THE MAXIMUM ENTROPY PRINCIPLE IN INFORMATION RETRIEVAL*

Paul B. Kantor & Jung Jin Lee
Tantalus, Inc., Cleveland, Ohio

ABSTRACT

Applications, assumptions and properties of the maximum entropy principle are discussed. The maximum entropy principle integrates prior estimates of relevance with the observed distribution of term combinations. The result may be a reordering of the segments of a database, compared to a naive estimate. Numerical examples obtained by solution of the non-linear equations for the dual variables are presented and discussed.

* Supported in part by the National Science Foundation under grant IST-8318630.

BACKGROUND

It is hardly necessary to state the importance of information retrieval. Whether for research, business or military purposes, the speedy and accurate retrieval of relevant information can mean the difference between success and failure. Information retrieval (IR) as a discipline has been primarily developed for the storage and retrieval of text materials. In the early days, it was easy to scoff at the emphasis given by research to the specific textual content of a document. [CALV81] However, with the onslaught of digital electronic storage of documents, it becomes clear that this emphasis was prophetic. All of the familiar clues -- size, shape, location within an office or a file cabinet and so forth are being taken away from us. Our ability to retrieve information quickly and accurately hinges on our ability to describe the items to be retrieved in terms of their textual content (and perhaps a few dozen additional descriptors).

The general outlines of term description of documents are given in Salton's monograph [Salt83]. Other kinds of application are becoming increasingly important. An example of non text information retrieval is provided by a network of weather observation stations. Each station compiles an extensive file of data on wind direction and speed, barometric pressure, precipitation and other meteorological data. It reports, to a central coordinating agency or computer, only a few summary statistics. These may be thought of as abstracts, or "key terms" describing the full file of information available at the weather station.

If the problem of weather prediction is to be handled in real time, it appears essential for the central agency or computer to be able to determine, from those abstracts, which detailed reports are essential to resolve ambiguities in short and medium range forecasting. The problem of coordinating the enormous quantity of available data is not likely to be solved by building a computer that can digest all of it at once. It is much more likely to be solved by a disaggregated scheme in which the central computer knows how to process abstracts, and has effective rules for determining when to request the "full document" from a reporting station.

The same hierarchical structure is present, with much more stringent real time constraints, in the design of a national defense system. Such a system, as presently conceived, would integrate vast quantities of data from observation stations, to estimate whether or not a threat is present, and if so to take appropriate responsive action. Again, it seems unlikely that this problem will be solved by simultaneous processing of all the available data. A much more likely solution path is the development of one or several centralized units, each capable of evaluating abstracts in real time and requesting appropriate further information.

What may also mention an intriguing relation to the field of expert systems (ES), and its more all-encompassing cousin, artificial intelligence (AI). It is widely recognized that both ES and AI have much to contribute to the area of document retrieval (GUID83). Expert systems can handle the complexity of dealing with a variety of databases and retrieval languages, once a query is cast into some standard "search language". Artificial intelligence, in its turn, can work to process natural language input, to reduce an unsophisticated expression of information need to a suitable standard form. What is less widely recognized is that information retrieval itself has much to contribute to both ES and AI.

Information retrieval is important in ES and AI because any nontrivial expert system, or intelligent program, must process new information or queries in terms of a substantial base of facts and rules. For all but the simplest queries, repeated scanning of the base of facts and rules can consume an enormous amount of time. If, on the other hand, the central processor is able to retrieve those facts and rules must likely to contribute to resolution of the query, the whole process will be enormously speeded. An initial acceleration can be achieved simply by a thorough inverted indexing of the base of facts and rules. However, the term indexing, by the logical equivalent of "document descriptors" is likely to offer substantial improvements.

In the present report, no specific application is made of the techniques discussed, but their general properties and their relation to some previous schemes is explored.

THE MAXIMUM ENTROPY PRINCIPLE

COOPER [COOP83] and COOPER and HUIZINGA [COOP82] have proposed that the maximum entropy principle can serve as a powerful tool to improve retrieval of information. The maximum entropy principle was proposed, as a means for unifying a number of physical problems, by JAYNES [JAYN57]. Independently, GOOD [GOOD63] develop maximum entropy as a technique for generalizing the notion of a "null hypothesis" from two dimensional contingency tables, to higher dimensional problems. An extensive discussion on the justification of maximum entropy methods in engineering has been given by JAYNES[JAYN82]. A technical discussion of the formulation of the maximum entropy problem for information retrieval, has been given by KANTOR [KANT84]. The present paper builds upon that reference, to which the reader is referred for more technical details.

The application of the maximum entropy principle to information retrieval has a good deal in common with Good's formulation. The entire set of "documents" (we shall use the term "documents" to refer in general to any retrievable report or item that might contain further information relevant to the problem at hand.) consist of items to which any of a number of descriptors may apply. If there are, say, 24 descriptors, one might try to represent this as a 24 dimensional contingency array. The number in each cell of this array would represent the number of documents having a particular combination of descriptions. We want to estimate the "relevance" of the documents in a given cell. This could be expressed by adding one more dimension to the contingency table. In order to express our results on paper, we resort to a two dimensional contingency table in which the rows represent all the cells of the very high dimensional table, and the columns represent the possible degrees of relevance. We call this the Grand Contingency Table.

In the simpliest case, the descriptor terms are given an order, fixed once and for all. The rows can then be labeled by a string of 0's and 1's indicating the absence or appearance of the term in the document. The rows are easily arranged lexicographically. In a more general formulation, the elements of the document vector could be integers greater than one, representing for example the number of times that a term appears in a document. The principle of formulating a two dimensional contingency table still applies. Descriptions could be even more general, such as "appears related to a document with description D". Such descriptions are likely to occur in the relevancy feedback models described by Salton. Although there might be some difficulty in defining a unique order for the rows of the table, this generalization also does not conflict with the essential two dimensionality of the problem.

The columns, as stated, correspond to varying possible degrees of relevance. In the remainder of this paper we shall make the classical assumption that the degree of relevance is either 0 or 1. However, we shall speak of the "expected relevance" of a document drawn from the set described by a particular row. This "expected relevance" is simply the number of relevant documents in that set divided by the total number of documents in that set. The notion of expected relevance can be easily generalized to a case where relevance assumes several discrete values, and even to the case where it may be continuously distributed between 0 and 1.

Stated in its simplest terms, the maximum entropy principle makes the following assertions:

(1) The distribution of documents into the rows of this table is given by the database. It is a fact, not open to speculation, and this distribution need not obey any specific assumptions.

(2) Some prior information on the expected distribution of relevant documents in this large table is available to us. In the most natural case, which will be used throughout this discussion, this information is given in terms of the probability that an item characterized by an individual descriptor (D_j) will be relevant. (It is common in retrieval systems to form the Boolean union of several terms expected to be nearly synonymous. This does not provide an essential complication as, in what follows, the word "descriptor" can also be interpreted to mean "union of closely related descriptors.")

(3) Finally, the relevant and non-relevant documents are assumed to be distributed, within this grand contingency table, as randomly as possible subject to the constraints of (1) and (2).

A constraint of the type (2) tells something not about a single row of the grand contingency table, but about half of the rows together -- that half to which the description (D_j) applies. Of course, the goal of an efficient retrieval system is to estimate which of the specific rows have the highest expected relevance.

As JAYNES and GOOD and KHINCHIN [KHIN67] have shown, the only defensible definition of the phrase "as randomly as possible" is "the distribution having the largest entropy."

The entropy function is defined up to a multiplicative constant by equation 1.

Eq. 1 $S(p_1,p_2,\ldots p_{2.2^N}) = -\sum_i p_i \ln p_i$

The choice of the base of logarithms is arbitrary, and for optimization analysis, it is most convenient to use logarithms to the base
The resulting constrained optimization problem is to maximize the entropy function subject to a set of limited constraints of the form given in equation 2.

Eq. 2 $\sum_{R\epsilon R(I)} n(R)v(R) = V_I \sum_{R\epsilon R(I)} n(R)$

The terms in this equation are defined as follows:

R *labels rows.* R=(00000.., 10000..., ..., 111111...)
I *labels constraints* $I = 1, \ldots k$
$R(I)$ = R *such that constraint* I *applies to* R
V_I = *expected relevance as in assumption (3)*
$v(R)$ = *expected relevance of row* R

Eq. 3

It can be shown (Kantor 1984) that if this problem has a solution, the solution may be expressed in terms of a set of Lagrange multipliers $\lambda_1,..,\lambda_k$, where k is the number of constraints provided. It is further shown that the richness of any row is given by Equation 4.

Eq. 4. $v(R) = v(\sum_{I\epsilon C(R)} \lambda_I)$

In this equation the set $C(R)$ is defined as in equation 5, and the universal value function v is defined as in equation 6.

Eq. 5 $C(R) =$ {I *such that* I *constrains* R}

Eq. 6 $v(x) = e^x/(1+e^x)$

The particular form of the function is determined by our assumption that relevance assumes only the values 0 and 1. For a more general discussions see [KANT84].

When Equation 4 is inserted into Equation 2, a set of coupled equations results. The number of these equations is equal to the number of independent constraints that have been provided. The equations are given in Equation 7

Eq.7 $\sum_{R\epsilon R(I)} n(R)v(\sum_{J\epsilon C(R)} \lambda_J) = V_I \sum_{R\epsilon R(I)} n(R)$

These equations are non-linear, and do not always have a real solution. When they do have a solution, the values of λ_J are determined by the constraint values V_I, AND the actual distribution of items among the rows of the grand contingency table {$n(R)$} The additional dependence on the actual distribution can produce surprising and significant results. In the remainder of this paper, to keep the discussion within bounds, we will consider only the case of two descriptive terms.

RELATIONSHIP TO TERM INDEPENDENCE

Suppose we know a partial table giving the distribution of relevant and non-relevant items in a very small data base. The numbers are as shown in Table 1.

Table 1. A hypothetical distribution

Row(R)	n(R)	not-Rel	Rel	z(R)
00	84	72	12	12:72
10	46	18	28	28:18
01	26	8	18	18:8
11				

In each row we show the total number of documents having the corresponding description, the number that are not relevant, the number that are relevant, and the odds ratio for relevance as apposed to not-relevance. For example, for items described by the first descriptor alone the odds ratio is 28 to 18. With reference to Equation 4 we see that the logarithm of the odds ratio is related to the Lagrange multipliers as shown in Equation 8.

Eq. 8 $\ln z(R) = \sum_{I\epsilon C(R)} \lambda_I$

The last row of this table has been left empty so that we may easily contrast the concept of binary term independence with two formulations of the Maximum Entropy principle

The principle of binary independence requires that the last row be as shown in Table 2.

Table 2. The completion under Binary Independence

Row (R)	n(R)	not-Rel	Rel	z(R)
11	44	2	42	42:2

The odds ratio for the appearence of both terms is the "natural" combination of the odds ratios for the three preceding terms as shown in Equation 9.

Eq. 9 $z(11) = z(10)z(01)/z(00)$

This equation states that addition of either term enhances the other in the same way that presence of the added term is an enhancement compared to the absence of either term. This result is plausible, and deserves to be preserved.

On the other hand, the binary independence model has also insisted that there be exactly 44 elements in the row 11. This insistence may or may not be a true description of the data base, and violates principle (1) of the maximum entropy formulation. It is not hard to see that in the extention to three or more descriptors the degree to which the problem is overconstrained increases rapidly.

TWO FORMULATIONS OF THE MAXIMUM ENTROPY PRINCIPLE

The maximum entropy principle may be applied to this problem in two different ways. In one way no prior assumption is made about the fraction of items (in the entire data base) that are relevant to the problem at hand. The resulting final row is shown in Table 3.

Table 3. Completion By the Minimal MEP

Row (R)	n(R)	not-Rel	Rel	z(R)
11	---	---	---	28:8

The odds ratio is substantially lower than in the binary independence model, but on the other hand no constraint is placed on the total number of items in this row.

A second formulation (which is the form first given by Cooper) supposes that there is some prior estimate of the probability that any item in the data base will be relevant. In this case the final row is as shown in Table 4.

Table 4. Single constraint added to the MEP

Row (R)	n(R)	not-Rel	Rel	z(R)
11	---	---	---	42:2

Note that the number of elements in the row again remains unconstrained, but the attractive rule for combination of odds ratios is preserved from the binary independence model. Thus, the transition to the maximum entropy principle, in this case loses nothing, except the unreasonable constraint that the total number of items described by both terms together be 44.

It should be noted that recent work by SARACEVIC and KANTOR (1986) (unpublished) provides a data base of some 6,000 retrieved items of known relevance for a variety of realistic research problems. Thus it is possible to test the odds ratios predicted by either of the maximum entropy models, and the distribution prediction of the binary independence model, against a substantial set of real data.

HOW DOES ALL THIS COME ABOUT?

The expression given as Equation 7 may be specialized to the two cases described above. In the case where there is no overall constraint on the items in the data base, there are two unknown multipliers and two coupled equations (Equation 10).

Eq. 10a $$n(10)v(\lambda_1)+n(11)v(\lambda_1+\lambda_2) = V_1[n(10)+n(11)]$$

Eq. 10b $$n(01)v(\lambda_2)+n(11)v(\lambda_1+\lambda_2) = V_2[n(01)+n(11)]$$

When there is also a constraint on the expected relevance of a document chosen entirely at random, there is one additional Lagrange multiplier λ_0, and correspondingly an additional equation.

Eq. 11a $$n(10)v(\lambda_0+\lambda_1)+n(11)v(\lambda_0+\lambda_1+\lambda_2) = V_1[n(10)+n(11)]$$

Eq. 11b $$n(01)v(\lambda_0+\lambda_2)+n(11)v(\lambda_0+\lambda_1+\lambda_2) = V_2[n(01)+n(11)]$$

Eq. 11c $$n(00)v(\lambda_0)+n(10)v(\lambda_0+\lambda_1)+n(01)v(\lambda_0+\lambda_2)+ n(11)v(\lambda_0+\lambda_1+\lambda_2)=V_0[n(00)+n(10)+n(01)+n(11)]$$

It is important to realize that the prior estimates V_0, V_1, V_2 represent "naive" estimates. The computed values of the Lagrange multipliers $\lambda_1, \lambda_2, \lambda_0$ are what determine the odds of relevance. The transformation between them is not simple. It depends in an essential way on the actual distribution of terms among items (or in a language more natural to our Grand Contingency Table, the distribution of items among term combinations.) This dependence is not surprising, since one key motivation for the maximum entropy principle is to use the information provided by the given distribution of terms.

The equations need not always have solutions. If the row sizes are as given in Equation 12

Eq. 12 $$n(10) = 10$$ $$n(01) = 10$$ $$n(11) = 80$$

then the initial conditions V_1=90%; V_2=10% are impossible. The first of these conditions requires that 71 of the 80 in n(11) be relevant. The second condition requires that 71 of those same items be not relevant.

Any neutral point at which all of the naive estimates are equal can always be solved. We use the neutral point with each estimate equal to 0.5 (the "vapid point") as the starting point in more difficult problems with higher dimensionality.

An example of the impact of the $\{n(R)\}$ in changing the rank order of the Lagrange multipliers is given in Table 5

We have not found an intuitive characterization of the permutations that occur. At this point it appears that any reordering of the naive estimates of relevance is possible in a suitably distributed data base.

SIMULTANEITY VS PATHWISE ANALYSIS

The shortcomings of term independence have been known for some time. [CHEES83], [Yu83]). The problem can be attacked by a pathwise approach to term combinations. These approaches can become quite complicated. They bear some similarity to the stepwise regression techniques used in linear modeling. Extending the same analogy, the maximum entropy approach is like multiple regression in that it deals with all of the information simultaneously. It would seem, whenever it can be implemented, that this principle is to be prefered.

Using available state of the art optimization programs (not to maximize the entropy, but to solve the coupled nonlinear equations) we have been able to handle solveable problems with 5 independent constraints in times of the order of 10 seconds, running on an IBM-PC compatible micro computer with an 8087 math chip, at a low clock speed.

TWO INTERPRETATIONS OF "FUZZINESS" OR "WEIGHT"

In Equation 11 we may introduce "weight" parameters in two different ways. Both of these arise from reinterpreting the incidence matrix defining the relationship between constraints and rows. This incidence matrix is shown as Table 6.

Table G: The incidence matrix

	1	2	0
00			✓
10	✓		✓
01		✓	✓
11	✓	✓	✓

A weight can be introduced into this matrix by interpretting each check mark as a 1 and changing one or more of the check marks to numbers less than 1. This is shown in Table 7.

Table 7: Weighted incidence

	1	2	0
00			1
10	g		1
01		1	1
11	1	1	1

In words, this transformation may be expressed as "we are not so sure about the relevance of term 1 when it occurs alone without term 2". The modifications effect only Equation 11a, which then takes the form shown in Equation 13.

Eq. 13. $$gn(10)v(\lambda_0+\lambda_1)+n(11)v(\lambda_0+\lambda_1+\lambda_2) = V_1\left[gn(10)+n(11)\right]$$

Another possible interpretation, which appears not to be equivalent to the first is shown in Equation 14.

Eq. 14. $$n(10)v(\lambda_0+g\lambda_1)+n(11)v(\lambda_0+\lambda_1+\lambda 2) = V_1\left[n(10)+n(11)\right]$$

Conceivably, two independent weight parameters g,h could be introduced simultaneously as in Equation 15.

Eq. 15. $$gn(10)v(\lambda_0+h\lambda_1)+n(11)v(\lambda_0+\lambda_1+\lambda_2) = V_1\left[gn(10)+n(11)\right]$$

The implications of this concept remain to be explored.

GENERALIZATION TO CONTINUOUS VALUE.

If value is not assumed to take only the values 0 and 1 but can assume several values in the unit interval, the effect is to build up a conditional entropy maximization problem. This is equivalent to a minimum cross-entropy model, with a prior distribution corresponding to 1/n at each of n allowed points in the unit interval. In this language, our present model has a Bayesian prior distribution (0.5, 0.5).

An accumulation of points can be used to approach a continuous distribution. In another point of view, the value may be assumed to vary continuously in the unit interval, with a uniform prior distribution. This case has been considered by KANTOR [KANT84]. The structure is essentially the same, and leads to Equation 7. However the definition of the universal value function is changed, to the Langevin function, Equation 16. The Langevin function is known from classical statistical mechanics.

Eq. 16. $$v(\lambda) = 0.5 + 0.5\left[coth\ \lambda - 1/\lambda\right]$$

In this model, which is a more realistic way to consider relevance, the odds ratio will not be simply multiplicative. The full implications of this model have not yet been explored.

Table 5. Examples of the change in ordering of intersection sets, induced by changes in the composition of the data base. This example is for three distinct terms, labeled A,B,C. The entire data base is refered to as R. Thus V1=0.1; V2= 0.2; V3= 0.2 and V0=0.5 in the first example. In the last 8 columns the Boolean atoms are shown in order of increasing richness.

V(A) = 0.100 V(B) = 0.200 V(C) = 0.200 V(R) = 0.500

n(000)	n(C)	n(B)	n(BC)	n(A)	n(AC)	n(AB)	n(ABC)	OBJ F.							RANK
0.50	0.10	0.10	0.05	0.10	0.05	0.05	0.05	0.0000	ABC	AC	AB	BC	A	C	B 000
0.50	0.05	0.05	0.10	0.05	0.10	0.10	0.05	0.0000	ABC	AC	AB	A	BC	C	B 000
0.30	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.0000	ABC	AB	AC	BC	A	B	C 000
0.30	0.05	0.05	0.10	0.05	0.10	0.10	0.25	0.0000	ABC	AB	AC	BC	A	B	C 000
0.10	0.10	0.10	0.20	0.20	0.10	0.10	0.10	0.0090	ABC	AB	AC	BC	A	B	C 000
0.10	0.10	0.05	0.30	0.30	0.05	0.05	0.05	0.0145	ABC	AB	AC	BC	A	B	C 000

V(A) = 0.600 V(B) = 0.800 V(C) = 0.500 V(R) = 0.100

n(000)	n(C)	n(B)	n(BC)	n(A)	n(AC)	n(AB)	n(ABC)	OBJ F.							RANK
0.50	0.10	0.10	0.05	0.10	0.05	0.05	0.05	0.0079	000	C	A	B	AC	BC	AB ABC
0.50	0.05	0.05	0.10	0.05	0.10	0.10	0.05	0.0147	000	C	A	B	AC	BC	AB ABC
0.30	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.0310	000	C	A	AC	B	BC	AB ABC
0.30	0.05	0.05	0.10	0.05	0.10	0.10	0.25	0.0467	000	C	A	B	AC	BC	AB ABC
0.10	0.10	0.10	0.20	0.20	0.10	0.10	0.10	0.0617	000	C	A	AC	B	BC	AB ABC
0.10	0.10	0.05	0.30	0.30	0.05	0.05	0.05	0.0660	000	C	A	AC	B	BC	AB ABC

V(A) = 0.300 V(B) = 0.500 V(C) = 0.400 V(R) = 0.600

n(000)	n(C)	n(B)	n(BC)	n(A)	n(AC)	n(AB)	n(ABC)	OBJ F.							RANK
0.50	0.10	0.10	0.05	0.10	0.05	0.05	0.05	0.0000	ABC	AC	AB	A	BC	C	B 000
0.50	0.05	0.05	0.10	0.05	0.10	0.10	0.05	0.0000	AC	ABC	A	AB	C	BC	000 B
0.30	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.0000	ABC	AC	AB	A	BC	C	B 000
0.30	0.05	0.05	0.10	0.05	0.10	0.10	0.25	0.0000	AC	ABC	A	AB	C	BC	000 B
0.10	0.10	0.10	0.20	0.20	0.10	0.10	0.10	0.0008	ABC	AC	AB	BC	A	C	B 000
0.10	0.10	0.05	0.30	0.30	0.05	0.05	0.05	0.0044	ABC	AC	AB	A	BC	C	B 000

ACKNOWLEDGEMENTS

It is a pleasure to acknowledge stimulating discussions with Dr. Richard Blankenbecler, Dr. Frederick Kantor, and the participants at the 1983 and 1985 Working Conferences on the Maximum Entropy Principle presented by Professor Ray Smith, at the University of Wyoming. Support from the Office of Naval Research made it possible for one of the authors (PBK) to attend the 1985 conference.

REFERENCES AND LITERATURE CITED

[CALV81] CALVINO, Italo "Se una notte d'inverno un viaggiatore" (If on a winter's night a traveler), Harcourt Brace Jovanovich, p260, c1981.

[CHEE83] CHEESEMAN, Peter "A method of computing maximum-entropy probability values for expert system", Proceedings of the 3rd Annual Workshop on Bayesian Methods in Applied Statistics; C. Ray Smith (ed). Laramie, WY: University of Wyoming; (1983)

[COOP83] COOPER, W. S. "Exploiting the maximum entropy principle to increase retrieval effectiveness", J. Am. Soc. Inf. Sci., v34n1pp31-39 (1983).

[COOP82] COOPER, W. S. and HUIZINGA, P. "The maximum entropy principle and its application to the design of probabilistic retrieval systems", Inf. Technol: Res. & Dev., v1n2pp99-112 (1982).

[GOOD63] GOOD, I. J. "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables", Ann. Math. Stat. v34pp911-932 (1963).

[GUID83] GUIDA, Giovanni and TASSO, C. "An Expert Intermediary System for Interactive Document Retrieval, Automatica, v19n6p759-766, November 1983. IR-NLI.

[JAYN82] JAYNES, Edwin T. "On the rational of maximum entropy methods", Proceedings of the IEEE, v70n9, September, 1982.

[KANT84] KANTOR, Paul B. "Maximum Entropy and the Optimal Design of Automated Information Retrieval Systems." Information Technology; v3n2pp88-94 (1984).

[KHIN67] KHINCHIN, A. I. "Foundations of statistical mechanics", Dover, USA (1967)

[SALT83] SALTON, Gerard "Introduction to modern information retrieval", McGraw-Hill, 448p, c(1983).

[YU83] YU, C. T., BUCKLEY, C., LAM, K., SALTON, G., "A generalized term dependence model in information retrieval", Technical report TR 83-543, Department of Computer Science, Cornell University (1983).