# A DEMPSTER-SHAFER MODEL OF RELEVANCE

by

**Shimon Schocken**
Leonard N. Stern School of Business
New York University
90 Trinity Place
New York, NY 10006

and

**Jisurk Pyun**
Leonard N. Stern School of Business
New York University
90 Trinity Place
New York, NY 10006

**December 1989**

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

**Working Paper Series**

CRIS #220
STERN #89-148

This paper was presented at **HICSS-23**(the Hawaiian International Conference on Systems Sciences) and was published in the conference proceedings.

## Abstract

We present a model for representing relevance and classification decisions of multiple catalogers in the context of a hierarchical bibliographical database. The model is based on the Dempster-Shafer theory of evidence. Concepts like *ambiguous* relevance, *inexact* classification, and *pooled* classification, are discussed using the nomenclature of belief functions and Dempster's rule. The model thus gives a normative framework in which one can describe and address many problematic phenomena which characterize the way people classify and retrieve documents.

# 1 Introduction

Every day, a few hundred articles are published in more than 20,000 scientific and professional journals. However, this vast literature is useless unless you know where to find what you are looking for. During the past decade, researchers and practitioners have developed new theories and techniques designed to organize, classify, and search bibliographical databases. And yet, many fundamental questions regarding classification and search are still open. How to resolve different *classification* decisions made by independent and equally qualified catalogers? How to measure the abstract notion of *relevance*? How to design effective *browsing* techniques which mimic the way people explore paper-based libraries?

A bibliographical database is a collection of textual records, consisting of documents, books, abstracts, or simply call numbers pointing to physical shelf locations. In this paper we refer to all these records as *documents*. The access method to a bibliographical database is based on a controlled lexicon of key-words, sometimes referred to as "subject headings". This lexicon is carefully balanced to achieve sufficient descriptive coverage, on the one hand, while minimizing duplication of similar terms, on the other. The basic assumption is that both documents and user queries may be represented through a fixed set of pre-defined key-words. In this paper we focus on the problem of *classification* – the activity through which a document is mapped on a subset of key-words which best describes it.

This definition of classification is tricky due to its reliance on the phrase "best describes it." What is the meaning of the word "best"? A given classification model may be justified on either normative or descriptive grounds. A normative analysis casts the classification process in a formal model within which the term "best" has an explicit meaning. A descriptive justification might involve an experiment in which human subjects search the indexed database. Depending on the experimental design, the goodness of the classification scheme can be gauged indirectly through the effectiveness of the search process.

In this paper we present a model of relevance and classification which is based on the Dempster-Shafer theory of evidence [8]. We argue that this theory is well-suited for the domain under investigtion for two reasons. First, the Depmster- Shafer theory lends itself naturally to hierarchical data structures. Second, by making proper use of belief functions and Dempster's rule, we give an axiomatic description of such problematic concepts as ambiguous relevance and inexact classification decisions. These issues are at the center of numerous *empirical* studies on document classification and retrieval. We contribute to this line of research by presenting a clear *normative* framework in which these phenomema can be clearly defined and investigated.

The Dempster-Shafer theory suffers from an obscure mathematical notation, and is quite abstract when detached from the context of a practical application. Therefore, we begin the paper with a description of the underlying domain – a hierarchical bibliographical database. This example sets the stage for Section 3, in which a domain-free synopsis of the Dempster-Shafer model is given. Section 4 presents the classification model and tracks its relationship to belief functions and Dempster's rule. A conclusion section comments on the normative validity of the model.

# 2 Hierarchical Bibliographical Databases

## 2.1 The Classification Scheme

The justification of a hierarchical architecture for textual databases stems from the belief that a hierarchical data structure gives a convenient means for cataloging and searching documents. Indeed, the concept of a hierarchy is simple, powerful, and cognitively appealing (Simon [10]). Most libraries are currently governed by hierarchical access methods, with the *Library of Congress* and the *Dewey Decimal Classification* being the two quintessential examples. The Dewey, for example, consists of ten main *classes* which are further broken into *divisions*. Each of these divisions consists of many *sections*, which are further divided into *subsections*, and so on and so forth.

The Library of Congress and the Dewey Decimal Classification are special cases of a more general architecture – a hierarchical bibliographical database. The remainder of this section gives a formal description of this architecture.

The lexicon of a hierarchical bibliographical database forms a hierarchy of *classes*: some key-words are high-level aggregations, or generalizations, of other key-words. Looking "down" the hierarchy, each class may be broken into one or more specific classes. The bottom boundary of the hierarchy consists of *terminal classes*. Looking "up" the hierarchy, each class can be generalized into at most one class. This one-to-many relationship forms a tree structure.

**Definition 1: Key-word Hierarchy:** A *key-word hierarchy* is a pair $< C, H >$. $C$ is a set of classes. $H$ is a relation defined over $C \times C$ with the following properties: (1) $\exists c_0[\forall c[\neg H(c, c_0)]]$, and (2) $\forall c \neq c_o[\exists x[H(x, c)] \wedge \forall y \neq x[\neg H(y, c)]]$ ($c_0$, $c$, and $x$ are members of $C$).

(1) identifies the root of the hierarchy – $c_0$ – the one class which cannot be generalized any further. (2) says that any non-root class has exactly one "parent".

If $H(x, y)$ holds, we say that $x$ is a *parent* of $y$ and that $y$ is a *child* of $x$. The *children set* of a class $c$ is the set $H(c) = \{x \in C | H(c, x)\}$. The set of *terminal classes* consists of all the classes which have no children: $TC = \{c \in C | \neg\exists x[H(c, x)]\}$. $x$ is said to be an *ancestor* of $y$ if $x$ is a parent of $y$ or $x$ is a parent of $z$ and $z$ is an ancestor of $y$. $x$ is said to be a *descendant* of $y$ if $y$ is an ancestor of $x$.

Figure 1 depicts a simple hierarchy describing a subset of a library on programming languages. Note that the hierarchy is completely defined in terms of the set $C = \{$*languages,procedural,4GL,C,Pascal,Cobol,Focus,dBASE*$\}$ and the relation $\{$*H(languages,procedural),H(languages,4GL),H(procedural,C), H(procedural,Pascal),H(procedural,Cobol),H(4GL,dBASE),H(4GL,Focus)*$\}$.

Note that the topology of the hierarchy depicted in Figure 1 is not unique. If we were to organize a body of documents on programming languages, it would make perfect sense to consider an alternative hierarchy in which the first level consists of *scientific*, *business*, and *general purpose* languages. The construction of an effective hierarchy is an important problem which concerns not only librarians but anybody who has to manage extensive amounts of textual information. For example, users of personal computers are faced with the problem of managing disk-based libraries consisting of hundreds if

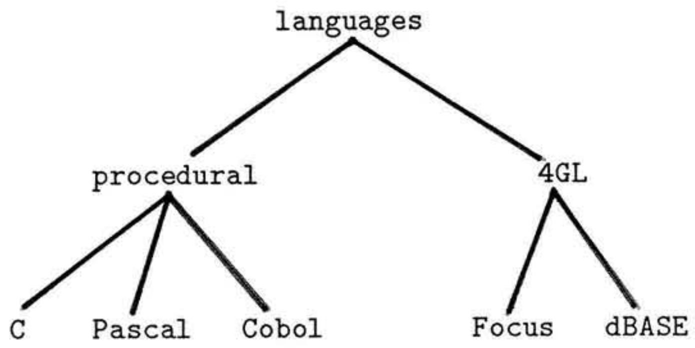


Figure 1: A simple classification scheme

not thousands of files. These libraries are organized in hierarchies of subdirectories, and it is up the users to determine and manage the topology of these hierarchies.

The decision to prefer one hierarchical organization on another depends heavily on the contents of the database, and requires a certain degree of domain expertise. There is extensive literature on the subject, e.g. Wason & Johnson [13] and Dumais & Landauer [3]. There are also a number of computer-based clustering methods (van Rijsbergen & Croft [12] and Salton & Wong [7]) designed to take an unstructured set of documents and partition them into homogeneous groups which form a hierarchy. In this paper, however, we assume that an experienced cataloger has already gone through the preliminary stage of constructing an effective hierarchy. The focus of the paper is thus the subsequent problem of indexing documents into a fixed hierarchy.

## 2.2 Binary and Fuzzy Classification

It is often said that classification is one of the most fundamental human activities. In the context of a bibliographical database, classification concerns the assignment of a document, $d$, to a certain class, $c$. The classification decision is made by a human cataloger who feels that the class $c$ is *relevant* to $d$. In its most primitive form, relevance is a binary relation, indicating categorically that a document is either relevant or irrelevant to a certain class. However, due to the vagueness of some documents and the fact that classes don't have crisp boundaries, the typical question is not whether $d$ belongs or doesn't belong to $c$, but, rather, what is the *degree of membership* associated with this relation. In other words, we wish to focus on a continuous relevance function $r : C \times D \rightarrow [0,1]$, rather than on the characteristic function $r : C \times D \rightarrow \{0,1\}$.

If we view the relevance measure $r(c,d)$ as a degree of membership, we can also 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 subjective probability. But what would be the meaning of an interim value, say, $r(c,d) = 0.7$, from a classification standpoint?

The probabilistic interpretation of $r(c,d) = 0.7$ depends on our choice of a sample space. If the sample space is taken to be *all the documents* in the database, $r(c,d) = 0.7$ means that if a document is pooled at random from the class $c$, the probability that this document will be relevant to $d$ (in view of a single, expert cataloger) is 0.7 . If, alternatively, we take the sample space to be a *set of catalogers*, $r(c,d) = 0.7$ means that 70% of the catalogers would say that $d$ is relevant to $c$. In what follows, we adopt the latter interpretation.

We distinguish between two types of classifications. A *binary classification* says that a certain document belongs to a certain class. A *fuzzy classification* describes the strength of this relationship through a number which varies from 0 to 1. The relationship between the two classification schemes is determined as follows. Let $d$ be a document, $< C, H >$ a hierarchy, and $\Omega$ a set of $n$ catalogers. Suppose that each cataloger is required to assign each document to *precisely one* class in the hierarchy.

**Definition 2: Binary Relevance:** If the cataloger $\omega$ assigns $d$ to $c$, we say that the relation $In_\omega(c,d)$ holds and that the characteristic function of $In_\omega$ is unity, i.e. $In_\omega(c,d) = 1$.

**Definition 3: Relevance:** The relevance of a document $d$ to a class $c$ is measured through the fraction of catalogers who thought that the two are relevant to each other:

$$r(c,d) = \frac{1}{n} \sum_{\omega \in \Omega} In_{\omega}(c,d)$$

**Definition 4: Cumulative Relevance:** The cumulative relevance of a document $d$ to a class $c$ is $r(c,d)$ plus all the relevance measures of $d$ to the descendants of $c$:

$$R(c,d) = r(c,d) + \sum_{x \in H(c)} R(x,d)$$

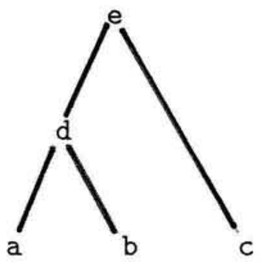
The relationship between *cumulative relevance* and *relevance* is subtle. A high $R(c,d)$ value does not necesarily imply a high $r(c,d)$ value. The only direct relationship between the two measures is $R(c,d) \geq r(c,d)$. In what follows, the $r$ function will be the key measure of relevance. The role of $R$ will become clearer when we discuss the linkage between relevance and Shafer belief functions.

To illustrate how $r$ and $R$ are derived from the primitive notion of a binary relevance ($In$), consider a setting in which a group of 6 catalogers are independently classifying the same document into the simple hierarchy $\{e\{d\{a,b\},c\}\}$ (Figure 2). The entries of the table record binary $In$ values. The $r$ and $R$ functions of each class are computed at the bottom of the table.

Recall that our definition of binary relevance requires that each cataloger classifies a document on precisely one class in the hierarchy. If the cataloger is unsure about the proper class of a particular document, as in the case of cataloger 4, this document is assigned by default to the root class. The rationale for this convention is that the root class represents the *entire library*. Thus, if the cataloger cannot classify a particular document into a specific class, it makes sense to classify this document into the non-discriminating class *library*.

An inspection of Figure 2 reveals the following observations:

1. For each $c \in C$, $r(c,d) \geq 0$



|   | a | b | c | d | e |
|---|---|---|---|---|---|
| 1 | 1 |   |   |   |   |
| 2 |   |   |   | 1 |   |
| 3 | 1 |   |   |   |   |
| 4 |   |   |   |   | 1 |
| 5 |   |   |   | 1 |   |
| 6 | 1 |   |   |   |   |
| r | 3/6 | 0 | 0 | 2/6 | 1/6 |
| R | 3/6 | 0 | 0 | 5/6 | 1 |

Figure 2: The Derivation of r and R

2. $\sum_{c \in C} r(c, d) = 1$

3. For each $c \in C$, $R(c, d) \geq 0$

4. The cumulative relevance of a document to the root class (the entire library) is always 1: $R(c_0, d) = 1$

Many researchers have observed high degrees of indexing inconsistencies across different catalogers (Stevens [11], Cooper [2], and Bates [1]). We take the position that this pluralism is a natural phenomenon which may serve to improve, rather than obscure, indexing decisions. Different catalogers represent different work experiences, multiple backgrounds, and versatile points of view; in a way, each cataloger represents a different body of potential users. If all the catalogers are considered equally-qualified experts, we ought to embrace this diversity of opinions. This, of course, requires formal techniques to resolve inconsistencies and pool individual classification decisions.

Of the three relevance measures, $In$, $r$, and $R$, the first is presented here strictly for analytical purposes. In order to credibly assess $r$ from $In$, we need a large sample of catalogers which is prohibitively expensive and unrealistic on practical grounds. However, in the event that we do have access to a small group of, say, three expert catalogers, we can ask each one of them to assess $r$, and then go on to pool the three opinions into a global classification decision. The validity of this approach hinges on the availability of a pooling mechanism which can be justified on normative grounds. We propose usage of Dempster's rule.

# 3 The Dempster-Shafer Model

Consider a finite and exhaustive set of mutually exclusive propositions, $\theta = \{q_1, ..., q_n\}$, exactly one of which is true. $\theta$ is called the *frame of discernment*, and the power-set which enumerates all of the subsets of $\theta$ is denoted $2^\theta$. The Dempster-Shafer theory of evidence concerns the representation and manipulation of degrees of belief rendered to various propositions in the frame of discernment. Degrees of belief are represented through *mass* and *belief* functions (Shafer [8]). When several independent belief functions lend credence to a particular proposition, the overall belief in that proposition is computed through Dempster's rule.

## 3.1 Mass and Belief Functions

Degrees of belief in propositions are expressed through two related functions, $m$ and $Bel$. These functions are defined over the power-set $2^\theta$. This is in contrast to a standard Bayesian design in which a subjective probability function is defined over $\theta$ only.

<u>**The Mass Function:**</u> The mass function $m : 2^\theta \rightarrow [0,1]$ has the following properties:

$$m(\emptyset) = 0$$

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

To illustrate, consider the simple frame of discernment $\theta = \{up, same, down\}$. The elements of $\theta$ represent three alternative directions of tomorrow's stock market. The power set of $\theta$ is $2^\theta = \{\{up,same\}, \{up,down\}, \{same,up\}, \{up\}, \{same\}, \{down\}, \{up,same,down\}, \emptyset\}$. Each of these subsets stands for a *disjunction* of propositions. For example, to say that the truth lies in $\{same, up\}$ is to say that tomorrow's market will either remain the same, or will go up.

The mass function assigns degrees of belief (which may be zero) to every element in $2^\theta$. These degrees of belief must sum up to 1. The *uncommitted belief*, or the mass which is left over after all the *proper* subsets of $\theta$ were assigned degrees of belief, is assigned by convention to $\theta = \{up, same, down\}$. The higher the uncommitted belief, the less information we have about the propositions in question. For example, consider a bullish expert (expert no. 1) who distributes his belief as follows: $m_1(\{up, same\}) = 0.6$, $m_1(\{down\}) = 0.1$, and $m_1(A) = 0$ for any other proper subset of $\theta$. Hence the *uncommitted* *belief* or the degree of "second order uncertainty" displayed by this expert is $m_1(\theta) = 1 - 0.6 - 0.1 = 0.3$.

Note that the mass function represents indivisible, or atomic, degrees of belief: knowledge of $m(\{up, same\})$ says nothing about $m(\{up\})$ and $m(\{same\})$. Likewise, knowledge of $m(\{up\})$ and $m(\{same\})$ says nothing about $m(\{up, same\})$.

**<u>The Belief Function:</u>** The total belief assigned to a set of propositions and to all of its subsets is given in terms of a "belief function" $Bel : 2^\theta \rightarrow [0, 1]$, defined as follows:

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

Like $m$, $Bel$ is defined on every subset $A \in 2^\theta$. Note that the two functions are tighly related: $Bel$ is completely determined by the mass function $m$, and, likewise, $m$ can be recovered from $Bel$ (Shafer [8]). This relationship (and $m$'s properties) implies that $Bel(\theta) = 1$ and $Bel(\emptyset) = 0$. The *core* of a belief function is the set of all subsets $X \in 2^\theta$ for which $m(X) > 0$. For example, the core of $Bel_1$ is $C_1 = \{\{up, same\}, \{down\}, \theta\}$.

## 3.2 Dempster's Rule

Suppose now that a second bullish expert (expert no. 2) is willing to express his optimistic prediction in terms of $m(.)$. In particular, the expert provides $m_2(\{up\}) = 0.8$ and $m_2(\theta) = 0.2$. Is there a credible way to combine the two expert opinions and generate a global prediction concerning the direction of tomorrow's market?

According to Shafer, once we cast degrees of belief in terms of belief functions, we ought to be able to combine them using Dempster's rule.

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 $C_1 = \{A_1, \ldots A_{n1}\}$ and $C_2 =$

$\{B_1, \ldots B_{n2}\}$, respectively. Dempster's rule computes the pooled mass function $m = m_1 \oplus m_2 : 2^\theta \rightarrow [0,1]$ as follows:

$$m(X) = \frac{1}{1-k} \sum_{A_i \cap B_j = X} m_1(A_i) m_2(B_j) \tag{1}$$

where

$$k = \sum_{A_i \cap B_j = \emptyset} m_1(A_i) m_2(B_j)$$

Note (from (1) and the definition of *core*) that the core of the pooled function is the intersection of the cores of its component functions: $C = C_1 \cap C_2$.

The following section demonstrates the application of Dempster's rule in a specific example involving multiple catalogers. We precede this example with a detailed discussion of the linkage between the abstract Dempster-Shafer model and the applied problem of classifying documents in a hierarchical database.

# 4 The Classification Model

## 4.1 The Classification Scheme and the Frame of Discernment

We take the frame of discernment $\theta = \{q1, \ldots, qn\}$ to represent a set of key-words, or a *lexicon*. The power-set $2^\theta$ represents all the possible ways to group together (or categorize) key-words in $\theta$. Figure 3-a depicts the power set (excluding $\emptyset$) of the simple lexicon $\{a, b, c\}$. Clearly, the size of $2^\theta$ is prohibitively large; at the same time, most of its elements represent meaningless groupings of key-words. We can thus focus our attention on a much smaller subset of $2^\theta$ consisting of meaningful categories. We restrict this subset further by focusing on *one* of the hierarchical subsets of $2^\theta$ (Figure

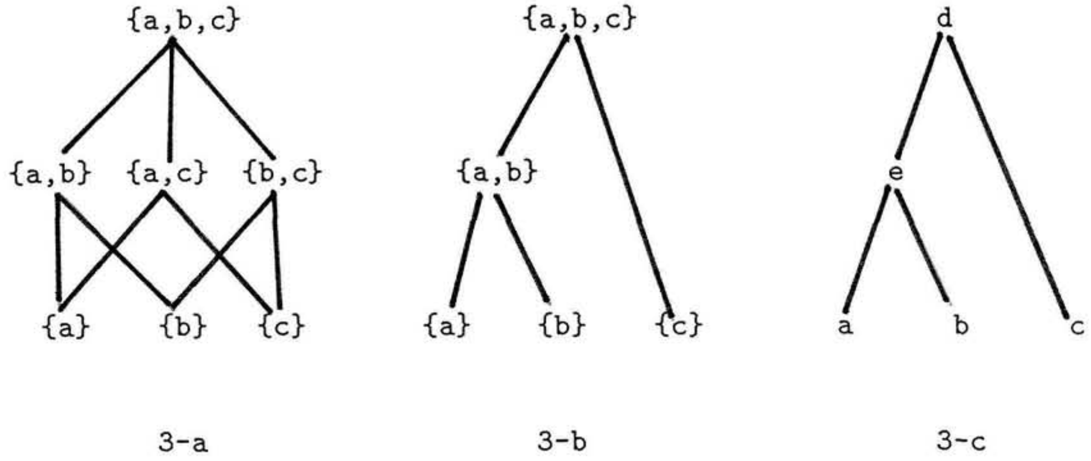


Figure 3: Derivation of the classification scheme

3-b)[2]. The last stage in setting up the classification scheme is the assignment of *names* to each class. In Figure 3-c, the class $\{a, b\}$ is named $e$, and the entire lexicon $\{a, b, c\}$ is named $d$.

To sum up, the relationship between a Shaferian frame of discernment $\theta$ and the architecture of a key-words hierarchy $< C, H >$ is as follows. $C$ is a subset of $2^\theta$. More specifically, the set of terminal classes of the hierarchy, $TC \subseteq C$, is simply $\theta$. The remaining elements of $C$ are classes, i.e. *named* subsets (or *categories*) of $\theta$ consisting of key-words which are semantically related to each other. The relation H ensures that this set of classes forms a hierarchy. Finally, the $r$ and $R$ relevance measures correspond to the $m$ and $Bel$ functions, respectively.

In the Dempster-Shafer model, the context of the frame of discernment $\theta =$

[2] The formal definition of a hierarchy requires the definition of a *level*, which, pictorially speaking, is a set of nodes which are equally distant from the root. A hierarchy in $2^\theta$ can now be defined as a set of levels in which all the nodes belonging to the same level are disjoint subsets of $\theta$.

$\{q_1, \ldots, q_n\}$ is logical: the underlying assumption is that the truth lies in precisely one of the $\{q_i\}$ propositions. What would be the equivalent meaning of this assumption in the context of a bibliographical database? Recall that the task of the cataloger is to classify a document in the "best" possible class. Maron [6] proposed that this task can be made easier if *"instead of asking what a document is about, a cataloger should ask: if people were looking for information of this sort, where would they look for it?"*

The Maron classification guideline suggests that there are two ways to think about the concept of a *class*. From a lexical standpoint, a class is a high-level aggregation of lower-level classes (e.g. $4GL$ is an aggregation of $dBASE$ and $FOCUS$). From a semantical standpoint, a class can be also viewed as *the set of documents belonging to that class and to all of its descendant classes* (e.g. $4GL$ is the set of all the documents in the library which are relevant to $4GL$, or to $Focus$, or to $dBASE$). Hence, when a cataloger who is presented with a document $d$ produces the relevance measure $R(c, d) = 0.8$, he is in fact saying that the likelihood of finding documents similar to $d$ *somewhere in the sub-tree rooted in* $c$ is 0.8. In contrast, the specification $r(c, d) = 0.8$ means that the likelihood of finding documents similar to $d$ *precisely in the class* $c$ is 0.8.

The concept of a *named class* presents a deviation from the original Dempster-Shafer model, in which a class is simply a disjunction of its lower-level elements. In the context of a hierarchal bibliographical database, however, a class is "larger" than the sum of its parts. For example, consider the sub-tree $\{4GL, \{dBASE, Focus\}\}$, and a document whose title is "Systems Analysis in a 4GL Environment". Where does this document belong? The document is clearly relevant to both $dBASE$ and $Focus$, and yet a better indexing decision would be to place it at the more general $4GL$ class. At the same time, every document which is relevant to either $dBASE$ or to $Focus$ is *also* relevant to $4GL$. We see that the set of documents relevant to the $4GL$ class is greater than the union of the sets of documents relevant to all of its children.

In order to fix this problem, we introduce the notion of a "net class". For each non-terminal class $c$ we attach a new child class called "net c" and denoted $n\_c$. This class consists of all the documents which are relevant to $c$ directly

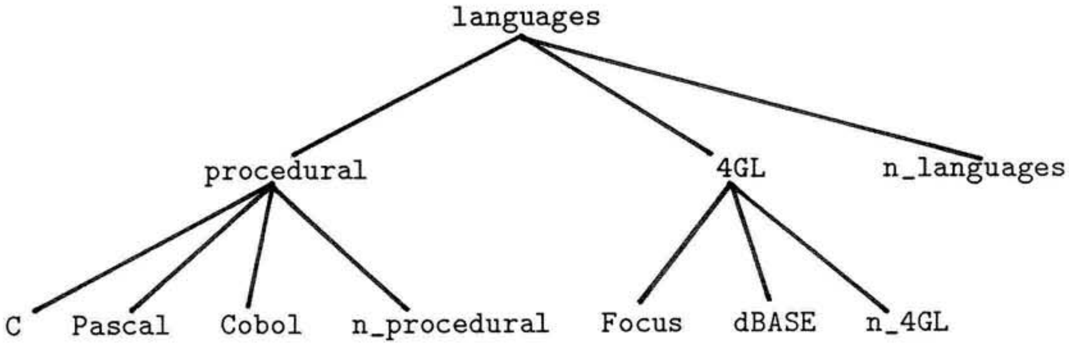


Figure 4: Extended classification scheme with net classes

but not to any of its children. We introduce this modification, which may be used in any situation involving *named classes*, is order to stay consistent with the original Dempster-Shafer framework. Specifically, once we add the class *net-c* to the children set of $c$, we get that $c$ is a disjunction of all the documents belonging to its children; thus, we are back in the familiar domain of a Shaferian frame of discernment. The extended classification scheme is presented in Figure 4.

## 4.2 Relevance Measures and Belief Functions

The process of classifying a document $d$ into a hierarchy $< C, H >$ by a single cataloger consists of two stages: screening and measurement. In the screening stage, the cataloger identifies classes in $C$ which seem to be relevant to the document $d$. The resulting subset, denoted $RC \subseteq C$, is called the index of $d$. In the measurement stage, the cataloger assesses the relevance of $d$ to each class $c \in RC$. If we choose to measure relevance through the $r$ function, we can present the cataloger with the following question: suppose that 100 catalogers were asked to assign this document to classes in $RC$. How many catalogers (a number between 0 and 100) would assign it to each class?

Now, some documents present straightforward classification decisions, while others are much harder to classify. For example, consider the following two documents and the hierarchy depicted in Figure 4:

$d_1$: *dBASE or C? The Choice is Getting Harder*

$d_2$: *The Rocky Road From Procedural Languages to 4GL's: a Case Study*

In the case of $d_1$, it is clear that $RC = \{dBASE, C\}$, and a plausible distribution of relevance would be $r(dBASE, d_1) = 0.5$ and $r(C, d_1) = 0.5$. $d_2$, on the other hand, presents a more challenging classification effort. The document is relevant to both *procedural* and to $4GL$, but the focus seems to be more on the latter; recalling Maron's guidelines, it is plausible that $d_2$ will be sought more by people who are interested in $4GL$ than by people who seek information about procedural languages. Suppose we were pressed to predict how many of 100 catalogers would classify $d_2$ into each of these classes. We may argue that of each 7 catalogers who would select $4GL$, 1 cataloger would select *procedural*; since this is a difficult question, we may qualify our judgement further by saying that of the 100 catalogers, we are willing to predict the decisions of only 60 catalogers.

One important feature of this elicitation procedure is an explicit representation of unassigned (or uncommitted) relevance. In the case of $d_1$, $r(dBASE, d_1) = 0.5$, $r(C, d_1) = 0.5$, and the unassigned relevance, $r(\theta, d_1)$, is 0. This means that the cataloger is 100% sure in his relevance assessment. In the less concrete case of $d_2$, $r(procedural, d_2) = 0.1$, $r(4GL, d_2) = 0.7$, and the unassigned relevance is $r(\theta, d_2) = 0.2$. Said otherwise, the relevance measures $r(4GL, d_2)$ and $r(procedural, d_2)$ lie in the intervals [0.7,0.9] and [0.1,0.3], respectively. The width of these intervals, 0.2, is the uncommitted relevance.

## 4.3 Pooled Relevance and Dempster's Rule

Suppose now that documents are classified by several equally-qualified catalogers. The catalogers work independently, and do not communicate their classification decisions to each other. Naturally, the classifications generated

| | $r_2(n\_4GL) = 0.8$ | $r_2(\theta) = 0.2$ |
|---|---|---|
| $r_1(4GL) = 0.6$ | $r(n\_4GL) = 0.48$ | $r(4GL) = 0.12$ |
| $r_1(Focus) = 0.1$ | $r(\emptyset) = 0.08$ | $r(Focus) = 0.02$ |
| $r_1(\theta) = 0.3$ | $r(n\_4GL) = 0.24$ | $r(\theta) = 0.06$ |

Figure 5: An intersection tableau

by these catalogers will vary in terms of focus ($RC$ – the output of the screening stage) as well as intensity ($r(c, d)$ for each $c \in RC$ – the output of the the measurement stage). For example, consider a setting in which two catalogers are independently classifying the following document:

$d3$: *Integrated Spreadsheet/Database Management in 4GL's*

Suppose that cataloger 1 feels quite strongly that the document is related to $4GL$, but is *unrelated* to either $Focus$ or to $dBASE$. In particular, the cataloger's relevance function is $r_1(n\_4GL) = 0.8$ and $r_1(\theta) = 0.2$. Cataloger 2 agrees that the document is relevant to the overall $4GL$ class; at the same time, he also thinks that some people who are interested in $Focus$ will find this document relevant. More specifically, his relevance function is $r_2(4GL) = 0.6$, $r_2(Focus) = 0.1$ and $r_1(\theta) = 0.3$.

The global classification of this document may be derived by pooling the two individual classification decisions through Dempster's rule [8]. The mechanics of this pooling operator are illustrated in Figure 5. To avoid clutter, we use the notation $r(c)$ instead of $r(c, d_3)$ throughout.

Following Dempster's rule [8], note that $k = 0.06$, and the global relevance measures $r$ is as follows:

$$r(n\_4GL) = 0.77, \quad r(4GL) = 0.13 \quad r(Focus) = 0.03, \quad r(\theta) = 0.07$$

# 5 Conclusion

Hierarchical classification is a challenging application from a probabilistic standpoint. There is no straightforward way to interpret the *hierarchical* nature of the key-words space, the concept of *relevance*, and the *pooling* of multiple classification decisions. The Dempster-Shafer theory, on the other hand, is especially suitable for situations involving belief-update in a hierarchical hypotheses space (Gordon and Shortliffe [4], Shenoy and Shafer [9]).

In this paper we gave a Dempster-Shafer interpretation of relevance and a classification model which can be implemented in a computer program. Aside of its practical merit, the model is interesting on theoretical grounds because it gives a canonical example in which the Dempster-Shafer theory "makes sense." This theory is still controversial on normative grounds, and there have been several attempts to give it a Bayesian interpretation. Hummel and Landy [5] have recently shown that if belief functions are viewed as statistics of experts opinions, then Dempster's rule [8] is isomorphic to a (complex) Bayesian design. The classification model presented in this paper has specific features which are necessary for the purpose of managing textual databases, and yet its underlying spirit is consistent with Hummel and Landy's analysis. Therefore, the model is practically appealing and, at the same time, enjoys a solid theoretical footing.

# References

[1] M. J. Bates. System meets user: problems in matching subject search terms. *Information Processing and Management*, 13:367–368, 1977.

[2] W. S. Cooper. Is interindexer consistency a hobgoblin? *American Documentation*, 7:268–278, 1969.

[3] S. T. Dumais and T. K. Landauer. Describing categories of objects for menu retrieval systems. *Behavior Research Method, Instrument and Computers*, 16(2):242–248, 1984.

[4] J. Gordon and E. H. Shortliffe. The dempster shafer theory of evidence. In B.G. Buchanan and E.H. Shortliffe, editors, *Rule-Based Expert Systems*, pages 272–294, Addison-Wesley, 1984.

[5] 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.

[6] M. E. Maron. Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing and Management*, 24(3):249–255, 1988.

[7] G. Salton and A Wong. Generation and search of clustered files. *ACM Transactions on Database Systems*, 3:321–346, 1978.

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

[9] P. Shenoy and G. Shafer. Propogating belief functions with local computations. *IEEE Expert*, 1:43–52, 1987.

[10] H. A. Simon. *The Sciences of the Artificial, 2nd ed.* MIT Press, 1981.

[11] M. E. Stevens. *Automatic Indexing: A State-of-the-art Report.* U.S. Government Printing Office, Washington D.C., 1965.

[12] C. J. van Rijsbergen and W. B. Croft. Document clustering: an evaluation of some experiments with the cranfield 1400 collection. *Information Processing and Management*, 11:171–182, 1975.

[13] P. C. Wason and P. D. Johnson-Laird. *Psychology of Structure and Reasoning: Structure and Content.* Harvard University Press, Cambridge, MA, 1972.