Prior Information and the Determination of Event Spaces in Probabilistic Information Retrieval Models

Corrado Boscarino and Arjen P. de Vries

Centrum Wiskunde & Informatica (CWI), Science Park 123,
1098 XG Amsterdam, The Netherlands
corrado@cwi.nl, arjen@acm.org

Abstract. A mismatch between different event spaces has been used to argue against rank equivalence of classic probabilistic models of information retrieval and language models. We question the effectiveness of this strategy and we argue that a convincing solution should be sought in a correct procedure to design adequate priors for probabilistic reasoning. Acknowledging our solution of the event space issue invites to rethink the relation between probabilistic models, statistics and logic in the context of IR.

1 Introduction

Information Retrieval (IR) can be distinguished from other Information Access (IA) classes of techniques, like that to which deterministic database access belongs, by being mainly concerned with uncertain knowledge, at least when assuming a notion of relevance of documents with respect to the subjective and unpredictable opinion of that particular human agent that is supposed to have issued the query. Acknowledging the presence of uncertainty leads naturally to probabilistic models as the chief mathematical description of uncertain information about reality. Probability theory is the framework within which we can make precise statements about imprecise features of the world, or, in slightly different terms, about features that admit multiple precisifications. As long as the human user is considered to provide the ultimate metric of success in information retrieval, alternative approaches differ mainly in how they represent uncertainty. These representations do not stand apart because they question the validity of probability theory, but in their usage of the theory. In a formally correct model there will be one or more stochastic variables representing the uncertain link between the output of the retrieval system and the end user's satisfaction. However, given the knowledge of the parameters and of the structure of two or more models that differ in non trivial features of their design or in important presuppositions, comparing the alternative representations can be a very difficult task.

Lafferty and Zhai [1, pp. 1-10] introduce a probabilistic framework that allows to compare both the ranking and the assumptions of two well known retrieval models, the RSJ (Robertson-Sparck Jones) [4] model and the more recent language models [3]. They show that, at least in this particular case, the probabilistic semantics, that is the general rules of modeling uncertainty, are acknowledged

L. Azzopardi et al. (Eds.): ICTIR 2009, LNCS 5766, pp. 257–264, 2009.

in both approaches. They differ, however, in how they apply those rules: they factorise the same probabilities in two different ways and they may also possibly estimate the model's components with varying techniques.

Robertson [6] warns against the risk of calculating *rank equivalence*, or any other relationship between two or more probabilistic models, without considering the event spaces, that is the set of the real world objects, which we suppose our probabilistic model is about and upon which probabilities ought to be calculated. The mistaken assessment of event spaces can easily generate paradoxes and theory seems to support two mutually exclusive models of reality. Luk [5] acknowledges that, not the rules of probability itself are to blame, but the accuracy of their application to a particular problem, giving more emphasis to another link to the world, which the probabilistic model claims to represent: the estimation of the probabilities by means of a set of statistical components. Luk seems also to endorse a hierarchy of a statistical components model that provides the empirical content to an upper-level probabilistic model. While in Robertson the model's interface is lumped in the event space, Luk expands it into a separate model.

In this paper we argue that the paradox, the existence of which both Robertson and Luk agree on, albeit they propose different ways to circumvent it, is equivalent to the same marginalisation paradox that is known to the general public as the Monty Hall paradox [2]. The accepted solution to this paradox, however, cannot be expressed at the level of the probabilistic model considered by Robertson, nor by additionally considering the computational level as Luk does, but only by adding a logical level on top of the probabilistic one. Acknowledging our solution of the event space issue invites then to rethink the relation between probabilistic models, statistics and logic in the context of IR.

2 Event Spaces and Probabilistic Models

Arguing against the rank equivalence put forward by Lafferty and Zhai in [1, pp. 1-10], Robertson sets out to explain why a mixture of different event spaces could be problematic by introducing an analogous setup that is easier to understand. While an event space is often loosely defined as the set of the possible results of an experiment, the analysis brought forth by Robertson in [6] shows a more precise and at the same time more profound significance of this notion, albeit stated rather implicitly. This alternative definition of event spaces arises from an attempt to determine the conditions of applicability of the marginalisation equation, which allows the distribution of a variable Y to be calculated from the knowledge of the distribution of another variable X and the conditional probability of Y given X, as $P(Y) = \sum_X P(X)P(Y|_X)$. It is just because event spaces are linked to this fundamental relation of probability theory that this concept receives a more rigorous specification. At the same time we will see that constraining the physical objects, which the probabilistic model is about, to the domain of the marginalisation equation amounts to the reduction of the power of probabilistic reasoning; this is the limitation that Luk in [5] tries to remove.

For both authors, however, the event space becomes an ontological notion as the interface between an abstract description of the world and its extension.

2.1 Robertson's Argument

The first claim in [6] is that even models for simple situations, where estimation seems to become a straightforward process, can generate paradoxical cases in which the marginalisation equation does not yield the expected results. One example is about a very simple universe made of only two stars and three planets, where some of the stars s have a magnetic field ($x_{s_1} = 1$, $x_{s_2} = 0$) and some of the planets t that orbit around the stars also do ($y_{t_{11}} = 1$, $y_{t_{12}} = y_{t_{21}} = 0$): the problem is to calculate the probability $P(Y = 1)$ to find a planet with a magnetic field. In this situation we presume to be able to calculate the probability $P(Y)$ to find a planet which has a magnetic field by marginalisation on the probabilities $P(X)$ of a star to have a magnetic field, provided that we know the marginal probabilities $P(Y|_X)$. However, when we compare the result obtained by marginalisation with that calculated by simply counting the occurrences of a planet with a magnetic field and dividing by the total number of planets, we obtain two different results.

This mismatch, so goes Robertson's argument, is clearly paradoxical. The cause relies on a lack of expressive power in the notation used to express the marginalisation equation: we are unable to specify the objects in the real world to which the different probabilities apply. The distribution $P(X)$ applies to the event space $\mathcal{S}$ of stars, the conditional distribution $P(Y|_X)$ applies to the event space $\mathcal{T}$ of planets, while we need information about the full event space given by the cross-product $\mathcal{ST}$.

Robertson applies this result to the claim by Lafferty and Zhai [1, pp. 1-10] that classic probabilistic approaches like the RSJ model in [4] are equivalent at the level of the probabilistic model to the language modeling approach, although they may differ at the lower level of their statistical components, that is in how the probabilities are estimated. According to Lafferty and Zhai these two approaches correspond to two different factorisations in the marginalisation equation; the RSJ model results from the factorisation $P(D, Q|_R) = P(D|_{Q,R})P(Q|_R)$ and the language modeling approach from the factorisation $P(D, Q|_R) = P(Q|_{D,R})P(D|_R)$, where Q, D and R are the stochastic variables associated to the set of queries, to that of documents and to relevance, respectively.

Now consider that queries and documents play the role of stars and planets in the metaphor. Relevance is mapped onto both variables X and Y in the following way: relevance corresponds to the magnetic field and the division of labour between the two variables is related to the two approaches to IR that have been reviewed in [1, pp. 1-10]. The case of considering relevant queries that generate documents, that is the RSJ model, corresponds to stars with a magnetic field and it is accounted for by the X variable; the same applies for the language model, the planets and the Y variable.

In order for two probabilistic models to avoid the marginalisation paradox, the event spaces upon which the probabilities ought to be calculated should be

the same; adapting the stars and planets example to the particular case of IR leads to the conclusion that the correct event space should be the most general $\mathcal{QD}$ obtained as the cross-product of the event spaces of queries and documents. Therefore, and this is Robertson's main claim, since Lafferty and Zhai do not apply the marginalisation equation to the full event space they may be easy pray of the marginalisation paradox.

The mismatch between event spaces appears to be a perfectly natural explanation for the lack of consistency between the results of applying marginalisation and those of direct frequency counting. In this particular case, however, the concept of event space is not derived from the applicability requirements of the marginalisation equation. Rather, less complete event spaces (e.g. $\mathcal{T}^+$ referred to in [6]) are identified *after* applying marginalisation in those cases where a paradoxical answer is claimed to be obtained. Event spaces are brought into life with the explicit purpose of resolving the marginalisation paradox and they are simply singled out by their being characteristic of two faulty applications of the marginalisation equation. Every time we find a minimum set of probability functions that, once employed in the marginalisation equation, let the results disagree with a frequentist interpretation of the same probabilities, we assign two different labels to the sets and we call the labels 'Event Space'.

2.2 Luk's Argument

Luk [5] understands the importance of the issue raised by Robertson, and he identifies the source of the problem mainly in the uniform probability assumption, which he considers not to bear any logical significance. The core issue is still how the probabilities that appear in the probabilistic model are coupled to worldly objects, but he shows to maintain that this link can be mapped, in addition to the event spaces, also onto the probability distributions: instead of letting the event space proliferate, we may as well keep the event space fixed and allow for multiple distributions. Luk shows that, in some configuration of distribution, the marginalisation paradox can be beaten and in this case rank equivalence holds, at least in a weak sense.

Luk's perspective specifies the distinction between the structural information carried by the event spaces, which can be graphically represented by the nodes of a tree structure and the information, which can be gathered about that structure and that is stored in the form of labels on the tree's branches (see Fig. 1 in Luk [5]). Given a direct probability measure onto the total event space, $\mathcal{ST}$ in Robertson's example, we can then construct an infinite number of distributions that result in the successful application of the marginalisation equation. Instead of focusing on one arbitrary configuration of distributions, e.g. a uniform probability distribution at each branch or, equivalently, at each marginalisation step, it is therefore far more instructive to assess, so goes Luk's argument, whether some distributions lead to data inconsistency, and avoid the paradox.

Luk follows a strategy which is somewhat anticipated in [1, pp. 1-10] where the authors make a distinction between rank equivalence at the probabilistic level and that at the lower statistical components level. Robertson only touched

upon this distinction when trying to determine to which of the many possible event spaces are the different probabilistic models meant to be applied. One of the most interesting insights that Luk provides in [5] is indeed related to his appeal for a modular design of IR applications in which the probabilistic models are thought of as populating a probabilistic functional block that interacts to the various data sets through a statistical block. The major advantage of this scheme is that the effects, in this case on rank equivalence, of employing different probabilistic models can be kept separated from the contributions from the statistical components. The primitive interface to the data sets represented by the event spaces does not allow to easily determine how different informal presuppositions relate to the mathematical form of the probabilistic model; at this stage, it is not clear how to determine for a given IR model which event space it actually refers to.

In the next section we discuss what Luk seems reluctant to pursue as a consequence of his design, that is to allow for adding multiple blocks to the probabilistic model, which will inevitably loose some of its prominence. We maintain that the probabilistic model receives its empirical content from the statistical components model, admitting, however, that other functional blocks may also have ontological significance.

3 The Marginalisation Paradox and the Determination of the Priors

We attempt to untangle the complex, and only implicitly defined notion of event space given in [6], using what we call three different 'event space expansions'. The first expansion makes more explicit who makes an observation and assigns probabilities to uncertain knowledge. Probability theory is a faithful model of the real world only to the extent to which our knowledge about matter of facts is accurate, that is the uncertainty modeled by probability theory resides in our comprehension of the world and not in the world itself. A probabilistic model always refers, although sometimes implicitly, to an epistemic process. The latter can, if one wishes, be personified by a hypothetical observer who describes the world from her particular point of view. According to this picture we can recast Robertson's example by positing an observer within a universe with just two stars and three planets, who is puzzled by two apparently innocuous, but mutually exclusive statements that she believes both should be true. She finds out that measuring the magnetic field of planets that belong to stars and then separately that of the stars in order to subsequently infer by applying the marginalisation equation, the probability to find a planet with a magnetic field gives a certain result. She also believes that simply detecting and counting the number of planets with a magnetic field and then dividing the result by the total number of planets should yield the same number. A paradox arises if we want to maintain that both marginalisation and frequency counting are two legitimate ways to calculate probabilities, *and that they can be applied simultaneously.*

Our first claim is that the marginalisation paradox in [6] does not arise because of a shortcoming of probability theory that does not allow to adequately

represent event spaces, but just because of the fact that the example presented by Robertson models an unrealistic process: in which realistic setting would an agent be interested into calculating $P(Y)$ by marginalisation when being in epistemic state $\mathcal{ST}$, which allows to simply count the occurrences of the Y variable?

We have thereby discovered a first link to elements of the real world, in this case agents characterised by certain epistemic states and processes in which they engage in order to modify their epistemic state, in a sense to evolve, which is curled into the notion of event space. The paradox reveals the presence of additional event spaces, besides the most intuitive full event space $\mathcal{ST}$, like $\mathcal{T}$ or $\mathcal{S}^+$ in [6], each labeling distinct epistemic states and processes.

This first expansion of event spaces in terms of an epistemic model resolves the paradox by inferring the presence of two different agents: one agent, being in the epistemic state $\mathcal{ST}$ calculates the probabilities by frequency counting; another agent, being in the lesser epistemic state $\mathcal{T}^+$ is forced to use marginalisation and she obtains results that disagree with the *other agent's* results. Needless to say, this happens all the time as agents may disagree on a lot of issues without generating paradoxes. Even stronger, the definition of agent as the collection of its information states, demands that agents cannot be in totally overlapping information states and still being distinguishable, hence no agent can totally agree with another agent.

This solution can also be explained in Luk's framework where an agent appears to use one IR model, while another uses a different model: in some configurations of the world the two agents may well derive the same conclusions, albeit when still in different epistemic states. This notion is indeed weaker than having one and the same agent using two models and ranking the documents in the same way, but it may be still worth investigating, as Luk does, which distributions lead to this result. The obvious advantage of the epistemic expansion, that we share with Luk's statistical expansion, is that we are able to distinguish the different gradations of rank equivalence and that we do not fall prey of the paradox; the advantage we have above Luk is that we are able to link the different cases to some procedural information, may that be made available by other components of the system like relevance feedback or other forms of user interaction.

Another expansion is particularly interesting for that it shows an unexpected connection between the marginalisation paradox in Robertson and another one, well known as the Monty Hall paradox [2]. One of its many versions involves three envelops, one of which contains a prize. After a contestant has chosen an envelop, the quiz-master opens one of the other two envelops, showing its empty content. She offers then the possibility to switch the envelops: the problem is to decide, based on objective reasons only, whether the contestant should accept the offer. At a first sight we may be tempted to reason like Lafferty and Zhai in [1, pp. 1-10], ignoring any event space and regarding relevance or the magnetic field as the prize. A naive conclusion would then be that, since the prize may be contained in either one or the other envelop, it does not really matter whether we switch or not; in fact, people that hear the paradox for the first time equally distribute their answers among the two possibilities. Once this metaphor is applied to IR,

rank equivalence follows. The correct solution of the Monty Hall paradox is well known: the way we are supposed to calculate the probabilities depends on who put the prize in the envelop, the quiz-master or someone else. If the quiz-master knows which envelop contains the prize, and she would never open that one, her knowledge of the right envelope to open must be included in the model to arrive at the correct answer for the quiz candidate: switch!

We understand now what Robertson exactly wanted to model in his stars and planets example. Although it may seem strange that he models one observable with two different variables, the magnetic field with both X and Y, we encounter the same need in the Monty Hall paradox. In order to make sense of the situation we must model one observable, the prize, with two variables: one variable, say X, which models the magnetic field of the stars, but also the relevance of the queries or the prize that the quiz-master put in the envelop, and another variable, say Y, which models the magnetic field of the planets, but also the relevance of the documents or the prize that someone else than the quiz-master put into the envelop. Robertson, who claims to provide a model of the physical world, by choosing this representation for the magnetic field, ends up in representing some knowledge about the world that is not immediately evident upon examination of the data set, but is nevertheless needed for a correct determination of the probabilities; like the information on who put the prize in the envelop, which is not provided by the problem's statement, but it is exactly the source of the Monty Hall paradox.

We also immediately see how this analysis leads to another event space expansion. To specify the event space upon which probabilities are considered to apply, amounts then to the determination of the background information that should be taken into account in order to correctly calculate the probabilities. Also in this case the paradox does not sustain a more attentive analysis: there is no contradiction in the fact that a probabilistic reasoning moving from two different sets of background information also leads to two different results. Discovering some local regularity is actually a quite interesting finding and therefore Luk's weak equivalence should not be underestimated; he did however fail to recognise that it is not the case that there are either different event spaces or different possible distributions, but there are different problems, for different observers, with different prior information, but possibly on the same data set. The event spaces show yet another face by playing the role, which the priors in Bayesian reasoning are usually charged with: they appear to be the core issue in the paradox resolution only because they are a primitive description of the prior information, which *de facto* resolves the paradox.

4 Drawing Conclusions and Consequences for IR

In this paper we show that the way Robertson in [6] questions the rank equivalence derived by Lafferty and Zhai in [1, pp. 1-10] is ineffective. The alleged paradox that would arise when the event spaces are not adequately taken into account, can be defeated in more than one way, each corresponding to one of

what we termed 'event space expansions'. Luk has proposed to extend the event space, understood as an interface with the data set only, by means of a statistical components module. While we welcome his attempt towards a more thoroughly understanding of the issues raised by Robertson by means of a functional hierarchy, we have argued that his claim is in itself also problematic because it does not address the most sensitive features of Robertson's argument: the existence of the paradoxical situation and the functional identity between the priors and the event spaces. Luk descends the hierarchy and attempts to solve this issue at the statistical level. We view the problem as one that needs more abstraction rather than less; if event spaces are really just priors, what we need is a method to select priors, which become sockets to interface upper-level functional blocks. For example, a probabilistic model can fetch, through its priors, the output of a dynamic epistemic logic module that formalises how observers change their information states upon which their relevance assessments highly depend.

Once we apply this conclusion to IR as a discipline, this is in essence an argument for revisiting the logical models of IR as first proposed in [7] for the solution to the paradox that has arisen calls an higher level of abstraction than that provided either by the probabilistic model or by the statistical components model.

References

1. Croft, B.W., Lafferty, J.: Language Modeling for Information Retrieval. The Information Retrieval Series. Springer, Heidelberg (1999)
2. Gardner, M.: Mathematical games column. In: Scientific American, October 1959, pp. 180–182 (1959)
3. Hiemstra, D.: Using language models for information retrieval. Ph.D. dissertation. University of Twente, Enschede (January 2001)
4. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. Inf. Process. Manage. 36(6), 779–808 (2000)
5. Luk, R.W.: On event space and rank equivalence between probabilistic retrieval models. Inf. Retr. 11(6), 539–561 (2008), http://dx.doi.org/10.1007/s10791-008-9062-z
6. Robertson, S.: On event spaces and probabilistic models in information retrieval. Inf. Retr. 8(2), 319–329 (2005), http://dx.doi.org/10.1007/s10791-005-5665-9
7. van Rijsbergen, C.J.: A new theoretical framework for information retrieval. SIGIR Forum. 21(1-2), 23–29 (1987)