Probabilistic Retrieval Revisited

C. J. VAN RIJSBERGEN
Department of Computing Science, University of Glasgow, Glasgow, Scotland G12 8QQ

The well-known probabilistic model for information retrieval based on Bayesian conditioning of probability functions is examined. It is extended to allow conditioning based on evidence derived from the 'passage of experience' which may be non-propositional in nature. This latter form of conditioning is derived from Jeffrey's work on probability kinematics and it is compared with the Dempster–Shafer approach to revising belief functions whilst motivating its appropriateness for adaptive information retrieval. This new form of conditioning is combined with a non-classical logic to define a new probabilistic model for information retrieval.

Received December 1991

1. INTRODUCTION

There are at least two different ways of viewing probability: the first is as a measure of the chance of an event, the second is as a measure of the degree of belief in a proposition. Each has its advantages and disadvantages, neither can be defined without paradox. When we define probability in terms of chance we often appear to be using circular definitions (e.g. equally probable cases), or appealing to a knowledge of infinite sets (e.g. a frequency value in the limit). In the case of degrees of belief, the definition seems to suffer from a lack of objectivity because it is based on a subjective opinion (e.g. the odds of a fair bet for *you*, the judge). In information retrieval we have never taken a strong position on the kind of probability that is appropriate for our work. However, when implementations are involved, the probabilities tend to be estimated as measures of chance events. This has led to a number of IR models that are difficult to generalise.

The need to measure uncertainty by, say, probability arises in IR in a natural way. This can best be seen by comparing IR with databases or question-answering systems. In the database context one seeks facts that satisfy a query; in question-answering one seeks specific answers to specific questions. In IR one seeks a 'document' that is likely to contain the information that the user seeks.

Let us take a simple example. Suppose amongst a set of documents we have a document about the Chernobyl disaster. A database query would seek answers to: Who wrote the document? Does it contain the string 's'? Does it cite a paper by Einstein? A question-answering system would respond to: Did the reporter photograph the scene of the disaster? But, an IR system does not seek answers but documents that are *about* expressed needs. So for example, an information retrieval system would react to: 'Give me a document concerned with the radioactive fallout during the Chernobyl disaster.'

A user will express the need for information about a topic, subject, or proposition. An IR system tries to locate a document that is likely to satisfy this need. Documents satisfying the need or the expressed need tend to be called *relevant*. At this stage we have started to use three interestingly undefined notions: 'aboutness', 'information' and 'relevance'. They are of course related. If a document contains information about X then it is likely to be relevant to X.

The process of locating relevant documents is inherently uncertain; it is also highly context dependent. The uncertainty enters in a number of ways. First through the 'aboutness' as it is only possible to determine that a document is about something to a degree. Secondly, whether a document is relevant to an expressed need is also a matter of degree. Finally, if a document is about X with probability α, it may or may not contain the information that X.

The way we express the uncertainty discussed above is by a belief function $\mathrm{Bel}(R, q, d)$ which measures the degree of belief that d is relevant in context q. The parameter R is normally assumed to have two values, relevant and non-relevant. The document d is assumed to have some kind of representation, it might be a set of index terms, a set of noun phrases, or a set of propositions. The context q in its simplest form is an expression of the user need, thus it is often in the form of a simple proposition, or a set of index terms.

Probably the most thoroughly researched belief function to capture the degree of belief in relevance is the conditional probability $P_q(R \mid d)$ within the Bayesian updating framework. The purpose of this essay is to show that there are other better approaches. Before discussing the other belief revision methods we need to look again at the Bayesian one for IR.

2. BAYESIAN BELIEF REVISION

The basis for Bayesian belief revision is of course the celebrated Bayes' Theorem:

$$P(H \mid e) \propto P(e \mid H)\, P(H)$$

or

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

since

$$\sum_H P(H \mid e) = 1.$$

The usual interpretation of the symbols is that H is a hypothesis (one of several) which is supported (or otherwise) by some evidence e. $P(H \mid e)$ is interpreted as the probability that H is true given certain evidence e. The reason Bayes' Theorem is a form of belief revision is that $P(e \mid H)$, $P(H)$, and $P(e)$ are probabilities associated with propositions (or events) *before* e is actually observed. $P(H \mid e)$ is the probability of H *after* e has been observed. The notation used to express this is especially opaque because the same $P(\cdot)$ is used for the prior and posterior

probabilities. In fact $P(H|e)$ would be better written as $P_e(H)$ indicating that we now have a *new* probability function P_e. A further difficulty with this approach is that the evidence e has to be certain at observation, that is, cannot be disputed once it is observed, which implies that $P_e(e) = 1$. The fact that in general a conditional probability can only take into account an exact and certain proposition, or event, as its basis for revising the probability in the light of observation, is a source of some difficulty. In IR we use the Bayesian approach by calculating $P_q(R|d)$ through $P(R|\boldsymbol{x})$ where $\boldsymbol{x}$ is some representation of the document assumed to be certain. This means that the description $\boldsymbol{x}$ is assumed to be true of the document, or *in* the document, depending on one's point of view. In many cases this is not fully appropriate, as the description, or representation $\boldsymbol{x}$ may be subject to uncertainty itself at the time of observation.

Let us examine the calculation of $P(R|\boldsymbol{x})$ in a little more detail: let $\boldsymbol{x}$ be a set of independent variables $x_1, \ldots, x_n$, then

$$\frac{P(R|\boldsymbol{x})}{P(\bar{R}|\boldsymbol{x})} = \frac{P(\boldsymbol{x}|R)\,P(R)}{P(\boldsymbol{x}|\bar{R})\,P(\bar{R})} \quad P(R) = 1 - P(\bar{R})$$

and

$$\log \frac{P(R|\boldsymbol{x})}{P(\bar{R}|\boldsymbol{x})} = \log \frac{p_1(x_1)}{q_1(x_1)} + \log \frac{p_2(x_2)}{q_2(x_2)} \ldots + \log \frac{p_n(x_n)}{q_n(x_n)} + \log \frac{P(R)}{P(\bar{R})}$$

where $P(\boldsymbol{x}|R) = p_1(x_1)p_2(x_2)\ldots p_n(x_n)$,
$P(\boldsymbol{x}|\bar{R}) = q_1(x_1)q_2(x_2)\ldots q_n(x_n)$.

Thinking now of the values of the individual variables x_i as applied to a particular document as bits of evidence in support of R or $\bar{R}$, the support will rise or fall as

$$\log \frac{p_j(x_i)}{q_j(x_i)}$$

increments or decrements the overall support. Remember that p_i and q_i have to be estimated; they are not given *ab initio*. The interpretation of one of the evidential weights, e.g.

$$\log \frac{p_j(x_i = 1)}{q_j(x_i = 1)},$$

is not entirely obvious. $p_j(x_i = 1)$ decodes to: if the document d *were* relevant, p_j is the probability that x_i *would* be true, or present. We are calculating the probability because we are observing a document with x_i true. There is room for confusion, x_i is true and yet we ask for the probability that x_i is true. It is like observing a '6' on a dice and asking what was the probability governing that chance event happening. But R is a hypothetical event, since if we knew R and observed x_i to be true then $P(x_i = 1|R)$ would be trivially 1.

This raises the question as to whether $P(\boldsymbol{x}|R)$ is a conditional probability or should be modelled by the probability of a (subjunctive) conditional. More about this later.

In addition it is very likely that the observation itself is uncertain. In IR we assign index terms with a degree of certainty. So for example, a component of a document representation, x_i, might only be true of, apply to, the document with a certain probability. This way of viewing indexing was explored earlier by Maron and Kuhns.[1] Adopting such a view makes the application of Bayes' Theorem less obvious, one should not identify the probability of relevance with $P(R|x_i)$ since x_i is now not certain, and its degree of uncertainty needs to be taken into account. Therefore other ways of conditionalising must be found.

3. PROBABILITY KINEMATICS

The problem of how to revise a probability measure in the light of uncertain evidence or observation was treated comprehensively by R. C. Jeffrey in his book *The Logic of Decision*.[2] Earlier similar discussions can be traced back to Donkin,[3] Boole[4] and Keynes.[5] Jeffrey introduced his method of conditionalisation through a now famous example. Imagine one inspects a piece of cloth by candlelight and one gets the impression that it is green, although it might be blue, or even violet. If, G, B, and V are the propositions involved, then the outcome of the observation might be that the degrees of belief in G, B or V are 0.7, 0.25 and 0.05, whereas *before* observation the degrees of belief were 0.3, 0.3, 0.4. In symbols we would write

$$P(G) = 0.3, \quad P(B) = 0.3, \quad P(V) = 0.4$$
$$P^*(G) = 0.7, \quad P^*(B) = 0.25, \quad P^*(V) = 0.05$$

Here P is a measure of the degree of belief before observation, and P^* the measure after observation. As Jeffrey puts it, the 'passage of experience' has led P to be revised to P^*. In Bayesian terms $P^*(x) = P(x|e)$ where e is a *proposition*, but Jeffrey rightly claims that it is not always possible to express the passage of experience as a proposition [see refs 6 (p. 46) and 7]. (Pearl has gone to some length to demonstrate that a Bayesian net formalism can make Bayesian conditionalisation appropriate so that Jeffrey's approach is not needed. Indeed, Turtle and Croft (see this issue) have implemented a version of Pearl's approach. It is the author's opinion that Pearl's presupposition of virtual evidence leads to infinite regress and that it is better to assume from the beginning that some evidence is based on the passage of experience leading to a direct revision of the probability functions.) Given that one has changed one's degree of belief in some propositions G, B, and V as shown above, how are these changes to be propagated over the rest of the structure of one's beliefs? For example, suppose saleability A of the cloth depends on the colour inspection in the following way:

$$P(A|G) = 0.4 \quad P(A|B) = 0.4 \quad P(A|V) = 0.8.$$

Prior to inspection

$$P(A) = P(A|G)\,P(G) + P(A|B)\,P(B) + P(A|V)\,P(V)$$
$$= 0.4 \times 0.3 + 0.4 \times 0.3 + 0.8 \times 0.4 = 0.56.$$

After inspection Jeffrey proposes:

$$P^*(A) = P(A|G)\,P^*(G) + P(A|B)\,P^*(B) + P(A|V)\,P^*(V),$$
$$= 0.4 \times 0.7 + 0.4 \times 0.25 + 0.8 \times 0.05 = 0.0485.$$

known as *Jeffrey's rule of conditioning*.

It is valid whenever $P^*(A|E_i) = P(A|E_i)$ where E_i is a partition of the sample space (although it can be generalised; see Williams[8]). This differs from Bayesian

conditioning which would use $P^*(G) = 1$, or $P^*(B) = 1$, or $P^*(V) = 1$ and so revise $P(A)$ to $P^*(A) = P(A \mid X)$ when $X = G$, B, or V. Thus Bayesian conditioning can be seen as a special case of Jeffrey conditioning.

Let us now try and interpret Jeffrey conditioning in terms of IR. Let A be the property of relevance and let us consider the effect of observing an index term (x_i), which is either present ($x_i = 1$) or absent ($x_i = 0$). Then,

$$P^*(\text{Relevance}) = P(\text{Rel} \mid x_i = 1)\, P^*(x_i = 1) + P(\text{Rel} \mid x_i = 0)\, P^*(x_i = 0)$$

(abbreviating 'Relevance' to 'Rel') which measures the probability of relevance for a particular document (we will come to the problem of observing further index terms later). We interpret the probabilities involved in the expression for P^*(Relevance) as follows: $P(\text{Rel} \mid x_i = 1)$ is the probability of the current document being judged relevant to an arbitrary query containing term x_i (commonly known as the probabilistic index term weight). Of course this conditional probability may also be estimated through the standard Bayesian inversion. Now consider $P^*(x_i = 1)$. This is a user estimate of the probability of $x_i = 1$ (cf. the user estimate of the probability of colour in the candlelight example). There need be no explicit expression of the evidence leading to $P^*(x_i = 1)$. The prior user estimate $P(x_i = 1)$ can be given by collection statistics, although what is important is the current value of $P^*(x_i = 1)$. Rewriting P^*(Rel) slightly we get

$$P^*(\text{Rel}) = \left[P(x_i = 1 \mid \text{Rel}) \frac{P^*(x_i = 1)}{P(x_i = 1)} + P(x_i = 0 \mid \text{Rel}) \frac{P^*(x_i = 0)}{P(x_i = 0)} \right] \times P(\text{Rel}),$$

showing how the change in P to P^* enters into the computation. It is important to keep in mind that the probability P^* is user generated based on the 'passage of experience'. Also we have assumed that

$$P(\text{Rel} \mid x_i = 1) = P^*(\text{Rel} \mid x_i = 1),$$

$$P(\text{Rel} \mid x_i = 0) = P^*(\text{Rel} \mid x_i = 0),$$

which I think are reasonable (but see, Ref. 6).

Although I have presented the formulation in terms of variables that are easily related to the classical probability retrieval models,[1,9–11] I will now show how easy it is to generalise the computation of P^*(Rel) to more general forms of evidence. Given the general expression

$$P^*(\text{Rel}) = P(\text{Rel} \mid E_1)\, P^*(E_1) + \ldots + (P(\text{Rel} \mid E_n)\, P^*(E_n),$$

where E_j is the conditioning event, or proposition, the examples are:

(1) $E_j = (x_i = 1), j = 1, 2$ and $(x_i = 1)$ is a proposition indicating absence/presence of index term i, i.e. $E_1 = \bar{E}_2$. (This is the example above repeated for convenient comparison.)

(2) $E_j = (x_i = j)\ j = 0, \ldots, k$ where j is the frequency with which term i occurs.

(3) $E_j = (x_i = f(j))\ \ j = 0, \ldots, k$, this time x_i is a transformation of j perhaps through some exponential function f (see later).

(4) $E_j = (\Gamma \rightarrow q)$, where $\Gamma \rightarrow q$ is a conditional statement, Γ a set describing the current document and q the query, $j = 1, 2$, and $E_2 = \bar{E}_1$.

Expressing the last example in more detail

$$P^*(\text{Rel}) = P(\text{Rel} \mid \Gamma \rightarrow q)\, P^*(\Gamma \rightarrow q) + P(\text{Rel} \mid \neg(\Gamma \rightarrow q))\, P^*(\neg(\Gamma \rightarrow q)).$$

To evaluate this probability we need to evaluate $P^*(\Gamma \rightarrow q)$, which is precisely the problem addressed in several earlier papers.[11,13] For example, $P^*(\Gamma \rightarrow q)$ could be calculated by the process of imaging[13,14] which uses information in *other* documents to calculate $P^*(\Gamma \rightarrow q)$. Two common retrieval models are special cases of this new model. In one case we could adopt a simpler expression for conditioning, e.g. $n(\Gamma \cap q)$, the extent to which the document shares terms with the query. Boolean retrieval is reached by a further simplification through conditioning on q only, that is,

$$P^*(\text{Rel}) = P(\text{Rel} \mid q)\, P^*(q) + P(\text{Rel} \mid \bar{q})\, P^*(\bar{q}).$$

In the classical case $P^*(q) = 1$ implies $P^*(\text{Rel}) = P(\text{Rel} \mid q)$ whereas $P^*(\bar{q}) = 1$ implies $P^*(\text{Rel}) = P(\text{Rel} \mid \bar{q})$. Boolean retrieval would judge that

$$P^*(q) = 1 \Rightarrow P^*(\text{Rel} \mid q) = P^*(\text{Rel}) = 1,$$
$$P^*(q) = 0 \Rightarrow P^*(\text{Rel} \mid \bar{q}) = P^*(\text{Rel}) = 0.$$

4. SUCCESSIVE UPDATE OR COMBINING EVIDENCE

We are all familiar with the process of updating a probability function under the Bayesian regime, i.e. $P(A \mid x_1) \propto P(x_1 \mid A)P(A)$, so that conditioning on x_1 leads to a new probability function $P^*(A)$, this in turn can be conditioned on x_2 giving $P^*(A \mid x_2) = P^{**}(A)$, etc. Moreover, the order of conditioning is irrelevant so that $P^{**}(A) = P(A \mid x_1, x_2) = P(A \mid x_2, x_1)$. But in general Bayesian conditioning is not reversible. That is if $P^*(A) = P(A \mid x_1)$ and we wish to condition P^* on y so that $P^*(A \mid y) = P(A)$ we cannot do it except in the trivial case when $P(x_1) = 1$. To see this consider $P^*(x_1 \mid y) = P(x_1 \mid x_1, y) = 1$, therefore if $P^*(x_1 \mid y) = P(x_1)$ it forces $P(x_1) = 1$. I can foresee that irreversibility could cause problems in IR applications. Jeffrey conditionalisation does not suffer from this particular problem.

To see how Jeffrey's conditionals fare under successive updating it is best to take an example. Let a, b, c, d be documents:

	x_1	x_2
a	1	1
b	1	0
c	0	1
d	0	0

Let P represent the prior probability of relevance, i.e. $P(a) = P(b) = P(c) = P(d) = \frac{1}{4}$. We have no reason to favour any particular document over another for relevance. Now consider two separate pieces of evidence, x_1 and x_2 which will affect the relevance of documents. Firstly x_1, let the user specify that the relevant document sought will have x_1 with probability 0.8, i.e.

$$P^*(x_1 = 1) = 0.8, \quad P^*(x_1 = 0) = 0.2.$$

In the light of this, a revised probability of relevance of each document is constructed by Jeffrey's rule, thus:

$$P^*(a) = P(a|x_1 = 1)\,P^*(x_1 = 1) + P(a|x_1 = 0)\,P^*(x_1 = 0)$$
$$= \tfrac{1}{2} \times 0.8 + 0 = 0.4,$$
$$P^*(b) = 0.4, \quad P^*(c) = 0.1, \quad P^*(d) = 0.1.$$

A further revision based on x_2 when

$$P^{**}(x_2 = 1) = 0.7, \quad P^{**}(x_2 = 0) = 0.3$$

gives

$$P^{**}(a) = P^*(a|x_2 = 1)\,P^{**}(x_2 = 1) + P^*(a|x_2 = 0)\,P^{**}(x_2 = 0)$$
$$= 0.8 \times 0.7 + 0 \times 0.3 = 0.56,$$

because $$P^*(a|x_2 = 1) = 0.4/0.5 = 0.8,$$

$$P^{**}(b) = 0.24, \quad P^{**}(c) = 0.14, \quad P^{**}(d) = 0.06.$$

So the ranking in the first iteration is $a = b, c = d$, in the second iteration: a, b, c, d. It should be noted immediately that in general it *does* matter in what order the evidence is presented (in this particular example it did not matter). Also, we have used a simple model for relevance and have calculated the conditional probabilities by counting.

There is a weak form of independence that will guarantee commutativity with respect to the update evidence. To explain this we need to consider the problem in slightly more general terms. Let the two sets of evidence with their associated probabilities be:

$$E = \{E_i, p_i\}_{i=1}^{e} \quad \text{and} \quad F = \{F_j, q_j\}_{j=1}^{f}.$$

The general case requires that $i > 2$ and $j > 2$. If P is the probability to be updated and P_E stands for update with regard to the E_i, P_F with regard to the F_j, and P_{EF} or P_{FE} stand for successive updates. The weak form of independence referred to above occurs when $P_E(F_j) = P(F_j)$ and $P_F(E_i) = P(E_i)$. (Diaconis & Zabell call this J-independence[15].) J-independence is necessary and sufficient for commutativity of evidence updating. Classical independence (let us call it P-independence) namely, $P(E_i | F_j) = P(E_i)$ and $P(F_j | E_i) = P(F_j)$ is stronger in the sense that P-independence implies J-independence but not vice versa. It must be noted that if either E or F is binary, then J-independence for any pair of probability measures p_i and q_i will imply P-independence and hence J-independence for all measures p_i and q_i; that is J- and P-independence are equivalent (see Ref. 15). It would seem that this weaker form of independence could be used to describe independence between multi-state variables describing within document frequencies.

The updating procedure just described is very sequential in nature; it may well be appropriate for such application as relevance feedback in IR. In moving away from evidence which is considered all at once and simultaneously, we may also wish to take pieces of evidence sequentially for the purpose of allowing a user to change her mind about the strength of the evidence. Also, under the Jeffrey procedure, a revision is reversible so that a user can change her mind about the significance of each piece of evidence pointing to relevance, as each revision step can be undone.

There are other theoretical properties of Jeffrey conditioning which need not concern us here, but one property is of interest as it bears on recent information theoretic considerations which have once again attracted attention.

Let us begin by defining the information in P relative to P^0 as

$$I(P, P^0) = \sum_{j=1}^{n} p_j \log (p_j/p_j^0),$$

where p_j^0 is the prior probability of the jth event, and we make the usual assumptions about the asymptotic properties of I. Fundamental is the property $I(P, P^0) \geqslant 0$, with equality iff $P = P^0$. A version of the principle of minimum information now goes as follows:

Given the prior distribution P^0, the probability distribution P appropriate to a new state of information is one that minimises $I(P, P^0)$ subject to whatever constrains the new information imposes.[8]

There are two cases to consider when revising probability functions, first when $P(E) = 1$, and second $P(E) < 1$. The first case, through the use of the principle, leads to Bayes' rule. Let E be the evidence in the domain of P^0 and let $P^0(E) \neq 0$ whilst $P(E) = 1$ then $I(P, P^0) = I(P_E, P_E^0) - \log P^0(E)$. This function is minimised when P_E is equal to P_E^0, the posterior probability given by Bayes' rule of conditioning. A similar argument shows that when the evidence does not reach certainty, i.e. $P(E) = q$ then $P(F) = P^0(F|E)\,q + P^0(F|\bar{E})\,(1-q)$ is the minimum information solution. Williams[8] extends these results to the more general situation where we have $E_1, \ldots, E_n$ as evidence whose probability is affected by the passage of experience and where the E_i are not necessarily mutually exclusive. There are other measures of closeness that could be used subject to constraints to express the closest probability function to a given one. One such is the Information Radius which was discussed by the author in 1979 in the context of IR.[11,16] However, it is likely that all measures of closeness will attain their minima on the same conditional functions. Van Fraassen[7] has shown how this notion of constraint, if accepted as reasonable for a probabilistic belief revision, will determine Jeffrey conditionalisation for a plausible class of constraints.

Let us briefly summarise the situation. Given that one has a sequence of pieces of evidence pointing to relevance, then the 'sensible' way of incorporating that evidence into the probability of relevance is either in terms of Bayesian conditioning for certain evidence or Jeffrey conditioning for uncertain evidence. Furthermore, the Bayesian approach is a special case of the Jeffrey approach. There is a third way of conditioning based on Dempster–Shafer theory to which we now turn.

5. DEMPSTER–SHAFER THEORY OF EVIDENCE

In 1976 Shafer published an influential book.[17] It contained a mathematical development and modification of earlier work by Dempster. In the book Shafer gives a detailed account of how to construct Belief functions based on an accumulation of evidence. In his later papers Shafer has gone to some trouble to establish what he calls 'canonical' examples to act as analogues for the evidential situation under consideration. In standard probability calculations such a canonical example would

be the throw of a dice. In the latter case, the *truth* of an event is generated by chance. The Shafer theory takes a different approach and uses canonical examples where the *meaning* of an event depends on chance.

Mathematically the Dempster–Shafer belief functions are quite simple. We have a belief function Bel defined on all subsets of a frame Θ (which can be assumed finite).

Bel is defined by:

$$\text{Bel}(A) = \sum_{B \subset A} m(B) \quad [\text{or} = \sum_{B \to A} m(B)],$$

where $m(B)$ are non-negative numbers satisfying

$$m(\varphi) = 0 \quad \text{and} \quad \sum_{B \subset \Theta} m(B) = 1.$$

Every Bayesian probability distribution is a belief function. But not every belief function is a Bayesian probability distribution. To quote Shafer: 'The theory of belief functions is based on a way of comparing our evidence to the scale of chances that is quite different from that of the Bayesian theory. Instead of comparing our evidence to a scale of examples where the truth is generated according to known chances, we compare it to a scale of examples where the reliability and meaning of a message [see below] depends on known chances.'[18]

The canonical example goes as follows: Suppose someone chooses a code at random from a list of codes, uses the chosen code to encode a message, and then sends us the result. Let $c_1, \ldots, c_n$ be the list of codes and p_i the probability of choosing c_i. We decode the encoded message using each of the codes and find that this always produces a message of the form 'the truth is in A' for some subset A of Θ. Let A_i denote the subset we get when we decode using c_i, and set

$$m(A) = \Sigma\{p_i \mid 1 \leqslant i \leqslant n; A_i = A\}$$

for each $A \subset \Theta$. The $m(A)$ can be interpreted as the total chance that the true message was A. And $\text{Bel}(A)$, given by $\Sigma_{B \to A} m(B)$, is the total chance that the true message implies A.

We are expected to realise that this is only an analogue example for generating an appropriate interpretation of Bel and m. Through this analogue example the famous Dempster rule of combination can be generated.

A simple illustration showing how the analogue could represent a retrieval example when *one* relevant document is sought is as follows: Let $c_1, \ldots, c_n$ be the n codes for the n documents in the system. The code is very simple and decodes to 'the relevant document is indexed by x'. Then $m_x(A) = \Sigma\{p_i \mid x \in d_i\}$ and p_i the probability of d_i being chosen. A subsequent index term y would generate m_y based on probabilities q_i. The p_i and q_i could depend on the distributional characteristics of x and y. The Dempster rule of combination would lead to m_{xy}, which would combine the evidence due to x and y. Without at this stage defining the rule of combination, it assigns an m value proportional to the product $m_x m_y$ for the intersection $A \cap B = C$. One can see that each subsequent index term in the query provides a further clue to the relevant document. It is useful to notice that within the representation chosen, any query can be handled by considering the subqueries that imply it, because

$$\text{Bel}(A) = \sum_{B \to A} m(B);$$

that is, the m are summed over the events B that imply A. (There is scope here for specifying the semantics of '$\to$'.)

My aim in the remainder of this section is to show how Dempster–Shafer theory can be used to establish the conditioning procedure due to Jeffrey as a special case of belief revision in terms of Bel and m. To do this we need to introduce some further machinery.

A simple support function focused on E is

$$\text{Bel}(A) = \begin{cases} 0 & \text{if } A \not\supset E \\ s & \text{if } A \supset E \quad \text{but } A \neq \Theta \\ 1 & \text{if } A = \Theta \end{cases}$$

and $m(E) = s$ and $m(\Theta) = 1 - s$. A subset E of Θ for which $m(E) > 0$ is called a *focal element* of Bel. Bel is carried by a partition $E_1, \ldots, E_n$ if and only if Bel's focal elements are all unions of the E_i. Given Bel_1 and Bel_2 with m-values m_1 and m_2, Dempster's rule of combination gives m:

$$m(C) = K\Sigma\{m_1(A)\, m_2(B) \mid A \subset \Theta, B \subset \Theta; A \cap B = C\},$$

where K is a normalising constant. The belief function resulting from m is called the orthogonal sum of Bel_1 and Bel_2 denoted $\text{Bel}_1 \oplus \text{Bel}_2$.

We are now in a position to define a form of *conditioning* based on belief functions. Let

$$\text{Bel}_E(A) = \begin{cases} 0 & \text{if } A \not\supset E \\ 1 & \text{if } A \supset E. \end{cases}$$

This belief function is used to represent evidence whose effect is to establish that the truth is in E. If Bel is a belief function satisfying $\text{Bel}(\bar{E}) < 1$, then $\text{Bel} \oplus \text{Bel}_E$ exists and it is natural to call $\text{Bel} \oplus \text{Bel}_E$ the result of conditioning Bel on E.

$$(\text{Bel} \oplus \text{Bel}_E)(A) = \text{Bel}(A \mid E) = \frac{\text{Bel}(A \cup \bar{E}) - \text{Bel}(\bar{E})}{1 - \text{Bel}(\bar{E})}.$$

If Bel happens to be an additive probability distribution P, then the above reduces to

$$\text{Bel}(A \mid E) = \frac{\text{Bel}(A \cap E)}{\text{Bel}(E)} = \frac{P(A \cap E)}{P(E)} = P(A \mid E).$$

This shows that Dempster–Shafer conditioning is a generalisation of Bayesian conditioning. The relationship between Dempster–Shafer belief functions and Jeffrey conditioning depends on the notion of weight of evidence. This notion has a long history, see for example Keynes[5] and Cohen.[19] Within the Dempster–Shafer theory, it is defined for simple support functions. If Bel_1 and Bel_2 are simple support functions focused on E with degrees of belief s_1 and s_2 for E, the $\text{Bel}_1 \oplus \text{Bel}_2$ is a simple support function focused on E with degree of belief s given by

$$-\log(1 - s) = [-\log(1 - s_1)] + [-\log(1 - s_2)].$$

The reader should consult Shafer's book[17] for a derivation of this result; essentially it ensures that the weight of evidence underlying the orthogonal sum of Bel_1 and Bel_2 will be $w_1 + w_2$. Shafer, in 1981,[20] goes on to show that an additive probability distribution P combined with a simple support Bel focused on E with weight w is given by

$$(P \oplus \text{Bel})(\theta) = \begin{cases} K\, e^w P(\theta) & \text{if } \theta \in E, \\ K\, P(\theta) & \text{if } \theta \notin E, \end{cases}$$

where K is the usual normalising constant. This generalises to combining P with a number of belief functions $\text{Bel}_1, \ldots, \text{Bel}_m$ where B_j is focused on F_j with finite weight w_j. The result is

$$(P \oplus \text{Bel}_1 \oplus \text{Bel}_2 \ldots \oplus \text{Bel}_m)(\theta) = K \exp[\Sigma\{w_j \mid \theta \in F_j\}] P(\theta).$$

The combined belief functions thus obtained are additive probability distributions. We can now show how Jeffrey's rule of conditioning follows from Dempster's rule of combination. First a couple of lemmas (taken directly from Shafer[20]):

Lemma 1. Suppose the belief function Bel over Θ is carried by a partition $E_1, \ldots, E_n$. If $\text{Bel} \oplus \text{Bel}_{E_i}$ exists then

$$\text{Bel} \oplus \text{Bel}_{E_i} = \text{Bel}_{E_i}.$$

Lemma 2. When we combine a belief function Bel with another belief function that is carried by a partition, we do not change the conditional values of Bel given elements of that partition. If Bel^1 is the second belief function carried by the partition $E_1, \ldots, E_n$, then

$$(\text{Bel} \oplus \text{Bel}^1)(A \mid E_i) = \text{Bel}(A \mid E_i).$$

The main result is now a Theorem due to Shafer.[20]

Theorem. Consider an additive probability distribution P and a belief function Bel defined in the same frame Θ. Suppose $E_1, \ldots, E_n$ is a partition such that Bel is carried by the E_i and $P(E_i) > 0$ for all i. Denote the orthogonal sum of P and Bel by $Q = P \oplus \text{Bel}$, and denote $Q(E_i)$ by q_i. Then

$$Q(A) = \sum_{i=1}^{n} q_i P(A \mid E_i)$$

for all $A \subset \Theta$.

Proof. Since Q is an additive probability distribution

$$Q(A) = \sum_{i=1}^{n} Q(E_i) Q(A \mid E_i).$$

Lemma 2 tells us that $Q(A \mid E_i) = P(A \mid E_i)$, which proves the theorem.

Notice that the $Q(E_i)$ are constructed by combing P and Bel, where Bel is carried by $E_1, \ldots, E_n$. For $n = 2$ this means

$$Q(A) = P(A \mid E) Q(E) + P(A \mid E) Q(\bar{E})$$

and we assign m values to E and $\bar{E}$ giving rise to Bel which combined with P gives Q. This is to be contrasted with the original Jeffrey approach which left the mechanism modelling the assignment of $Q(E_i)$ unspecified.

The construction of a Bel carried by $E_1, \ldots, E_n$ given that P and Q are related by Jeffrey's rule is also possible, although not unique. The analysis is enlightening from an IR point of view. This time I shall do the constructing only for E_1 and E_2 (but see Shafer[20] for the general case).

Given two additive probability distributions P and Q that are related by Jeffrey's rule relative to E_1 and E_2, can we find a Bel such that $Q = P \oplus \text{Bel}$? Let $p_i = P(E_i)$ and $q_i = Q(E_i)$, $i = 1, 2$ and $Q(A) = q_1 P(A \mid E_1) + q_2 P(A \mid E_2)$. Remember that p_i are the degrees of belief before observations, and that the direct effect of observation is to change those to q_i. Also the observation is assumed to leave the probabilities conditional on E_i unchanged, i.e. $P(A \mid E_i) = Q(A \mid E_i)$. Without loss of generality we can assume

$$\frac{q_1}{p_1} \leqslant \frac{q_2}{p_2},$$

which means that the evidence from the observation favours E_2 over E_1. In IR parlance this means that if $E_1 = (x = 1)$ for index term x, then $E_2 = (x = 0)$, which in turn means that

$$\frac{Q(x=1)}{P(x=1)} \leqslant \frac{Q(x=0)}{P(x=0)}$$

The appropriate Bel function carried by E_1, E_2 such that $P \oplus \text{Bel} = Q$ is constructed as follows: Choose Bel as the simple support function focused on E_2 with weight of evidence

$$w = \log \frac{q_2}{p_2} - \log \frac{q_1}{p_1}.$$

Earlier we showed how to combine a simple support function with an additive probability distribution, giving

$$(P \oplus \text{Bel})(\theta) = K \frac{q_i p_1}{p_i q_1} P(\theta) \qquad \begin{matrix} i = 1, 2, \\ \theta \in E_i. \end{matrix}$$

Thus

$$(P \oplus \text{Bel})(E_i) = K \frac{q_i p_1}{p_i q_1} P(E_i) = K \frac{p_1}{q_1} q_i$$

$$(P \oplus \text{Bel})(E_1) + (P \oplus \text{Bel})(E_2) = 1,$$

which implies that

$$K = \frac{q_1}{p_1}, \quad \text{or} \quad (P \oplus \text{Bel})(E_i) = q_i.$$

By construction Bel is carried by E_1 and E_2, therefore $P \oplus \text{Bel}$ has the same conditional probabilities given E_i as P does. Hence $P \oplus \text{Bel}$ must be the same as Q.

A construction like this is not unique. Instead of focusing Bel on E_2 with weight w, Bel could have been made up of $\text{Bel}_1 \oplus \text{Bel}_2$ each a simple support function with weight, w_1 and w_2 such that the combined weight was

$$\log \frac{q_2}{p_2} - \log \frac{q_1}{p_1}.$$

The construction of Bel to modify the probability distribution P generalises to multi-state evidence $E_1, \ldots, E_n$. The belief function Bel is made up of Bel_j, each Bel_j focused on a F_j constructed out of the $E_1, \ldots, E_n$ with weights chosen similarly to the binary case.

The construction of Bel leading to $Q = P \oplus \text{Bel}$ is a convenience. It allows us to view the modification of P in the light of user specified weights w in a straight forward way. In the original formulation of the Jeffrey rule, the change from $P(E_i)$ to $Q(E_i)$ was left unexplained. The observations were assumed to have a direct effect leading to a change in $P(E_i)$. In the case of modification through belief functions, the user is asked to specify a weight (or weights) associated with simple support functions focused on appropriate sets. It should not be assumed that the choice of these sets is always obvious, although in IR a

natural choice is the set of documents indexed by a given term. There is in general some guidance available for this choice (see the excellent paper by van Fraassen[7]).

Shafer's analysis of the derivation of the q_i is not the only one. A closely related one is due to Field.[21,22] It is a simple matter to define Field's specification of the q_i now that we have Shafer's. Field insisted that in the Jeffrey rule the q_i should be derived from the p_i by a function dependent on a parameter α 'that represents the effects that the sensory stimulation has *by itself* (independently of the value of p)', and then that q should be represented as a function of α and p together.[21] He chose

$$q = \frac{pe^{\alpha}}{pe^{\alpha}+(1-p)e^{-\alpha}} \quad \alpha = \tfrac{1}{2}\log\frac{q/p}{(1-q)/(1-p)}.$$

Jeffrey's rule now becomes for the binary case:

$$Q(A) = [e^{\alpha}P(A \wedge E)+e^{-\alpha}P(A \wedge \bar{E})]/(e^{\alpha}P(E) + e^{-\alpha}P(\bar{E})).$$

Notice that α needs to be estimated only once in a series of revisions. So, in a feedback situation, α could be established on the first round and then used in subsequent feedback iterations, but see Garber[22] for some counter-intuitive results.

Apart from the factor of $\frac{1}{2}$, the Field weight α is the same as the Shafer weight of evidence w. In both cases it is:

$$w = 2\alpha = \log\left(\frac{q}{1-q}\right)\Big/\frac{p}{1-p} = \log\left[\frac{O(q)}{O(p)}\right] \quad \text{where } O \text{ is odds.}$$

6. IMAGING AND GENERALISED CONDITIONALISATION

In the analysis of conditionals so far we have assumed a simple conditioning even E enabling us to write

$$P^*(A) = P(A \mid E)\,P^*(E)+P(A \mid \bar{E})\,P^*(\bar{E}).$$

This is appropriate when we are dealing with simple keyword or index term based retrieval, that is, when $E = (x = k)$ for $k = 1$ usually indicating absence or presence, and $P^*(E)$ indicates the importance the user attaches to E's satisfaction. In some earlier papers, I have described some tentative ways of dealing with conditional events.[12,13] For example, it is possible to view information retrieval as a form of inference where it becomes necessary to evaluate $P(d \rightarrow q)$, the probability of a conditional event, namely the event of the current document implying the query. As is well known from experimental and practical work, it is rare for a document to imply a statement with certainty which leaves us with the problem of evaluating $P(d \rightarrow q)$. As argued in my earlier work, I believe that the correct way to proceed is to specify a semantics of the conditional '$\rightarrow$' connecting arbitrary propositions in a given language.

To make sense of the evaluation of the probability of conditional statements, we have to move beyond restricting our event space to Boolean Algebra. It is not at all clear how this should be done (but see Ref. 23). The intention is that we wish to introduce a new connective '$\rightarrow$' into our language which somehow captures the kind of inferences we accept, that is, the kind we make in natural language. Or more to the point, which rejects ones we do not accept, e.g. $A \rightarrow B \mid = A \;\&\; C \rightarrow B$ (weakening).

We are attempting here to move meta-level reasoning into the object-level. In a sense the Dempster–Shafer theory does that when it interprets $\mathrm{Bel}(A)$ as the probability that the proposition A is provable given the evidence. Shafer is not alone in proposing that the impact of evidence on a hypothesis be measured by the probability of provability. Cohen[24] in his early work proposed a similar idea, one that probability statements be evaluated in terms of their inferential soundness. This proposal becomes especially attractive if the event space is non-Boolean.

Let us now examine the impact of complex evidence on conditionalisation. Consider, for example:

$$P^*(\mathrm{rel}) = P(\mathrm{rel} \mid d \rightarrow q)\,P^*(d \rightarrow q) + P(\mathrm{rel} \mid \neg(d \rightarrow q))\,P^*(\neg(d \rightarrow q))$$

If '$\rightarrow$' is a connective in a non-standard logic, evaluation of $P^*(d \rightarrow q)$ might prove difficult, the properties of '$\rightarrow$' would certainly play a role. Moreover, $P(\mathrm{rel} \mid d \rightarrow q)$ cannot now be interpreted as simple probabilistic indexing weight.

However, it is possible to rewrite the equation so that it can be evaluated. Let us take

$$P^*(\mathrm{rel}) = P((d \rightarrow q) \rightarrow \mathrm{rel})\,P^*(d \rightarrow q) + (P\neg(d \rightarrow q) \rightarrow \mathrm{rel})\,P^*(\neg(d \rightarrow q))$$

What we have just done is replaced the conditional probability with the probability of the conditional. This transformation will not work in general, that is it will not work for any P and continue to allow arbitrary nesting of arrows such that $P(A \rightarrow B) = P(A \mid B)$ (Stalnaker Thesis) where A and B are now themselves conditional statements. The triviality results of Lewis[25] block the transformation. However, Lewis also showed that if Bayesian conditionalisation is replaced by imaging then we can indeed introduce a connective '$\rightarrow$' into our language such that $P(A \rightarrow B) = P_A(B)$ where P_A is P imaged on A. The semantics of '$\rightarrow$' in this case is the one for the Stalnaker conditional given in terms of possible worlds in an earlier paper.[13] Van Fraassen[26] showed that by assuming the Stalnaker Thesis with the usual definition of $P(A \mid B)$ and allowing a limited amount of nesting we arrive at an underlying non-classical logic which rejects:

$$A \rightarrow B \vdash A \wedge C \rightarrow B \qquad \text{(weakening)}$$
$$A \rightarrow B, B \rightarrow C \vdash A \rightarrow C \qquad \text{(transitivity)}$$

but accepts the conditional excluded middle:

$$\vdash (A \rightarrow B) \vee (A \rightarrow \neg B).$$

This logic is commonly called C2 and is conveniently summarised in van Fraassen.[26]

I conjecture that this is the appropriate non-classical logic for IR, and that the appropriate probabilistic revision process is imaging. For, according to this logic we can rewrite the last equation for $P^*(\mathrm{rel})$ as:

$$P^*(\mathrm{rel}) = P((d \rightarrow q) \rightarrow \mathrm{rel})\,P^*(d \rightarrow q) + P((d \rightarrow \neg q) \rightarrow \mathrm{rel})\,P^*(d \rightarrow \neg q)$$

making use of the conditional excluded middle. To compute $P^*(\mathrm{rel})$ one would now apply the process of imaging to the probabilities in the equation. Potentially a very powerful result.

7. CONCLUSIONS

The theory developed in this paper is an attempt to introduce an epistemic view of probability into IR. The proposal is that Bayesian conditionalisation be replaced by Jeffrey conditionalisation where appropriate, and that we base our logic on $C2$. It is interesting that Wong and Yao[27] have been motivated to tackle probabilistic inference for IR by similar considerations. The results in my paper are entirely theoretical and need thorough experimental exploration, but that is the next task!

REFERENCES

1. M. E. Maron and J. L. Kuhns, On relevance, probabilistic indexing and retrieval. *Journal of the ACM* **7**, 216–244 (1960).
2. R. C. Jeffrey, *The Logic of Decision*, 2nd edn. University of Chicago Press, Chicago (1983).
3. W. F. Donkin, On certain questions relating to the theory of probabilities. *Philosophical Magazine*, 4th series **1** (5), 353–368, 458–466 (1851).
4. G. Boole. *An Investigation of The Laws of Thought*. Dover (1854).
5. M. Keynes, *A Treatise on Probability*. Macmillan, London (1929).
6. J. Pearl, Probabilistic reasoning in intelligent systems. In *Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California (1988).
7. B. C. van Fraassen, Relational belief and probability kinematics. *Philosophy of Science* **47**, 165–187 (1980).
8. P. M. Williams, Bayesian conditionalisation and the principle of minimum information. *British Journal of the Philosophy of Science* **31**, 131–144 (1980).
9. S. E. Robertson, The probability ranking principle in IR *Journal of Documentation* **33**, 294–304 (1977).
10. S. E. Robertson and K. Sparck Jones, Relevance weighting of search terms. *Journal of the American Society for Information Science* **27**, 129–146 (1976).
11. C. J. van Rijsbergen, *Information Retrieval*, 2nd edn. Butterworths, London (1979).
12. C. J. van Rijsbergen, A non-classical logic for Information Retrieval. *Computer Journal* **29**, 481–485 (1986).
13. C. J. van Rijsbergen, Towards an information logic. In *Proceedings of the Twelfth Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, edited N. Belkin and C. J. van Rijsbergen, pp. 77–86. ACM, New York (1989).
14. N. Fuhr, Probabilistic models in information retrieval, this issue (1992).
15. P. Diaconis and S. L. Zabell, Updating subjective probability. *Journal of the American Statistical Association* **77** (380), 822–830 (1982).
16. R. Sibson, Information radius. *Z. Wahrscheinlickkeitstheorie verw. Geb.* **14**, 149–160 (1969).
17. G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, Princeton (1976).
18. G. Shafer, Constructive probability. *Synthese* **48**, 1–60 (1981).
19. L. J. Cohen, *Probability – The One and the Many*. Annual Philosophical Lecture, Henrietta Hertz Trust. OUP, London (1975).
20. G. Shafer, Jeffrey's rule of conditioning. *Philosophy of Science* **48**, 337–362 (1981).
21. H. Field, A note on Jeffrey conditionalization. *Philosophy of Science* **45**, 362–367 (1978).
22. D. Garber, Discussion: Field and Jeffrey conditionalization. *Philosophy of Science* **47**, 142–145 (1986).
23. I. R. Goodman, H. T. Nguyen and E. A. Walker, *Conditional Inference and Logic for Intelligent Systems: A Theory of Measure-free Conditioning*. North Holland, Amsterdam (1991).
24. L. J. Cohen, *The Probable and the Provable*. Clarendon Press, Oxford (1977).
25. D. Lewis, Probabilities of conditionals and conditional probabilities. *Philosophical Review* **85**, 297–315 (1976).
26. B. C. van Fraassen, Probabilities of conditionals. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, edited W. L. Harper and C. A. Hooker, pp. 261–308. Reidel, Dordrecht (1976).
27. S. K. M. Wong and Y. Y. Yao, A probabilistic inference model for information retrieval. *Information Systems* **16**, 301–321 (1991).

Correspondence

Sir,

The recent article[1] describes a supposedly new algorithm for finding elements that occur more than $n \div k$ times in a bag of n items in time $O(nk)$ (for a given k). For the case $k = 2$, the algorithm was first developed by Boyer and Moore some ten years ago[2]. Two algorithms for general k were published nine years ago in Ref. 3, and the second of these two algorithms is essentially the one appearing in Ref. 1. Ref. 3 also shows that if the objects in the bag are totally ordered, the time of the algorithm can be reduced from $O(nk)$ to $O(n \log k)$. The algorithm was also discussed in Ref. 4.

Yours faithfully,

DAVID GRIES
Computer Science,
Cornell University,
and
J. MISRA
Computer Sciences,
University of Texas at Austin

References

1. D. Campbell and T. McNeill, Finding a majority when sorting is not available. *The Computer Journal* **34** (2), 186 (1991).
2. R. S. Boyer and J. S. Moore, *MJRTY – a Fast Majority Vote Algorithm*. Institute for Computing Science and Computer Applications, University of Texas at Austin, Technical Report ICSCA-CMP-32 (1982). Also, *MJRTY – a Fast Majority Vote Algorithm*. In *Automated Reasoning: Essays in Honor of Woody Bledsoe*, edited R. S. Boyer, Dordrecht, The Netherlands: Kluwer Academic Publishers (1991).
3. J. Misra and D. Gries, Finding repeated elements. *Science of Computer Programming* **2**, 143–152 (1982).
4. D. Gries, A hands-in-the-pocket presentation of a k-majority vote algorithm. In: *Formal Development of Programs and Proofs*, edited E. W. Dijkstra, pp. 43–45. Addison-Wesley, Reading, (1990).

Author's comment

The article was developed from class members' responses to an assignment given in a graduate class on algorithms. Both authors appreciate the bibliographical information, of which they were unaware.