Information Technology: Research and Development (1982), *1* (99–112)

THE MAXIMUM ENTROPY PRINCIPLE AND ITS APPLICATION TO THE DESIGN OF PROBABILISTIC RETRIEVAL SYSTEMS

W. S. COOPER AND P. HUIZINGA

School of Library and Information Studies, University of California, Berkeley, CA 94720, USA

(Received 7 September 1981, revised 23 November 1981)

ABSTRACT

The Maximum Entropy Principle is a body of statistical theory addressed to the problem of how to make probability estimates in the face of an apparent insufficiency of data, without introducing unrealistic independence postulates. It is potentially important for the design of 'probabilistic' information retrieval systems — systems based on the premise that the essential task of a retrieval system is to order the items in the search universe by the estimated probability of their usefulness to the user. The utility of the maximum entropy principle lies in its capacity for providing rational probability estimates even in seemingly underdetermined situations, thus obviating the need for artificial simplifying assumptions. One possible application of the maximum entropy formalism involves the use of a request language in which each request term is weighted by the user's subjective estimate of its precision in the collection. Other applications are also possible. Precise maximum entropy calculations are computationally inefficient under some conditions but the prospects for finding fast approximations seem good.

1. BACKGROUND

In recent years there has been a growing interest in 'probabilistic' and 'utility-theoretic' approaches to the design of document retrieval systems and similar information search systems. The papers of Tague (1973), Bookstein and Swanson (1974), Harter (1975), Robertson and Sparck Jones (1976), Salton *et al.* (1976), Bookstein and Kraft (1977), van Rijsbergen (1977), Cooper and Maron (1978), Harper and van Rijsbergen (1978), and van Rijsbergen *et al.* (1980) are representative. A thoughtful review of many of the ideas has been provided by Robertson (1977a). Under the probabilistic approach it is assumed that the purpose of a search system is to rank the documents in the collection (or more generally, the items in the search universe whatever it may be) in decreasing order of their estimated probability of satisfying the searcher's need. A probability ranking of this sort

allows the searcher to examine the items one at a time, likeliest ones first, until either his need is satisfied or for some other reason he decides to terminate the search.

The probabilistic approach is based on the so-called 'Probability Ranking Principle' (Cooper, 1976; Robertson, 1977b). As originally stated, this principle — actually more of a design heuristic than a provable law — asserts that the overall effectiveness of a retrieval system to its users will be on average the best obtainable on the basis of the data available to the system if the system's response to each request is to rank the documents of the collection for the user in order of decreasing probability of usefulness to him, where the probability estimates are the best that can be made on the basis of those data. Counterexamples to this assertion can be constructed, which is why it cannot be regarded as an absolute principle (Cooper, 1976). Nevertheless it is generally conceded to be a helpful guide and it underlies, implicitly at least, much of the research in information retrieval that exploits probability theory in a nonsuperficial way.

Conceptually the probabilistic approach is immensely appealing because, with the help of the probability ranking principle, it succeeds in reducing most of the theory of information searching to a problem in probability estimation. For instance, the question of what the request language should be like reduces to a question of which probabilistic clues to elicit from the user as a basis for making probability-of-usefulness estimates. Similarly the problem of how to index the items to be searched reduces to the challenge of providing the most helpful data possible for the estimation of these same probabilities. And the choice of retrieval rule — that is, the strategy for ordering the output on the basis of the request, the indexing, and so on — becomes a matter of finding the most appropriate probability estimation formula for exploiting the available clues.

On the other hand, the usefulness of probabilistic analysis in information retrieval has been hindered by an apparent technical obstacle having to do with what we shall call 'underdetermination'. Simply described, the difficulty is that usually too few probabilistic data are available to allow probability-of-usefulness estimates to be made in a straightforward way using only the classical calculus of probabilities. The challenge is a serious one. The problem of underdetermination, if not met, would in many situations force the information retrieval theorist either to resort to unrealistic independence assumptions or to abandon the probabilistic approach altogether.

In what follows we wish to propose and discuss a theoretical solution to the problem of underdetermination and to illustrate its application. Our proposed solution amounts to the application of a body of statistical theory known as the *Maximum Entropy Principle*, also described sometimes as the 'Maximum Entropy Formalism'. Techniques closely related to it have already been applied to particular aspects of the retrieval problem (van Rijsbergen, 1977). Our aim here is to present the principle in a broader setting as a general tool for making probability-of-usefulness estimates under a variety of conditions, and even as a formalism capable of serving in and of itself as the central retrieval strategy of a search system.

2. INDEPENDENCE ASSUMPTIONS

As already suggested, one of the things that makes the task of estimating usefulness probabilities challenging is the problem of underdetermination — the presence in an analysis of too many unknowns — so that in practice there are too few data at hand

to calculate the desired probabilities in a straightforward way. True, it is often possible to use the classical probability calculus to derive formal expressions for the probabilities of interest, but in nontrivial cases such formulae typically turn out to be inapplicable because they involve quantities whose values cannot all be determined from available data. This is hardly surprising, for when one is concerned in the analysis not only with the event that a given document will be found useful, but also with N other events associated with the document properties mentioned in an N-element request, then one has to deal with an event space of size 2^{N+1}. This number is ordinarily much larger than the number of available pieces of relevant probabilistic evidence. There are just too many degrees of freedom to allow the probabilities of interest to be determined directly.

A common response to the problem of underdetermination has been to introduce special independence assumptions into the theory. One such set of independence assumptions asserts that (i) all searchable document properties are statistically (i.e., stochastically) independent given that a document is in fact useful (or 'relevant'), and that (ii) they are also independent given that it is not useful. By invoking strong independence assumptions such as these it is often possible to determine the desired usefulness probabilities from available data.

Unfortunately, such assumptions are usually grossly inaccurate. They are apt to be too simple, too strong, and are sometimes in direct conflict with available data about term dependencies. As a case in point, the independence postulated by the assumptions (i) and (ii) just described simply does not exist, even approximately, for many combinations of clues likely to be encountered in search situations. The reader can readily convince himself of this by thinking through a few particular examples. One might under some circumstances be inclined to forgive serious oversimplification in particular cases if the assumptions were in some sense correct on the average, or if they constituted a best guess in some cogent statistical sense, but no convincing arguments have been advanced showing that the assumptions are supportable even in this weak sense.

It would appear, then, that arbitrarily adopting special independence assumptions is not a wholly desirable approach to the problem of obtaining sound probability-of-usefulness estimates in information search systems. Indeed such assumptions are usually recognized to be crude even by those who employ them, their use being justified more or less as a desperation measure. Might there be a way to avoid the introduction of such artificial independence assumptions entirely?

3. THE MAXIMUM ENTROPY PRINCIPLE

Suppose a probability distribution is known to satisfy certain constraints, but that these are insufficient to determine it completely. Suppose too that nothing beyond these constraints is known about the distribution of interest. *Some* distribution has to be assumed because practical decisions must be made on the basis of the probabilities it assigns. What is the most natural distribution to adopt under these perplexing circumstances?

The *Maximum Entropy Principle* amounts to an attempt to answer this question. The principle specifies what has been described as the 'minimally prejudiced' or 'maximally vague' probability distribution consistent with the known constraints (Tribus, 1969). It is a method of translating fragmentary probability information into a complete probability assignment. As such it offers a possible solution to the

problem of underdetermination.

The principle is easy to state. Let the *entropy* E of a probability distribution $(p_1, \cdots, p_K)$ over an event space containing K mutually exclusive and exhaustive events be defined by the formula

$$E = -\sum_{i=1}^{K} p_i \log p_i \tag{1}$$

Now suppose that a probability distribution of interest is known only to satisfy a few miscellaneous relationships (e.g., it might be known that $p_2 = 0.3$, $p_5 - 0.4$ $p_6 = 0$, etc.) but that these constraints are insufficient to specify it completely. Then the maximum entropy principle may be stated, roughly and intuitively, as follows:

> *The minimally prejudiced probability distribution is that which maximizes the entropy subject to the given constraints* (Tribus, 1969, p. 120).

As a guide to practical decision-making this translates into the advice that a rational way of dealing with partial ignorance about a probability distribution is to find the distribution of maximal entropy consistent with whatever may be known, and then to act in accordance with the probabilities derived from it.

The information scientist will naturally associate the entropy formula (1) with the concept of entropy used in the information theory of Hartley, Shannon and Wiener. Indeed, one intuitive interpretation of the maximum entropy principle is that it provides a way of minimizing the information, or surprise value, of the probability distribution in the information-theoretic sense. It tells how to go about constructing a set of probabilities in which no information is implicit beyond that already contained in the known constraints. At one point Shannon himself, in a brief remark made in connection with the problem of assigning probabilities to English messages, proposed the use of what is now called the maximum entropy principle by suggesting that one 'consider the source with the maximum entropy subject to the statistical conditions we wish to retain' (Shannon, 1948). However, most of the literature on information theory *per se* has not been specifically concerned with the role of entropy in specifying probability distributions, and it would lead to confusion to think of the maximum entropy principle as part of garden-variety information theory.

The original and more convincing justification of the maximum entropy principle predated Shannon's information theory by several decades. It was given in 1871 by L. Boltzman in connection with a problem in statistical mechanics (see Jaynes, 1979). Boltzman's reasoning, translated out of the context of physics and into the terminology of information retrieval, might be encapsulated somewhat as follows. One asks: In how many ways could a collection of M documents be partitioned into K nonoverlapping sets containing respectively $M_1, \ldots, M_K$ documents? The answer is the multinomial coefficient

$$C = \frac{M!}{M_1!\, M_2! \ldots M_K!} \tag{2}$$

Now suppose that certain relationships are known to constrain the numbers

$M_1, \ldots, M_K$, but that these constraints are so weak that there remain many ways of choosing the K numbers in such a way as to satisfy them. Out of these many possible distributions, which is most likely to be realized in an M-document collection about which nothing is known beyond the stated constraints? Boltzman's answer is essentially that the most probable distribution is the one that can be realized in the most ways, i.e., it is the distribution that maximizes the multinomial coefficient C given by (2), subject to the known constraints.

If M and the M_i's are not too small, the Stirling approximation for factorials allows (2) to be rewritten (after taking the log of both sides) as

$$\text{Log } C = -M \sum_{i=1}^{K} \frac{M_i}{M} \log \frac{M_i}{M} \tag{3}$$

To choose $M_1, \ldots, M_K$ so as to maximize this expression subject to known constraints is of course just to maximize the entropy E of the probability distribution for the associated properties of a randomly drawn document under those constraints. Hence Boltzman's reasoning provides a rationale for regarding the probability distribution of maximum entropy as the best guess that can be made about what the actual distribution is like, given the lamentable state of partial ignorance of a guesser who knows nothing beyond the given constraints.

For probabilistic information search systems the maximum entropy principle supplies a convenient way of combining miscellaneous pieces of input evidence into probability-of-usefulness estimates. The various statistical input clues — request weights, indexing statistics, etc. — are the 'constraints' and the maximum entropy principle specifies the likeliest probability distribution satisfying those constraints. Once this probability distribution has been computed the desired usefulness probabilities can be derived from it and the collection ranked in accordance with them.

4. CONSEQUENCES OF THE PRINCIPLE

To gain a full appreciation of the meaning and power of the maximum entropy principle it is necessary not only to ponder Boltzman's reasoning but also to become acquainted with some of the other supporting considerations. An excellent historical review of these has been provided by Jaynes (1979). In addition it is helpful to become familiar with a few of the principle's particular consequences. The following are illustrative.

1. *The maximum entropy principle subsumes the principle of indifference.* Jacob Bernoulli's 'Principle of Indifference' (also sometimes called the 'Principle of Insufficient Reason') asserts that if one knows of no evidence bearing on the question of which of two complementary events is more probable than the other, then both events should be assigned a probability of 0.5, this being the only honest way to describe one's state of ignorance. In the absence of any constraints whatever, the maximum entropy principle leads to the same result, since $-(0.5 \log 0.5 + 0.5 \log 0.5)$ exceeds $-(p \log p) + (1-p) \log (1-p)$ whenever $p \neq 0.5$. In the general case of n mutually exclusive and exhaustive events, both the principle of indifference and the maximum entropy principle assign

the events equal probabilities of $1/n$ when no evidence to the contrary can be brought to bear.

2. *In the absence of dependency-inducing constraints, the maximum entropy principle makes events statistically independent.* For instance, it is readily verified that if the only known constraints are $P(A)=0.5$ and $P(B)=0.4$, then in the distribution of maximum entropy $P(A,B)$ will be assigned the value $0.5\times0.4=0.2$, making A and B independent. Figure 1 displays this distribution in the form of a Venn diagram. The reader might wish to check with a pocket calculator that the probabilities in the diagram cannot be changed without either lowering the entropy of the distribution or violating the constraints $P(A)=0.5$ and $P(B)=0.4$.

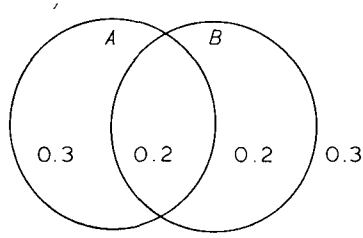

FIG. 1. Venn diagram displaying the probability distribution of maximum entropy satisfying the constraints $P(A)=0.5$ and $P(B)=0.4$. The two events turn out to be statistically independent in this distribution

More generally, the maximum entropy principle will tend to preserve as much independence overall as the constraints allow. As I. J. Good has remarked, 'in some sense it (the m.e.p.) pulls out the hypothesis in which the amount of independence is as large as possible' (1963). This then is how the principle removes the need for special independence postulates or other artificial simplifying assumptions. The maximum entropy formalism can be viewed as supplementing the available empirical data with what amounts to a single, residual, highly generalized, assumption of the greatest degree of independence consistent with the known constraints.

3. *The maximum entropy principle gives rise to a mutual reinforcement effect whereby individually weak pieces of evidence are combined into stronger composite evidence.* This property, which is crucially important for the search system application, is best explained by way of an example. Suppose the probability of the event U that a randomly drawn document will be useful to the requestor is $P(U)=0.1$; that is, it is known that one tenth of the collection is responsive to the information need. But when it is learned that a randomly selected document has a certain property A, its probability of usefulness goes up to $P(U/A)=0.3$, i.e., one's estimate of its usefulness probability would triple upon learning that it has property A. Similarly for a second property B, $P(U/B)=0.3$. For concreteness assume further that one tenth of the collection has property A, one tenth has B, and the two attributes are statistically independent in it so that $P(A)=P(B)=P(A/B)=P(B/A)=0.1$. Nothing beyond these few facts is known, we shall suppose.

The distribution of maximum entropy satisfying these conditions is shown in

Figure 2. The reader can easily check that this distribution does indeed fulfill all the constraints, and with a little more work it can also be verified that it is the highest-entropy distribution capable of doing so (c.f. Appendix). Now the value of $P(U/A,B)$ in this distribution is 0.67. Thus, while learning that a randomly drawn document has property A would triple one's estimate of its probability of usefulness from 0.1 to 0.3, and the same would be true for B, learning that the document has *both* properties A and B should according to the maximum entropy principle increase one's estimate from 0.1 to 0.67, a much larger factor. Thus the probability-enlarging effect of the two clues taken together is, according to the principle, much greater than that of either clue in isolation, as seems reasonable. It is this reinforcement phenomenon that allows the principle to be used to combine many pieces of interacting evidence into a single probability-of-usefulness estimate.

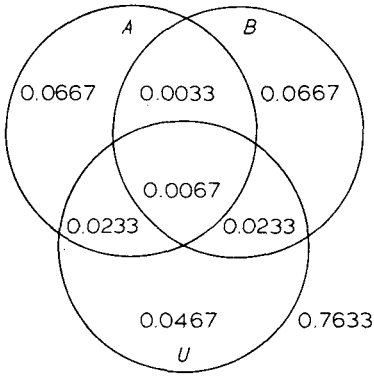

FIG. 2. The probability distribution of maximum entropy satisfying certain known constraints (see text). There is a 'reinforcement' effect in that $P(U/A,B)$ turns out to be larger than either $P(U/A)$ or $P(U/B)$

4. *The reinforcement effect is sensitive to statistical dependencies among clues.* For instance, if the properties A and B of the foregoing example had not been independent but instead highly correlated — say, $P(A/B)=P(B/A)=0.9$ rather than 0.1 — the maximum entropy computation would have yielded the result that $P(U/A,B)$ is only 0.32. The probability of usefulness inferred from the presence of both favorable clues is in this case only barely greater than that deducible from just one, for the news that a document has the second property is unsurprising when it is known already that it has the first. Because maximum entropy results are well behaved in this respect they offer a basis for the exploitation of descriptor co-occurrence data and other collection statistics, and more generally, for any scheme involving the construction of a retrieval ranking on the basis of a number of different and not necessarily independent clues.

5. APPLICATIONS

To illustrate the application of the maximum entropy principle to search system

design we consider first the case of a 'weighted request' system in which each document is indexed with a set of unweighted descriptors and each request consists of a set of weighted descriptors drawn from the same indexing/requesting vocabulary. The numeric weight associated with a given term in the request is, we shall suppose, to be interpreted as the user's subjective estimate of the *precision* of that term — that is, of the proportion of documents bearing the term that would be found useful. Thus if a term A were to appear in a request with weight 0.3, that value would be interpreted as a guess by the requestor that around 30% of the documents indexed by A would be relevant to his need (more formally, that $P(U/A)=0.3$). Various interactive aids might be used to help users make such estimates more accurately, but these will not be discussed here.

Now suppose a request of form 'A:0.3, B:0.3' is received by the system, signifying that the user guesses the term precisions of both term A and term B to be around 30%. Assume that it is known from the indexing statistics that A and B each index 10% of the collection, and that 1% of it is indexed jointly by both of them. Suppose too that it is known or estimated that the proportion of useful documents in the collection does not exceed 10%. The probability distribution of maximum entropy satisfying all these constraints was determined earlier and is shown in Figure 2. From that figure it is readily computed that $P(U/A,B)=0.67$, $P(U/A,\bar{B})=P(U/\bar{A},B)=0.26$, and $P(U/\bar{A},\bar{B})=0.06$. The probability ranking of the collection produced in response to this request therefore places all documents bearing both A and B in the top rank with an estimated probability of usefulness of 0.67, all documents with only one of the two terms next with a probability of 0.26, and the remainder of the collection at the bottom of the ranking with a usefulness probability of only 0.06.

The treatment of requests containing more than two terms is an obvious extension of this example. As many term co-occurrence data as are conveniently available can be used in such a maximum entropy computation. For instance, if indexing statistics are readily obtainable on the frequency of posting of any given request term, and also on the joint posting frequency of any pair of request terms, but to supply the joint posting frequency for any triple (or quadruple, etc.) of terms is deemed too cumbersome or not worthwhile, then the maximum entropy principle can be applied using as constraints only the data pertaining to single terms and term pairs. In this way accuracy of estimate can, to whatever extent may be desired, be traded off for computational simplicity.

The request weights need not necessarily be interpreted as user's subjective estimates of term precisions as in the example just discussed. Possible alternative interpretations of request weights that might be worth considering include the user's subjective estimate of the probability increase factor $P(U/A)/P(U)$, and the estimated term recall $P(A/U)$. Indeed, any probabilistically interpretable quantity which in the presence of other available input information determines the value of the joint probability $P(A,U)$ will do. Estimates of these and other relevant input quantities need not necessarily be subjective estimates by humans; they might instead be statistical estimates based on past experience in other systems, experience with past requests in the present system, or user feedback about preliminary output offered in response to the current request.

In systems receiving sets of unweighted terms as requests, it might be effective to employ a maximum entropy computation as though some standard weight with an associated standard interpretation were attached to all of the request terms, thus treating the unweighted requests as degenerate forms of weighted requests in which

the weights are all assumed to be alike. Under this arrangement the benefit of the user's judgement about the magnitudes of the probabilities in question would be lost, but the indexing statistics about term breadths and co-occurrences in the collection would still be exploited to improve the ranking over what would be obtainable from, say, a simple coordination-level retrieval rule using the same unweighted requests. It has been suggested that some of the advantages of a Boolean request language may be obtainable in this way without imposing the complexities of Boolean logic on casual users (Cooper, 1981).

Other sorts of maximum entropy systems might deal in very different kinds of probabilistic clues. Specifically, corresponding to the foregoing schemes involving probabilistically interpreted request weights, there are entirely analogous schemes of probabilistically interpreted document index term weights and ways of exploiting them via maximum entropy computations. It is even possible to construct a 'unified' probabilistic theory of retrieval involving both requestors' probability estimates for events involving document properties and document indexers' probability estimates for events involving information need properties, all such clues being combined into probability-of-usefulness estimates via a maximum entropy calculation. A unified theory of this sort has been proposed elsewhere by Robertson *et al.* (1982). The essential point here is that the maximum entropy principle will always be capable of serving as the probability estimation technique so long as the available data can be interpreted as a series of constraints on a probability distribution over some event space involving the event U and a family of other events associated with document and/or information need properties.

6. COMPUTATIONAL CONSIDERATIONS

It should be clear by now that the maximum entropy principle is flexible enough to serve, in theory at least, as the inductive engine for any of a large class of probabilistic search systems. There remains, however, the question of computational feasibility.

Computational methods for calculating maximum entropy distributions are well known (see, for example, Gokhale and Kullback, 1978). A number of them have been implemented computationally, and a few such programs are even available commercially. The methods in present use are iterative and fall into two categories. The first are of the iterative scaling type. Such algorithms proceed from an initial distribution, which need not satisfy any of the given constraints, and perform a series of iterations each of which adjusts the distribution so that it satisfies one of the constraints. The adjustments of one iteration may undo the work of another, but the algorithm eventually converges to the distribution of maximum entropy. The Deming-Stephan algorithm (Ireland and Kullback, 1968) is of this type.

The other kind of iterative solution exploits the techniques of numerical analysis. One such algorithm is the Newton-Raphson iteration procedure, which searches for values of certain parameters (Lagrange multipliers) from which the distribution of maximum entropy can be calculated. Alhassid *et al.* (1978) describe a program based on such an algorithm, and offer to make a listing of their program together with flow charts and a user's guide available for free to interested researchers. The 'geometric programming' approach (Beightler and Phillips, 1976; Avriel, 1980) suggests similar algorithms.

The iterative techniques generally converge quickly when the event space and

number of constraints are small, but computation time goes up rapidly as these grow larger. Experience with these iterative programs suggests that searches involving the manipulation of only a few descriptors (less than five, say) could probably be processed within a second or so with reasonable accuracy, assuming careful programming for a fast computer. Somewhat more complex searches could presumably be handled quickly if not all of the collection statistics were used (if, say, only single attribute probabilities and pair-wise joint probabilities were exploited). But for searches involving, say, ten or more information need or document properties, search times measured in minutes rather than seconds are to be expected. Thus if we assume that users of future on-line systems will be willing to wait no more than a few seconds for their output, the standard iterative approach seems viable for searches involving few clues but not for searches involving many.

Searches involving many clues call for measures which drastically reduce computation time while still retaining an acceptable degree of accuracy. Various tricks might be contemplated. Here we shall content ourselves with a brief discussion of the uses of the especially versatile concept of a *product approximation* (Lewis, 1959). A product approximation is a way of approximating higher order joint probabilities rapidly by taking products of lower order probabilities. For example, the formula

$$P(U, A_1, A_2, A_3, A_4) \approx P(A_1, U)P(A_2/U)P(A_3/U)P(A_4/U) \tag{4}$$

is a product approximation of a 5-event joint probability in terms of pair-wise joint and conditional probabilities. By replacing events with their complements, the formula can be made to yield a probability distribution over the entire 32-element event space generated by U, A_1, A_2, A_3, and A_4. Similarly

$$\begin{aligned} P(U, A_1, \ldots, A_{12}) \approx\ & P(A_1, \ldots, A_4, U)P(A_5, \ldots, A_8/U) \\ & P(A_9, \ldots, A_{12}/U) \end{aligned} \tag{5}$$

is one of the product approximations of a 13-ary joint probability in terms of 5-ary joint probabilities. The formula

$$\begin{aligned} P(U, A_1, \ldots, A_{12}) \approx\ & P(A_1, A_2, U)P(A_3/A_2, U)P(A_4/A_3, U) \\ & \ldots P(A_{12}/A_{11}, U) \end{aligned} \tag{6}$$

is one of the ways of approximating a 13-ary joint probability in terms of 3-ary joint probability data; others may be obtained from it by lowering the subscript on any event to the right of a slash.

Lewis proved that a product approximation formula yields a distribution that has maximal entropy relative to the probability data appearing in it. In other words, the only kind of error introduced by approximation formulae such as Eqs. (4)–(6) is the kind produced by not using all of the available collection statistics. Lewis also provided a remarkable way of telling, without knowledge of the true maximum entropy distribution, which product approximation formula yields the best approximation to the true distribution in a plausible information-theoretic sense of 'best'. Roughly speaking, the best product approximation is the one in which there is the greatest tendency for strongly dependent events to be linked together within individual probability expressions that are factors in the product. Using Lewis' method one can custom-design a product approximation formula to a request in

such a way as to take into account just the dependencies that matter the most among the request clues.

In circumstances where product approximations are too crude to be relied upon, their use can profitably be combined with other approaches. One way is to use product approximations as starting values to speed up a standard iterative process. To illustrate, Eq. (4) could be used to obtain, using term-precision estimates obtained from a four-element request, an initial distribution which would then be refined by an iterative process; the result would be an accurate estimate of $P(U, A_1, A_2, A_3, A_4)$ and hence of $P(U/A_1, A_2, A_3, A_4)$. Another way of exploiting product approximations is to use them as a way of breaking down the processing of long requests into shorter chunks. Suppose for instance that the standard iterative procedure would be out of the question for 12-clue searches but is rapid for 4-clue searches. Then a 12-clue computation could be broken down into three 4-element parts, each part handled iteratively in the standard way, and the results recombined using Eq. (5). If some attempt were made to choose the 4-element sets in such a way as to avoid putting strongly dependent clues in different sets, the resulting estimates should be accurate enough for practical purposes.

Carrying the idea a step further, a 12-element search could if necessary be processed using only 3-ary joint probability data by applying formula (6) or a variant thereof. Use of (6) is especially attractive computationally because there is a rapid non-iterative solution to the problem of determining from weighted request or index term data the maximum entropy distribution over an event space generated by just three events. The details are provided in the Appendix. Under this scheme the standard iterative process would be eliminated altogether, yet thanks to Lewis' method of selecting the best approximation the most important 3-way interactions could still be taken into account.

Evidently then the problem of searches involving many clues can be met by elaborating on the approach of choosing a product approximation formula which captures as many as possible of the stronger dependencies — that is to say, which takes into account as high a degree of clue interaction as there is time to deal with. Alternatives to the product approximation approach are possible, and of course such routine programming techniques as screening and preprocessing can be brought to bear in various ways. We conclude that the prospect of basing a practical search algorithm on maximum entropy computations is far from hopeless, and that with a little ingenuity rough but tolerable accuracy of estimation at on-line speeds is probably attainable.

7. CONCLUDING REMARKS

The maximum entropy principle is still an object of controversy, presenting philosophical difficulties to statisticians of the traditional frequentist school. Some fine points could also be raised about the particular way in which the principle has been applied here, which is admittedly crude in some respects. On the other hand, many of the objections commonly raised against the principle and its applications stem mainly from ignorance and dissolve when it is understood in greater depth. Moreover, a substantial body of successful experience gained in applying the method in such diverse fields as physics, geology, and quality control now supports it. It would appear to provide a sound and practical answer to certain heretofore severe theoretical obstacles to rational search system design, especially the problem

of underdetermination and the challenge of combining miscellaneous probabilistic clues into rational probability-of-usefulness estimates.

APPENDIX
A Maximum Entropy Formula for Combining the Evidence of Two Retrieval Clues

In the special case of an event space generated by only three properties there is an exact non-iterative method for determining the probability distribution of maximum entropy provided enough data are available to leave only one degree of freedom (Good, 1963). For example, for search systems accepting precision-weighted requests this means that for any 2-term request involving terms A and B, there is a closed formula for $P(U/A,B)$ using input estimates for $P(U)$, $P(A)$, $P(B)$, $P(A,B)$, $P(U/A)$, and $P(U/B)$. Notice that with the help of the identity $P(U,A) = P(U/A)\,P(A)$ and the similar identity for B, this input data can be transformed into the simpler set of values $P(U)$, $P(A)$, $P(B)$, $P(A,B)$, $P(U,A)$, and $P(U,B)$. We wish to find $P(U,A,B)$, from which $P(U/A,B)$ will be immediately obtainable by dividing by $P(A,B)$.

Adopting subscripts expressed in the binary number system, let $p_{000} = P(\bar{U},\bar{A},\bar{B})$, $p_{001} = P(\bar{U},\bar{A},B)$, $\cdots$, $p_{111} = P(U,A,B)$. It is required to find values for these probabilities for which $-\sum_{i=000}^{111} p_i \log p_i$ is maximal and the input constraints are satisfied. Let $x = p_{111}$. Then all the probabilities of interest can be expressed in terms of x and the known input data as follows:

$$p_{111} = x \tag{7a}$$
$$p_{110} = P(U,A) - x \tag{7b}$$
$$p_{101} = P(U,B) - x \tag{7c}$$
$$p_{100} = P(U) - P(U,A) - P(U,B) + x \tag{7d}$$
$$p_{011} = P(A,B) - x \tag{7e}$$
$$p_{010} = P(A) - P(U,A) - P(A,B) + x \tag{7f}$$
$$p_{001} = P(B) - P(U,B) - P(A,B) + x \tag{7g}$$
$$p_{000} = 1 - P(U) - P(A) - P(B) + P(U,A) + P(U,B) + P(A,B) - x \tag{7h}$$

Setting the first derivative of $-\sum_{i=000}^{111} p_i \log p_i$ with respect to x equal to 0 and simplifying, one eventually obtains a cubic equation of the form

$$x^3 + a_2x^2 + a_1x + a_0 = 0. \tag{8}$$

Letting q_{jkl} be the probability expression to which x is added or from which it is subtracted to get p_{jkl} in Eqs. (7a-h), the constants in Eq. (8) can be written

$$a_2 = q_{010}q_{001} + q_{100}q_{001} + q_{100}q_{010} - q_{011}q_{000} - q_{101}q_{000} - q_{101}q_{011} - q_{110}q_{000} - q_{110}q_{011} - q_{110}q_{101}$$
$$a_1 = q_{100}q_{010}q_{001} + q_{101}q_{011}q_{000} + q_{110}q_{011}q_{000} + q_{110}q_{101}q_{000} + q_{110}q_{101}q_{011}$$
$$a_0 = q_{110}q_{101}q_{011}q_{000}$$

Equation (8) is readily soluble for x by standard analytic methods. Computer and pocket calculator programs for extracting the roots of cubic equations are widely available. The appropriate (positive, real) root of Eq. (8) is the desired maximum entropy value for $p_{111} = P(U,A,B)$. The values of other probability expressions, in particular conditional probabilities of form $P(A/U,B)$ for use in Eq. (6), are obtainable from this value with the help of appropriate members of Eqs. (7a–h).

Some examples of maximum entropy estimates are shown in Table 1. The examples may be

Table 1. Examples of maximum entropy estimates

Prior probability	*Indexing statistics*			*Request weights*		*Ranking coefficient*
$P(U)$	$P(A)$	$P(B)$	$P(A,B)$	$P(U/A)$	$P(U/B)$	$P(U/A,B)$
0.1	0.1	0.1	0.01	0.3	0.3	0.67
0.1	0.1	0.1	0.09	0.3	0.3	0.32
0.1	0.1	0.1	0.001	0.3	0.3	0.77
0.1	0.1	0.2	0.02	0.3	0.3	0.71
0.1	0.1	0.1	0.01	0.3	0.05	0.17
0.001	0.02	0.005	0.0006	0.025	0.01	0.07

interpreted as the probabilities that would be assigned in response to a 2-clue precision-weighted request to documents possessing both clue properties. The first two rows correspond to two situations described earlier as illustrations of consequences [3] and [4] of the maximum entropy principle.

ACKNOWLEDGEMENTS

We are indebted to M. Maron and S. Robertson for stimulating discussions of the maximum entropy principle and for comments on earlier drafts of this paper. We would also like to thank Professor Gokhale of the University of California, Riverside, and Dr. J. C. Keegel of Sokuke Associates, Inc., Washington, D.C. for their valuable advice on computational aspects of the maximum entropy approach. The research was supported under NSF Grant No. IST-7917566.

REFERENCES

Alhassid, Y., Agmon, G. and Levine, R. D. (1978) An upper bound for the entropy and its applications to the maximal entropy problem. *Chemical Physics Letters 53*, 22–26.

Avriel, M. (Ed.) (1980) *Advances in Geometric Programming.* New York: Plenum Press.

Beightler, C. S. and Phillips, D. T. (1976) *Applied Geometric Programming.* New York: John Wiley.

Bookstein, A. and Swanson, D. R. (1974) Probabilistic models for automatic indexing. *Journal of the American Society for Information Science 25*, 312–319.

Bookstein, A. and Kraft, D. (1977) Operations research applied to document indexing and retrieval decisions. *Journal of the Association for Computing Machinery 24*, 418–427.

Cooper, W. S. (1976) *The suboptimality of retrieval rankings based on probability of usefulness.* Berkeley, California: School of Library and Information Studies, University of California (Technical Report).

Cooper, W. S. (1981) Exploiting the Maximum Entropy Principle to achieve probabilistic information retrieval. *Journal of the American Society for Information Science* (in press).

Cooper, W. S. and Maron, M. E. (1978) Foundations of probabilistic and utility-theoretic indexing. *Journal of the Association for Computing Machinery 25*, 67–80.

Gokhale, D. V. and Kullback, S. (1978) *The Information in Contingency Tables.* New York: Marcel Dekker Inc.

Good, I. J. (1963) Maximum entropy for hypothesis formulation. *Annals of Mathematical Statistics 34*, 911–934.

Harper, D. J. and van Rijsbergen, C. J. (1978) An evaluation of feedback in document retrieval using cooccurrence data. *Journal of Documentation 34*, 189–216.

Harter, S. P. (1975) A probabilistic approach to automatic keyword indexing (Parts I and II). *Journal of the American Society for Information Science 26*, 197–206, 280–289.

Ireland, C. T. and Kullback, S. (1968) Contingency tables with given marginals. *Biometrika 55*, 179–188.

Jaynes, E. T. (1979) Where do we stand on maximum entropy? In *The Maximum Entropy Formalism.* (R. D. Levine and M. Tribus, eds.) Cambridge, Mass: M.I.T. Press.

Lewis, P. M. (1959) Approximating probability distributions to reduce storage requirements. *Information and Control 2*, 214–225.

Robertson, S. E. (1977a) Theories and models in information retrieval. *Journal of Documentation 33*, 126–148.

Robertson, S. E. (1977b) The probability ranking principle in IR. *Journal of Documentation 33*, 294–304.

Robertson, S. E., Maron, M. E. and Cooper, W. S. (1982) Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development 1*, 1–21.

Robertson, S. E. and Sparck Jones, K. (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129–146.

Salton, G., Wong, A. and Yu, C. T. (1976) Automatic indexing using term discrimination and term precision measurement. *Information Processing and Management 12*, 43–51.

Shannon, C. E. (1948) The mathematical theory of communication. In *The Mathematical Theory of Communication.* (C. E. Shannon and W. Weaver, eds.) Urbana: University of Illinois Press, 1962.

Tague, J. M. (1973) A Bayesian approach to interactive retrieval. *Information Storage and Retrieval 9*, 129–142.

Tribus, M. (1969) *Rational Descriptions, Decisions, and Designs.* Oxford: Pergamon Press.

Van Rijsbergen, C. J. (1977) A theoretical basis for the use of cooccurrence data in information retrieval. *Journal of Documentation 33*, 106–119.

Van Rijsbergen, C. J., Robertson, S. E. and Porter, M. F. (1980) *New Models in Probabilistic Information Retrieval.* Cambridge: Computer Laboratory, University of Cambridge (British Library R & D Report No. 5587).