A Study of Probabilistic Information Retrieval Systems in the Case of Inconsistent Expert Judgments

Jung Jin Lee*
Department of Business Information Systems and Quantitative Analysis, Mississippi State University
P.O. Drawer DB, Mississippi State, MS 39762

Paul B. Kantor
Tantalus, Inc., 2140 Lee Road, Suite 218, Cleveland, OH 44118

The maximum entropy principle may be applied to the design of probabilistic retrieval systems. When there are inconsistent expert judgments, the resulting optimization problem cannot be solved. The inconsistency of the expert judgments can be revealed by solving a linear programming formulation. In the case of inconsistent judgment, four plausible schemes are proposed in order to find revised judgments which are consistent with the true data structure but still reflect the original expert judgment. These schemes are the Interactive, Minimum Distance, Minimum Cross-Entropy, and Path methods.

Background and Purpose of the Study

The maximum entropy principle (MEP) based on Shannon's measure (Shannon, 1928) has been used with great success in many areas. Cooper and Huizinga (1982) and Cooper (1983) have applied the MEP to the design of probabilistic information retrieval systems. Specifically, we consider a collection of documents which are categorized into Boolean components by attributes. The MEP estimates the probability of "relevance" of each Boolean component by integrating expert judgments about the "relevance" of attributes with the observed distribution of the Boolean components. The MEP retrieval system, in response to a user's request, provides an ordering of the Boolean components using this estimated probability of "relevance." Cooper (1983) has noted the potential of the MEP retrieval system as follows

> It overcomes several objections to the traditional Boolean design including its output size problems and some of its expressive limitations. It also improves in several ways on the traditional weighted-request design that uses uninterpreted weights and retrieval coefficients.

Several refinements of the MEP retrieval system design have been developed by Kantor (1982, 1984), Kantor and Lee (1986) using what is called the dual problem to the original constrained optimization.

In an MEP retrieval system, the probabilities that documents having certain attributes are relevant should be estimated by an expert (posssibly several experts) or an expert system. However, these estimates of the probabilities may be inconsistent with the actual data structure and hence the optimization problem of the maximum entropy principle may have no solution. Especially when the number of attributes is increased, it is easy to generate inconsistencies of the expert judgments. We anticipate that this difficulty will persist in the practical application of the MEP retrieval systems.

The central optimization problem of the MEP retrieval system is a nonlinear optimization with linear constraints. In this article, a linear programming problem with the same linear constraints is formulated to check consistency of the expert judgments. In the case of inconsistent expert judgments, four plausible schemes are proposed in order to find revised estimates which are consistent with the data structure but still reflect the original expert judgment. These are the Interactive, Minimum Distance, Minimum Cross-Entropy, and Path methods.

In the next section, we briefly review the features of a probabilistic retrieval system, and establish notations.

Probabilistic Information Retrieval System

We shall use the term "documents" to refer in general to any retrievable report or item that might contain

*To whom all correspondence should be addressed. Permanent address: Dept. of Statistics, Soong Jun University, Seoul, Korea.

Received November 21, 1988; revised March 6, 1989; accepted April 3, 1989.

further information relevant to the problem at hand. In the simplest example situation the entire set of "documents" may have neither, either, or both of two index terms A and B. We represent the four possible Boolean components of the entire set of "documents," R, using the notations of set operations as following

$$R = \bar{A}\bar{B} \cup \bar{A}B \cup A\bar{B} \cup AB$$

Let f_i, $i = 1,2,3,4$, be the fraction of all documents lying in each Boolean component. Suppose that the value of a document is either 0 or 1 which represents "not relevant" or "relevant" respectively. Let $p(i,v)$ denote the probability of document value v in the Boolean component i where $i = 1,2,3,4$ and $v = 0,1$. The situation may be described by Table 1. Since $p(i,v)$'s are the joint probabilities, we have the following probability constraints

$$\sum_{i=1}^{4}\sum_{v=0}^{1} p(i,v) = 1$$

$$p(i,v) \geq 0 \qquad i = 1,2,3,4 \quad \text{and} \quad v = 0,1 \qquad (1)$$

We want to provide an ordering of the Boolean components, in response to a user's request, by estimating the conditional probability of "relevance" of the documents in each component. Note that the conditional probability that, for example, an item in Boolean component 3 be relevant is $p(3,1)/f_3$. Since the fraction f_i lying in each Boolean component can usually be determined using a computer, the question is how to estimate the joint probability $p(i,v)$, the "relevance decomposition" of the documents in each component. If we know all the $p(i,v)$'s, then the Boolean components can be ranked by the order of the conditional probability $p(i,1)/f_i$, $i = 2,3,4$. (Since we are not interested in the documents which have neither of two index terms A and B, the Boolean component 1, $\bar{A}\bar{B}$, is removed from consideration.) The component which has the highest conditional probability will be the first candidate for "relevant" information retrieval.

It is impossible for an expert (or an expert system) to estimate all the $p(i,v)$'s. However, an expert might be able to provide an opinion in the form "the chance that documents indexed by the term A (or B) are relevant is V_A (or V_B). For example, in a system processing the query "Where are terrorists trained?" the index term A might be the word "terrorists." An expert may estimate that 30% of the documents indexed by the term "terrorists" will be relevant. An expert system processing a query "Should we bid the SDIO C^3/BM architecture program?" may contain some rules with conclusion "BID" or "DON'T BID," and others with conclusion "BUY" or "DON'T BUY." The person using the system may estimate that the former have a higher expected relevance than the latter.

The expert estimates of V_A and V_B provide partial information on the data structure and can be used to estimate the $p(i,v)$'s using the maximum entropy principle. The resulting constrained optimization problem is to maximize the entropy function of the probabilities $p(i,v)$ subject to V_A and V_B. Note that V_A and V_B can be represented using $p(i,v)$ and f_i as follows

$$V_A = \frac{p(3,1) + p(4,1)}{f_3 + f_4}$$

$$V_B = \frac{p(2,1) + p(4,1)}{f_2 + f_4} \qquad (2)$$

Hence the MEP optimization problem becomes

Find $p(i,v)$, $i = 1,2,3,4$ and $v = 0,1$ which

Maximize $-\sum_{i=1}^{4}\sum_{v=0}^{1} p(i,v) \ln p(i,v)$

Subject to

$$p(3,1) + p(4,1) = V_A(f_3 + f_4)$$

$$p(2,1) + p(4,1) = V_B(f_2 + f_4)$$

$$\sum_{i=1}^{4}\sum_{v=0}^{1} p(i,v) = 1$$

$$p(i,v) \geq 0 \qquad i = 1,2,3,4 \quad \text{and} \quad v = 0,1. \qquad (3)$$

This is a nonlinear programming problem which has linear constraints. If we use the conditional probability q_i

$$q_i = \frac{p(i,1)}{f_i}, \qquad i = 1,2,3,4, \qquad (4)$$

TABLE 1. Notation used for four Boolean components.

Component Number i	Boolean Component	Probability of Document Value Not Relev. $p(i,0)$	Relevant $p(i,1)$	Fraction of Boolean Component $f_i = p(i,0) + p(i,1)$
1	$\bar{A}\bar{B}$	$p(1,0)$	$p(1,1)$	$f_1 = p(1,0) + p(1,1)$
2	$\bar{A}B$	$p(2,0)$	$p(2,1)$	$f_2 = p(2,0) + p(2,1)$
3	$A\bar{B}$	$p(3,0)$	$p(3,1)$	$f_3 = p(3,0) + p(3,1)$
4	AB	$p(4,0)$	$p(4,1)$	$f_4 = p(4,0) + p(4,1)$

then the MEP optimization problem (3) can be written as follows

Find q_i, $i = 1,2,3,4$, which

Maximize

$$-\sum_{i=1}^{4}[q_i f_i \ln(q_i f_i) + (1 - q_i) f_i \ln(1 - q_i) f_i]$$

Subject to

$$\begin{aligned} f_3 q_3 + f_4 q_4 &= V_A(f_3 + f_4) \\ f_2 q_2 + f_4 q_4 &= V_B(f_2 + f_4) \\ 0 \le q_i \le 1, \quad & i = 1,2,3,4. \end{aligned} \tag{5}$$

Note that the probability constraint in the optimization problem (3), i.e.,

$$\sum_{i=1}^{4}\sum_{v=0}^{1} p(i,v) = 1$$

becomes redundant in this transformation because the sum of the f_i is equal to 1. If we have the solution of the optimization problem (5), then the Boolean component which has the highest value of q_i, $i = 2,3,4$, will be the best candidate for "relevant" information retrieval.

If there are N index terms, then there are 2^N Boolean components. The optimization problem (5) is generalized with N index terms in the Appendix.

Checking the Feasibility of the Expert Judgments

The expert estimates of V_A and V_B may easily become inconsistent with the data structure because they are estimated without detailed knowledge of the contents of the data base. For example, suppose there are 1,000 documents which may have neither, either, or both of two index terms A and B as shown in Table 2. There are 400 documents indexed by the term A (documents in $A\bar{B}$ and AB) and 300 documents indexed by the term B (documents in $\bar{A}B$ and AB). Suppose the expert estimate of the chance that documents indexed by A are relevant is

$$V_A = 10\%$$

and the chance that documents indexed by B are relevant is

$$V_B = 60\%.$$

TABLE 2. Computation of fixed fractions of documents.

Component Number i	Boolean Component	Number of Documents	Fraction f_i
1	$\bar{A}\bar{B}$	500	.50
2	$\bar{A}B$	100	.10
3	$A\bar{B}$	200	.20
4	AB	200	.20
	Total	1,000	1.00

This is not possible because only 40 (10% of 400) documents indexed by A are relevant, while 180 (60% of 300) documents indexed by B are relevant and 80 (180–100) of them (at least) also have index term A.

It is easy for such contradictions to arise when an expert must assign the estimates without detailed knowledge of the sizes of the Boolean components. If the number of index terms is increased, there is more chance that an expert judgment becomes inconsistent. When the maximum entropy principle is applied to such an inconsistent judgment, the MEP optimization problem (5) has no solution and hence the retrieval system fails to answer the user's request. Thus inconsistent expert judgment is one of the obstacles to practical application of the MEP retrieval systems.

A linear programming (called a Phase 1 LP) problem, using the linear constraints of the MEP optimization problem (5), can be formulated in order to check whether the expert judgment is consistent with those constraints or not. The Phase 1 LP of the MEP optimization problem (5) is as follows

Find q_1, q_2, q_3, q_4, y_1, and y_2 which

Minimize $y_1 + y_2$

Subject to

$$\begin{aligned} f_3 q_3 + f_4 q_4 + y_1 &= V_A(f_3 + f_4) \\ f_2 q_2 + f_4 q_4 + y_2 &= V_B(f_2 + f_4) \\ 0 \le q_i \le 1, \quad i = 1,2,3,4, \quad & y_1, y_2 \ge 0. \end{aligned} \tag{6}$$

Basically, y_1 and y_2 (called slack variables) represent the amount by which the expert judgments (V_A and V_B) are inconsistent with the data structure (f_1, f_2, f_3, f_4). Hence, if the expert judgments are consistent with the data structure, then the solutions for the slack variables should be equal to zeros. It can be shown, e.g., Solow (1984), that if $q_1^*, q_2^*, q_3^*, q_4^*, y_1^*$, and y_2^* are the optimal solution of the Phase 1 LP (6) and $y_1^* = y_2^* = 0$, then the MEP optimization problem (5) is feasible and hence the expert judgment is consistent with the data structure. In this case the values q_1^*, q_2^*, q_3^*, and q_4^* can be used as a starting point for solving the MEP optimization problem (5).

The optimal solution of the Phase 1 LP (6) using the above example is

$$q_1^* = 0, \quad q_2^* = 1, \quad q_3^* = 0, \quad q_4^* = .2,$$
$$y_1^* = 0, \quad \text{and} \quad y_2^* = 0.04.$$

This confirms that the expert judgments, $V_A = .10$ and $V_B = .60$, are inconsistent with the data structure because $y_2^* \neq 0$.

Methods to Find a Feasible Estimate of (V_A, V_B)

When we find that an expert estimate (V_A, V_B) is inconsistent, we need to revise the estimate so that it is consistent with the true data structure but still re-

flects the expert judgment. In order to do this we must study the region of estimates (V_A, V_B) (the "feasible region") that are consistent with a given data structure (f_1, f_2, f_3, f_4).

The feasible region of (V_A, V_B) is the set of (V_A, V_B), $0 \le V_A, V_B \le 1$, such that there exists some q_1, q_2, q_3, q_4 which satisfy the constraints of the MEP optimization problem (5):

$$\begin{aligned} f_3q_3 + f_4q_4 &= V_A(f_3 + f_4) \\ f_2q_2 \quad + f_4q_4 &= V_B(f_2 + f_4) \\ 0 \le q_1, q_2, q_3, q_4 &\le 1 \end{aligned} \tag{7}$$

In this case the feasible region of (V_A, V_B) is all the possible linear combinations (with coefficients $0 \le q_1, q_2, q_3, q_4 \le 1$) of the following three points

$$\begin{pmatrix} 0 \\ \dfrac{f_2}{f_2 + f_4} \end{pmatrix} \quad \begin{pmatrix} \dfrac{f_3}{f_3 + f_4} \\ 0 \end{pmatrix} \quad \begin{pmatrix} \dfrac{f_4}{f_3 + f_4} \\ \dfrac{f_4}{f_2 + f_4} \end{pmatrix}$$

It can be shown that this feasible region is a convex set. (See the Appendix for a general proof.) Figure 1 shows this convex feasible region of (V_A, V_B).

Therefore, the problem of finding an "acceptable and feasible" estimate is equivalent to finding reasonable criteria that select a point in the convex feasible region. We propose four attractive methods to select a point in the feasible region. These are the Interactive, Minimum Distance, Minimum Cross-Entropy, and Path methods.

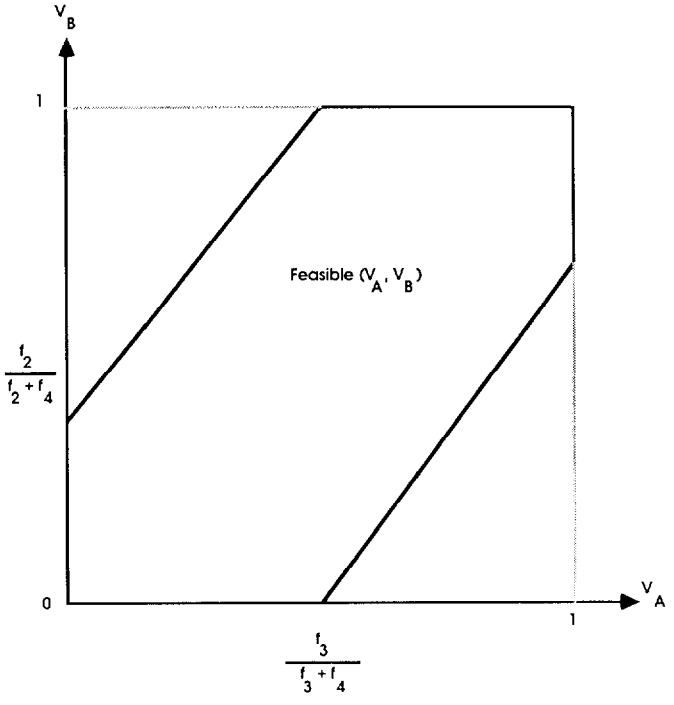

FIG. 1. The set of feasible value estimates determined by Table 2.

Interactive Method

One possible simple approach (described to one of us by A. Charnes) to revise the inconsistent judgment is interactive feedback with the expert. In this approach the expert is asked to broaden the estimates into ranges until they finally reach the convex feasible region such as

$$\begin{aligned} a_L \le V_A \le a_U, &\quad \text{where } 0 \le a_L, a_U \le 1 \\ b_L \le V_B \le b_U, &\quad \text{where } 0 \le b_L, b_U \le 1 \end{aligned}$$

The revised estimate and the optimal ordering of the Boolean components can be found by solving the MEP optimization problem (5) with additional constraints for the range estimates as follows

Find q_i, q_2, q_3, q_4, V_A, and V_B which

Maximize

$$-\sum_{i=1}^{4} [q_i f_i \ln(q_i f_i) + (1 - q_i) f_i \ln(1 - q_i) f_i]$$

Subject to

$$\begin{aligned} f_3q_3 + f_4q_4 &= V_A(f_3 + f_4) \\ f_2q_2 \quad + f_4q_4 &= V_B(f_2 + f_4) \\ 0 \le q_i \le 1, &\quad i = 1,2,3,4. \\ a_L \le V_A \le a_U, &\quad \text{where } 0 \le a_L, a_U \le 1 \\ b_L \le V_B \le b_U, &\quad \text{where } 0 \le b_L, b_U \le 1 \end{aligned} \tag{8}$$

The relative rates at which V_A and V_B are broadened will directly reflect the firmness of those prior expert estimates. Of course, we may presume that any true judgment lies between the lower and upper bounds.

Minimum Distance Method

If the interactive feedback with the expert is not practical, we need an algorithm which generates a feasible problem by making the estimates V_A and V_B consistent with the data. We presume that the inconsistent expert judgment is not completely wrong and has some valuable information. One intuitive method for finding an "acceptable and feasible" estimate is to find the "closest" feasible point to the inconsistent expert judgment $(\hat{V}_A, \hat{V}_B)$ using the Euclidean distance metric as follows

Find q_i, q_2, q_3, q_4, V_A, and V_B which

Minimize $(\hat{V}_A - V_A)^2 + (\hat{V}_B - V_B)^2$

Subject to

$$\begin{aligned} f_3q_3 + f_4q_4 &= V_A(f_3 + f_4) \\ f_2q_2 \quad + f_4q_4 &= V_B(f_2 + f_4) \\ 0 \le q_1, q_2, q_3, q_4, V_A, V_B &\le 1 \end{aligned} \tag{9}$$

This is a quadratic programming problem which has a unique global solution because the objective function

is strictly convex and the feasible region is a convex set (Bazaraa, 1979). Note that this method of revising $(\hat{V}_A, \hat{V}_B)$ makes no further use of the entropy concept.

Minimum Cross-Entropy Method

The inconsistent expert judgments $\hat{V}_A$ (hence $1 - \hat{V}_A$) and $\hat{V}_B$ (hence $1 - \hat{V}_B$) can be considered as the prior estimates of the true probabilities of "relevance" in each index term. That is, they are estimates prior to knowing the facts about the database and its components. Note that $(\hat{V}_A, 1 - \hat{V}_A)$ and $(\hat{V}_B, 1 - \hat{V}_B)$ are probability distributions. Another method to revise the expert judgment $(\hat{V}_A, \hat{V}_B)$ is to find posterior probability distributions which are consistent with the data structure and reflect the prior probabilities. The cross entropy function (or directed divergence) can find "the most difficult to discriminate" posterior probabilities from the prior probabilities (1968). In other words, it finds a feasible (V_A, V_B) that has nearly the same information as the given inconsistent expert judgment $(\hat{V}_A, \hat{V}_B)$.

The Minimum Cross-Entropy method is as follows:

Find q_1, q_2, q_3, q_4, V_A, and V_B which

Minimize

$$V_A \ln\left(\frac{V_A}{\hat{V}_A}\right) + (1 - V_A) \ln\left(\frac{1 - V_A}{1 - \hat{V}_A}\right) + V_B \ln\left(\frac{V_B}{\hat{V}_B}\right) + (1 - V_B) \ln\left(\frac{1 - V_B}{1 - \hat{V}_B}\right)$$

Subject to

$$f_3 q_3 + f_4 q_4 = V_A(f_3 + f_4)$$

$$f_2 q_2 + f_4 q_4 = V_B(f_2 + f_4)$$

$$0 \le q_1, q_2, q_3, q_4, V_A, V_B \le 1 \qquad (10)$$

Since the cross-entropy function is strictly convex, this nonlinear programming problem (10) also has a unique solution.

Path Method

Suppose we don't have any information about the data structure. Then the natural estimate of the conditional probabilities of "relevance" in each index term is $(V_A, V_B) = (0.5, 0.5)$. This estimate is always feasible and has the maximum possible entropy. The path method is to find the "closest" feasible point (V_A, V_B) to the inconsistent original judgment $(\hat{V}_A, \hat{V}_B)$, on a straight line from the point (0.5, 0.5), which has the maximum possible entropy. If we let $\mathbf{V}_o = (0.5, 0.5)$ and $\hat{\mathbf{V}} = (\hat{V}_A, \hat{V}_B)$, then the path method is equivalent to finding a number c' such that the MEP optimization problem (6) is feasible for $(1 - c)\mathbf{V}_o + c\hat{\mathbf{V}}$ where $0 < c \le c'$ and infeasible otherwise. Hence the feasible estimate of the Path method is $(1 - c')\mathbf{V}_o + c'\mathbf{V}$. Since the set of feasible (V_A, V_B) is a convex set, it includes all boundary points. Hence there exists unique positive $c' < 1$. Practically, it may be hard to find the exact c', but the following simple algorithm can find an approximate value of c'.

[Algorithm] Path Method

Let δ be an increment between 0 and 1.

Step 1: Initialize $c = 0$.

Step 2: Let $c = c + \delta$.
Solve the Phase 1 LP (6) with $(1 - c)\mathbf{V}_o + c\hat{\mathbf{V}}$.

Step 3: If $(1 - c)\mathbf{V}_o + c\hat{\mathbf{V}}$ is feasible, go to step 2 otherwise go to step 4

Step 4: $c' = c - \delta$.

Of course, the small value of the increment δ will find an accurate c'.

Numerical Example

Figure 2 shows the convex feasible region of (V_A, V_B) using the example in the previous section where the fractions are $f_1 = .50$, $f_2 = .10$, $f_3 = .20$, and $f_4 = .20$. Note that the inconsistent expert judgment, $(\hat{V}_A, \hat{V}_B) = (.10, .60)$, is not in the feasible region.

Suppose the expert is firm about the estimate $\hat{V}_B$ but not sure about $\hat{V}_A$. Then the interactive feedback with the expert might give the following ranges

$$.05 \le V_A \le .30$$

$$.55 \le V_B \le .65$$

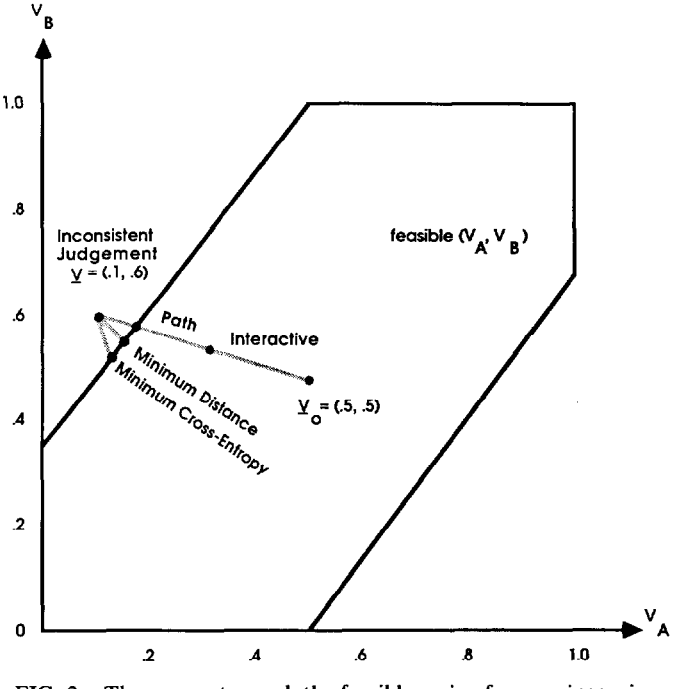

FIG. 2. Three ways to reach the feasible region from an inconsistent prior (see Table 3).

In this case the solution of the optimization problem (8) is as follows

$$q_1 = .50, \quad q_2 = .79, \quad q_3 = .17, \quad q_4 = .43,$$
$$V_A = .30, \quad V_B = .55$$

Using the inconsistent expert estimate, $(\hat{V}_A, \hat{V}_B) = (.1, .6)$, of the example discussed in the previous section, the revised feasible estimates of the Minimum Distance, Minimum Cross-Entropy, and Path methods are as shown in Table 3. Figure 2 shows these revised feasible estimates. Note that, in this example, the revised estimates are different from each other depending on the method, but the rank orders of q_i's in each method are the same. Hence, all methods provide the same ordering of the Boolean components, in descending order of "relevance" excluding $\bar{A}\bar{B}$, for information retrieval as

$$\bar{A}B, AB, A\bar{B}$$

As a practical matter, small differences of the revised estimates are probably unimportant and any choice would be acceptable. However, in more complex examples, each method may lead to different rankings of the Boolean components, and it becomes necessary to choose a method. One possibility is to choose on the basis of computational simplicity. The Path method requires only a linear programming code which is available easily, while the other methods require a nonlinear programming code. A second is to provide all four possibilities to the user with a message such as shown in Table 4. A user who is firm about the low relevance of A would choose 2 or 3. One who is firm about the high relevance of B would prefer 4.

Conclusions

The MEP retrieval system still shows more promise than performance. As introduced by Cooper and Huizinga (1982) and Cooper (1983), it provides a unique solution to an under-specified problem. In the dual formulation by Kantor [4][5] it becomes more mathematically tractable. Kantor and Lee [6] have shown that its effect is to impose the reality of the data base on the expert's (or end-user's) naive estimates of conditional probabilities of relevance. In addition to the problem of high dimensionality, noted by Cooper and Huizinga (1982) and resolved by Kantor (1984), Cooper and Huizinga (1982) noted that the expert judgment may be incompatible with the data structure.

TABLE 3. Comparison of several revisions of inconsistent prior estimates.

Method	Revised Estimate (V_A, V_B)	Solution of the MEP optimization (5) (q_1, q_2, q_3, q_4)
Min Distance	(.16, .55)	(.50, .99, .00, .32)
Min Cross-Entropy	(.15, .53)	(.50, .98, .01, .30)
Path	(.20, .57)	(.50, .96, .03, .37)

TABLE 4. The observed use of the terms in your question suggests that the relevance estimates should be revised. Which of the four revisions looks most reasonable to you?

	Your Estimates (%)	Choice 1 (%)	Choice 2 (%)	Choice 3 (%)	Choice 4 (%)
A	10	30	16	15	20
B	60	55	55	53	57

The present article shows that the region of usable judgments is a convex set and examines the Interactive, Minimum Distance, Minimum Cross-Entropy and Path methods to give revised estimates that are consistent with the data structure but still reflect the original expert judgment. Thus the problem of inconsistent judgment may be resolved interactively or algorithmically and it will not prevent the development of the MEP retrieval system "front ends" for information retrieval. Each method gives a different revision and may lead to different rankings of the Boolean components. The choice among methods may be made, at least initially, on the basis of design convenience.

It is important to note, however, that all of the work to date has been theoretical, and there are no MEP retrieval systems in actual operation. It is possible to do some off-line testing of the MEP retrieval system, for example, using the Rutgers Information Retrieval Database developed by Saracevic (1988). However, the best way to appraise and improve the MEP retrieval system, as for any other proposed information retrieval systems, is to install it in heavy traffic and monitor its performance (Kantor, 1988). While it is attractive, in principle, to combine user judgments and term frequency data via the MEP, the ultimate test lies in end user evaluation of the ranked retrieval material.

Acknowledgments

This article was supported in part under National Science Foundation Grant IST-8318630. It is a pleasure to acknowledge discussions with W. Cooper, A. Charnes, and participants at several maximum entropy principle workshops.

Appendix

If there are N index terms, then there are 2^N Boolean components. The MEP optimization problem (5) can be generalized in this case as follows:

$$\text{Find} \quad q_i,\ i = 1, 2, \ldots, 2^N, \quad \text{which}$$

$$\text{Maximize} - \sum_{i=1}^{2^N} [q_i f_i \ln(q_i f_i) + (1 - q_i) f_i \ln(1 - q_i) f_i]$$

Subject to

$$\sum_{i \epsilon A(k)} q_i f_i = V_k \left(\sum_{i \epsilon A(k)} f_i \right), \qquad k = 1, 2, \ldots, N$$

$$0 \leq q_i \leq 1, \qquad i = 1, 2, \ldots, 2^N$$

where V_k, $k = 1, 2, \ldots, N$, represents the probability of relevance for index term k and $A(k)$ is the set of the Boolean components which are constrained by the index term k. Note that the number of unknowns in each of these grows exponentially with N.

The following theorem proves that the feasible region of $(V_1, \ldots, V_N)$ in the MEP optimization problem (11) is a convex set.

Theorem

Let M be the coefficient matrix of the constraints of the MEP problem (11) and D be the set of $V = (V_1, \ldots, V_N)$ which is feasible to the MEP problem (11). Then D is a convex set.

Proof

The N-vector $\mathbf{V}_o = (0, \ldots, 0)$ is always feasible because there exists an N-vector $\mathbf{q} = (0, \ldots, 0)$ and $M\mathbf{q} = \mathbf{V}_o$. Hence D is not an empty set. Let $\mathbf{V}_1$ and $\mathbf{V}_2$ be the elements in D. Then there exists $\mathbf{q}_1$ and $\mathbf{q}_2$ such that $0 \leq \mathbf{q}_1 \leq 1$, $0 \leq \mathbf{q}_2 \leq 1$, $M\mathbf{q}_1 = \mathbf{V}_1$ and $M\mathbf{q}_2 = \mathbf{V}_2$. We must show that, for every $0 < c < 1$, $c\mathbf{V}_1 + (1 - c)\mathbf{V}_2$ is also an element in D. Let $\mathbf{q} = c\mathbf{q}_1 + (1 - c)\mathbf{q}_2$ where $0 < c < 1$. Then $0 \leq \mathbf{q} \leq 1$ and

$$\begin{aligned} M\mathbf{q} &= M(c\mathbf{q}_1 + (1 - c)\mathbf{q}_2) \\ &= cM\mathbf{q}_1 + (1 - c)M\mathbf{q}_2 \\ &= c\mathbf{V}_1 + (1 - c)\mathbf{V}_2. \end{aligned}$$

Hence $c\mathbf{V}_1 + (1 - c)\mathbf{V}_2$ is also an element in D and therefore D is a convex set.

References

Bazaraa, M. S., & Shetty, C. M. (1979). *Nonlinear programming.* New York: John Wiley & Sons, Inc.

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

Cooper, W. S. (1983). Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science, 34,* 31–39.

Kantor, P. B. (1982). Evaluation of the feedback in information storage and retrieval systems. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology.* (pp. 99–120).

Kantor, P. B. (1984). Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Development, 3,* 88–94.

Kantor, P. B. (1988). National language specific evaluation sites for retrieval systems and interfaces. RIAO 88 Program. In *User-oriented content based text and image handling.* Cambridge, Massachusetts.

Kantor, P. B., & Lee, J. J. (1986). The maximum entropy principle in information retrieval. *The proceedings of the 1986 ACM conference on research and development in information retrieval.* Pisa, Italy (pp. 269–274).

Kullback, S. (1968). *Information theory and statistics.* New York: Dover Publications, Inc.

Saracevic, T. (1988). Retrieval abstracts, keywords and end user evaluations have been compiled for 360 searchers on 40 questions. Unpublished paper. Rutgers School of Communication, Library and Information Science.

Shannon, C. E. (1928). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379–432, 623–656.

Solow, D. (1984). *Linear programming — a finite improvement algorithm.* North Holland Publishing Company.

Zangwill, W. I., & Garcia, C. B. (1981). *Pathways to solutions and equilibria.* Englewood Cliffs, NJ: Prentice-Hall.