turns out, this conclusion is quite naïve. For example, belief intervals are not as flexible a representation as we would like them to be. That is, when one elicits [Bel, Pl] intervals from a source of evidence, it is not true that the only restriction is that $0 \leq \text{Bel} \leq \text{Pl} \leq 1$. Again, a full understanding of these constraints requires a semantic interpretation, which we now proceed to present.

4 A Canonical Indexing Model

As figure 1 illustrates, the key theme of this paper is the interplay of the theory and practice of the Dempster Shafer model, as viewed through the 'lens' of a particular application. The previous section was structured around the key constructs of the *theory*: the frame of discernment, mass and belief functions, and Dempster's rule. Coming from the other extreme, this section is structured around the key constructs of the *application*: taxonomies, relevance functions, and index aggregation operators. This leads to the development of a canonical indexing model, around which the remainder of the paper evolves. In building this model, our intention is to articulate an indexing mechanism which is simple, intuitive, and, most importantly, has a straightforward probabilistic interpretation.

The main result that we are aiming at is this: notwithstanding its domain-specific origin and its strict probabilistic foundation, the canonical model that we will expound is isomorphic to the DS model. This result is based on previous work by Hummel and Landy (1988), who defined an opinions pooling mechanism which plays an important role in our canonical model. The isomorphism has three important implications. First, the canonical model addresses all the questions that were raised about the theoretical fit between the DS theory and information indexing and retrieval applications. Second, because the limitations of the former will be explicit, implicit limitations of the latter will become apparent. Third, because the canonical model makes no use of extra probabilistic arguments, it also provides a simple probabilistic interpretation to the DS theory, which is often claimed to be an extension of probability theory.

4.1 The Taxonomy

So far, we have assumed that the keyword-list is a 'flat' set of index terms. In most IR applications, though, the keyword-list has a rich semantic structure that can be used to improve and refine the indexing process. Typically, the structure can be described in terms of an `is-a` network that imposes a generalization/specialization relation on the keyword-list. Taken together, a keyword-list along with its underlying structure will be henceforth referred to as a *taxonomy*. From the user's perspective, the taxonomy is a structured set of *classes*, or categories, designed to facilitate access to a body of material in a particular subject of interest. For example, consider the taxonomy depicted in figure 3, which organizes art-related documents according to major artists and artistic movements.

Put figure 3 around here

Taxonomies are constructed by domain experts — in this case art scholars — who provide two types of information: (i) a set of classes; and (ii) a taxonomical structure. For example, the set of classes in figure 3 is $C = \{$`Art`, `Braque`, `Cubism`, `Dada`, `Impressionist`, `Janco`, `Modern`, `Picasso`$\}$. The taxonomical structure can be represented as a set of ordered pairs, where (x, y) codes the assertion 'class x is a direct generalization of class y'. With this notation, the structure of figure 3 is completely defined by the set $H = \{$(`Art`,`Modern`), (`Art`,`Impressionists`), (`Modern`, `Cubism`), (`Cubism`,`Braque`), (`Cubism`, `Picasso`), (`Dada`, `Picasso`), (`Dada`, `Janco`)$\}$. As the figure illustrates, the resulting topology is a directed graph consisting of the nodes set C and the edges set H. Note that each node $x \in C$ can be associated with a set of terminal nodes which we denote $X =$ TERM(x). For example, TERM(`Cubism`) $= \{$`Braque`, `Picasso`$\}$. The notion of terminal sets plays a role in our definition of a taxonomy, which is as follows:

Definition: A taxonomy is a directed, acyclic, and finite graph $T =< C, H >$, where C is a set of classes (nodes) and $H \subset C \times C$ is a *direct-subclass* relation (set of edges), under three constraints:

- C1: For every class $y \in C$, save for one exception, there is at least one other class $x \in C$ such that $(x, y) \in H$. The one class in C that is no a subclass of any other class is called the *root* of the taxonomy.
- C2: If y is a subclass of x, i.e. $(x, y) \in H$, then there is no other chain of subclasses beginning with x and terminating with y in H. That is, H designates a minimal set of subclass relations.
- C3: For each $x, y \in C, \text{TERM}(x) = \text{TERM}(y)$ if and only if $x = y$. That is, in a taxonomy graph every class is associated with a unique set of terminals.

A taxonomy is much like a tree, except that children nodes can have multiple parent nodes at the level above. For example, in figure 3, {Picasso} is a subclass of both {Cubism} and {Dada}. Thus there can be multiple paths from the root to any given node. Further, those paths can have different lengths, so that unlike a tree, the depth of a node is ambiguous. (Unless it is defined as the minimum path length from the root, in which case the length is not necessarily monotonically increasing when traversing subclass relations.) Using tree terminology, we refer to the classes that can be reached by traversing subclass relations from x 'downward' as the *descendants of* x, and to the classes of which x is a descendant as the *predecessors of* x. The *root* of the taxonomy is the only class that (i) has no predecessors, and (ii) is the predecessor all other classes, e.g., art in figure 3. Since the taxonomy graph is finite and acyclical, it contains a 'boundary,' or a set of *terminal classes*, which we denote $\mathcal{K}$. A class $k \in C$ is said to be terminal if it has no descendants, i.e. if no edges of the form (k, x) exist in H. The set $\mathcal{K} = \{k | \neg \exists x \in C \; with \; (k, x) \in H\}$, which is completely determined by C and H, plays an important role in the subsequent analysis. Therefore, we will sometimes use it to subscript the name of its respective taxonomy, as in $T_{\mathcal{K}} =< C, H >$.

Given a taxonomy $T_{\mathcal{K}} =< C, H >$, we will further characterize each class $x \in C$ by two sets of classes that we denote $\mathtt{LIB}(x)$ and $\mathtt{VOL}(x)$, and refer to as the *library* rooted at x and the *volumes* of x, respectively. $\mathtt{LIB}(x)$ contains all the descendants of x, including x itself, representing the entire set of classes into which c may be decomposed. $\mathtt{VOL}(x)$ represents a larger set, consisting of $\mathtt{LIB}(x)$ as well as all the predecessors of keywords in $\mathtt{LIB}(x)$, i.e. all the classes that have something in common with x. For example, $\mathtt{LIB(Cubism)} = \{\mathtt{Cubism}, \mathtt{Braque}, \mathtt{Picasso}\}$, and $\mathtt{VOL(Cubism)} = \{\mathtt{Cubism}, \mathtt{Braque}, \mathtt{Picasso}, \mathtt{Modern}, \mathtt{Art}\}$. These sets can be given a recursive definition, as follows:

$$\mathrm{LIB}(x) = \{x\} \cup \{y \in C | \exists z \in \mathrm{LIB}(x) \text{ with } (z, y) \in H\} \tag{17}$$

$$\mathrm{VOL}(x) = \mathrm{LIB}(x) \cup \{y \in C | \exists z \in \mathrm{VOL}(x) \text{ with } (y, z) \in H\} \tag{18}$$

Definitions (17-18) imply that (i) $\mathrm{LIB}(root) = C$, so that the root library contains all the classes in the taxonomy; (ii) $\mathrm{LIB}(k) = \{k\}$ if and only if $k \in \mathcal{K}$, so that terminal classes are characterized by libraries that contain singleton classes only, and (iii) for each $x \in C$, $\mathrm{TERM}(x) = \mathrm{LIB}(x) \cap \mathcal{K}$, a convenient definition of the terminals set. The definition of $\mathtt{LIB}$ and the last assertion imply that $(x, y) \in H \rightarrow \mathrm{TERM}(x) \supset \mathrm{TERM}(y)$.

The indexing process: The act of indexing a document within a taxonomy can be described as a top-down, depth-first search process. To illustrate, suppose that an art-related document is to be indexed within the art taxonomy from figure 3. Without loss of generality, assume that the document is relevant to modern art. Beginning at the first level under `Modern` and proceeding left to right, we test if the document is relevant to `Cubism`.

If the answer is 'yes,' we step down one level and test if it is relevant to Braque. If the answer is 'yes,' we index the document in Braque. If the answer is either 'no' or 'unsure,' we test if it is relevant to Picasso. If the answer is either 'no' or 'unsure,' and assuming that Picasso is the last class below Cubism, we backtrack one level, index the document in Cubism, and proceed to explore Dada. If the document is not relevant to any of the classes thus visited, we backtrack one level and index the document in Modern. This would reflect the notion that even though the document is related to modern art, the existing taxonomy fails to discern the exact category to which it belongs. Thus the indexing process involves a depth-first search which is cut off at any class that is deemed irrelevant to the indexed document. In practice, the process can be considerably shortened by using the domain knowledge of a human cataloger.

We see that the notion of relevance that is consistent with this process is defined over *subsets* of, rather than *individual*, keywords. That is, if a document is indexed under, say, Cubism, it implies that the document is relevant to LIB(Cubism), like other documents about Cubism, Braque, or Picasso. Beyond this interpretation, however, the indexing decision does not imply any specific logical relationship involving the keywords in LIB(Cubism), except that it says that the document should not be indexed under any one of the keywords alone. This definition of relevance is convenient because it allows us to be as specific as we wish in our relevance statements. If we are sure that a document is relevant to a certain class, we index it under that class. If we are not sure, we can backtrack and index the document in a library that contains the class. We can do this all the way up to the root of the taxonomy, at which point the indexing decision root would express the opinion that the document belongs somewhere in the library, without specifying exactly where.

As we will see shortly, the above indexing process maps well on the standard DS model. However, the process as described has one critical shortcoming: it does allow us to differentially specify a document's relevance to *conjunctions* of keywords, a situation which occurs frequently in IR applications. It turns out that conjunctive indexing can be handled quite effectively by a simple extension of the standard DS model. Since the extension entails a diversion from the main theme of the paper, we will describe it in a separate appendix.

Relationship to the theory of evidence: In IR applications, relevance judgements are expressed and manipulated over a taxonomy – a structured set of classes. In the DS model, belief functions are elicited and combined over a set-theoretic frame of discernment. Hence, in order to implement relevance judgements as belief functions, we must specify a unique mapping from the semantic domain of classes to the set-theoretic domain of lexical-subsets. The linkage will be established through the keyword-list $\mathcal{K}$, which is a shared property of both domains.

Beginning with the DS domain, let $\mathcal{K}$ be a collection of keywords, or a lexical frame of discernment, and let $2^{\mathcal{K}}$ be the power-set of $\mathcal{K}$, excluding the empty-set. We define the *subset graph* of $\mathcal{K}$ to be $G_{\mathcal{K}} =< 2^{\mathcal{K}}, S >$, where $2^{\mathcal{K}}$ (nodes set) enumerates all the subsets of $\mathcal{K}$ excluding the empty set, and $S \subset 2^{\mathcal{K}} \times 2^{\mathcal{K}}$ (edges set) is a *minimal subset* relation. That is, $(X, Y) \in S$ if and only if $X \supset Y$ and there are no other subsets that can fit in between, i.e. there is no $Z \in 2^{\mathcal{K}}$ such that $X \supset Z$ and $Z \supset Y$. It follows that (i) subset X has exactly one more element than subset Y, and (ii) the root node of G is the maximal set in $2^{\mathcal{K}}$, i.e. $\mathcal{K}$ itself.[5] We will also speak of the *transitive closure* $G^*_{\mathcal{K}}$, which is the graph that

[5] The subset graph has a layered structure. Specifically, if $\mathcal{K}$ consists of n elements, then the subset graph will consist of n levels. Labeling the root level 1 and the terminals level n, each level $i = 1, \ldots, n$ consists of $\binom{n}{n-i+1}$ subsets, each of cardinality $n - i + 1$.

is obtained from G by adding new edges from subset X to subsets Y whenever there is a path from X to Y in G. Hence, $G^*_{\mathcal{K}}$ contains a directed edge (X, Y) for every two subsets $X, Y \in 2^{\mathcal{K}}$ such that $X \supset Y$.

Given these constructs, it is not difficult to see that every taxonomy $T_{\mathcal{K}} =< C, H >$ can be embedded into its related $G^*_{\mathcal{K}}$ graph. We sketch the proof by construction, as follows. Given a taxonomy $T_{\mathcal{K}} =< C, H >$, we use C and H to extract the set of terminal classes $\mathcal{K}$. Next, we use the power-set $2^{\mathcal{K}}$ to construct the graph $G^*_{\mathcal{K}}$. Now, let us fix a pair of connected nodes in $T_{\mathcal{K}}$ so that $x, y \in C$ and $(x, y) \in H$. We have to show that both the nodes and the edge map uniquely on $G^*_{\mathcal{K}}$. Since $T_{\mathcal{K}}$ is a taxonomy, every node $x \in C$ is associated with a unique set of terminals $\text{TERM}(x) = X \in 2^{\mathcal{K}}$. By construction, every lexical subset in $2^{\mathcal{K}}$ is associated with a single node in $G^*_{\mathcal{K}}$. Thus, both x and y map uniquely on the nodes $X = \text{TERM}(x)$ and $Y = \text{TERM}(y)$ in $G^*_{\mathcal{K}}$. Now, $(x, y) \in H$ implies that $\text{TERM}(x) \supset \text{TERM}(y)$. This, in turn, implies that there is an edge between X and Y in $G^*_{\mathcal{K}}$. Thus, both x, y, and (x, y) map uniquely on $G^*_{\mathcal{K}}$, implying that $T_{\mathcal{K}}$ can be embedded in $G^*_{\mathcal{K}}$. An example for the case of $\mathcal{K} = \{\texttt{Braque}, \texttt{Picasso}, \texttt{Janco}\}$ is given in figure 4.

Put figure 4 around here

Note that the transitive closure is necessary for establishing the relationship. In the IR taxonomy domain, $(x, y) \in H$ implies that $\text{TERM}(x) \supset \text{TERM}(y)$. However, the semantic structure of the taxonomy may well be such that the number of elements in $\text{TERM}(x)$ is more than one plus the number of elements in $\text{TERM}(y)$. Because the taxonomy's subset graph $G_{\mathcal{K}}$ represents only minimal subset relationships, it will contain no edge associated with $(\text{TERM}(x), \text{TERM}(y))$. However, such an edge *would* exist in the transitive closure $G^*_{\mathcal{K}}$, enabling the embedment.

From a practical standpoint, it is convenient to distinguish between two types of taxonomies: static and adaptive. A *static taxonomy* consists of a fixed set of classes, like the Dewey decimal system or the Library of Congress index. An *adaptive taxonomy* is a dynamic data structure that evolves from the indexing process itself. Such a taxonomy consists of an open-ended set of classes, each class being a different grouping of keywords from $\mathcal{K}$. That is, when a new document is deemed relevant to a subset of keywords that do not make up an existing category, one simply announces this subset a new class and adds it to the taxonomy. Hence, a document titled *"A letter from Braque to Janco"* may well be indexed in the class {Braque, Janco}, something that would have been impossible in a static taxonomy that does not contain such a predefined category. The only restriction that is placed on an adaptive taxonomy is that it must contain at least all the elements in $\mathcal{K}$ (as singletons, or classes that are made up of single keywords), as well as $\mathcal{K}$ itself. Hence, we begin with the initial set of classes $C = \{\{k_1\}, \ldots \{k_n\}, \mathcal{K}\}$, and add more classes to it as we go along. In the extreme case, the taxonomy might end up consisting of 2^n classes, one for each indexing possibility in $\mathcal{K}$. Of course, such a taxonomy will become prohibitively large even with only a few dozen keywords. However, note that once the *semantics* of the keyword-list is taken into consideration, many if not most of the classes in $2^{\mathcal{K}}$ will become irrelevant, since they represent arbitrary grouping of keywords that are not likely to arise in the indexing process.

In sum, we have shown that relevance judgements made over a taxonomy of classes can be pegged into and manipulated over a standard frame of discernment. In order to assert this relationship, we had to constrain the definition of a taxonomy by disallowing certain graph forms that do not map well on the set-theoretic notion of a frame of discernment. The practical implications are as follows. The first constraint (C1) simply requires that every taxonomy will have a single entry point. According to the second constraint (C2), if class x generalizes class y and class y generalizes class z, there is no need to hardwire into the taxonomy the fact that x also generalizes z. Since this relationship can be inferred automatically, the constraint serves to minimize duplication and inconsistencies. Constraint (C3) rules out taxonomies in which non-terminal classes converge on the same set of terminals through different paths. This constraint can be restrictive, because the different paths might have a distinct taxonomical interpretation that a cataloger may wish to preserve. If we have to deal with such a taxonomy, (C3) could be enforced structurally by adding auxiliary nodes to $\mathcal{K}$ in such a way that makes all the terminal sets of the non-terminal classes unique. Purists may find this solution crude, but the adjustment is necessary if one wants to apply the DS model to information indexing and retrieval applications without violating, or misinterpreting, the set theoretic premise of the model.

We now turn to question Q1, which asked whether the DS concept of a lexical power-set provides an adequate 'skeleton' for indexing documents in IR applications. The answer to this question is 'yes,' but there are two caveats, regarding structural complexity and interpretation. Ideally, we would have liked the DS theory to apply to the whole gamut of keyword structures – from simple hierarchical architectures (e.g. the Dewey Decimal System) to the loose networks that result from 'hot-word' Hypertext indexing. However, according to our analysis only a certain family of these topologies – namely those that conform to our definition of a taxonomy – yield to a standard DS interpretation. This rules out certain keyword structures that might arise in practice, although in some cases non-taxonomical structures can be converted into taxonomies by adding auxiliary keywords. The second caveat is that the standard DS model interprets lexical subsets (non-terminal classes) as *exclusive disjunctions* of their constituent keywords. Yet when a cataloger states that a document is relevant to, say, $\{k_1, k_2, k_3\}$, he may well mean that the document is relevant to *all* the keywords, i.e. to the *conjunction* $k_1 \wedge k_2 \wedge k_3$. In fact, the cataloger may wish to index the document in other logical connectives involving those keywords, e.g. $k_1 \vee (k_2 \wedge k_3)$. It turns out that even though the standard DS model does not support such indexing decisions, a simple extension of the model can go a long way toward solving the problem, as we treat in Appendix A.

4.2 Relevance Functions

The fundamental rule of indexing is that a document should be indexed using certain keywords if prospective users of the document would find it *relevant* to these keywords. In its most primitive form, then, relevance is a Boolean and subjective relation, indicating categorically that a document $d \in D$ is relevant to a lexical subset $X = \{k_1, \ldots, k_m\}$ in the view of a particular library patron. However, due to the fact that bibliographical classes do not have crisp boundaries, and due to the multitude of relevance opinions expressed by different catalogers and library patrons, a more reasonable question is not whether d is relevant to X, but rather what is the *intensity* of this relation. In other words, we seek to represent relevance in terms of a mapping $r : 2^{\mathcal{K}} \times D \rightarrow [0,1]$, rather than in terms of a characteristic function $r : 2^{\mathcal{K}} \times D \rightarrow \{0,1\}$.

There have been many efforts to interpret relevance on probabilistic grounds, Maron and Kuhns (1960) being the defining article. One of the fundamental problems in this area has been the proper definition of the *sample space* from which relevance propositions are drawn. This point was alluded to by Maron, as follows:

> "The notion of probability of relevance can be interpreted in two different perspectives: of the *document*, as the proportion of patrons of a given type who would judge that document relevant, and of the *patron* himself, as the proportion of documents of a given type which he would judge relevant. The first

> model leads to a theory of probabilistic indexing; The second model leads to a theory of probabilistic query formulation (Maron, 1982)."

In what follows we will focus on Maron's first perspective, in which multiple patrons form relevance opinions about a fixed document. Consistent with Maron's observation, this perspective yields a model of inexact indexing. Unlike Maron, though, the uncertainty associated with the indexes will lead in our model not to probability functions, but rather to Dempster Shafer mass functions, i.e. functions that conform to definition (3).

Let $U = \{u_1, \ldots, u_n\}$ be a set of catalogers, and let $\mathcal{K}$ be a keyword-list. Suppose that each cataloger in U is asked to index the same document using $\mathcal{K}$, i.e. to specify one or more keywords from $\mathcal{K}$ that are relevant to the document. Suppose that cataloger u_i supplies the opinion that the document is relevant to the lexical subset $X \subseteq \mathcal{K}$; we then record this opinion by means of the following Boolean function:

$$v_i(X) = \begin{cases} 1 & \text{if } u_i \text{ indexed the document using } X \\ 0 & \text{otherwise} \end{cases} \qquad (19)$$

$$i = 1, \ldots, n$$

Since each cataloger u_i supplies one set of relevant keywords, there will be exactly one subset $X \in 2^{\mathcal{K}}$ such that $v_i(X) = 1$. Also, the empty set is not allowed to be a valid indexing opinion. If a cataloger is unwilling to give an opinion or is unsure about the proper classification of the document, the document is indexed by default in the root class $\mathcal{K}$, which is also an element of $2^{\mathcal{K}}$. This convention makes sense because the root class represents the *entire library*, and is therefore the natural place to store documents whose specific class membership is undiscernible.

After all n catalogers have cast their indexing opinions regarding *the same document* d, we compute for each lexical subset $X \in 2^{\mathcal{K}}$ three 'relevance counters,' as follows:

$$r(X) = \sum_{i=1}^{n} v_i(X, d) \tag{20}$$

$$r_{\text{LIB}}(X) = \sum_{Y \in \text{LIB}(X)} r(Y, d) \tag{21}$$

$$r_{\text{VOL}}(X) = \sum_{Y \in \text{VOL}(X)} r(Y, d) \tag{22}$$

In words, $r(X)$, $r_{\text{LIB}}(X)$, and $r_{\text{VOL}}(X)$ count the number of catalogers who classified the document in X, in the library rooted in X, and in libraries that intersect (or in a hierarchical taxonomy, *contain*) X, respectively. (When d is fixed in our analysis, we will suppress it from the notation, writing $r(X)$ instead of $r(X, d)$.)

Relationship to the theory of evidence: Suppose now that the Boolean indexing opinions of the catalogers are averaged over the space of catalogers U as follows:

$$m(X) = \frac{1}{n} \cdot r(X) = \frac{1}{n} \sum_{i=1}^{n} v_i(X) \tag{23}$$

Then the resulting function $m(X)$ is a DS mass function over the lexical space $\mathcal{K}$. Formally, we have the following proposition (the proofs are given in a separate appendix):

Proposition 1: Let $U = \{u_1, \ldots, u_n\}$ be a set of catalogers with their Boolean indexing opinions $v_1, \ldots, v_n : 2^{\mathcal{K}} \rightarrow \{0, 1\}$. The real function $m : 2^{\mathcal{K}} \rightarrow [0, 1]$ defined by $m(X) = \frac{1}{n} \cdot r(X) = \frac{1}{n} \sum_{i=1}^{n} v_i(X)$ is a mass function, satisfying definition (3).

The consequence of Proposition 1 is that DS mass functions arise *naturally* when we view the relevance functions as derived from averages of multiple Boolean indexing opinions. We begin with a space U of n catalogers who are asked to index the same document using the same keyword-list $\mathcal{K}$. Each cataloger supplies an individual opinion that specifies which keywords are relevant to the document. Note that the cataloger's indexes are not restricted, and that they are free to choose any keyword or combination of keywords that, in their opinion, are relevant to the document. Next, shifting our attention from the catalogers space U to the keyword space $\mathcal{K}$, we compute for each lexical subset $X \subseteq \mathcal{K}$ a measure of 'average relevance,' $\frac{1}{n} \cdot r(X)$, which represents the fraction of catalogers who thought that the document was relevant to X. Disregarding the lexical subsets that no cataloger has chosen, we obtain a set of pairs of the form $\{(K_1, r_1), \ldots, (K_n, r_n)\}$ in which $K_i \in 2^{\mathcal{K}}$ and $0 \leq r_i = \frac{1}{n} \cdot r(K_i) \leq 1$.

We are now in a position to answer question Q2, regarding the 'type' of relevance that DS mass functions represent, given the context of multiple indexing opinions. First, note that the canonical model has yielded the type of relevance numbers that are at the center of any probabilistic indexing model. Second, according to Proposition 1, these numbers form a mass function, consistent with the standard DS model. Finally, in our multi-player interpretation, the meaning of the mass $m(X)$ is simply the fraction of catalogers who think that the document is relevant to the lexical subset X.

Following the same line of reasoning, we can also provide an answer to question Q4, that sought an IR interpretation of the meaning and relationship of mass functions, belief functions, and belief intervals. Given the IR context in which $\theta \equiv \mathcal{K}$, it is easily seen that the relevance counters (20-22) are proportional to the mappings that represent degrees of belief in the DS model. Specifically, dividing each counter by n — the number of catalogers — yields the mass, belief, and plausibility functions defined in (3), (11), and (12), respectively:

$$m(X) = \frac{1}{n} \cdot r(X) \tag{24}$$

$$\mathrm{Bel}(X) = \frac{1}{n} \cdot r_{\mathrm{LIB}}(X) \tag{25}$$

$$\mathrm{Pl}(X) = \frac{1}{n} \cdot r_{\mathrm{VOL}}(X) \tag{26}$$

If we combine these observations with the interpretation of the power-set of the keyword-list as a taxonomy, we see that the mass on a lexical subset X is given by the fraction of catalogers who indexed the document using X directly. Similarly, the belief in X is the fraction of catalogers who indexed the document in libraries within X, and the plausibility of X is the fraction of catalogers who indexed the document in libraries that intersect (in a hierarchical taxonomy, *contain*) X.

The key component of the canonical model that enables this interpretation of the DS functions is the assumption of multiple catalogers and the $v_i(\cdot)$ functions that keep track of their individual indexing opinions. Although this assumption is not part of the DS model, we see that once posited, it provides a foundation on which the model can be given (one) plausible interpretation, without having to invoke extra-probabilistic arguments.

4.3 Aggregating Relevance

So far, we have assumed that (i) relevance is a two-place function $r(X, d)$ between a document d and a lexical subset X, and that (ii) all the catalogers from whom $r(X, d)$ was elicited were of the same 'type,' using Maron's terminology (see quote in Section 4.2). In this section we retract both assumptions. Specifically, we argue that relevance, in its most elementary form, is a three-place relation $r(X, d, q)$ in which q is the *classifier* dimension, or *context*, in which d was judged to be relevant to X. With that in mind, $r(X, d)$ can be viewed as a measure of *aggregate relevance* that runs over the various contexts in which d's relevance to X was judged. We now turn to describe a pooling mechanism that implements such an aggregation.

Let $U_1 = \{u_1, \ldots, u_{n_1}\}$ be a group of n_1 catalogers who are asked to index a document d using a keyword-list $\mathcal{K}$ based on a certain classifier, or source of information, denoted q_1. Similarly, let $U_2 = \{u'_1, \ldots, u'_{n_2}\}$ be a group of n_2 catalogers who are asked to index the same document, based on another classifier, denoted q_2. The semantics of the classifiers depends on the indexing scenario. For example, q_1 might be the document's title, whereas q_2 might be the document's abstract. To illustrate, let $\mathcal{K} = \{a, b, c\}$ and let U_1 and U_2 consist of 4 and 3 catalogers, respectively. Assume that within the U_1 group, two catalogers index the document in $\{a, b\}$, one in $\{a\}$, and one in $\{b\}$. Within the U_2 group, one cataloger indexes the document in $\{b, c\}$, one in $\{a, b\}$, and one in $\{b\}$. These indexing opinions are tabulated in the two tables on the left side of figure 5. The columns of each table represent the common keyword-list $\mathcal{K} = \{a, b, c\}$. The ith tuple in each table represents the indexing opinion elicited from the ith cataloger in the respective group as a binary vector. Hence, 1 in the (i, j)th table entry indicates that cataloger i has included the jth keyword in his indexing opinion and 0 indicates that he did not.

Put figure 5 around here

In what follows, we denote the binary vector that represents the indexing opinion of cataloger u_i by w_i. Similarly, the set of all opinions of a group of catalogers will be denoted $W = \{w_i | u_i \in U\}$. Finally, the group of catalogers U together with their indexing opinions W will be denoted $Z =< U, W >$ and referred to as a *model.* With this notation, consider two groups of catalogers U_1 and U_2 together with their indexing opinions W_1 and W_2. If all the catalogers in both groups are considered equally qualified to cast indexing opinions, then a variety of different pooling mechanisms may be used to compute the aggregate index induced by *all* the catalogers. Symbolically, we seek an operator $\otimes$ to compute the model $< U, W >=< U_1, W_2 > \otimes < U_1, W_2 >$.

One such operator – denoted hereafter by $\otimes$ – is illustrated in figure 5. This operator implements a pooling mechanism that can be described as "a consensus opinion formed by the committees of two" (Hummel and Landy, 1988). We have chosen to focus on this particular operator for two reasons. First, $\otimes$ enables an intuitive interpretation of Dempster's rule, as we will see shortly. Second, the operator has a straightforward meaning in terms of combining expert opinions, although it is certainly not the *only* plausible technique for pooling indexing decisions in an IR context. At the same time, the operator provides a simple point of departure from which more sophisticated IR pooling techniques can be derived, as we discuss toward the end of the section.

Going back to figure 5, note that $U = U_1 \times U_2$ enumerates all the $n_1 \cdot n_2$ possible two-member committees of catalogers that can be drawn from U_1 and from U_2. Within each committee (unique pair of catalogers) $(u_i, u'_j) \in U$, the committee's indexing opinion is defined to be the *binary conjunction* of the individual opinions of u_i and u'_j, which we denote $w_{i,j} = w_i \cdot w'_j$. For example, consider the first tuple in the U table in figure 5. This tuple gives the opinion of the committee (u_1, u'_1), i.e. $w_{1,1'} = (0,1,0)$. This opinion is the binary conjunction of the individual opinion $w_1 = (1,1,0)$ and $w'_1 = (0,1,1)$ as given by catalogers u_1 and u'_1 respectively.

The pooling operation $\otimes$ is completed by treating U as a new group of catalogers and using (23) to compute the mass function that it induces:

$$\begin{array}{lclcl} m'(\{a\}) & = & m'(1,0,0) & = & 1/12 \\ m'(\{b\}) & = & m'(0,1,0) & = & 7/12 \\ m'(\{a,b\}) & = & m'(1,1,0) & = & 2/12 \\ m'(\emptyset) & = & m'(0,0,0) & = & 2/12 \end{array} \tag{27}$$

Note that m' is not necessarily a mass function. since $\obslash$ can yield a result like $m'(\emptyset) > 0$. This happens when there is a pair of opinions (e.g. u_2 and u'_1 in our example), such that the conjunction of the opinions gives the empty set even though neither opinion gives the empty set individually. To resolve the problem, we normalize $m'(\cdot)$ as follows:

$$\begin{array}{lclcl} m(\{a\}) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(X) & = & 1/10 \\ m(\{b\}) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(B) & = & 7/10 \\ m(\{a,b\}) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(AB) & = & 2/10 \\ m(\emptyset) & \overset{\text{def}}{=} & 0 & & \end{array} \tag{28}$$

In words, for each lexical subset $X \in \mathcal{K}$, $m'(X)$ is the fraction of the (paired) catalogers who classified the document in that subset. Next. the fraction of the catalogers who agreed on *nothing* – $m'(0,0,0)$ – is distributed proportionally among the fractions of catalogers who agreed on *something*, yielding a new mass that sums up to unity. This function is now taken to be the 'aggregate index' of the document d, implying the taxonomy depicted at the top right of the figure. We may also view $m(X)$ as the fraction of (paired) catalogers who index the document in X among those paired catalogers who do not index the document in the empty set $\emptyset$. That is, if we discard pairs that agree on no relevant keywords, then the remaining pairs can compute their pooled relevance and then yield a mass function m.

Relationship to the theory of evidence: In order to discuss the relationship of the multiple catalogers/multiple classifiers scenario to the DS model, we first have to step back and say a few words about the role of 'sources of evidence' in the latter. Basically, the DS theory models a situation in which a finite set of 'pieces' or 'sources' of evidence $E = \{e_1, \ldots, e_n\}$ is used to discern the likelihoods of various possibilities X drawn from a common frame of discernment. Yet the *identity* of the sources of evidence is rather implicit in the model's language. That is, the common notation $m_i(X)$ and $\text{Bel}_i(X)$ is meant to be shorthand of the mass and belief functions $m(X|e_i)$ and $\text{Bel}(X|e_i)$, where e_i is the source of evidence whose 'support' of the possibility X we are trying to capture. The total support that the body of evidence E lends to X is computed through Dempster's rule (7-8), which yields a new function of the form $m(X|e_1, \ldots, e_n) = m(X|e_1)\oplus, \ldots, \oplus m(X|e_n)$.[6] For simplicity's sake, we denote the latter function $m(X)$, which reads 'the mass that the possibility X attains after all the available evidence has been taken into consideration.'

With that, the relationship between the canonical model and the DS model is as follows: possibilities correspond to lexical subsets, and sources of evidence correspond to classifiers, i.e. to different aspects of the document (title, abstract, author, etc.) that help discern the document's proper classification. The missing piece in the analogy is the set of catalogers who inspect each classifier individually and cast Boolean indexing opinions based on that information. In the DS model, the notion of multiple catalogers does not exist. In the canonical model, they are the driving force of the entire analysis. Specifically, it is assumed

[6] Like Bayes rule, Dempster's rule is commutative and associative, so its extension from 2 to n operands (sources of evidence) is straightforward.

that each group of catalogers is given access to one source of evidence – a classifier – and proceeds to cast its indexing opinions in view of that evidence (Whether or not this multi-player scenario makes sense in practice will be discussed in the next section. For now, the reader is asked to treat it as a theoretical metaphor). Hence, the overall classification decision is characterized by two types of uncertainty: (i) the uncertainty which is attributed to the fact that different sources of evidence might suggest different indexing decisions; and (ii) the uncertainty that is attributed to the variance of indexing opinions *within* each group of catalogers who are given access to the same information (classifier). How should we combine this multitude of indexing opinions into an aggregate index? In the canonical indexing model, the opinions are combined at the catalogers level, through the cartesian consensus operator $\otimes$. In the DS model, where the cataloger spaces do not exist, the opinions are combined at the classifiers level, via Dempster's rule $\oplus$. The key point, as illustrated in figure 5, is that both combination methods yield precisely the same result. Formally, we have the following proposition:

Proposition 2: Let $Z_1 =< U_1, W_1 >$ and $Z_2 =< U_2, W_2 >$ be two sets of catalogers together with their Boolean indexing opinions, and let $Z =< U, W >$ be the outcome of $Z = Z_1 \otimes Z_2$, as follows: (i) $U = U_1 \times U_2$; and (ii) $W = \{w_{i,j} = w_i \cdot w'_j | w_i \in W_1 \text{ and } w'_j \in W_2\}$. Let $\oplus$ be Dempster's rule as it is applied to mass functions. Let m_{z_1}, m_{z_2}, and $m_{z_1 \otimes z_2}$, be the mass functions induced by the models Z_1, Z_2, and $Z_1 \otimes Z_2$. Then we have the following: $m_{z_1 \otimes z_2} = m_{z_1} \oplus m_{z_2}$.

We are now in a position to return to question Q3, which sought a plausible interpretation of Dempster's rule in the context of IR applications. First, note that the cartesian product operator $\oplus$ implements a pooling mechanism which may or may not make sense in the applied context of IR. From a theoretical perspective, though, $\otimes$ is quite unique because of Proposition 2. That is, once we accept the fact that Dempster's rule is isomorphic to $\otimes$, a whole set of questions emerges: (1) why are the individual catalogers forced to specify only

Boolean, and not probabilistic, indexing opinions? (2) why are the groups of catalogers joined using a *set product* operator, as opposed to other set combination operators, e.g. union? (3) why committees of *two*, and not, say, committees of three? (4) why are the individual opinions combined using a binary *conjunction* rule? (5) why are all cataloger opinions given the *same weight*, where in practice some opinions may be more informed or worthy than others?

A proper answer to these questions requires an elaborate research program, involving both theoretical and empirical work. Also, the exact nature of the combination rule can vary from one situation to another. In the specific context of information indexing and retrieval, one can think of a family of index elicitation and aggregation models, designed to operate under different sets of assumptions. For example, if the catalogers prefer to express binary indexing opinions, we can use Dempster's rule (or the equivalent $\otimes$) to combine them. If they wish to express relevance by selecting a number between 0 and 1, we can modify the combination rule to accommodate this language as well (this will be similar to the way Yen (1989) extended Dempster's rule in the GERTIS system). If the catalogers wish to use a discrete language such as 'remotely relevant,' 'somewhat relevant,' etc., we can develop a fuzzy version of the rule. The key point here is that the precise definition of $\otimes$, along with Proposition 2, provide clear guidelines as to (i) which aspect of the combination rule has to be modified, and (ii) what will be the normative relationship between the modified rule, Dempster's rule, and probability theory.

5 Conclusion

The major implications of the research were already discussed in the body of the paper. We conclude with several comments regarding (i) efforts to apply the DS model to information indexing and retrieval applications; and (ii) efforts to interpret the theory of evidence on logical or probabilistic grounds.

Information Indexing and Retrieval: One objective of the paper was to articulate a concrete relationship between the Dempster Shafer model and information indexing and retrieval applications. The relationship that we have expounded can be summarized as follows:

IR application	Dempster Shafer model
keyword-list ($\mathcal{K}$)	frame of discernment (θ)
taxonomy ($< C, H >$)	subset of 2^θ
classification criteria (q_i)	sources of evidence (e_i)
groups of catalogers (U_j)	not part of the model
individual indexing opinions (W_i)	not part of the model
relevance measure to class (r)	mass function (m)
relevance measure to library ($r_{\text{LIB}}(X)$)	belief function (Bel)
relevance measure to volume ($r_{\text{VOL}}(X)$)	plausibility function (Pl)
relevance aggregation operator ($\otimes$)	Dempster's rule ($\oplus$)

We hope that the details of this interpretation, as discussed in the paper, will promote a better understanding of the proper way to apply the DS model to IR applications. In addition, the interpretation provides a practical foundation for building a variety of different indexing algorithms. These algorithms can use the $\otimes$ combination rule, or versions thereof, as called by the application. Ultimately, the success of one relevance calculus or another will depend on face validity and on field performance considerations.

In view of the fact that economic considerations hardly permit documents to be classified by even *one* human cataloger, the multiple catalogers setting that we have postulated seems to be rather unrealistic. There are two ways to address this point. First, a chief objective of this paper was to shed light on the DS theory; in that context, the notion of multiple catalogers serves as a theoretical artifact that enables a probabilistic interpretation of the theory. Second, with the advent of wide-area information services (such as Thinking Machine's WAIS system), the notion of multiple catalogers is no longer a remote academic exercise. In such systems, indexing opinions can be dynamically elicited from qualified library searchers and then used to automatically refine (or even create from scratch) index vectors for the benefit of other searchers. The library patrons who act as catalogers reach the same documents with different backgrounds and interests, each highlighting a different facet of the composite relation that we call 'relevance.' The canonical model that was described in this paper can be viewed as a first step toward implementing an interactive, multi-player, system for pooling such indexing opinions and converting them into composite relevance measures.

The Dempster Shafer theory of evidence: Several authors provided canonical examples that explain the rationale of the DS model in the way of analogy. Zadeh (1986) illustrated how mass functions and Dempster's rule can be mapped on fuzzy queries about *interval-valued*, rather than point-valued, attributes, in a relational database. Gordon and Shortliffe (1985) gave a compelling interpretation of how a DS calculus can be used to represent and combine the degrees of belief that clinical symptoms (pieces of evidence) render to classes of bacterial organisms (disjunctions of hypotheses), whose set relationships form a hierarchy. Coming from a different, domain-independent, direction, Hummel and Landy (1988) analyzed the statistical foundation of the theory of evidence *in general*, without making any assumptions on the underlying domain or the logical structure of the hypotheses. In contrast to other researchers who attempted to interpret high-level constructs of the DS model *directly* (e.g. Baron, 1987, Kyburg, 1987, and Schocken and Kleindorfer, 1989), Hummel and Landy took a more fundamental viewpoint that showed how the theory's constructs were implicitly linked to statistics of hypothetical opinions. However, their abstract mathematical analysis made no use of canonical examples, and thus it is difficult to interpret its implications on specific domain of application.

With that in mind, one objective of this paper was to illustrate how constructs of the DS theory that up until now defied simplistic interpretations have a plausible semantics in the context of a multi-classifier/multi-cataloger model. We have seen, in propositions 1 and 2, that the canonical model leads to exactly the same set of functions and formulae of the DS model. Hence, from a mathematical perspective, the canonical model is isomorphic to the DS model. Yet from a semantic perspective, it invokes the notion of multiple catalogers, consistent with several previous analyses of probabilistic relevance (Maron and Kuhns, 1960, Maron, 1982, Thompson, 1990).

To what extent are we forced to accept the multi-player premise of our canonical interpretation? One can simply reject the notion, avoiding the isomorphism by denying the possibility of multiple opinions, and relying simply on the DS theory as presented in Section 3. In that case, however, one is left with philosophical questions like Q1 through Q4. There could, of course, be other interpretations. However, in a real sense, *all* valid interpretations must be accepted or explained. That is, either the interpretation is accepted as is, or one must show how another set of semantic constructs provides a plausible interpretation of the theory. One advantage of our approach is that new calculi can be developed, different from the DS combination rule, that might better suit particular applications, based on modifications of the canonical model. It is precisely the unsatisfactory elements of this model that permit us to systematically seek improved methods for managing uncertainty.

Since our analysis was strictly probabilistic, it seems to support Lindley's observation that *"Anything that can be done with belief functions can better be done with probability theory"* (Lindley, 1987,p. 20). However, we believe that this argument misses an important point. To use a crude but useful analogy, it will be unreasonable to write off a programming language like Pascal simply because every Pascal program can be rewritten in machine language. Just like high-level languages provide complex structures for dealing with specialized problems, the DS model provides non-elementary functions and operators that lend themselves nicely to certain domains of application, information indexing and retrieval being one such example.

We conclude that the Dempster Shafer theory of evidence provides an attractive framework for supporting information indexing and retrieval applications, and that these applications, in turn, serve to highlight the internal validity and limitations of the theory. Dempster's rule remains a controversial operator for combining degrees of beliefs, but this paper has illustrated that it is just one member in a parametric family of combination rules, and that the question of whether to use this rule or another is more a matter of reasoned choice than a matter of adhering to a fixed set of formulae.

Appendix A: Conjunctive Frames of Discernment

IR taxonomies (left of figure 4) differ from their respective lexical DS power-sets (right of figure 4) in a subtle but profound way. In the former, the classes often have meaningful *names*, like `Cubism`; in the latter, the classes correspond to *anonymous* lexical subsets, like {`Braque,Picasso`}. As a result, indexing a document in a *named class* might mean something quite different than the implication that the document should be indexed in one of the class's constituent keywords.

For example, suppose that a cataloger decides to index a document in `Cubism`. In the standard DS model, this indexing opinion would be interpreted as "the document is relevant to exactly one of the following: Picasso, Braque, or any other (single) Cubist artist." Although this interpretation may be logically correct, it clearly entails a loss of concrete information about the document's relevance to Cubism proper, a relationship that can take many different forms in terms of the keywords that make up that class. In particular, the DS model does not support indexing opinions about *conjunctions* of keywords. That is, there is no provision for expressing the opinion that a document is relevant to k_1 *and* to k_2, only that it is relevant to k_1 or to k_2. In reality, of course, it is quite common for catalogers to specify conjunctive indexing decisions.

To partly solve the problem, we propose the following extension of the standard notion of a DS frame of discernment. Given a taxonomy graph $T_{\mathcal{K}} =< C, H >$, we augment the graph with a new set of terminal nodes that we call *net_classes*: for each non-terminal class $x \in C$, we add a new terminal class *net_x* to C and a new edge (x, net_x) to H. We interpret each class *net_x* to be the conjunction of all the keywords in $\text{TERM}(x)$. Since the newly-added *net_classes* are all terminal, we are essentially extending the lexical frame of discernment $\mathcal{K}$ (which is a subset of C) to include conjunctions of keywords as well as elementary keywords. At the same time, we do not add any new interior nodes (non-terminal classes) to the taxonomy graph. Instead, each existing non-terminal class is interpreted as the disjunction of all its terminal classes, which now include the net_classes as well. The extension is illustrated in figure 6, where $\mathcal{K} = \{a, b, c\}$. In the figure and hereafter, the notation ab stands for the conjunction $a \wedge b$.

Compared to the original DS model, the extended model implies a less restrictive interpretation of lexical subsets. Whereas in the former the indexing decision $\{a, b, c\}$ meant that the document is relevant to $a \vee b \vee c$, in the extended model the decision is tantamount to saying that the document is relevant to $a \vee b \vee c$, or to abc (which once again stands for $a \wedge b \wedge c$), or to any disjunction of conjunctions in which each one of the subset's keywords appears at least once, e.g. $a \vee bc$, or $ab \vee ac$, or $a \vee b \vee bc$. Disjunctions that involve *less* than all of the subset's keywords, e.g. $a \vee b$, or $b \vee bc$, represent more focused indexing opinions that should be pegged lower in the taxonomy, namely under the lexical subsets $\{a, b\}$ and $\{b, c\}$, respectively.

In the extended model, the decision to index a document in a class like `Cubism` implies that the document is relevant to a logical combination of Cubist artists whose exact form is intentionally left unspecified by the cataloger (and note that the indexing decision `Cubism` implies that the cataloger did not select a more focused class such as `synthetic Cubism` or `constructive Cubsim`). This interpretation is consistent with the notion of 'uncommitted belief,' as it allows catalogers to be as vague as they wish about their indexing opinions, without restricting them to exclusive disjunctions of keywords, as is done in the standard DS model. At the same time, the model supports very concrete conjunctive decisions. For example, if a cataloger thinks that the title `a letter from Braque to Picasso` is relevant to *both* artists, he can choose to index the document on the *net_class* associated with `Picasso` ∧ `Braque`. This will represent a very informed opinion, because the *net_class* is a terminal node in the taxonomy. As a rule, the less committed one's opinion, the higher one's mass is assigned in the taxonomy.

One advantage of a conjunctive frame of discernment is that it does not necessitate corresponding changes in the standard algebra of the DS model. In the extended model, when a source of evidence assigns a mass $m(X)$ to a lexical subset X, it not only makes no distinction between the elementary keywords in X, but also between any disjunction of conjunctions involving all of X's keywords. When *two* sources of evidence assign masses to subsets X and Y, the combined mass (as computed by Dempster's rule) will focus on

$X \cap Y$, as usual, but the intersection will once again be interpreted as *any* disjunction of conjunctions involving *all* the elements in the intersection. Hence, our extended logical interpretation of subsets is closed under Dempster's rule. For example, the common ground of the statements "the truth lies in a logical connective involving $\{a, b, c\}$," and "the truth lies in a logical connective involving $\{b, c, d\}$," is: "the truth lies in a logical connective involving $\{b, c\}$." The support in this proposition will be computed by Dempster rule from $m_1(\{a, b, c\})$ and $m_2(\{b, c, d\})$, as usual, but the exact form of the logical connective is left unspecified at any stage of the analysis.

In sum, the revised interpretation that we propose implies three generic indexing options with respect to a keyword list $X = \{k_1, \ldots k_m\}$. If a cataloger thinks that the document is relevant to a particular keyword $k_i \in X$, the document is indexed directly in the terminal class $\{k_i\}$. If he thinks that the document is relevant to an elementary conjunction $k_{i_1} \wedge, \ldots, \wedge k_{i_m}$ where $\{k_{i_1}, \ldots, k_{i_m}\} = Y \subseteq X$, the document is indexed directly in the terminal *net_class* associated with Y. Finally, if the cataloger thinks that the document is relevant to the keywords in Y, but he is unsure about the exact logical form of that relevance, then the indexing decision is taken to be the lexical subset Y itself.

Recall that in a standard DS frame of discernment, indexing decisions consist of either individual keywords (singletons), or exclusive disjunctions of keywords (non-terminal classes). A conjunctive frame of discernment represents a step forward in terms of specificity, because it supports conjunctive indexing decisions as well. At the same time, neither frame supports direct assignment of degrees of support to any other logical keyword connectives. For example, there is no provision for indexing a document on, say, $a \vee (b \wedge c)$ *directly.* The only way to deal with such an indexing decision is to index the document by default in the sweeping class $\{a, b, c\}$, implying that the document is relevant to a logical combination of the three keywords which the model fails to represent directly. Another disadvantage of conjunctive frames is that they do not support exclusive disjunctions, as an indexing decision like $\{a, b\}$ always implies that the document is relevant to a or to b or to $a \wedge b$. Thus, the design of relevance calculi that support *any* logical combination of keywords without violating the basic philosophy of the DS model remains an important area of future research.

We conclude with some observations about complexity. Augmenting a taxonomy with conjunctive *net_classes* obviously enlarges the indexing space, but the increase is not exponential. In the worst-case situation, one would have to add a new *net_class* element for every non-singleton subset $X \in 2^{\mathcal{K}}$. For example, let $\mathcal{K}$ contain n keywords, and consider a maximal taxonomy that contains $2^n - 1$ classes (we exclude the empty class). If we extend this taxonomy with all possible *net_classes*, the size of the augmented taxonomy would be $2^n - 1 + (2^n - 1 - n) = 2^{n+1} - n - 2$, less than twice the size of the original taxonomy. Note that in reality, the IR semantics will significantly reduce the model's complexity, because the fully augmented taxonomy contains numerous conjunctions that make little or no sense on bibliographical grounds. Accordingly, if conjunctive keywords are introduced only dynamically, as they occur in indexing scenarios, then the actual taxonomy will be significantly smaller than the worst-case (fully-augmented) taxonomy.

The extension that we have proposed - augmenting the frame of discernment yet leaving the non-terminal classes set intact - represents a compromise between two extremes. On one hand, the standard DS model supports only disjunctive indexing opinions. On the other hand, we can envision a model in which indexing opinions can focus *directly* on any disjunction of conjunctions of keywords, such as "the document is directly relevant to $a \vee (b \wedge c)$". The problem with such a model is that it will contain a formidable number of non-terminal classes: if we let $\mathcal{K}$ contain n keywords, such a model will be based on a frame of discernment consisting of 2^n elements, upon which 2^{2^n} classes would be formed.

Obviously, managing such an indexing space would be a daunting task. The intermediate solution that we propose is a reasonable compromise, because (i) it allows indexing opinions to focus directly on any elementary conjunction of keywords, and (ii) it yields a taxonomy which is at most twice as large as a standard DS-based taxonomy.

Appendix B: Proofs

Proposition 1: Let $U = \{u_1, \ldots, u_n\}$ be a set of catalogers with their Boolean indexing opinions $v_1, \ldots, v_n : 2^{\mathcal{K}} \rightarrow \{0, 1\}$. The real function $m : 2^{\mathcal{K}} \rightarrow [0, 1]$ defined by $m(X) = \frac{1}{n} \cdot r(X) = \frac{1}{n} \sum_{i=1}^{n} v_i(X)$ is a mass function, satisfying definition (3).

Proof: For each class $X \in 2^{\mathcal{K}}$, either *all*, *some*, or *none* of the catalogers indexed the document in X. Hence, $r(X) = n$, or $r(X) < n$, or $r(X) = 0$, respectively, implying that $0 \leq m(X) \leq 1$. Hence, $m(\cdot)$ is a mapping from $2^{\mathcal{K}}$ to $[0, 1]$, satisfying the first requirement of being a mass function. The second requirement is that the function will sum up to 1 over all the subsets of $\mathcal{K}$. This is proved as follows. For each cataloger u_i, exactly one of the subsets $X \subseteq \mathcal{K}$ is such that $v_i(X) = 1$. For all other subsets Y, $v_i(Y) = 0$. Thus $\sum_{X \in 2^{\mathcal{K}}} v_i(X) = 1$. We thus have the following:

$$\sum_{X \in 2^K} m(X) = \sum_{X \in 2^K} \frac{1}{n} \sum_{i=1}^{n} v_i(X)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{X \in 2^K} v_i(X)$$

$$= \frac{1}{n} \sum_{i=1}^{n} 1 = 1$$

Further, since no cataloger gives $\emptyset$ as his opinion, it is always true that $v_i(\emptyset) = 0$. Therefore, the third requirement of definition (3) is satisfied. Thus m is a mass function.

<u>**Definition of the $\otimes$ combination rule:**</u> Let $U = \{u_1, \ldots, u_n\}$ be a set of catalogers with their Boolean indexing opinions $v_1, \ldots, v_n : 2^{\mathcal{K}} \rightarrow \{0,1\}$. To denote the fact that the keyword $k \in \mathcal{K}$ was included in the indexing opinion of the ith cataloger, we use the following notation:

$$w_i(k) = \begin{cases} 1 & \text{if } v_i(X) = 1 \text{ and } k \in X \\ \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

If $\mathcal{K} = \{k_1, \ldots, k_n\}$, the binary vector obtained by $w_i(k_1), \ldots, w_i(k_n)$ is denoted w_i and called the Boolean indexing opinion of u_i. The collection of all such opinions of members of U is denoted $W = \{w_i | u_i \in U\}$. To combine the indexing opinions of two sets of catalogers $< U_1, W_1 >$ and $< U_2, W_2 >$, we use the following formulae ($\otimes$):

$$U = U_1 \times U_2, \tag{30}$$

$$W = \{w_{i,j}(\cdot) | u_i \in U_1, u_j \in U_2\}, \tag{31}$$

$$w_{i,j}(k) = w_i(k) \cdot w'_j(k). \tag{32}$$

Where $w_i(k)$ and $w'_j(k)$ are as defined in (29) for $u_i \in U_1$ and of $u'_j \in U_2$.

In section 4.2 we have shown how a mass function can be constructed from a set of catalogers (Proposition 1). Specifically, recall that the mass function induced by the model $Z =< U, W >$, denoted hereafter $m_T(X)$, gives the fraction of catalogers in U, among those catalogers who express an opinion (i.e. $w_i \neq \vec{0}$), whose indexing opinion exactly matched X. This is the same as those catalogers for whom $w_i(k_j) = 1$ if and only if $k_j \in X$. For $Z =< U, W >$, This fraction can be written down exactly:

$$m_z(X) = \frac{\#\{u_i \in U | w_i(k_j) = 1 \text{ if } k_j \in X \text{ and } w_i(k_m) = 0 \text{ if } k_m \notin X\}}{\#\{u_i \in U | w_i \neq \vec{0}\}}, \tag{33}$$

for $X \neq \emptyset$. Of course, $m_T(\emptyset) = 0$. We are now in a position to prove the following.

Proposition 2: Let $Z_1 =< U_1, W_1 >$ and $Z_2 =< U_2, W_2 >$ be two sets of catalogers together with their Boolean indexing opinions, and let $Z =< U, W >$ be the outcome of $Z = Z_1 \otimes Z_2$, as follows: (i) $U = U_1 \times U_2$; and (ii) $W = \{w_{i,j} = w_i \cdot w'_j | w_i \in W_1 \text{ and } w'_j \in W_2\}$. Let $\oplus$ be Dempster's rule as it is applied to mass functions. Let m_{z_1}, m_{z_2}, and $m_{z_1 \otimes z_2}$, be the mass functions induced by the models Z_1, Z_2, and $Z_1 \otimes Z_2$. Then we have the following: $m_{z_1 \otimes z_2} = m_{z_1} \oplus m_{z_2}$.

Proof: This proposition asserts a relationship between the general Dempster Shafer model and the canonical indexing model presented in section 4. The fact that the mapping from one model to the other is homomorphic follows from Hummel and Landy (1988), but we will supply an independent argument here in the context of the indexing model.

Let us assume that there are n_1 catalogers in U_1 and n_2 catalogers in U_2, and let us fix a particular nonempty lexical subset X of the keyword-list $\mathcal{K}$. We wish to show that

$$m_{z_1 \otimes z_2}(X) = (m_{z_1} \oplus m_{z_2})(X) \tag{34}$$

Beginning with the right hand side of (34) and using the definition of Dempster's rule $\oplus$, $(m_{z_1} \oplus m_{z_2})(X)$ is equivalent to

$$\frac{\sum_{A\cap B=X} m_{z_1}(A)\cdot m_{z_2}(B)}{\sum_{A\cap B\neq\emptyset} m_{z_1}(A)\cdot m_{z_2}(B)}. \tag{35}$$

Multiplying top and bottom by $n_1 \cdot n_2$ and distributing, we obtain

$$\frac{\sum_{A\cap B=X} n_1 m_{T_1}(A)\cdot n_2 m_{T_2}(B)}{\sum_{A\cap B\neq\emptyset} n_1 m_{T_1}(A)\cdot n_2 m_{T_2}(B)}. \tag{36}$$

Recalling how mass functions are induced from the opinions of groups of catalogers (Eqn. 23 in Section 4.2), we may interpret this expression as follows. The value $n_1 m_{z_1}(A)$ counts the number of catalogers in U_1 who have indexed the document in the lexical subset A. Likewise, $n_2 m_{z_2}(B)$ counts the number of catalogers in Z_2 who have indexed the document in the lexical subset B. Hence, the product $n_1 m_{z_1}(A) \cdot n_2 m_{z_2}(B)$ counts the number of distinct pairs of catalogers $(u_i, u_j{}')$ in $U_1 \times U_2$ where $u_i \in U_1$ has indexed the document in A and $u_j{}' \in U_2$ has indexed the document in B. Now, according to the way $\otimes$ is defined, if u_i has indexed in A and $u_j{}'$ has indexed in B, then the pair of catalogers $(u_i, u_j{}')$ end up indexing the document in $A \cap B = X$. Thus, the numerator of expression (36) counts *all* the cataloger pairs that end up indexing the document in X.

Precisely the same argument can be used to show that the denominator of (36) counts all the pairs of catalogers who do not index the document in $\emptyset$. Thus (36) gives the fraction of cataloger pairs in $U_1 \times U_2$ that have indexed the document in X out of the pairs of catalogers in $U_1 \times U_2$ who have indexed the document in some non-empty lexical subset, which is exactly the definition of $m_{z_1 \otimes z_2}$, the left hand side of (34).

References

[1] J. Baron. Second-order probabilities and belief functions. *Theory and Decision*, 22, 1987.

[2] G. Biswas, J.C. Bezdek, M. Marques, and V. Subramanian. Knowledge assisted document retrieveal. *Journal of the American Society for Information Science*, 38(2):83–110, 1987.

[3] A. Bookstein. Outline of a general probabilistic retrieval model. *Journal of Documentation*, 39(2):63–72, 1983.

[4] R. Buxton. Modleing uncertainty in expert systems. *International Journal Man-Machine Studies*, 31:415–476, 1989.

[5] W.S. Cooper and P. Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development*, 1(2):99–112, 1982.

[6] W.S. Cooper and M.E. Maron. Foundations of probabilistic and utility-theoretic indexing. *Journal of the ACM*, 25(1):67–80, 1978.

[7] W.B. Croft and D.J. Harper. Using probabilistic models of document retrieval without relevenace information. *Journal of Documentation*, 45(4):285–295, 1979.

[8] A.P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals Mathematics Statistics*, 38:325–339, 1967.

[9] A.P. Dempster. Upper and lower probability inferences based on a sample from a finite univariate population. *Biometrica*, 54:515–528, 1967.

[10] J. Gordon and E.H. Shortliffe. A method for managing evidntial reasoning in a hierarchical hypothesis space. *Artificial Intelligence*, 26:323–357, 1985.

[11] S.P. Harter. A probabilistic approach to automatic keyword indexing: Part i. *Journal of the American Society for Information Science*, 26(4):197–206, 1975.

[12] R.A. Hummel and M.S. Landy. A statistical viewpoint on the theory of evidence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(2):235–247, 1988.

[13] J. Jacoby and V. Slamecka. Indexing consistency under minimal conditions. Bethesda, MD: Documentation, Inc., 1962.

[14] H.E. Kyburg. Bayesian and non-bayesian evidential updating. *Artificial Intelligence*, 31:271–293, 1987.

[15] D.V. Lindley. The probability approach to the treatment of uncertainty in artificial intelligence and expert systems. *Statistical Science*, 2(1):17–24, 1987.

[16] M.E. Maron. Associative search techniques versus probabilistic retrieval models. *Journal of the American Society for Information Science*, pages 308–310, 1982.

[17] M.E. Maron and J.L. Kuhns. On relevance, probablistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7(3):216–244, 1960.

[18] T. Radecki. Trends in research on information retrieval - the potential for improvements in conventional boolean retrieval systems. *Information Processing & Management*, 24(3):219–227, 1988.

[19] S.E. Robertson and K. Sparc Jones. Relevance weighing of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[20] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523, 1988.

[21] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, 1983.

[22] T. Saracevic. Relevance: a review of and a framework for thinking on the notion in information science. *Journal of the American Society for Information Science*, Nov-Dec:321–342, 1975.

[23] S. Schocken and P. R. Kleindorfer. Artificial intelligence dialects of the bayesian belief revision language. *IEEE Transactions on Systems, Man, and Cybernetics*, 19:1106–1121, 1989.

[24] S. Schocken and J. Pyun. A dempster shafer model of relevance. In *Proc. of the 23rd Hawaii Int. Conf. on System Sciences*, pages 544–551, 1990.

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

[26] G. Shafer. Probability judgement in artificial intelligence and expert systems. *Statistical Science*, 2(1):3–16, 1987.

[27] M.E. Stevens. Automatic indexing: A state of the art report. Washington, DC: U.S. Government Printing Office, 1965.

[28] P. Thompson. A combination of expert opinion approach to probabilsitic information retrieval: Part i: the conceptual model. *Information Processing & Management*, 26(3):371–382, 1990.

[29] R.M. Tong and D.G. Shapiro. Experimental investigations of uncertainty in a rule-based system for informtaion retrieval. *International Journal of Man-Machine Studies*, 22:265–282, 1985.

[30] H. Turtle and W.B. Croft. Evaluation of an inference network-based retrieval.model. *ACM Transactions on Information Systems*, 9(3):187–222, 1991.

[31] J. Yen. Gertis: a dempster-shafer approach to diagnosing hierarchical hypotheses. *Communications of the ACM*, 32(5):573–585, 1989.

[32] C.T. Yu and G. Salton. Precision weighting - an effective automatic indexing method. *Journal of the ACM*, 23(1):76–88, 1976.

[33] L.A. Zadeh. A simple view of the dempster-shafer theory of evidence and its implication on the rule of combination. *The AI Magazine*, pages 85–90, 1986.

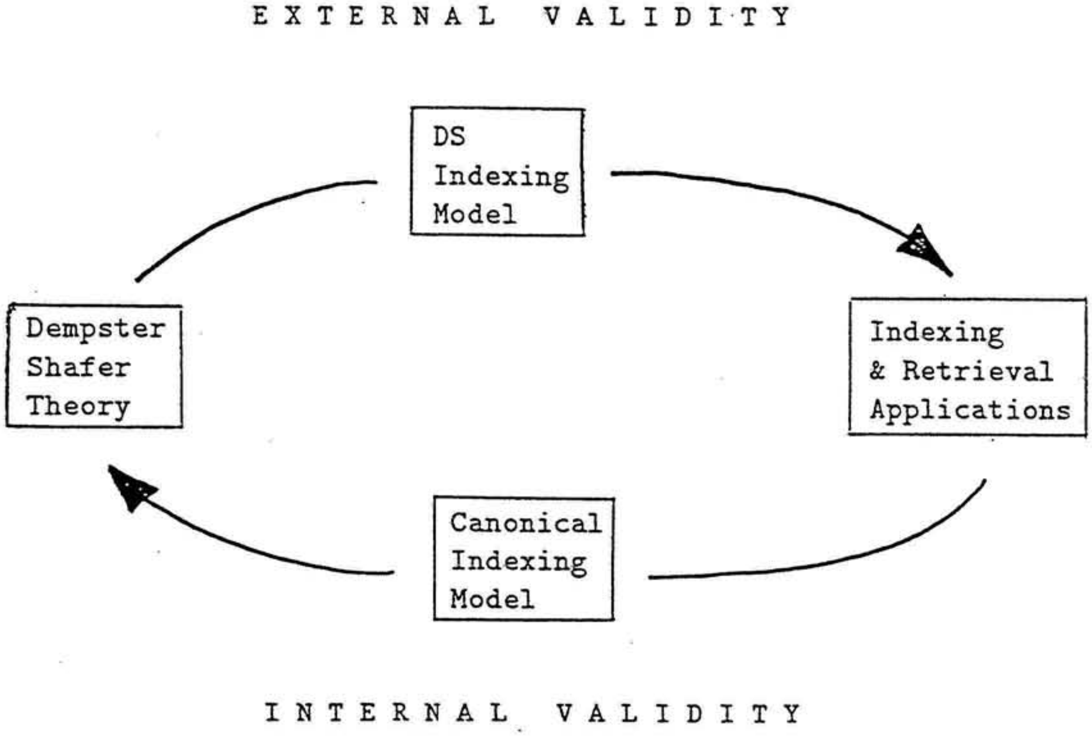


Figure 1: A pictorial description of the paper's methodology. Section 3 uses the terminology and rationale of the Dempster Shafer theory to derive a DS indexing model for IR applications (top arrow). Taking the opposite direction, Section 4 presents a canonical indexing model that uses a particular method to combine cataloger opinions. Toward the end of the paper, we show that the canonical model provides a probabilistic and domain-independent interpretation of the Dempster Shafer model (bottom arrow).

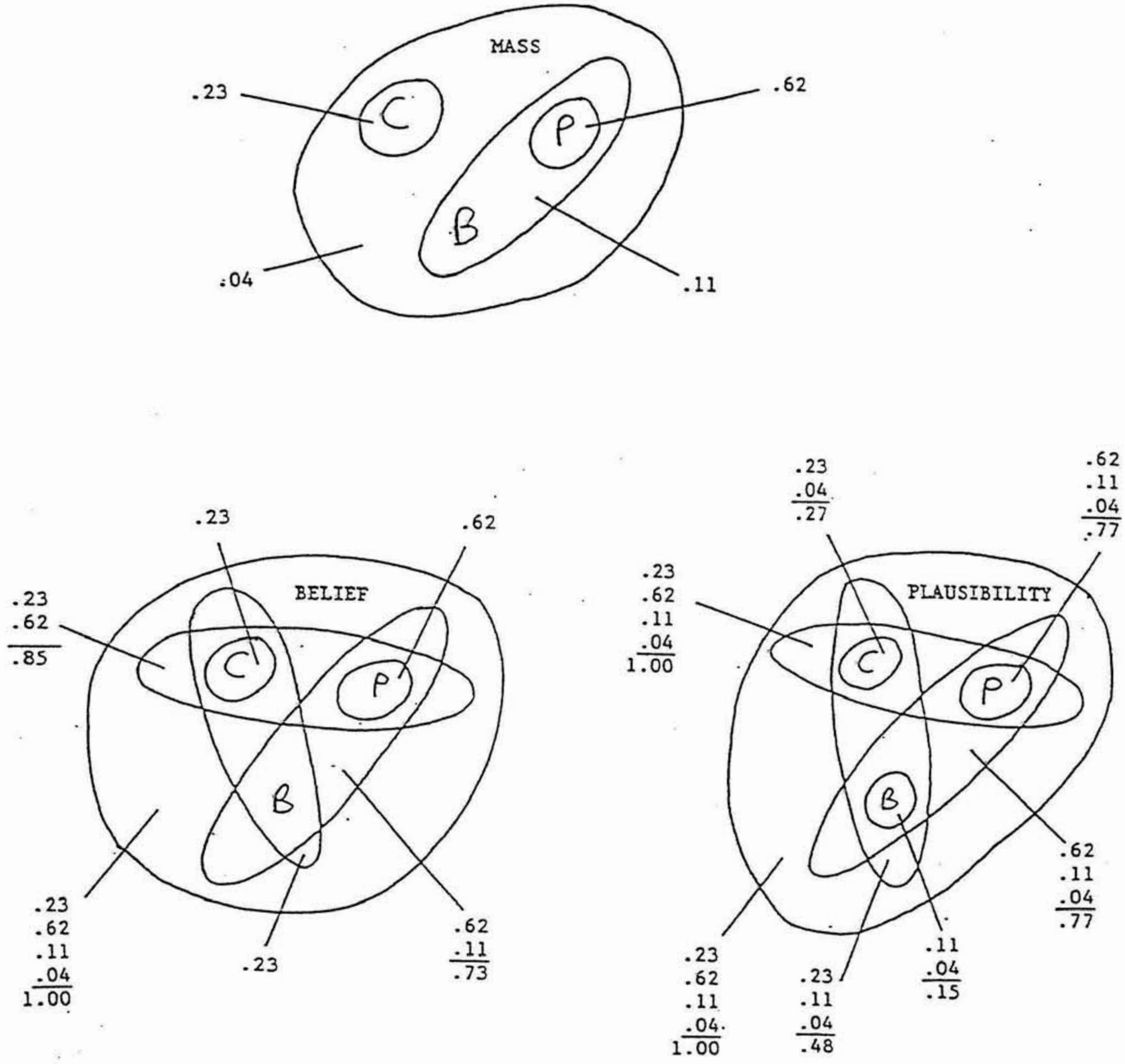


Figure 2: An illustration of the relationship that exists among mass (top), belief (left), and plausibility (right) functions that represent the same set of primitive degrees of support.

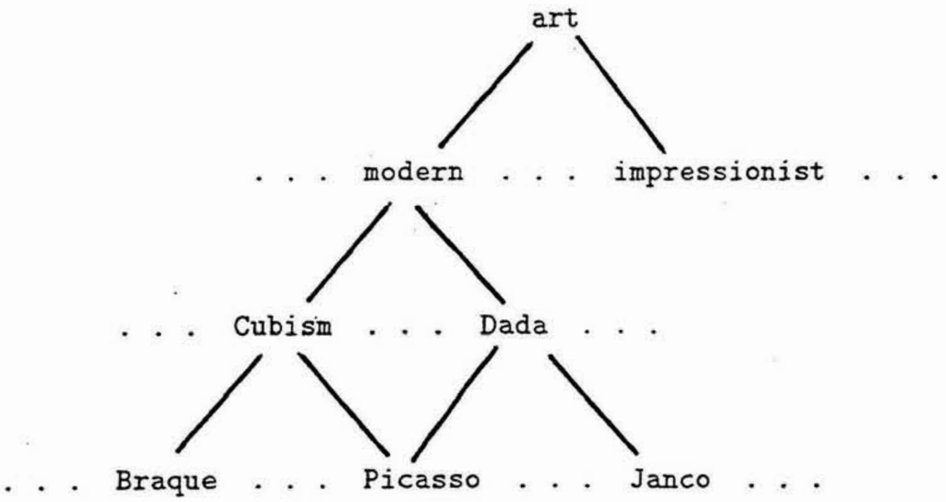

art
. . . modern . . . impressionist . . .
. . . Cubism . . . Dada . . .
. . . Braque . . . Picasso . . . Janco . . .

Figure 3: An excerpt from an art-related taxonomy designed to classify documents on major artists and artistic movements.

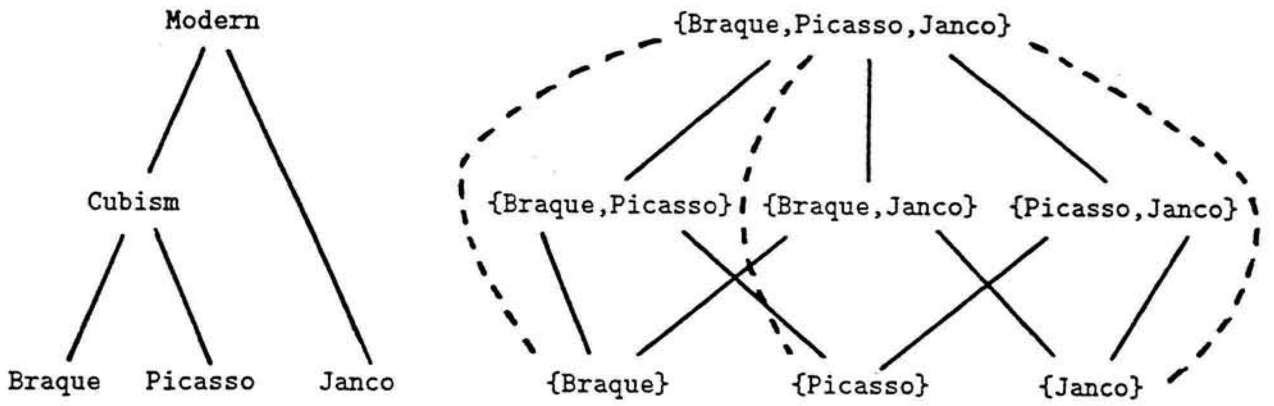
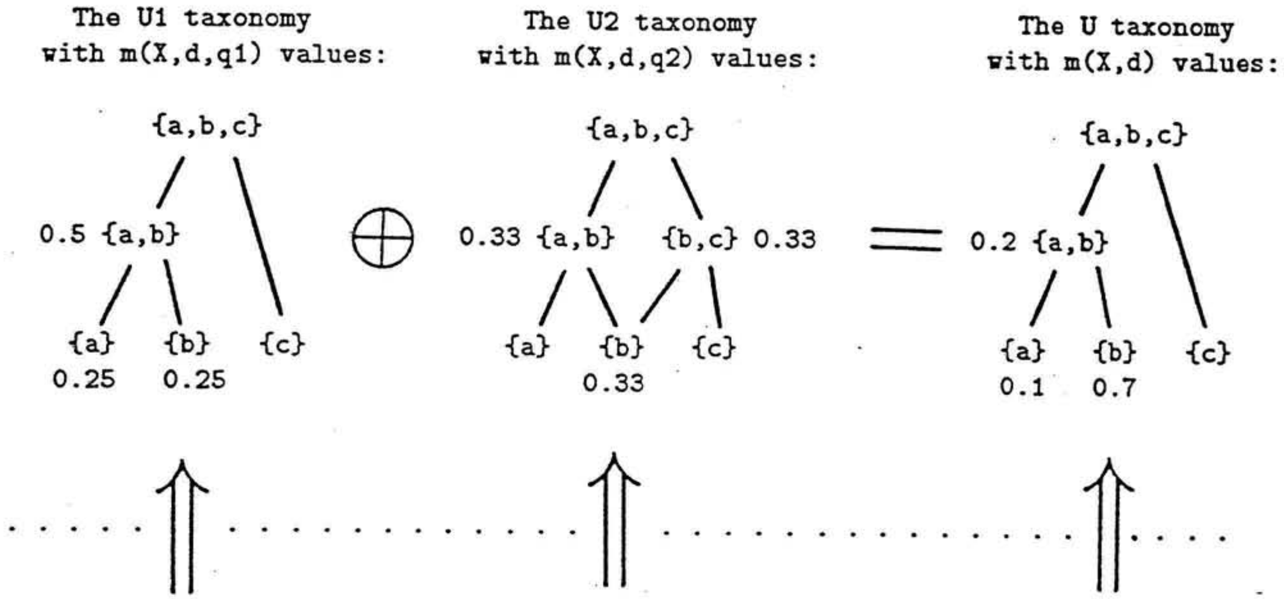


Figure 4: The relationship between a taxonomy (left) and a lexical frame of discernment (right). Using the terminals set $\mathcal{K} = \{\texttt{Braque,Picasso,Janco}\}$ of the taxonomy, one constructs the transitive closure of the subset graph $G^*_{\mathcal{K}}$, where broken lines represent the edges added by the transitive closure operation. The embedment is established by mapping each node x in the taxonomy on its terminals set $\text{TERM}(x)$ in $G^*_{\mathcal{K}}$. In this particular case we have (using the first letter of each keyword) $\texttt{b} \mapsto \{\texttt{b}\}$, $\texttt{p} \mapsto \{\texttt{p}\}$, $\texttt{j} \mapsto \{\texttt{j}\}$, $\texttt{c} \mapsto \{\texttt{b}, \texttt{p}\}$, and $\texttt{m} \mapsto \{\texttt{b}, \texttt{p}, \texttt{j}\}$. The edges map in a corresponding fashion.



U1	a	b	c
u1:	1	1	0
u2:	1	0	0
u3:	0	1	0
u4:	1	1	0

⊗

U2	a	b	c
u1':	0	1	1
u2':	1	1	0
u3':	0	1	0

=

U1xU2	a	b	c
u1,u1':	0	1	0
u1,u2':	1	1	0
u1,u3':	0	1	0
u2,u1':	0	0	0
u2,u2':	1	0	0
u2,u3':	0	0	0
u3,u1':	0	1	0
u3,u2':	0	1	0
u3,u3':	0	1	0
u4,u1':	0	1	0
u4,u2':	1	1	0
u4,u3':	0	1	0

Figure 5: The DS model (top) and its probabilistic interpretation (bottom) in one particular indexing scenario. The individual indexing opinions of two groups of catalogers (U_1 and U_2) induce two different taxonomies and two different relevance functions – $m(X,d,q_1)$ and $m(X,d,q_2)$. The combination of the relevance functions via Dempster's rule $\oplus$ at the *classifiers level* and the combination of the opinions via the cartesian consensus rule $\otimes$ at the *catalogers level* lead to the same pooled index depicted at the top right of the figure.

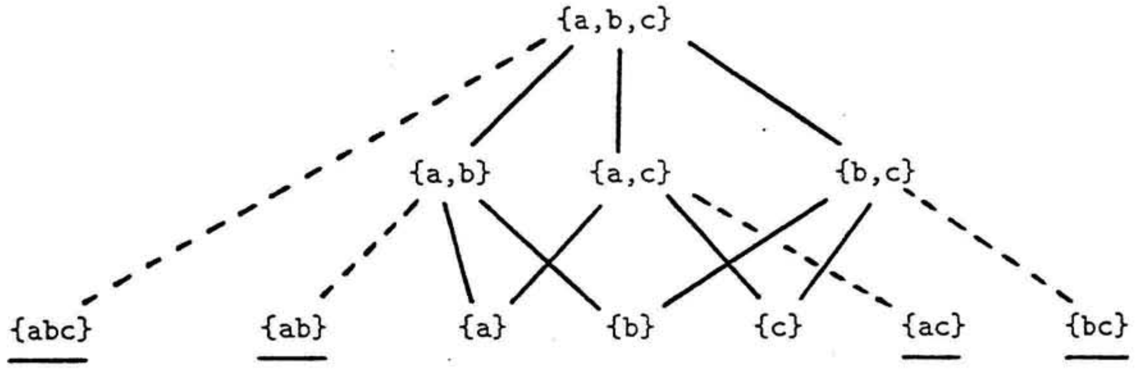


Figure 6: A conjunctive frame of discernment. Nodes of type $\{x, y\}$ represent disjunctions, as usual, whereas nodes of type $\{xy\}$ represent newly-added conjunctions. The nodes and edges added by the extension are underlined and broken, respectively. Note that while the extended model contains many more terminal classes than the original model, it has the same number of non-terminal classes (albeit with a new disjunctive interpretation). As a result, the size of the extended taxonomy is less than twice the size of the original taxonomy.