Presented at MF/IR Workshop, SIGIR 2002, Tampere, Finland

On Bayesian models and event spaces in information retrieval

Stephen Robertson

Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB
UK
ser@microsoft.com

August 20, 2002

1 Introduction

There have been several attempts recently to reconcile, or at least to understand the relationship between, traditional probabilistic models of information retrieval and the newer language models. Since both treat the retrieval problem probabilistically, it might be expected that they can be formulated in comparable terms. However, this has proved difficult. One question concerns the role of relevance, which takes a central position in some traditional models (such as Robertson and Sparck Jones [1976], referred to as RSJ), but does not appear explicitly in at least the early language models (e.g. Ponte and Croft [1998]).

The present author and others [Sparck Jones et al. 2002] have recently claimed that the early language models assume that there is only one relevant document per query. This claim is based on the observation that language models ask the question of each document: What is the probability that this document, or rather the model which generated this document, also generated the query? Since each document is taken to have its own language model, if it turns out that a particular document is relevant (that is, its model did indeed generate the query), it would seem that no other model could have done.

Lafferty and Zhai [2002], on the other hand, in a recent paper, develop a basic probabilistic model from which they derive both the RSJ model and the simple language model. They claim in conclusion that (a) RSJ and the simple language model are equivalent; and (b) that the language model requires no such assumption as that there is only one relevant document per query.

The present paper discusses an issue underlying all probabilistic models, that of the event space assumed, and draws in part from a pair of old papers [Robertson et al. 1982; Robertson et al. 1983]. I discuss possible views of the event space in case of documents, queries and relevance judgements, and come to some different conclusions about the relationship between RSJ and the simple language models. However, in order to illustrate the event space issues, the paper first introduces a rather different example from the IR one, with different structural characteristics.

2 Random variables, conditional probabilities and event spaces

Suppose we have two random variables, X and Y, with some assumed relation between them. We can imagine (though this is not necessary) that there is a causal relation $X \to Y$. Then we might consider a model which models the following quantities:

Model A: $P(X), P(Y|X)$

Can we now, without asking *any* further questions, apply such equations as:

$$P(Y) = \sum_X P(X)P(Y|X) \text{ ?} \tag{1}$$

Equation 1 is one of the basic relationships in probability theory. These relationships imply that Model A provides a full description of the event space involving these two random variables: that if we have Model A, then we can infer any other quantity involving just these variables.

However, the following example will show that we *cannot* blindly apply equation 1 to a situation in which we have all the information for Model A. There is, of course, a simple explanation for this apparent contradiction of the laws of probability; however, the explanation needs to be investigated.

Example: we have stars $\mathcal{S}$, and planets $\mathcal{T}$. Stars either have ($X = 1$) or do not have ($X = 0$) magnetic fields. Planets either have ($Y = 1$) or do not have ($Y = 0$) magnetic fields. We have a (complete) universe consisting of 2 stars and 3 planets. Star s_1 has $x_1 = 1$; it has two planets t_{11} and t_{12} with $y_{11} = 1$ and $y_{12} = 0$. Star s_2 has $x_2 = 0$; it has one planet t_{21} with $y_{21} = 0$. In this universe, the following probabilities may be calculated (not estimated, since the universe is complete, but calculated exactly):

$P(X = 1) = \frac{1}{2}$
$P(Y = 1|X = 1) = \frac{1}{2}$
$P(Y = 1|X = 0) = 0$

From these we would infer using equation 1 that $P(Y = 1) = \frac{1}{4}$. But we have three planets, one of which has a magnetic field, so actually we have $P(Y = 1) = \frac{1}{3}$.

We could construct a similar example using parents and children and some genetically-determined property (such as eye colour). We could have a model that specified (for a population) the probability of each combination of the relevant genes (X); also the probability of each eye colour in a child conditional on the parents' genes ($Y|X$). But if we wanted to infer the probability of the parents' gene combination on the basis of the observed eye colour of a child ($X|Y$), we would run into the same problem.

2.1 Brief specification of the problem

What is the problem here? In short, it is the event space. The laws of probability are written in terms of a single event space with a single probability measure defined on it; for historical reasons (which I believe to be unfortunate), the standard notation $P(.|.)$ does not provide for the denotation of the event space. If we denote a probability for a particular event space $\mathcal{E}$ as $P_{\mathcal{E}}(.|.)$, then I should rewrite the data I have for the example as:

$P_{\mathcal{S}}(X = 1) = \frac{1}{2}$
$P_{\mathcal{T}}(Y = 1|X = 1) = \frac{1}{2}$
$P_{\mathcal{T}}(Y = 1|X = 0) = 0$,
referring to the event spaces of stars and planets. It is immediately obvious that we cannot apply equation 1 to this data, because the probabilities are defined in different event spaces.[1]

So the answer to my question above is: We emphatically *cannot* apply the equation without asking any questions.

But this situation deserves much more detailed analysis. The combination of event spaces I have exemplified, involving stars and planets, has a slightly complex but not particularly unusual structure (that of every many-to-one relation in every relational database in the world). It is worth asking questions about what we can say about such combinations.

2.2 Overview of event spaces

The traditional view is that the event space is the set of all possible outcomes of an experiment, that the probability measure is a measure satisfying certain properties on this event space, that a random variable is a deterministic function of the outcome of an experiment. One could discuss this at length, but it will do for the present discussion. If we want to define a probability $P_{\mathcal{E}}(X)$, we need to assume that we do indeed have a well-defined event space $\mathcal{E}$, probability measure P, and random variable X defined on $\mathcal{E}$. For $P_{\mathcal{E}}(Y|X)$, we need both Y and X to be defined on $\mathcal{E}$; the values of X are used to induce a partition on $\mathcal{E}$.

I will also observe that in a finite event space, the usual simplest probability measure assigns equal probability to each elementary event. However, there are many circumstances in which that is simply inappropriate, to the extent that one would not even consider it a candidate. For example: suppose the experiment consists in tossing two coins, with outcomes HH,HT,TT. Our knowledge of the structure of this event space is such that we would (probably without thinking) reject the simple probability measure (1/3,1/3,1/3) and instead use (1/4,1/2,1/4). Our understanding of the structure of this event space is enough to convince us that the simple one is simply bad.

3 Detailed analysis of the example

The full event space of the stars and planets example is a set of stars, a set of planets, and a one-to-many relation between them. We have this knowledge of the structure, and we need to work out the implications of this knowledge for any probabilistic statements about our models for the event space. I will refer to this full event space as $\mathcal{ST}$. To help focus the discussion, I will imagine that one question I might want to ask is the following:

> What is the probability that a star has a magnetic field, given that I know that it has two planets with magnetic fields?

This is a perfectly reasonable question to ask, though it may need some refining. But first I will explore some possible ways of looking at the event space $\mathcal{ST}$. We can consider several simpler event spaces. The first is just $\mathcal{S}$. This is easy: it stands on its own (does not need $\mathcal{T}$ to define it), has X

[1] What does it mean to refer to a set of objects as an event space? The simplest interpretation is that the basic event is to choose one of the objects at random. In this case, the probability $P_{\mathcal{S}}(X = 1) = \frac{1}{2}$ means "if we choose a star at random from this universe, this is the probability that it has a magnetic field". For the other two probabilities, we have to choose a *planet* at random.

defined on it, and has no internal structure. It makes perfect sense to define a uniform probability measure on it; this yields the above value for $P(X = 1) = P_{\mathcal{S}}(X = 1)$ in the specified universe.

The second is $\mathcal{T}$. This is also moderately straightforward, if we look at Y alone; however, it does have some internal structure (planets are siblings of each other or not), which treating it as a straightforward uniform-probability event space will simply ignore. More on this later. But there is another slight complication: I want to consider X as a condition on this space. Is this valid? Well, with a very slight extension, yes: we can ask of any planet, as well as "Do you have a magnetic field?", "Does your star have a magnetic field?". This question is clear and unambiguous, so it is perfectly reasonable to assert that X is defined on $\mathcal{T}$ as well as on $\mathcal{S}$. However, in order to be strictly accurate, I should treat this as involving an extension of both the event space and the variable; we might refer to these as $\mathcal{T}^+$ and X'. In $\mathcal{T}^+$, we associate with each planet not only its own properties but those of the star to which it belongs, and X' is the property *of the planet* of belonging to a star with or without a magnetic field. Then I should rephrase my data about the specified universe as:

$P_{\mathcal{S}}(X = 1) = \frac{1}{2}$
$P_{\mathcal{T}^+}(Y = 1|X' = 1) = \frac{1}{2}$
$P_{\mathcal{T}^+}(Y = 1|X' = 0) = 0$

– but it also becomes clear that I need more data to specify the probabilistic properties of this event space more fully. For example, I may need $P_{\mathcal{T}^+}(X')$, which is not deducible from the above and which is different from $P_{\mathcal{S}}(X)$. This would then allow me to use equation 1.

Note that I *cannot* do the same trick the other way around: I cannot simply ask a star "Does your planet have a magnetic field?", because the question is ill-defined. I could define an $\mathcal{S}^+$ in a different way, e.g. by defining a new random variable Y' as the proportion of the star's planets that have magnetic fields, and construct a probabilistic model with this combination of event space and random variables. As before, a simple uniform probability measure (on stars) is quite appropriate.

I now have five event spaces: $\mathcal{ST}$, $\mathcal{S}$, $\mathcal{T}$, $\mathcal{S}^+$, $\mathcal{T}^+$.[2] There are differences between them, some minor, some significant. I have simple, straightforward probability measures on each of the last four of these spaces. I do not, however, have a probability measure of any kind on $\mathcal{ST}$. Nor is it possible to define one which (a) makes sense, and (b) allows me to express all probabilistic aspects of the space.

Does this mean that we cannot make probabilistic statements about $\mathcal{ST}$? Of course not. Any statement about any of the other event spaces is also about $\mathcal{ST}$. However, none of the other spaces captures all that we might want about $\mathcal{ST}$. Given that we need event spaces with probability measures, it follows that we need more than one event space to make probabilistic sense of $\mathcal{ST}$.

Now we may return to the question with which I started this section. Of the event spaces I have considered, the one which comes closest to helping us with this question is $\mathcal{S}^+$ as defined; but this does not quite do the trick. Answering the question would actually need a good understanding of the full structured event space $\mathcal{ST}$, and a combination of models which explicity took this structure into account. It simply does not make sense in $\mathcal{T}^+$, because in this event space, *there is no such thing as an individual star*. (In fact $\mathcal{S}^+$ does not have individual planets, either.)

4 A small observation

None of the above problems arise if X and Y are initially defined on the same event space – in this case Model A and equation 1 go together perfectly well. I believe this is the situation most theorists

[2] It would be possible to define other event spaces in addition to these five. For example, I could take each star and define an event space of its own planets, giving me two more in this particular universe. I do not pursue this line for the present example.

have in mind when they automatically assume equation 1. I also believe this assumption is dangerous.

5 Further analysis of $\mathcal{T}^+$

I said above that in $\mathcal{T}^+$ there is no such thing as an individual star. This statement deserves further analysis.

$\mathcal{T}^+$ consists of a set of elementary events $\{(t, s)\}$ where $t \in \mathcal{T}$ is a planet and $s \in \mathcal{S}$ is a star (together with a uniform probability measure on these events). That is, we take each planet t and its properties, and its associated star s and its properties, as the single event $\{(t, s)\}$ which gives us the values of our random variables. Each such event is distinct as an event from every other; this is the nature of a simple event space. So in some sense there are individual stars; the concept that is missing from this space is that of different planets sharing an individual star. To put it another way: if we take $\mathcal{T}^+$ as our complete probabilistic model, we are in effect assuming that *every star has just one planet*.

At least, it is clear that the model $\mathcal{T}^+$ is consistent with such an assumption. Furthermore, I cannot see any way of generating this model under a weaker assumption.

6 Queries and documents: the cross product structure

The structure of the $\mathcal{ST}$ example is not uncommon, but it is not the same as the query / document / relevance case. I would like to discuss some aspects of this case.

We start with the set of all queries $\mathcal{Q}$. As with stars, this is the set of all actual queries, representing information needs, not the set of all values of some random variable. We may define one or more random variables on this set, including (following Lafferty) the text of the query. Similarly we have a set $\mathcal{D}$ of documents. Ignoring relevance for the moment, queries and documents have at first glance no logical relationship. What this means is that we can pair *any* document with any query. The logical structure of this space ($\mathcal{QD}$) is a cross product of the two individual spaces $\mathcal{Q}$ and $\mathcal{D}$.

In this case, there is a simple and fairly obvious probability measure on the space, namely one that is uniform over pairs $(q, d) \in \mathcal{Q} \times \mathcal{D}$. I shall refer to the probabilistic event space defined by this measure over this set of events as $\mathcal{QD}_0$; the cross product on its own, without a probability measure, I shall call $\mathcal{QD}$ on its own. In the probabilistic event space $\mathcal{QD}_0$, each pair is regarded as a single event, unrelated to any other, and every pair has the same probability.

It might be assumed that this uniform probability measure on pairs ($\mathcal{QD}_0$) is in some sense equivalent to treating the two separate spaces $\mathcal{Q}$ and $\mathcal{D}$ uniformly. However, this is not the case (or at least, such equivalence has severe limitations). A consequence of choosing pairs as the basis for the probability measure in $\mathcal{QD}_0$ is that it loses part of the structure of the full space $\mathcal{QD}$, in the same way that $\mathcal{T}^+$ loses part of the structure of $\mathcal{ST}$. Consider the following structural aspects:

- $\mathcal{QD}$ is 'striped' – any property that a particular query has is shared across all pairs involving this same query with any document (and vice-versa); $\mathcal{QD}_0$ has no such striped character.

- in $\mathcal{QD}$ every pair has one set of q-siblings (all the pairs sharing the same query-event) and another set of d-siblings; there are no siblings in $\mathcal{QD}_0$.

- the question 'what can I say probabilistically about a query, if I know something about two of the pairs to which it belongs?' has meaning in $\mathcal{QD}$, but none in $\mathcal{QD}_0$.

- the concept of an individual query or document is apparently meaningless in $\mathcal{QD}_0$ (but see the following section).

6.1 Individuality of documents and queries

If we take $\mathcal{QD}_0$ as our complete probabilistic model of the query-document situation, we are assuming a uniform space of unit events which are query-document pairs. Every such event is distinct, and there is no concept that two such events may share the same individual query event (say). This is equivalent to assuming that every document (individual document event) has exactly one query and every query has exactly one document. Note that this is independent of the relevance variable, which we have not yet introduced.

6.2 Random variables

We may define as random variables in the space of query-document pairs the text of the query and the text of the document. However, to be strictly accurate, we must acknowledge that (for example) the random variable which is the text of the query, defined on the event space of all queries, is different from the random variable which is the text of the query, defined on the event space of all query-document pairs. Query-siblings in $\mathcal{QD}$ share the same query event and therefore necessarily the same query-text. No such relationship can exist in $\mathcal{QD}_0$ – two query-document pairs may share the same query text but only accidentally. Thus by reinterpreting the query-text variable as a random variable in $\mathcal{QD}_0$, we are at the same time changing its nature significantly.

A binary relevance variable may be defined as a random variable on the space of individual query-document pairs (but not, clearly, on individual documents or queries). Thus this variable naturally resides in $\mathcal{QD}_0$. However, since neither the text of the query nor the text of the document naturally reside there, we have to be careful about models involving all three variables.

6.3 Samples

If we were to sample query-document pairs, the resulting sample would have the characteristics of $\mathcal{QD}_0$. On the other hand, if we were to sample queries and documents separately, and then take all the resulting pairs, we would have a space with all the above characteristics which $\mathcal{QD}_0$ lacks (stripes, siblings, etc.). (This is close to what we actually do in experiments – for the very good reason that it preserves aspects of the structure of the full space $\mathcal{QD}$ in which we are interested.) This is a perfectly good form of sampling, but one which is simply not described by the probabilistic event space $\mathcal{QD}_0$.

We might consider constructing a composite space for a probabilistic model from a uniform probability distribution on queries and a separate uniform probability distribution on documents. However, this would not constitute a single probabilistic event space in the usual sense. Can we define a single probabilistic event space which preserves any of the above aspects? As with the $\mathcal{ST}$ example, it is not possible to find one which preserves everything, but different spaces preserve different things.

7 Possible event spaces for IR

7.1 The event space of the RSJ model

The RSJ model [Robertson and Sparck Jones 1976] is formulated for a single query. All probabilities are about documents in relation to this single query. We can thus see the event space as the space of documents (with a uniform prior distribution); but this event space is reinterpreted for each query. In terms of the discussion above, we have q-siblings as an implicit part of the model; however, the model cannot see d-siblings. The model thus assumes that there is just one query for every document but not vice versa. In effect, the document collection is reinvented for every query. This means that we can learn about the specific query (relevance feedback in the usual sense), but not about the specific document over successive queries.

7.2 The event space of the simple language model

This is a little more difficult to see. Since the language model expresses a probability of a query given a document, it is tempting to see it as the dual of the RSJ model. This would mean that the event space was the space of queries, considered only in relation to this document. This would be consistent with the discussion in Robertson, Maron, and Cooper [1982], and would make the language model equivalent to the original probabilistic IR model of Maron and Kuhns [1960].

However, it is clear that this is not the interpretation placed on the language model by its proponents. The language model is commonly used to derive a score by which documents are ranked for a given query, in the usual fashion. But this requires that the scores for different documents, but for the same query, are directly comparable. Under the above interpretation of the event space, the scores cannot be comparable, since they come from probability distributions in different event spaces.

It is unclear to me what should be taken as the event space for the simple language model. Possible solutions would be $\mathcal{QD}_0$ or that proposed by Lafferty and Zhai. Either of these would imply that the simple language model is not capable of supporting per-query-event relevance feedback (that is, relevance feedback in the usual sense), though it would support relevance feedback across all queries (from different users) sharing the same text.

7.3 The Lafferty/Zhai model

The model proposed by Lafferty and Zhai [2002] is even simpler than $\mathcal{QD}_0$. They consider only the cross-product of *values* of Q and D (i.e. the texts as above), not of individual events, with a uniform probability distribution on the pairs in this cross-product. This has several implications.

One is that replicated queries (multiple query-events with the same text) are regarded as having the same probability, irrespective of their frequency of replication. This probably is of no importance, for the same reason that the RSJ approach is reasonable: all comparisons which the model is intended to allow are between documents for the same query. However, it would be important for inter-query-event feedback.

A more serious implication is the following. If relevance is to be taken as a random variable on this event space, it means that we must assume that relevance is determined only by the *values* (texts) of Q and D. This means that (we assume) any two people who ask the same question and see the same document will make the same relevance judgement. Since we know very well that many queries are highly ambiguous, quite apart from subjective differences, this is a strong assumption.

Despite these qualifications, in many respects the Lafferty/Zhai model seems to be similar to $\mathcal{QD}_0$. The event space refers to no individual events (either query-events or document-events). In this space, it is not possible to distinguish between query-document pairs which share the same query text because they share the same query-event, and those which share the same query text by accident (because a sampling process has happened to throw up the same value). The space cannot therefore have any of the structural characteristics of $\mathcal{QD}$ discussed in section 6 above (striping, siblings etc.). Thus it seems to assume implicitly with $\mathcal{QD}_0$ that for every document-event there is exactly one query-event and vice versa. The assumption is implicit because it is not possible to express either this assumption or its negation in terms of the event space, but it is required because of the absence of structural characteristics.

The assumption of a one-to-one relationship between document and query events is even stronger than the assumption attributed to the simple language model earlier, that each query has only one relevant document.

The version of RSJ which Lafferty and Zhai derive from their model is a special case of RSJ, since the assumption that each query has just one document is not necessary for RSJ. Therefore the conclusion that RSJ and the simple language model are equivalent is not a valid general inference from this model.

8 Conclusions

1. Model A does not carry equation 1 as a necessary consequence under all conditions under which it (Model A) can be defined.
2. When a probabilistic model is being constructed, the structure of the event space cannot be ignored.
3. Sometimes an event space is of sufficient structural complexity that a single probabilistic model (based on a single event space with a single probability measure) would be unable to capture all important statistical knowledge about it.
4. This last is the case with queries, documents and relevance judgements.
5. RSJ is not equivalent to a model based on uniform probabilities over query-document pairs.

Bibliography

LAFFERTY, J. AND ZHAI, C. 2002. Probabilistic relevance models based on document and query generation. To appear in book on Language Modelling and IR, ed. Bruce Croft.

MARON, M. AND KUHNS, J. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM 7*, 216–244.

PONTE, J. AND CROFT, W. 1998. A language modeling approach to information retrieval. In W. CROFT ET AL. Eds., *Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)* (New York, 1998), pp. 275–281. ACM.

ROBERTSON, S., MARON, M., AND COOPER, W. 1982. Probability of relevance: a unification of two competing models for information retrieval. *Information Technology - Research and Development 1*, 1–21.

ROBERTSON, S., MARON, M., AND COOPER, W. 1983. The unified probabilistic model for ir. In G. SALTON AND H.-J. SCHNEIDER Eds., *Research and development in information retrieval* (Berlin, 1983), pp. 108–117. Springer-Verlag.

ROBERTSON, S. AND SPARCK JONES, K. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science 27*, 129–146.

SPARCK JONES, K., ROBERTSON, S., ZARAGOZA, H., AND HIEMSTRA, D. 2002. Language modelling and relevance. To appear in book on Language Modelling and IR, ed. Bruce Croft.