Some Inconsistencies and Misnomers in Probabilistic Information Retrieval

William S. Cooper*

Abstract

The probabilistic theory of information retrieval involves the construction of mathematical models based on statistical assumptions of various sorts. One of the hazards inherent in this kind of theory construction is that the assumptions laid down may be inconsistent with the data to which they are applied. Another hazard is that the stated assumptions may not be the real assumptions on which the derived modelling equations or resulting experiments are actually based. Both kinds of error have been made repeatedly in research on probabilistic information retrieval. One consequence of these lapses is that the statistical character of certain probabilistic IR models, including the so-called 'binary independence' model, has been seriously misapprehended.

Introduction

Probability theory provides a powerful platform on which to construct theories of information retrieval and inductive searching. It is of course desirable that a formalism be logically powerful; however, such power comes at the price of a certain risk of accidental misuse and abuse. One of the hazards that an IR system designer should be aware of is that it is possible to become ensnared in statistical simplifying assumptions that are logically inconsistent. Another danger is that the fundamental assumptions underlying a theory might be incorrectly stated, and the merits of the theory misjudged for that reason. I should like to discuss both of these hazards – inconsistency and misidentification of underlying assumptions – in the light of the history of probabilistic IR.

*School of Library and Information Studies, University of California, Berkeley, California.

The Independence Assumptions

The fundamental modelling assumptions most commonly adopted in probabilistic retrieval theory have consisted of various combinations of the following three statistical independence assertions. The first is an assumption of absolute independence of document or information need properties; for the special case of just two properties A and B it is

$$I1. \quad P(A, B) = P(A)\,P(B).$$

The second and third assumptions are assertions of conditional independence, given relevance or its absence. Letting R denote the event of relevance, their formal statements are

$$I2. \quad P(A, B \mid R) = P(A \mid R)\,P(B \mid R);$$
$$I3. \quad P(A, B \mid \sim R) = P(A \mid \sim R)\,P(B \mid \sim R).$$

These three assumptions are interpreted differently in different contexts, depending upon whether the clues A and B are regarded as properties of documents or of users ('information needs'). Assumption $I2$ was first introduced in the pioneering paper by Maron & Kuhns (1960) as the basis of their proposed system of probabilistic indexing. They interpreted properties A and B as user properties. Assumptions $I2$ and $I3$ were used in combination by Yu and Salton (1976), and also by Robertson & Sparck Jones (1976) to develop what later became well known as the 'binary independence' IR model. In that application, A and B were regarded as document properties. All three assumptions – $I1$, $I2$ and $I3$ – were adopted by Robertson, Maron & Cooper (1982) as the assumptions underlying their 'unified' probabilistic IR model. There one of the clues was taken to be a document property and the other a user property.

The '$I\,1 + I\,2$' Inconsistency

When $I\,1$ and $I\,2$ are used together, logical inconsistencies can and do arise. The problem is not that $I\,1$ and $I\,2$ directly contradict each other. However, as a pair, they are jointly inconsistent with much of the empirical data to which one would want to apply them. They overconstrain the probability measure in such a way as to exclude the kinds of probability distributions that arise in the IR application.

To see the nature of the inconsistency, suppose the empirical data that is available concerning a certain pair of properties A and B indicates that

$$P(A) = P(B) = P(R) = 0.1$$
$$P(R\,|\,A) = P(R\,|\,B) = 0.5$$

These hypothetical data are perfectly consistent among themselves. However, there is no probability distribution in which $I\,1, I\,2$, and the data could all be true simultaneously. As a result, when one attempts to use $I\,1$ and $I\,2$ to draw inferences from such data, strange things can happen. For instance, probability values much larger than 1.0 are calculated for certain conditional events.

The logical inconsistency that lurks here can be demonstrated formally by deriving an absurdity. Assume $I\,1, I\,2$, and the foregoing empirical data. Then

$$\begin{aligned} P(A,\,B) &= P(A)\,P(B) && [\text{by } I\,1] \\ &= .1 \times .1 && [\text{from the data}] \\ &= .01 \end{aligned}$$

On the other hand

$$\begin{aligned} P(A,B,R) &= P(A,B|R)\,P(R) && [\text{identity}] \\ &= P(A|R)\,P(B|R)\,P(R) && [\text{by } I\,2] \\ &= P(R|A)\,P(A)\,P(R|B) \\ &\quad\; P(B)/P(R) && [\text{identity}] \\ &= .5 \times .1 \times .5 \times .1/.1 && [\text{from the data}] \\ &= .025 \end{aligned}$$

So we have $P(A,\,B,\,R) > P(A,\,B)$. But this contradicts the elementary laws of probability theory, for the probability of a conjunction of several events can never be made smaller – only larger – by removing one of the events from the conjunction. This anomaly was first noticed, and correctly classified as a logical inconsistency, by Robertson (1974).

The logical indiscretion that is at issue here – I shall call it the '$I\,1 + I\,2$ inconsistency' – has been committed by a number of investigators, including myself. An early instance of it occurs in the work of Miller (1971), as was pointed out by Robertson shortly thereafter (1974). A further example of the same type of inconsistency was cited by Robertson and Sparck Jones (1976). The properties A and B were in these instances document properties. The $I\,1 + I\,2$ inconsistency would appear to be present also in the interesting work reported recently by Fuhr and Buckley (1990). In that case the properties were user properties (query terms). Still another instance of the $I\,1 + I\,2$ inconsistency, with a mixture of document and user properties involved this time, appeared in the already-mentioned unified theory of Robertson, Maron, & Cooper (1982). These authors were aware of the inconsistency, pointed it out explicitly, and discussed an alternative development of their theory that would get around it. Nevertheless they let stand a modelling formula derived with the help of these inconsistent premises.

The '$I\,1 + I\,3$' Inconsistency

Just as $I\,1$ and $I\,2$ are inconsistent in the presence of typical data, so $I\,1$ and $I\,3$ can yield inconsistencies. Robertson, Maron, and Cooper (1982), who adopted all of $I\,1, I\,2$, and $I\,3$, thereby simultaneously committed not only the $I\,1 + I\,2$ inconsistency but also the $I\,1 + I\,3$ inconsistency. However, their double infraction was mitigated somewhat by the fact that they admitted to it, and discussed a way of removing the offending inconsistencies.

Effective Retrieval From Faulty Theory?

Interestingly enough, in most of these historic cases of inconsistent premises the faulty theory was used as the basis of experimental work that was on the whole successful. The early $I\,1 + I\,2$ inconsistencies noted by Robertson & Spark Jones occurred in the course of the design of probabilistic systems that produced acceptable retrieval results. The experimental methods of Fuhr and Buckley, though also derived from a theoretical basis contaminated by the $I\,1 + I\,2$ inconsistency, resulted in high levels of retrieval effectiveness. And the double inconsistency in the theory of Robertson, Maron, & Cooper did not prevent a later experimental system based on that theory from performing adequately (Maron, Curry, & Thompson 1986).

This apparent experimental success in the face of inconsistent premises is curious. A logically inconsistent theory is supposed to be worse than no theory at all. It is a well-known metatheorem of logic, provable for all standard systems of deductive inference, that an inconsistent theory implies any proposition whatever. Since one can derive all possible assertions from an inconsistent theory, such a theory must be meaningless –

entirely lacking in significance or predictive power. It makes no sense that good experimental results could come out of an inconsistent theory.

It is tempting to explain this conundrum by suggesting that the inconsistencies in question were only minor ones. However, there is no such thing as a theory that is just 'a little bit' inconsistent. A theory cannot be just a little bit inconsistent, any more than the scientist proposing it can be just a little bit pregnant. A logical inconsistency, if its implications are followed out, destroys a theory utterly. It is a disaster, and is totally unacceptable. If rationality is to be preserved, inconsistencies simply cannot be tolerated.

How then are we to explain the fact that the experimental systems worked, and worked well, even though their stated theoretical basis was inconsistent? I think the answer to this puzzle lies in the other kind of logical error mentioned earlier, namely, the misidentification of fundamental assumptions. Although the authors in question laid down assumptions that were logically inconsistent, their experiments were in fact founded upon somewhat different assumptions – upon unstated theories that were consistent after all. The investigators were saved from their self-contradictory postulates by the fact that they did not actually apply them as they stood, but instead did something else that made more logical sense. Their precepts were flawed, but they were astute enough not to follow out all the implications of these precepts in practice.

Repairing the Inconsistencies

In the case of at least some of the $I1 + I2$ inconsistencies, the explanation of the successful experimental results would seem to be that assumption $I1$ was not actually needed. The required probability rankings of the document collection can be achieved on the basis of $I2$ alone (plus of course the empirical data). A mathematical basis for probabilistic indexing using $I2$ alone was set forth by Maron & Kuhns (1960) when they showed how, by working with proportionalities instead of equalities, one could obtain from $I2$ all the needed probability comparisons. So we may view the later authors who used $I1 + I2$ as having added assumption $I1$ gratuitously and inconsistently, but in such a way that it affected only their formal theory and not their actual output rankings.

In the case of the $I1 + I2$ and $I1 + I3$ double-inconsistency occurring in the Robertson, Maron, & Cooper theory, the explanation of the successful followup experiment is different. Assumption $I1$ really is essential to their unified model, but neither $I2$ nor $I3$ is needed in full strength. On examining the details of their derivation (p. 14), it becomes clear that $I2$ and $I3$ can be replaced by the single assumption

$$I4. \quad \frac{P(A, B \mid R)}{P(A, B \mid \sim R)} = \frac{P(A \mid R)}{P(A \mid \sim R)} \frac{P(B \mid R)}{P(B \mid \sim R)}$$

Their modelling equation (Eq. (8), p. 13) is easily derived from this considerably weaker assumption. Since $I4$ is consistent with $I1$, using it in place of $I2$ and $I3$ gets rid of the $I1 + I2$ and $I1 + I3$ inconsistencies. So their fundamental modelling formula was not really contaminated by the $I1 + I2$ and $I1 + I3$ inconsistencies after all.

This is an interesting case because the authors recognized the presence of the inconsistencies and admitted to them in print. What they failed to notice at the time was that they had at the same time misidentified their underlying assumptions, so the inconsistencies they confessed to did not really have anything to do with their end result.

'Binary Independence' a Misnomer

The Robertson, Maron, & Cooper theory is not the only context in which $I2$ and $I3$ should be replaced by $I4$. The same substitution is also needed to clarify the basis of the popular 'binary independence' model. Although that model is ordinarily presented as though it required $I2$ and $I3$, on examining its derivation one finds that $I4$ serves just as well, and more economically. $I2$ and $I3$ are indeed sufficient for the model, but they are stronger than necessary and so are misleading.

The mathematical point involved here is worth spelling out. The binary independence model, so-called, can be derived very simply from the odds form of Bayes Rule of Inference. (The 'odds' in favor of an event X, denoted by the expression $O(X)$, is by definition $P(X) / P(\sim X)$.) For the case of two document properties A and B, Bayes Rule of Inference states that

$$O(R \mid A, B) = \frac{P(A, B \mid R)}{P(A, B \mid \sim R)} O(R).$$

That is to say, the posterior odds $O(R \mid A, B)$ in favor of relevance given the two clues A and B is equal to the prior odds of relevance $O(R)$ times the fractional expression in the right hand side, known as the 'likelihood ratio' (Eels 1982). This identity follows immediately from the definition of odds and of conditional probability.

In the 'binary independence' IR model, Bayes Rule

is modified so that in application it becomes

$$O(R \mid A, B) = \frac{P(A \mid R)}{P(A \mid \sim R)} \frac{P(B \mid R)}{P(B \mid \sim R)} O(R).$$

In other words, a simplifying assumption is invoked in order to break up the single likelihood ratio appearing in the original Bayesian identity into the product of two separate likelihood ratios for the two properties A and B. Traditional accounts of the model have claimed that the simplifying assumptions needed to effect this are $I2$ and $I3$. These assumptions are indeed sufficient, but obviously they are not necessary, for the weaker assumption $I4$ will do.

Now, $I4$ is not an independence assumption at all. It is more of an assumption of linked dependence, in which the degree of statistical dependence between A and B in the relevant set is asserted to be associated in a certain way with their degree of dependence in the nonrelevant set. Since $I4$ expresses the essence of the model more accurately than $I2 + I3$, the term 'binary independence model' is a misnomer.

This is no mere academic quibble about nomenclature. In attempting to evaluate the validity of a probabilistic IR model, one is naturally led to study the plausibility of the assumptions on which it is based. As everyone seems to agree, the independence assumptions $I2 + I3$ are not especially plausible, so it is important to understand that these are not the assumptions that actually underlie the model. The simplifying assumption on which it is really founded is assumption $I4$, which is weaker and more plausible than $I2 + I3$.

Less Need for Dependency Data

The belief that the so-called 'binary independence model' is founded on strong independence assumptions is an error that has wrought mischief. One issue into which it has injected an unfortunate element of confusion is the question of how important it is to make use of empirical term co-occurrence data. Under the misimpression that they were otherwise faced with the necessity of relying on highly artificial independence assumptions, researchers have been led to reason that it is important to make use of empirical dependency information. In this vein Robertson and Sparck Jones wrote (1976, p. 140) "The use of any independence assumptions at all is suspect, since they certainly do not hold universally. The alternative would be to look for term co-occurrence information ..."

But as we have seen, this is a false dichotomy. The true alternative to introducing term co-occurrence data is not the use of the strong independence assumptions $I2 + I3$, but only the weaker and more acceptable balanced-dependence assumption $I4$. It is fallacious to argue that, because the independence assumptions are very bad, we very badly need empirical term dependency information to get around them. In fact, the independence assumptions do not really bear on the matter at all, having insinuated themselves into the discussion only as a result of a misidentification of modelling postulates.

Considerable research effort has been expended on the question of how best to exploit term co-occurrence information (van Rijsbergen 1977, 1979; Harper & van Rijsbergen 1978; Cooper & Huizinga 1982; Robertson & Bovey 1982; Cooper 1983; Yu, Buckley, Lam, & Salton 1983; Kantor 1984; Lee & Kantor 1990). But when it is recognized that it is not independence, but linked dependence, that the so-called binary independence model really assumes, the need for introducing empirical dependency information is less keenly felt. This is not to say that using co-occurrence data is a bad idea. The point is merely that co-occurrence data is not so important as had been generally thought, that the improvement in retrieval effectiveness to be won from its use could well be slight, and that the strength of the theoretical argument in favor of introducing it needs to be re-evaluated.

Conclusions

The various inconsistencies and misidentified modelling assumptions that we have perpetrated on ourselves as IR researchers have not halted the progress of probabilistic IR. However, they have surely obscured to some extent the true character of our theories. On clearing away some of the confusion, we find that the standard models are different, and in some cases actually better, than we had thought; for our real modelling assumptions are more plausible than the ones we thought we had adopted. And our logical sins, black as they may be, lay only in what we said our theories were, not in what they really were.

Acknowledgments

The points made in this paper arose from discussions with a number of colleagues, whose contribution it is a pleasure to acknowledge: J. Allan, C. Buckley, M. E. Maron, S. E. Robertson, and G. Salton. The Department of Computer Science at Cornell University provided the hospitable environment in which the work was undertaken.

References

Cooper, W. S. Exploiting the maximum entropy principle to increase retrieval effectiveness. *Journal of the American Society for Information Science*, 34(1): 31-39; 1983.

Cooper, W. S.; Huizinga, P. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development*, 1(2): 99-112; 1982.

Eels, E. *Rational Decisions and Causality.* Cambridge University Press. 1982.

Fuhr, N.; Buckley, C. Probabilistic document indexing from relevance feedback data. *Proceedings Thirteenth International Conference on Research and Development in Information Retrieval*, Brussels: 345-61. Sept. 1990.

Harper, D. J.; Van Rijsbergen, C. J. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3): 189-216; 1978.

Kantor, P. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Development*, 3(2): 88-94; 1984.

Lee, J. J.; Kantor, P. A study of probabilistic information retrieval systems in the case of inconsistent expert judgement. *Journal of the American Society for Information Science*, 42(3), 1990.

Maron, M. E.; Curry, S.; Thompson, P. An Inductive Search System: Theory, Design, and Implementation. *IEEE Transactions on Systems, Man, and Cybernetics*, 16(1): 21-28; 1986.

Maron, M. E.; Kuhns, J. L. On relevance, probabilistic indexing, and information retrieval. *Journal of the Association for Computing Machinery*, 7(3): 216-244; 1960.

Miller, W. H. A probabilistic search strategy for Medlars. *Journal of Documentation*, 27: 254-266; 1971.

Robertson, S. E. Specificity and weighted retrieval. *Journal of Documentation*, 30: 41-46; 1974.

Robertson, S. E.; Bovey, J. D. Statistical problems in the application of probabilistic models to information retrieval. *British Library Research and Development Department*, Report No. 5739; 1982.

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

Robertson, S. E.; Sparck Jones, K. Relevance weighting of search terms.*Journal of the American Society for Information Science*, 27(3): 129-146; 1976.

van Rijsbergen, D. J. *Information Retrieval*(2nd ed.). London: Butterworth & Co. Ltd; 1979.

Yu, C.T.; Buckley, C.; Lam, H.; Salton, G. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2: 129-154; 1983.

Yu, C.T.; Salton, G. Precision Weighting – An Effective Automatic Indexing Method. *Journal of the Association for Computing Machinery*, 23(1): 76-88; 1976.