AN INTUITIVE INTERPRETATION OF THE THEORY OF EVIDENCE IN THE CONTEXT OF BIBLIOGRAPHICAL INDEXING

by

Shimon Schocken

AN INTUITIVE INTERPRETATION OF THE THEORY OF EVIDENCE IN THE CONTEXT OF BIBLIOGRAPHICAL INDEXING

by

Shimon Schocken
Information Systems Department
Leonard N. Stern School of Business
New York University
New York, NY 10003

April 1992

Center for Research on Information Systems
Information Systems Department
Leonard N. Stern School of Business
New York University

Working Paper Series

STERN IS-92-18

The author is indebted to Bob Hummel for useful comments on earlier versions of this paper.

An Intuitive Interpretation of the Theory of Evidence in the Context of Bibliographical Indexing

Models of bibliographical Indexing concern the construction of effective keyword taxonomies and the representation of relevance between documents and keywords. The theory of evidence concerns the elicitation and manipulation of degrees of belief rendered by multiple sources of evidence to a common set of propositions. The paper presents a formal framework in which adaptive taxonomies and probabilistic indexing are induced dynamically by the relevance opinions of the library's patrons. Different measures of relevance and mechanisms for combining them are presented and shown to be isomorphic to the belief functions and combination rules of the theory of evidence. The paper thus has two objectives: (i) to treat formally slippery concepts like probabilistic indexing and average relevance, and (ii) to provide an intuitive justification to the Dempster Shafer theory of evidence, using bibliographical indexing as a canonical example.

Keywords: Probabilistic indexing, measures of relevance, Dempster Shafer theory of evidence, evidential reasoning.

Center for Digital Economy Research
Stern School of Business
Working Paper IS-92-18

1 Introduction

Consider a finite and exhaustive set of n 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 – non-negative numbers that sum up to unity (over the mutually-exclusive propositions) and obey the axioms of subjective probability – and use Bayesian inference rules to revise them. One problem with this approach is that it doesn't offer a clear way to model the various degrees of "uncommitted beliefs," or "second order uncertainties," that characterizes most realistic reasoning problems. For example, consider the extreme case of "insufficient reason," in which one knows absolutely nothing about the n propositions. In the face of this total ignorance, the common Bayes-LaPlace solution (as well as the unconstrained maximum entropy solution) is to assign a degree of belief of $1/n$ to each of the propositions under consideration.

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 quasi-probabilistic models that attempt to capture the elusive notion of uncommitted belief explicitly. Perhaps the best known model in this category is the so-called "theory of evidence," originated by Dempster's 1967 work on upper and lower probabilities [3],[4]. Dempster's ideas, which were based on a frequentist view of probability, were refined and extended by Shafer [13], resulting in an elaborate theory for representing and revising subjective beliefs as well as chance likelihoods.

When the work of Dempster and Shafer was "discovered" by the artificial intelligence community, it immediately stirred a considerable interest among researchers and practitioners of expert systems – an area in which normative models of belief formation play a key role. In particular, the model holds promise for supporting rule-based inference under uncertainty, an aspect of expert systems that was traditionally dominated by *ad-hoc* belief revision calculi whose relationship to probability theory was quite murky. In contrast, the Dempster Shafer model rests on a solid and defensible mathematical foundation. Yet the probabilistic roots of the model remain controversial: whereas Shafer argues that the theory of evidence is a natural extension of probability theory [14], critiques of the theory, like Lindley, view it

as a reformulated version of a specialized, albeit interesting, case of classical probabilistic reasoning [8]. The debate is not helped by the theory's nomenclature, which is expansive and somewhat obscure.

Several authors have tried to give plausible interpretations of the theory of evidence, using familiar metaphors. For example, Zadeh illustrated how the Dempster Shafer model can be used to support fuzzy queries about *interval-valued*, rather than point-valued, attributes, in a relational database [18]. In a similar vein, this paper attempts to justify the underlying rationale of the theory of evidence in the way of analogy, using a familiar *bibliographical indexing* metaphor. Therefore, it should be emphasized that the paper is not intended to propose a new Dempster Shafer approach to document storage and retrieval, as was done for example by Tong and Shapiro [15] and by Biswas et al [2], among others. The present paper deviates from this line of research in two ways. First, it concerns not retrieval, but *classification*, of information. Second, the paper's style is axiomatic, not prescriptive: our chief objective is to present a canonical example which supports the internal, rather than external, validity of the Dempster Shafer model.

Previous work: The research reported here is relevant to two lines of previous research: (i) efforts to interpret the theory of evidence on logical or probabilistic grounds; and (ii) efforts to apply the theory to hypotheses spaces that have specific structures. To the best of our knowledge, Gordon and Shortliffe were the first to recognize the potential utility of the theory of evidence to artificial intelligence applications [5]. In particular, they showed how a Dempster Shafer belief 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 forms a hierarchy. However the degrees of belief that their approach generated did not conform to a standard probabilistic interpretation. The problem was taken up by Yen, who built an expert system shell called GERTIS in which uncertainty was managed by an extended Dempster Shafer calculus whose degrees of belief yielded to a probabilistic interpretation [17]. Yen also made the critical observation that the theory of evidence is based on mappings from an evidence space to an hypotheses space, and that these mappings can be viewed as a collection of conditional probabilities whose values are either zeroes or ones. He then proceeded to propose combination formulae that can handle "regular" probabilities as well.

Gordon and Shortliffe and Yen were motivated by a practical objective – trying to build

a credible belief calculus for clinical reasoning in a hierarchical hypotheses space. Coming from a totally different direction, Hummel and Landy analyzed the probabilistic foundation of the theory of evidence *in general*, without making any assumptions on the underlying domain or the logical structure of the hypotheses [6]. In contrast to other researchers who attempted to interpret high-level constructs of the theory of evidence *directly* (e.g. Baron [1], Kyburg [7], and Schocken and Kleindorfer [12]), Hummel and Landy took a more fundamental viewpoint that showed how the theory's belief functions were implicitly linked to a hypothetical space of boolean expert opinions. They described a logical mechanism for pooling opinions in that boolean space, and showed that the mechanism was isomorphic to Dempster's rule. Concluding that Dempster's rule was limited to tracking only boolean opinions, Hummel and Landy proposed alternative combination formulae that are sensitive to probabilistic opinions as well.

Hummel and Landy's probabilistic interpretation of the theory of evidence is complete and satisfying. However, their abstract mathematical analysis made no use of canonical examples, and it is therefore difficult to map their approach on real-life problems like clinical diagnosis or rule-based inference. One objective of the present paper is to "operationalize" Hummel and Landy's insights in the context of a practical example – bibliographical indexing – and demonstrate that their boolean spaces are useful not only on normative grounds, but also in modeling practical reasoning problems that are characterized by several "layers" of inference.

The plan of the paper is as follows. §2 gives a brief overview of the theory of evidence, including definitions and illustrations of the frame of discernment, mass, belief, and plausibility functions, and Dempster's rule. This sets the stage for §3 and §4, which present the bibliographical model around which the paper evolves. §3 describes the keyword taxonomy and relates it to the frame of discernment, and §4 discusses various measures of relevance that are then shown to be equivalent to Shafer's mass, belief, and plausibility functions. §5 illustrates a plausible pooling mechanism for combining multiple relevance opinions elicited from distinct groups of library patrons. §6 is a discussion section.

2 Overview of the Theory of Evidence

The theory of evidence concerns the representation and manipulation of degrees of belief 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 belief are normally assigned to elements of θ, the theory of evidence assigns degrees of belief to *subsets* of propositions, i.e. to members of the power set 2^θ. In this paper subsets of propositions are referred to as "possibilities."

The theory of evidence offers several complementary ways to express degrees of belief in possibilities. In particular, the theory defines several mappings from 2^θ to $[0,1]$ among which we focus in this paper on what is 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, the three mappings can be viewed as alternative means to keep score of the same primitive set of degrees of belief. When several sources of evidence lend credence to a common set of possibilities (the evidential support may be expressed in any one of the three mappings), the overall belief in the possibilities can be computed through Dempster's rule. The remainder of this section reviews the main constructs of this model as they unfold in a the context of a simple example. Throughout this exposition, our goal is to emphasizes clarity and play down complex notation. For a complete and formal treatment of the theory the reader is referred to Chapter 2 of Shafer's monograph on the "Mathematical Theory of Evidence" [13]

The Frame of Discernment: Let $\theta = \{q_1, \ldots, q_n\}$, be an exhaustive set of mutually exclusive propositions (or hypotheses, or simply labels). The power-set which enumerates all the subsets of θ is denoted 2^θ. For example, consider the simple frame of discernment $\theta = \{up, same, down\}$, representing three alternative directions of tomorrow's stock market. The power set is $2^\theta = \{\{up\}, \{same\}, \{down\}, \{up, same\}, \{up, down\}, \{same, up\}, \{up, same, down\}, \emptyset\}$. Each of these subsets represents a possibility, which is essentially a disjunction of propositions. Thus, the statement "the truth lies in $\{up, same\}$" implies one's belief that tomorrow's market will either remain the same, or will go up (which is the same as saying "the market will not go down"). Although the number of possibilities in 2^θ grows exponentially with the cardinality of θ, the semantics of the frame of discernment will usually render most of these possibilities nonsensical. Such possibilities can be effectively

eliminated from the model by setting the belief in them to zero.

Mass Functions: An assignment of degrees of belief to possibilities $m : 2^\theta \rightarrow [0,1]$ with the properties:

$$m(\emptyset) = 0 \tag{1}$$

$$\sum_{A \subseteq 2^\theta} m(A) = 1 \tag{2}$$

is called a *mass function.* As a rule, the so-called *uncommitted belief* – the mass which is left over after all the *proper* subsets of θ have been assigned degrees of belief – is assigned to θ. Operationally, then, after all the non-zero degrees of belief have been elicited from the expert or the source of evidence that provides $m(\cdot)$, the mass $m(\theta)$ is set to $1 - \sum_{x \neq \theta} m(x)$. To illustrate, consider a bullish expert (expert no. 1) who distributes his belief among the various stock market directions as follows: $m_1(\{up, same\}) = 0.6$ and $m_1(\{down\}) = 0.1$. The complete mass function of the expert is:

$$\begin{aligned} m_1(\{up, same\}) &= 0.6 \\ m_1(\{down\}) &= 0.1 \\ m_1(\theta) &= 0.3 \\ m_1(A) &= 0 \ \textit{for all other proper subsets of } \theta \end{aligned} \tag{3}$$

The "uncommitted belief" displayed by the expert is set by convention to $m_1(\theta) = 1 - 0.6 - 0.1 = 0.3$. The rationale for assigning the uncommitted belief to θ is as follows. If an expert knows absolutely nothing about the stock market, we can represent his or her ignorance by the belief function $m(\{up, same, down\})=1$ and $m(\cdot) = 0$ elsewhere, implying the (not very useful) conviction that the market will either go up, down, or stay the same. Other experts might display smaller levels of $m(\theta)$, representing more educated guesses about the market's direction. Hence, unlike a standard probabilistic design, in which the notion of uncommitted belief is not well-defined, the theory of evidence offers explicit means to quantify and manipulate uncommitted beliefs[2].

[2]Uncommitted beliefs or "second-order uncertainties" can also be expressed with standard probabilistic or statistical tools (e.g. Baron [1]), but there is no *simple* way to do it. The theory of evidence is unique in that it treats the notion of uncommitted belief at the axiomatic level.

It's important to observe that the mass function represents indivisible, or atomic, degrees of belief. For example, the magnitudes of $m(\{up, same\}), m(\{up\})$, and $m(\{same\})$ are unrelated, and a belief assignment, say, $m(\{up, same\}) = 0.9$, $m(\{up\}) = 0$, and $m(\{same\}) = 0$ is not inconsistent with the theory. This particular function represents an expert who strongly believes that the market will not go down, although he is not willing to say anything specific beyond this prediction.

The Core: The *core* of a mass function m is the set of possibilities $X \in 2^\theta$ for which $m(X) > 0$. For example, the core induced by (3) is $C_1 = \{\{up, same\}, \{down\}, \theta\}$. In other words, the core of a mass function is the subset of all likely possibilities, in the view of one particular expert. Suppose now that we have access to a second expert (expert no. 2), whose belief in the market direction is captured by the following mass:

$$\begin{aligned} m_2(\{up\}) &= 0.8 \\ m_2(\theta) &= 0.2 \\ m_2(A) &= 0 \ \textit{for all other proper subsets of } \theta \end{aligned} \tag{4}$$

With the core $C_2 = \{\{up\}, \theta\}$. Is there a credible way to combine the two expert opinions (3)-(4) and generate a global prediction concerning the direction of tomorrow's stock market? As a first approximation, one can focus on the set of possibilities which both experts agree are likely. In particular, if expert 1 thinks that X is likely and expert 2 thinks that Y is likely, then *both* experts agree that $X \cap Y$ is likely (recall that a possibility is a disjunction of propositions). This leads to the following definition of a "pooled core": Let $m_1, m_2 \rightarrow 2^\theta$ be two mass functions with cores C_1 and C_2. The pooled core $C = C_1 \oplus C_2$ is defined as follows:

$$C_1 \oplus C_2 = \bigcup_{X \in C_1, Y \in C_2} X \cap Y \tag{5}$$

For example, the pooled core of (3) and (4) is $C = \{\{up\}, \{up, same\}, \{down\}, \theta\}$. In general, then, the pooled core can be viewed as a first approximation of the degree of consensus or disagreement between two expert 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 experts agree on nothing; If $C_1 \oplus C_2$ is not empty, we have an overlap of some opinions. Of course

the problem with such boolean observations is that they merely *identify* areas of mutual agreement (or lack thereof) between two experts. In order to compute the *intensity* of such agreements, a more sensitive pooling mechanism is called for. Dempster's rule provides one such mechanism.

Dempster's Rule: The most fundamental (and debateable) pillar of the theory of evidence is the convention that once degrees of belief 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 $C_1 = \{A_1, \ldots, A_n\}$ and $C_2 = \{B_1, \ldots, B_m\}$, respectively. Dempster's rule computes the pooled mass function $m = m_1 \oplus m_2 : 2^\theta \rightarrow [0,1]$ as follows:

$$m'(X) = \sum_{A_i \cap B_j = X} m_1(A_i) \cdot m_2(B_j) \tag{6}$$

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

$$m(X) = \frac{1}{1-k} \cdot m'(X) \tag{8}$$

The rationale behind (6-8) can be explicated through an intersection table [13]. In the two-experts stock market example, the table has the following form:

	$m_1(up, same) = 0.6$	$m_1(down) = 0.1$	$m_1(\theta) = 0.3$
$m_2(up) = 0.8$	$m(up) = 0.48$	$m(\emptyset) = 0.08$	$m(up) = 0.24$
$m_2(\theta) = 0.2$	$m(up, same) = 0.12$	$m(down) = 0.02$	$m(\theta) = 0.06$

The top margin of the table records the mass function of the first expert excluding its zero elements, i.e. the set of values $m_1(A_1), \ldots, m_1(A_n)$. The left margin of the table records the mass function of the second expert excluding its zero elements, i.e. the set of values $m_2(B_1), \ldots, m_2(B_n)$. (The curly brackets are dropped for the sake of brevity, e.g. $m(up, same)$ stands for $m(\{up, same\})$, etc.). Inside the table, the (i,j)'th cell records the pooled mass $m_1 \oplus m_2(A_i \cap B_j)$, which is taken to be the product $m(A_i) \cdot m(B_j)$. Using these entries and following (6-7) one obtains:

$$\begin{aligned} m'(up) &= 0.48 + 0.24 = 0.72 \\ m'(up, same) &= 0.12 \\ m'(down) &= 0.02 \\ m'(\theta) &= 0.06 \\ k &= 0.08 \end{aligned} \tag{9}$$

And, after multiplying by $\frac{1}{1-k}$ one obtains:

$$\begin{aligned} m(up) &= 0.78 \\ m(up, same) &= 0.13 \\ m(down) &= 0.02 \\ m(\theta) &= 0.07 \\ m(\emptyset) &\stackrel{\text{def}}{=} 0 \end{aligned} \tag{10}$$

Since the $m(\cdot)$'s sum up to unity and $m(\emptyset) = 0$, the mapping $m = m_1 \oplus m_2$ that emerges from Dempster's rule is also a mass function. We now turn to give a preliminary interpretation of this pooling mechanism in light of the above example. In essence, the rule computes a measure of agreement among two experts who express independent opinions about the likelihoods of various possibilities drawn from a common set of propositions. The rule is conservative in the sense that it focuses only on those possibilities which *both* experts agree are likely. The magnitude of the pooled agreement is computed through the product of the two individual masses, which can be interpreted as the joint probability that *both* experts agree on the possibility under consideration. This explains the product operator in (6). Now, because the experts express their opinions over 2^θ rather than over θ, a joint agreement on any one possibility can occur in more than one way, i.e. whenever the experts agree that a superset of that possibility is likely. This explains the summation

operator in (6). Finally, when a pairing of two expert opinions results in a null possibility (empty set), the multiplication of their individual masses may still be positive. This is an anomaly, since the definition of a mass function (1) requires that the mass of the null possibility be zero. This explains the role of (7-8), in which (i) the pooled masses rendered to null possibilities is summed into k, and (ii) k is eliminated from the total mass and the remaining mass is divided by $1 - k$ to ensure that it will sum up to unity.

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

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

Whereas $m(A)$ measures the belief rendered to A (a subset of propositions) directly, $Bel(A)$ measures the total belief rendered to A and to all its subsets (each being a more specific proposition). For example, the belief function implied by the mass function (10) is as follows:

$$\begin{array}{lcl} Bel(up) & = & 0.78 \\ Bel(same) & = & 0 \\ Bel(down) & = & 0.022 \\ Bel(up, same) & = & 0.13 + 0.78 = 0.91 \\ Bel(up, down) & = & 0.78 + 0.02 = 0.8 \\ Bel(same, down) & = & 0.02 \\ Bel(up, same, down) & = & 0.78 + 0.13 + 0.02 + 0.07 = 1 \\ Bel(\emptyset) & = & 0 \end{array} \tag{12}$$

Note that (1-2) and (11) imply that $Bel(\emptyset) = 0$ and $Bel(\theta) = 1$ always. In general, the Bel function is completely determined by the mass function m, and, likewise, m can be recovered from Bel's definition ([13], p. 39). Since the two functions are mathematically equivalent, the question of whether to use m or Bel to elicit and manipulate degrees of belief depends on cognitive and on efficiency considerations. The theoretical "need" for a separate Bel function can be motivated on logical as well as on probabilistic grounds. The logical argument is based on the observation that if X and A are taken to be disjunctions

of propositions, the set-theoretic statement $X \subseteq A$ is equivalent to the "rule" $X \rightarrow A$. Therefore, the sum of all the degrees of belief in propositions X that *imply* A can be viewed as a measure of total support in A. The probabilistic argument for keeping track of $m(A)$ as well as $Bel(A)$ is that the former can be interpreted as the probability that the truth lies in A and the latter as the probability that the truth lies in a subset of A.

Whereas $Bel(A)$ measures the *total* degree of belief rendered to a possibility A, the *plausibility* function, denoted $Pl(A)$, measures the *maximal* degree of belief that A can possibly attain under a given mass function m. Specifically:

$$Pl(A) = \sum_{X \cap A \neq \emptyset} m(X) \tag{13}$$

In words, $Pl(A)$ records the total mass allocated to all the possibilities with which A intersects. Since a possibility is a disjunction of propositions, the mass m rendered to it can "float" freely to any one of its subsets. In the extreme case, a single subset of the possibility may inherit its entire mass. Hence, $Pl(A)$ is the upper bound of $Bel(A)$.

To do justice to the theory of evidence, it should be noted that the derivation of Bel and Pl from m is only one way to define these functions. Shafer provided direct definitions of Bel, Pl, and several other related functions, as well as mappings from one function to another. He has also emphasized the key role that subaditivity plays in the theory, a point which we now turn to illustrate. Denoting the complement set $\theta \setminus A$ by $\overline{A}$, (11) and (13) imply two important properties, as follows:

$$Pl(A) = 1 - Bel(\overline{A}) \tag{14}$$

$$0 \leq Bel(A) \leq Pl(A) \leq 1 \tag{15}$$

If a certain Bel_b were a *Bayesian* representation of degrees of belief, the additvity axiom of Bayesian inference ($A \cap B = \emptyset \rightarrow Bel(A \cup B) = Bel(A) + Bel(B)$) would imply that

$$Bel_b(A) = 1 - Bel_b(\overline{A}) \tag{16}$$

However, (14) and (15) imply that in the general case $Bel(A) \leq 1 - Bel(\overline{A})$, leading to the famous subaditivity property of the theory of evidence:

$$Bel(A) + Bel(\overline{A}) \leq 1 \tag{17}$$

One implication of subaditivity is that the belief that one holds in a possibility does not automatically imply one's disbelief in the negation of that possibility. For example, a physician's belief that a patient suffers from a certain disease should not necessarily rule out the possibilities of other diseases, *especially if the physician is not sure about his prognosis.* In particular, the difference $1 - Bel(A) - Bel(\overline{A})$ is called the uncommitted belief w.r.t. to A. If Bel were a Bayesian representation of degrees of belief, the uncommitted belief would be 0 by definition. This is best illustrated in the "state of insufficient reason," in which one knows absolutely nothing about a set of n propositions θ. Whereas the common solution is to set $Bel(q) = 1/n$ for all $q \in \theta$, the theory of evidence would set $Bel(\theta) = 1$ and $Bel(A) = 0$ for all the other proper subsets of θ. This is the case when the uncommitted belief is at maximum.

The interpretation of Bel and Pl 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 of evidence provides means to express the belief in every hypothesis A through the interval $[Bel(A), Pl(A)]$, which may be updated as new evidence about A is brought to bear. Note that the width of the interval, $Pl(A) - Bel(A)$, is by definition $1 - Bel(A) - Bel(\overline{A})$, or the uncommitted belief w.r.t. A. If the uncommitted beliefs induced by a certain mass function m were 0 for all the hypotheses under consideration, the intervals would degenerate to point beliefs 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 about the hypotheses under consideration, the degree of belief in possibilities A drawn from θ will be allowed to "float" between $Bel(A)$ and $Pl(A)$. One benefit of such a model is that it is more robust and less prone to human errors in assessing subjective probabilities.

The purpose of this section was to present the key features of the Dempster Shafer theory

of evidence and to illustrate the theory's power in modeling aspects of evidential reasoning that seem to defy simplistic Bayesian solutions. We also sought to demonstrate that with the theory's alternative axiomatic system, specialized mappings, and controversial combination rules, this power does not come free: "It is appropriate to examine the formal relations between various Bayesian and non-Bayesian approaches to what has become to be called evidence theory, in order to explore the question of whether the new techniques are really more powerful than the old, and the question of whether, if they are, this increment of power is bought at too high a price (Kyburg [7])." It is in that spirit that we now turn to explore the theory's implicit foundations in an attempt to understand whether it truly extends classical Bayesian inference, or merely reformulates it. This question will be explored in the context of a canonical example, taken from the domain of bibliographical databases.

3 The Taxonomy

Let D be a set of documents about a certain subject and let $\mathcal{S}$ be a structured set of keywords, or a *taxonomy*, designed to facilitate rapid access to the documents in D. The act of *indexing* or *classification* amounts to assigning each document $d \in D$ a subset of keywords $S_d \subseteq \mathcal{S}$, denoted "the index of d," which is supposed to serve as a pointer to the document's contents. We distinguish between two types of taxonomies: static and adaptive. A *static taxonomy* consists of a fixed and unmodifyable 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 classification process itself.

A Static Taxonomy: A static taxonomy is a fixed set of classes, or categories, designed to organize documents in a particular subject of interest. For example, consider an unordered collection of documents about major artists and the movements to which they belonged. Suppose that a domain expert (e.g. an art scholar) proposes to organize the documents according to the following classes: $S = \{\texttt{Art}, \texttt{Braque}, \texttt{Cubism}, \texttt{Dada}, \texttt{Impressionist}, \texttt{Janco}, \texttt{Modern}, \texttt{Picasso}\}$. Suppose further that the expert also indicates the taxonomical relationship of the classes. Specifically, if we let $H(x, y)$ code the assertion "y is a direct sub-class of x", the expert might specify a relation like
$H = \{(\texttt{art}, \texttt{modern}), (\texttt{art}, \texttt{impressionists}), (\texttt{modern}, \texttt{Cubism}), (\texttt{Cubism}, \texttt{Braque}),$

(Cubism, Picasso), (Dada, Picasso), (Dada, Janco)}. The resulting taxonomy is depicted in figure 1.

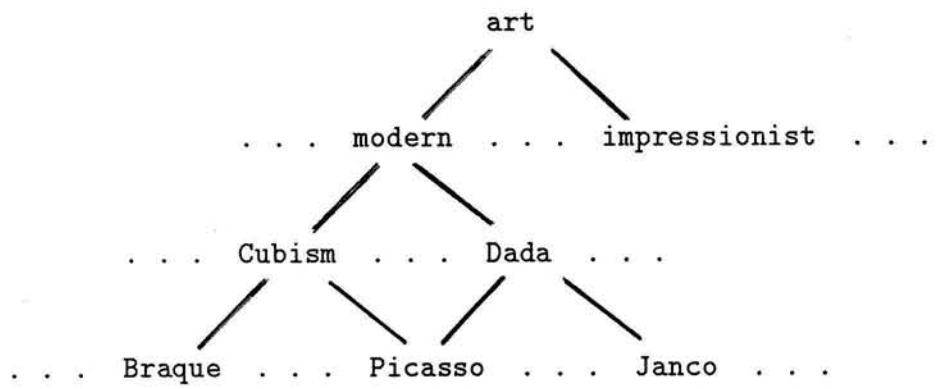


Figure 1: An excerpt from an art-related taxonomy designed to classify documents on major artists and artistic movements.

Formally, then, a taxonomy is a pair $\mathcal{S} =< S, H >$. S is a set of classes, and H is an acyclic relation on $S \times S$ with two restrictions: (a) there exist one $r \in S$ for which there exist no other $x \in S$ with $H(x, r)$, and (b) there exist one or more $k's \in S$ for which there exist no $x \in S$ with $H(k, x)$. In what follows, r is called the *root* of the taxonomy (e.g. art), and the k's are called *terminal classes.* The union of the terminal classes, denoted K, is thus $K = \{k \in S | \neg \exists x (x \in S \text{ and } H(k, x)\}$.

As figure 1 illustrates, it's convenient to view the taxonomy as an acyclical directed network. The nodes of the network are the classes, and a directed edge (x, y) represents the relation $H(x, y)$. Looking "down" the taxonomy, each non-terminal class may be broken into one or more specific classes, all the way down to the network's boundary, where terminal nodes represent terminal classes. Looking "up" the taxonomy, each class can be generalized into one or more other classes, with the exception of the root class, that can't be generalized any further. To complete the construction of a taxonomy $\mathcal{S} =< S, H >$, we characterize each class $c \in S$ by two sets of classes which are defined recursively, as follows:

$$S_{\downarrow}(c) = \{c\} \cup \{x \in S | H(c,x) \; or \; (H(c,y) \; and \; x \in S_{\downarrow}(y)\} \tag{18}$$

$$S_{\uparrow}(c) = \{c\} \cup \{x \in S | H(x,c) \; or \; (H(y,c) \; and \; x \in S_{\uparrow}(y)\} \tag{19}$$

For example, $S_{\downarrow}(\texttt{Cubism}) = \{\texttt{Cubism}, \texttt{Braque}, \texttt{Picasso}\}$ and $S_{\uparrow}(\texttt{Cubism}) = \{\texttt{Cubism}, \texttt{Modern}, \texttt{Art}\}$. Hence, if we view the taxonomy as a family tree, the sets $S_{\downarrow}(x)$ and $S_{\uparrow}(x)$ contain the descendants and ancestors of x, respectively. Unlike a common family tree, though, each class in the taxonomy can have as many parents as we desire. Also note in passing that definition (18) implies that (i) $S_{\downarrow}(root) = S$, i.e. the root library contains all the other classes; and (ii) $S_{\downarrow}(k) = \{k\}$ for all $k \in K$, i.e. the terminal classes are all singletons. In what follows, we'll sometimes refer to the set $S_{\downarrow}(x)$ as "the library rooted in x".

Suppose now that we are asked to index an art-related document on the taxonomy depicted in figure 1. We'll do this in a top-down depth-first fashion, as follows. Beginning at the first level under the root and proceeding left to right, we first test if the document is relevant to `modern` art. If the answer is 'yes,' we step down one level and test if it's relevant to `Braque`. If the answer is 'yes,' we index the document on `Braque`. If the answer is either 'no' or 'unsure,' we test if it's 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 and index the document on `Cubism`. If the document is deemed irrelevant to all the classes thus visited, we backtrack all the way to the root of the taxonomy and index it on `art`. This would reflect the notion that even though the document is art-related, the existing taxonomy fails to discern the exact category to which it belongs.

We see that the notion of "relevance" that emerges from this classification process is defined over *subsets* of classes, not over *individual* keywords: if we index a document on, say, `Cubism`, it implies that the document belongs to the library $S_{\downarrow}(\texttt{Cubism})$, i.e. to the collection of documents about `Cubism`, `Braque`, or `Picasso`. This definition of relevance is convenient because it allows us to be as specific as we wish in our relevance statements. If we're sure that a document is relevant to a certain class, we index it on that class. If we're not sure, we can step back and index the document on a library that contains that class. We can do this all the way up to the top 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. This simplistic model has several caveats, most of which are resolved by the move from static taxonomies to adaptive taxonomies.

An adaptive Taxonomy: An adaptive taxonomy consists of a fixed set of keywords, denoted K, and an "open-ended" set of *classes*, each class being a different grouping of keywords. As before, the act of indexing amounts to assigning each document $d \in D$ to a subset of classes S_d. Unlike the static case, though, the indexes are not drawn from a predefined set of classes. Rather, the index S_d may be any desired subset of lexical keywords. Hence, a document titled *"A letter from Braque to Janco"* may well be indexed on the class {`Braque`, `Janco`}, something that would have been impossible in a static taxonomy that doesn't contain such a predefined category.

In the extreme case, an adaptive taxonomy is simply the union of all the indexes of the documents in the library, i.e. $\mathcal{S} = \cup_{d \in D} S_d$. Hence, the taxonomy is a flexible data structure that evolves dynamically from the classification process itself. When a new document is deemed relevant to a subset of keywords that don't make up an exiting category, we simply announce this subset a new class and add it to the taxonomy. The only restriction that we place on the taxonomy is that it will contain at least all the elements in K (as singletons, or classes that are made up of single keywords), as well as K itself. Thus, we begin with the initial taxonomy $\mathcal{S} = \{\{k_1\}, \ldots \{k_n\}\}, K\}$, and we add more classes to it as we go along.

Figure 2 depicts two adaptive taxonomies that evolved from two different hypothetical classification processes. The key difference between the two taxonomies is that the one on the right is a tree. Using the notation $|X|$ to represent the cardinality of a set X, we can characterize each class $C \in \mathcal{S}$ by the set $L(C) = \{X \in \mathcal{S} || X| = |C|\}$. A taxonomy $\mathcal{S}$ is said to be a *tree taxonomy* if and only if for every class $C \in \mathcal{S}$, $L(C)$ contains only disjoint sets.

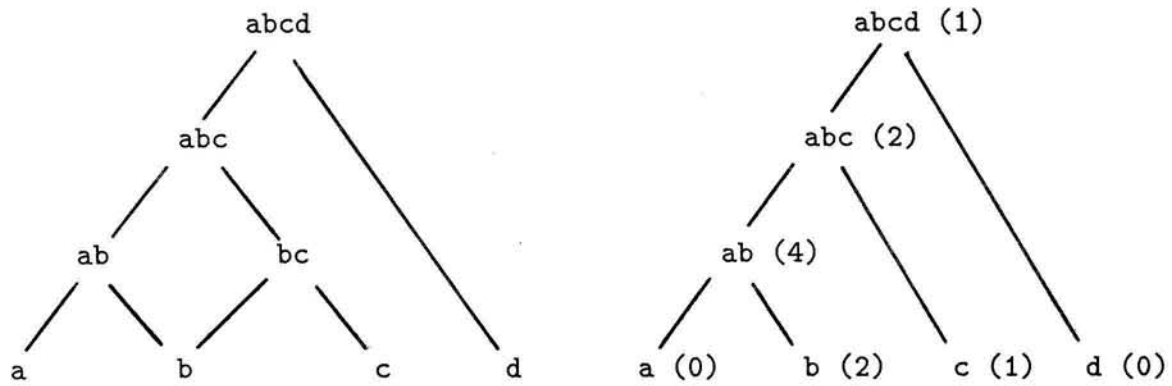


Figure 2: A network taxonomy (left) and a tree taxonomy (right). The numbers in parentheses record indexing decisions, as explained later in the paper.

Relationship to the Theory of Evidence: The linchpin that connects the taxonomical model expounded here and the theory of evidence described in §2 is the treatment of the keywords lexicon K as the frame of discernment and the observation that 2^K enumerates all the possible ways to group together, or categorize, keywords into classes. With that in mind, it is easy to see that any static taxonomy is conceptually a "frozen" and "named" version of some adaptive taxonomy, and that any adaptive taxonomy, in turn, is a subset of the lexical power set 2^K. This relationship is illustrated in figure 3.

Figure 3-a depicts the power set (excluding $\emptyset$) of the simple lexicon $\{a, b, c\}$. Clearly, with only a few dozens key words, the set of all possible classes becomes prohibitively large. Note however that once the *semantics* of the lexicon K is taken into consideration, many if not most of the classes in 2^K become arbitrary grouping of keywords that can be excluded from the taxonomy for all practical purposes. If we choose to focus on *tree* taxonomies only, the power set can be restricted further by disregarding all its non-hierarchical subsets. Figure 3-b depicts a specific adaptive taxonomy that emerged from a hypothetical classification process. Finally, figure 3-c depicts a "frozen" version of that taxonomy, after its classes

were given unique identifiers.

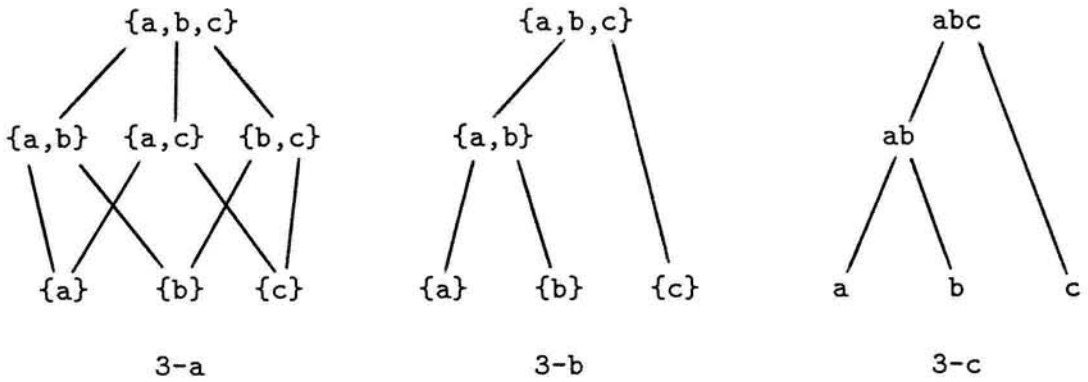


Figure 3: The evolution of a taxonomy from a lexical power set

For simplicity's sake, the identifier of a class is taken here to be the concatenation of the names of its members. For example, $\{a, b\}$ is named `ab`, $\{a\}$ is named `a`, etc. If a certain class "makes sense" on semantic grounds, it can be given an alias name that reflects its contents. For example, the class `ab` $= \{a, b\}$ can be named `Cubism`, the class `abc`$=\{a, b, c\}$ can be named `art`, etc. It turns out that this naming scheme presents a subtle theoretical problem. In the logical context of the theory of evidence, the subset $\{a, b\}$ is interpreted as the disjunction of a and b. Thus, to say that a document is relevant to $\{a, b\}$ is to say that the document is relevant to either a or to b. In the bibliographical context of a taxonomy, however, subsets of keywords have meaningful names, like `Cubism` and `Dada`, just like the elementary keywords that make up their contents. Hence, a cataloger may well wish to index a title like *"Cubist Landscapes"* directly on the class `Cubism`. However, the present model will interpret this indexing decision as "the document is relevant either to `Picasso`, to `Braque`, or to `Cubism` at large." Although such an interpretation would not be erroneous, it would clearly entail loss of concrete information about the document's *direct* relevance to `Cubism`. Also, this will lead to a situation in which the set of documents relevant to any one class would be *larger or equal* than the union of the sets of documents relevant to all of its children.

The problem may be resolved by augmenting the taxonomy with a new set of what might be called "net classes". For each non-terminal class $c \in S$ we add (i) a new class named *net_c* to S, and (ii) a new tuple $H(c, net\_c)$ to H. The new class *net_c*, which is a direct descendant of c (for every non-terminal class), can now serve as the index of the documents that are relevant to c *directly*. With this modification, c becomes a mere tag, or a pointer, and the statement "the document is relevant to the class c" is equivalent to saying "the document is relevant to the library rooted in c", and we are back in the familiar disjunctive stance of a Shaferian frame of discernment.

4 Relevance Functions

Having described the taxonomy that provides the skeleton of the classification process, we now turn to discuss in detail the classification process itself. In general, a document should be classified in a certain class if the users of the document perceive it *relevant* to that class. In its most primitive form, then, relevance is a boolean and subjective relation, indicating categorically that a document d is relevant to a class c in the view of a certain user. However, due to the fact that bibliographical classes don't have crisp boundaries, and due to the multitude of relevance opinions expressed by different library patrons, a more reasonable question is not whether d *belongs* to c, but rather what is the *intensity* of this relation. In other words, we seek to represent relevance in terms of a mapping $r : \mathcal{S} \times D \rightarrow [0, 1]$, rather than in terms of a characteristic function $r : \mathcal{S} \times D \rightarrow \{0, 1\}$.

There have been many efforts to give bibliographical relevance a probabilistic interpretation, the defining article being Maron and Kuhns [10]. One of the fundamental problems in this area has been the proper definition of the *space* from which probabilistic statements are drawn: "The notion of probability of relevance can be interpreted in two different perspectives: of the documents, as the proportion of searchers of a given type who would judge that document relevant, and of the patron himself, as the proportion of document of a given type which he would judge relevant (Maron, [9])." Indeed, if we view the relevance measure $r(c, d)$ as a degree of class membership, we can attempt to interpret it as a probability, i.e. a non-negative and additive set function which ranges on the interval $[0, 1]$ and obeys the axioms of probability. If such a probabilistic interpretation is undertaken, the frequentist meaning of an expression like $r(c, d) = 0.9$ would depend on our choice of a sample space. If the sample space is taken to be *all the documents* in the library,

$r(c,d) = 0.9$ might mean that if a document is pooled at random from the class c, the probability that the document is relevant to d is 0.9. If, alternatively, we take the sample space to be a *set of library patrons*, $r(c,d) = 0.9$ might mean that 90% of the users of d perceive it relevant to c. It turns out that the theory of evidence is consistent with the latter interpretation.

Formally, let $d \in D$ be a document, $\mathcal{S}$ a taxonomy, and $U = \{u_1, \ldots, u_n\}$ a set of library patrons who act as catalogers. Suppose that each cataloger u_i is given (1) the taxonomy $\mathcal{S}$, (ii) a copy of the same document d, and (iii) a directive to select a class $c \in S$ where d should be classified. We record the individual indexing decision of the catalogers through the following functions:

$$v_i(c,d) = \begin{cases} 1 & \text{if } u_i \text{ classified } d \text{ in } c \\ 0 & \text{otherwise} \end{cases} \qquad (20)$$
$$i = 1, \ldots, n$$

If a cataloger is unsure about the proper classification of a document, we assign the document by default to the root class. 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 made their classification decisions regarding d, we compute for each class c in the taxonomy three "relevance counters," as follows:

$$r(c,d) = \sum_{i=1}^{n} v_i(c,d) \qquad (21)$$

$$R_{\downarrow}(c,d) = \sum_{x \in S_{\downarrow}(c)} r(x,d) \qquad (22)$$

$$R_{\uparrow}(c,d) = \sum_{x \in S_{\uparrow}(c)} r(x,d) \qquad (23)$$

In words, $r(c)$ counts the number of catalogers who classified the document d in c (Since the document d is fixed in the following analysis, we'll sometimes eliminate it from the function notations to avoid clutter.) $R_{\downarrow}(c)$ counts the catalogers who classified the document in the library rooted in c, whereas $R_{\uparrow}(c)$ counts the catalogers who classified the document in libraries that contain c. To illustrate, suppose we ask ten catalogers to classify the same document on the tree taxonomy depicted on the right of figure 2. The numbers in parentheses give the results of one such hypothetical classification process, recording the number of catalogers who indexed the document in each class. Focusing for example on the class `ab`, we get $r(ab) = 4$, $R_{\downarrow}(ab) = 6$, and $R_{\uparrow}(ab) = 7$. We see that the (21-23) counters are merely different means to keep track of *the same* set of individual indexing decisions made by a group of n catalogers.

Relationship to the Theory of Evidence: If the keyword lexicon is taken to be the frame of discernment, it can be easily shown that the relevance counters (21-23) are proportional to the mappings that represent degrees of belief in the theory of evidence. Specifically, dividing each counter by n – the number of catalogers – yields the mass, belief, and plausibility, functions defined in (2),(11), and (13), respectively:

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

$$Bel(c) = \frac{1}{n} \cdot R_{\downarrow}(c) \tag{25}$$

$$Pl(c) = \frac{1}{n} \cdot R_{\uparrow}(c) \tag{26}$$

Hence, the only difference between the standard constructs of the theory of evidence and the classification scenario expounded here is the proportionality constant $\frac{1}{n}$. Operationally speaking, this constant is of little interest, as it only serves to translate counting functions to fraction functions (defined over a space of catalogers). One key difference, though, is the $v_i(\cdot)$ functions (20) that keep track of individual indexing decisions. In the classification model, this function is the foundation on which everything else rests; In the theory of evidence, that function is implicit.

5 Probabilistic Indexing

So far, we assumed that (i) relevance is a two-place function $r(c, d)$ between a document d and a class c, and that (ii) all the library patrons from whom $r(c, d)$ was elicited expressed their relevance opinions in the context of a uniform information need. In this section we retract both assumptions. Specifically, we argue that relevance, in its most elementary form, is a three-place relation $r(c, d, q)$ in which q is the *information need*, or the *context*, in which d is perceived relevant to c. With that in mind, $r(c, d)$ can be viewed as a measure of *average relevance* that runs over all the possible information needs in the context of which the document might be used. We now turn to describe a pooling mechanism that estimates such an average.

Let Q be a group of n_q patrons, each with the same information need q, and let R be a group of n_r patrons, each with the same information need r. Suppose we ask each one of the $n_q + n_r$ patrons to classify the same document d on a common keywords lexicon K. Since the patrons are not confined to a static taxonomy, it's entirely possible that the two groups will yield two different adaptive taxonomies. Figure 4 depicts examples of two such taxonomies, as well as the pooled taxonomy that emerges from combining them. The figure raises two immediate questions: (i) how to construct the pooled taxonomy $\mathcal{S}_{qr}$ from the individual taxonomies $\mathcal{S}_q$ and $\mathcal{S}_r$; and (ii) how to compute the average index $m(c, d)$ from the individual indexes $m(c, d, q)$ and $m(c, d, r)$. The remainder of this section addresses these questions.

To construct $\mathcal{S}_{qr}$, begin by setting it to the empty set. Next, apply the following admission test to each class $x \in \mathcal{S}_q \cup \mathcal{S}_r$: if at least one patron in *both* groups has classified the document in a library that contains x, include x to $\mathcal{S}_{qr}$. Otherwise, exclude it. In the above example, this procedure will yield the pooled taxonomy depicted on the right of figure 4. The computation of the average relevance $m(c, d)$ for every class in the pooled taxonomy is based on two conceptual steps. First, recalling how the functions $m(c, d, q)$ and $m(c, d, r)$ are derived from individual indexing decisions, one can "step back" and construct the groups of patrons that yielded these functions. For example, the $m(c, d, q)$ function depicted in figure 4 is consistent with a group of 4 patrons in which two members classifed the document on the class `ab`, one member classified it on `a`, and one member on `b`. Denoting this group Q, one can use the same rationale to construct the group R whose indexing decisions are consistent with the function $m(c, d, r)$. The resulting decisions are

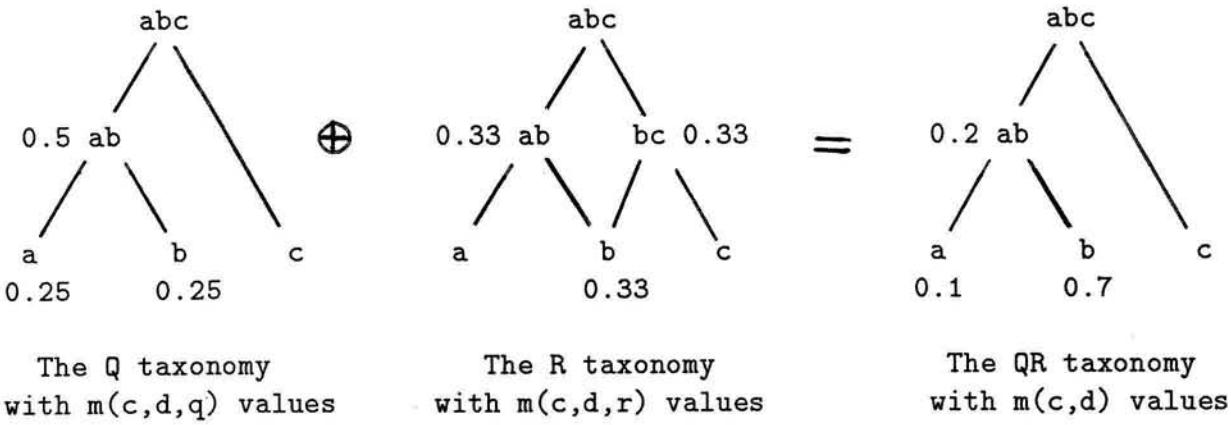


Figure 4: The result of combining two taxonomies and pooling their respective indexes. The $m(c, d, x)$ values record the fraction of patrons in group X who classified the document d on the class c, i.e. $r(c, d, x)$ divided by n_x. For example, in group Q one quarter of the patrons classified d in a, one quarter in b, and one half in ab.

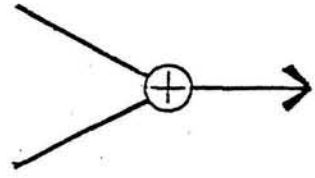
tabulated in the left hand side of figure 5. The columns of each table represent the common lexicon which in this example is $K = \{a, b, c\}$. The ith tuple in each table represents the indexing decision elicited from the ith patron in the respective group as a binary vector in which 1 in the jth column codes the fact that the patron has classifed the document on the jth keyword and 0 otherwise.

Having constructed all the individual indexing decisions of the two groups Q and R, one can choose a variety of different pooling mechanisms to compute the index induced by the combined pool of patrons. The pooling mechanism depicted in figure 5 is a special case of a combination scheme described by Hummel and Landy in [6], who called it "a consensus opinion by the element of the product set of experts formed by the committees of two." Specifically, the QR table is made up of $n_q \cdot n_r$ tuples, one for each unique pair of patrons drawn from Q and from R. The combined index associated with the pair (q_i, r_j) is defined

to be the binary *conjunction* of the individual indexing decisions of q_i and r_j. For example, the pooled tuple $(q_1, r_1) = (0, 1, 0)$ is the conjunction of the individual tuples $q_1 = (1, 1, 0)$ and $q_2 = (0, 1, 1)$.

Q	A	B	C
q_1	1	1	0
q_2	1	0	0
q_3	0	1	0
q_4	1	1	0

R	A	B	C
r_1	0	1	1
r_2	1	1	0
r_3	0	1	0



$Q \times R$	A	B	C
q_1, r_1	0	1	0
q_1, r_2	1	1	0
q_1, r_3	0	1	0
q_2, r_1	0	0	0
q_2, r_2	1	0	0
q_2, r_3	0	0	0
q_3, r_1	0	1	0
q_3, r_2	0	1	0
q_3, r_3	0	1	0
q_4, r_1	0	1	0
q_4, r_2	1	1	0
q_4, r_3	0	1	0

Figure 5: An example of using the cartesian consensus operator to pool the indexing decisions implied by figure 4.

The pooling operation can now be completed by treating QR as a new group of patrons and computing the new m function that it induces:

$$\begin{array}{lllll} m'(A) & = & m'(1,0,0) & = & 1/12 \\ m'(B) & = & m'(0,1,0) & = & 7/12 \\ m'(AB) & = & m'(1,1,0) & = & 2/12 \\ m'(\emptyset) & = & m'(0,0,0) & = & 2/12 \end{array} \qquad \begin{array}{lllll} m(A) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(A) & = & 1/10 \\ m(B) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(B) & = & 7/10 \\ m(AB) & = & \frac{1}{1-m'(\emptyset)} \cdot m'(AB) & = & 2/10 \end{array} \tag{27}$$

In words, for each class $c \in \mathcal{S}_{qr}$, $m'(c)$ is the fraction of the (paired) patrons who classifed

the document on that class. Next, the fraction of the patrons who agreed on *nothing* – $m(0,0,0)$ – is distributed evenly among the fractions of patrons who agreed on *something*, yielding a new mass that sums up to unity. This function is now taken to be the "average index" of the document d. The reader is asked to withhold judgement about the external validity of this method until the discussion section.

Relationship to the Theory of Evidence: The theory of evidence concerns the assignment and manipulation of degrees of belief rendered to subsets of propositions in light of a certain *piece of evidence* which, surprisingly, is rather implicit in the theory's notation. That is to say, $m(X)$ and $Bel(X)$ are meant to be short-hands of $m(X|e)$ and $Bel(X|e)$, e being a fixed piece of evidence that helps discern the likelihood of various subsets $X \subseteq \theta$. When two pieces of evidence e_1 and e_2 render evidence to a common frame of discernment, the notation $m_i(X)$ and $Bel_i(X), i = 1,2$ is used as a short-hand of $m(X|e_i)$ and $Bel(X|e_i), i = 1,2$. The combined impact of the body of evidence $\{e_1, e_2\}$ is computed through Dempster's rule (6-8), which yields a new function $m(X|\{e_1, e_2\}) = m(X|e_1) \oplus m(X|e_1)$ (or the equivalent $Bel(X|\{e_1, e_2\}) = Bel(X|e_1) \oplus Bel(X|e_1)$).

In our indexing model, the relevance of a document to a class is viewed as a single point summary of the relevance opinions of many library patrons. The patrons approach the library with different information needs (or queries) in mind, each corresponding to a piece of evidence that highlights one facet of the complex relation that we call "relevance." The model computes this composite relevance by fusing the individual relevance opinions through the cartesian consensus rule illustrated in figure 5. As Hummel and Landy have shown, this pooling operation corresponds exactly to Dempster's rule (6-8). In other words, had we applied (6-8) to the functions $m(c,d,q)$ and $m(c,d,r)$ from figure 4, we would obtain the $m(c,d)$ function in (27). This is not a coincidence, but rather a corollary of Dempster's original approach of inducing "lower" and "upper" probabilities from multivalued mappings and combining them through his combination formulae.

6 Discussion

This section summarizes the relevance of our research to bibliographical models and to the theory of evidence.

Bibliographical Models: Although our indexing model was not intended to serve as the basis of a working application, it's instructive to envision how versions of the model could be used to augment document storage and retrieval systems. In essence, the indexing methods that we described can be implemented in two different ways. If the library is to be indexed "once" by a group of professional catalogers (other than by its ordinary users), the model can be used to record the individual indexing decisions of the catalogers and to transform them into a composite probabilistic index. Alternatively, we can envision a learning scenario in which indexes are continuously revised to reflect the actual use of the library's documents. Such a dynamic indexing model would require an information system that keeps track of (i) how many patrons sought each document, (ii) the patron's opinions regarding the relevance of the document to various classes encountered in the search process, and (iii) the information need, or query, that launched the patron's search process. Relevance opinions can be elicited by probing the patrons randomly, building a cumulative database of individual indexing decisions. Information needs can be detected either explicitly, by asking direct questions, or implicitly, by analyzing the language that the patron uses to describe his or her query [2].

The notion of relevance between a document and a class is closely related to the notion of relevance between an information need and a class, as both documents and information needs can be characterized by subsets of keywords [16]. Therefore, one area of future research is to extend the indexing model described in this paper to a browsing model that helps searchers pursue the most promising class of documents, given a (possibly fuzzy) information need. Such a model will maintain a vector of the form $S_q = \langle (c_1, r_i), (c_2, r_2), \ldots, (c_n, r_n) \rangle$, where q is the information need, the c_i's are all the classes in the taxonomy, and the r_i's are dynamic measures of relevance of the patron's information need to the ith class. As the patron browses the library, he or she may be asked to provide relevance feedback concerning the classes that were vistaed thus far; This feedback can then be used to update the vector S_q, which serves as the road map that guides the search direction.

Depending on the parlance that is chosen to represent relevance, the r_i's in S_q can be either point masses, point beliefs, or belief intervals. For example, if we take r_i to stand for the belief function $Bel(c_i)$, we can interpret this number as the likelihood that the information need q can be satisfied somewhere in the library rooted in c_i. Together with the topology of the taxonomy (the relation H that defines the structure of the c_i's), this information can be effectively used to guide bibliographical searches and fine tune browsing techniques. Some of these ideas were already implemented in an experimental browsing system designed by

Pyun [11].

The Theory of Evidence: Recent criticism of the theory of evidence has centered around the argument that "Anything that can be done with belief functions can better be done with probability (Lindley [8], p. 38)." We believe that this argument, although correct, misses the 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 feature complex structures for dealing with generic programming tasks, the theory of evidence provides non-elementary functions and operators that lend themselves nicely to certain inferential problems. This paper has illustrated how one such example – bibliographical indexing – maps very well on the various constructs of the theory of evidence. The details of this "mapping" are listed in table 6.

indexing model	Dempster-Shafer model
keyword lexicon (K)	frame of discernment (θ)
set of classes (S)	subset of 2^θ
index of a document (S_d)	core (C)
group of patrons (U)	implicit
individual indexing decisions (v_i's)	implicit
relevance counter (r)	mass function (m)
total-relevance counter ($R_\downarrow$)	belief function (Bel)
maximal-relevance counter ($R_\uparrow$)	plausibility function (Pl)
information need (q)	body of evidence (e)
average relevance operator	Dempster's rule

Table 1: A summary relationship between the indexing model and the theory of evidence

Viewed as a canonical example, the bibliographical model that we described serves to justify the logical backdrop and the key constructs of the theory of evidence. Yet the example also exposes the theory's shortcomings, and, in particular, those of Dempster's rule. These limitations manifest themselves quite clearly in the cartesian consensus operator depicted in figure 5. Since the operator is limited to tracking boolean opinions only, it amounts to a conservative pooling mechanism that is insensitive to dissenting views of individual experts. That is to say, in order for an expert opinion to "survive" the pooling operation, at least one more expert must concur with it. Furthermore, the operator does not offer means to assign different weights to different experts, as would be desirable in many applications and as indeed is done in most pooling mechanisms. Now, both limitations can be easily fixed by using an "improved" version of the cartesian consensus operator. However, our intention here is to interpret the Dempster Shafer theory, not to modify or extend it. With that in mind, we wish to underscore the fact that any criticism of the cartesian consensus operator must be interpreted as a criticism of the logical assumptions that underlie Dempster's rule, as the operator and the rule are completely isomorphic.

We conclude that the theory of evidence provides an attractive framework for building models of documents storage and retrieval, and that these models, in turn, serve to highlight the theory's internal validity. Dempster's rule remains a controversial operator for combining beliefs, and modified versions of it may be used to parameterize the theory and make it more plausible in certain applications. Whichever form the theory will take, though, a dual analysis that focuses on an hypotheses space, on the one hand, and on an experts space, on the other, holds the key for understanding the probabilistic roots of the theory.

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

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

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

[6] 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.

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

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

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

[10] 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.

[11] J. Pyun. *An Investigation of Inexact Classification and Browsing Techniques in Hierarchical Document Retrieval Systems.* PhD thesis, Stern School of Business, New York University, 1990.

[12] 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.

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

[14] G. Shafer. Probability judgement in artificial intelligence and expert systems. *Statistical Science*, 2(1):3–44, 1987.

[15] 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.

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

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

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

Center for Digital Economy Research
Stern School of Business
Working Paper IS-92-18