

# On event space and rank equivalence between probabilistic retrieval models

Robert W. P. Luk

Received: 2 May 2006 / Accepted: 21 May 2008 / Published online: 14 June 2008  
© Springer Science+Business Media, LLC 2008

**Abstract** This paper discusses various issues about the rank equivalence of Lafferty and Zhai between the log-odds ratio and the query likelihood of probabilistic retrieval models. It highlights that Robertson's concerns about this equivalence may arise when multiple probability distributions are assumed to be uniformly distributed, after assuming that the marginal probability logically follows from Kolmogorov's probability axioms. It also clarifies that there are two types of rank equivalence relations between probabilistic models, namely strict and weak rank equivalence. This paper focuses on the strict rank equivalence which requires the event spaces of the participating probabilistic models to be identical. It is possible that two probabilistic models are strict rank equivalent when they use different probability estimation methods. This paper shows that the query likelihood,  $p(q|d, r)$ , is strict rank equivalent to  $p(q|d)$  of the language model of Ponte and Croft by applying assumptions 1 and 2 of Lafferty and Zhai. In addition, some statistical component language model may be strict rank equivalent to the log-odds ratio, and that some statistical component model using the log-odds ratio may be strict rank equivalent to the query likelihood. Finally, we suggest adding a random variable for the user information need to the probabilistic retrieval models for clarification when these models deal with multiple requests.

**Keywords** Probabilistic models · Event space · Information retrieval

## 1 Introduction

Robertson's (2005) instructive paper cautions some of the mathematical derivations of language models for information retrieval using marginal probabilities (in Eq. 1 of Robertson 2005) because the event space of these probabilities is different from the event space of the conditional probability. He further questions the event spaces of the earlier language model by Ponte and Croft (1998) and by Lafferty and Zhai (2003). To answer these

---

R. W. P. Luk (✉)

Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong  
e-mail: csrluk@comp.polyu.edu.hk

questions, Roelleke and Wang (2006) show how event spaces of the different retrieval models are related via parallel derivations with a novel language model that uses a query likelihood ratio for ranking. In this paper, we believe that Robertson's assumption about the probability distributions being uniform is not a logical necessity. It is instructive to develop our understanding on the basis of Kolmogorov's probability axioms and revisit the rank equivalence of Lafferty and Zhai (2003) in this perspective.

The rest of this paper is organized as follows. Section 2 derives the marginal probability based on Kolmogorov's probability axioms and highlights the necessary condition that needs to be true for the marginal probability calculation to be correct. Section 3 clarifies Robertson's example of stars and planets. We provide an interpretation of how the marginal probabilities can be obtained from random sampling tuples in a cross-product event space. In addition, this section clarifies the rank equivalence of Lafferty and Zhai (2003). Section 4 discusses rank equivalence between two retrieval models, clarifies that there are two types of rank equivalence relations (i.e., strict and weak types) in this paper, and discusses the impact of using different probability estimations on rank equivalence. Section 5 shows that the language model by Ponte and Croft (1998) and the language model by Lafferty and Zhai (2003) will be (strict) rank equivalent if assumptions 1 and 2 of Lafferty and Zhai (2003) are true. Section 6 discusses these assumptions. It shows that some statistical component language models are strict rank equivalent to the log-odds ratio, and some statistical component models are strict rank equivalent to the query likelihood. Section 7 introduces the topic random variable in the probabilistic retrieval models for clarification and for (pseudo) relevance feedback. Finally, Sect. 8 summarizes this paper.

## 2 Marginal probability derivation

In this paper, probabilities are defined according to Kolmogorov's axioms, namely:

- (a) *Normalization axiom*:  $p(\Lambda) = 1$  where  $\Lambda$  is the event space and;
- (b) *Non-negativity axiom*:  $p(e \in \wp(\Lambda)) \in [0,1]$  where  $e$  is an event and  $\wp(\cdot)$  is the powerset of its argument and;
- (c) *Finite additivity axiom*:  $p(a \cup b) = p(a) + p(b)$  for all  $a, b$  in  $\wp(\Lambda)$  such that  $a \cap b = \emptyset$ .

For ease of discussion, we assume that all probabilities are discrete in this paper. We use uppercase letters to denote a set and lowercase letters to denote an element or an event. We further place a suffix after  $p$  to indicate the event space  $\Lambda$  of the probability  $p$  as follows  $p_\Lambda(\cdot)$ . If  $\Lambda = D \times Q$ , where  $\times$  is the Cartesian product, then  $p_\Lambda(\cdot)$  will be the same as  $p_{D \times Q}(\cdot)$ . Since the Cartesian product is commutative,  $p_{D \times Q}(\cdot)$  and  $p_{Q \times D}(\cdot)$  are probabilities of the same event space. A marginal probability can be identified by observing that its argument does not specify all the values of the event space. For example,  $p_{D \times Q}(d)$  is a marginal probability of  $d$  in  $D$  because this probability does not specify a particular query  $q$  in  $Q$ . If a subset  $C$  of events in  $\Omega$  has the probability of one (i.e.,  $p_\Lambda(C) = 1$ ), then it follows that any event  $e$  in  $\Lambda - C$  has zero probability (i.e.,  $p_\Lambda(e) = 0$ ). Therefore, we can write an alternative probability of events that are defined in  $C$  as  $p_C(a)$  where  $a$  is an event in  $C$ . This shows that it is possible to equate probabilities of different event spaces provided one event space is a subset of the other. The event space  $\Lambda$  and event subspace  $C$  can be denoted by two random variables, say  $R1$  and  $R2$ , that take values in those event (sub)spaces, respectively. Hence,  $p_\Lambda(a)$  and  $p_C(a)$  can be interchangeably written as  $p_{R1}(a)$  and  $p_{R2}(a)$ , respectively. In general, one can apply Bayes' theorem to equate two event

spaces where one event space is a subset of the other. However, Bayes’ theorem rests on the marginal probability calculation which is shown to be derived from the probability axioms as follows.

By the normalization axiom, we have  $p_X(X) = 1$ , and  $p_Y(t) = p_X(X) \times p_Y(t)$ . Since the occurrence of  $X$  is independent of  $t$  (because  $X$  always occurs independent of  $t$  as  $p_X(X) = 1$ ),  $p_X(X) \times p_Y(t)$  is the joint probability  $p_{X \times Y}(X, t)$  of  $t$  and  $X$ , i.e.,

$$p_Y(t) = 1 \times p_Y(t) = p_X(X)p_Y(t) = p_{X \times Y}(X, t).$$

Let us assume that  $X$  is exhaustively partitioned into pair-wise disjoint events, say  $\{s\}$ . Using the finite additivity axiom, the (marginal) probability  $p_Y(t)$  of event  $t$  can be expressed as:

$$p_Y(t) = p_{X \times Y}(X, t) = p_{X \times Y} \left( \bigcup_{s \in X} s, t \right) = \sum_{s \in X} p_{X \times Y}(s, t). \tag{1}$$

Based on Eq. 1, the marginal probability calculation does not depend on any uniform probability distribution assumption. Instead, Eq. 1 depends on the following necessary condition to be true:

**Necessary condition 1** For any two different events,  $s_1$  and  $s_2$  in  $X$  (that is exhaustively partitioned),  $s_1 \cap s_2 = \emptyset$  where  $X$  is used as in Eq. 1.

Equation 1 always holds for any event space  $X$  that cross-products with the event space  $Y$  provided that  $p_X(X) = 1$ , which is always true by the normalization axiom. Hence, the only condition of concern is necessary condition 1 when developing retrieval models using Eq. 1.

From the point of view of probabilistic modeling (Robertson 2003), it is possible that some hidden variable (say  $R_1$ ) is unaccounted for in the current model. The event space  $X$  of this hidden variable can combine with the event space  $Y$  of the current probabilistic model in order to form a new probabilistic model that is defined over the event space  $X \times Y$ . The probabilities  $\{p_{X \times Y}(\cdot)\}$  of the existing model become the marginal probabilities of the new probabilistic model. Robertson (2005) found that if the (marginal) probabilities of the current and the new models are uniformly distributed (subject to some structural constraint), then inconsistencies arise when using Eq. 1 to obtain the marginal probabilities, compared with assigning uniform marginal probabilities. However, as we explain later, assigning uniform probabilities is not a logical necessity where as event spaces can always be structured as lists of tuples or equivalently as trees. Therefore, we believe that the problem occurs because multiple probability distributions are assumed to be uniform and not solely because of different event spaces.

It should be noted that the concern of possible problematic ranking caused by the uniform probability distribution assumption as discussed in (Robertson 2005) is not unjustified because some past language models do assume that prior probabilities are uniformly distributed. For example, Berger and Lafferty (1999) considered  $p_{D \times Q}(d)$  as the prior probabilities which were thought to be uniform. However, they also indicated that this probability might vary depending on document length. Therefore,  $p_{D \times Q}(d)$  is not necessarily uniform for some language models. Similarly, one version of the relevance language model by Lavrenko and Croft (2001) also assumed that a prior for the document models is uniform (i.e., Method 2: conditional sampling). However, the other version of the relevance language model does not make such a uniform prior assumption (i.e., Method 1: independent and identically distributed sampling). Therefore, there is a need to investigate this issue in a more thorough manner.

### 3 Clarification

This section clarifies the issues related to the event space (Robertson 2005) and the rank equivalence of Lafferty and Zhai (2003).

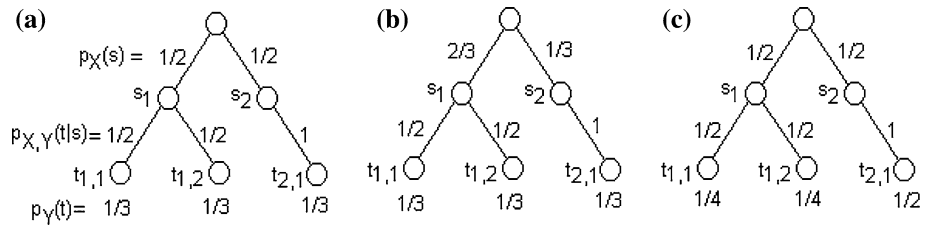
#### 3.1 Event space

Robertson (2005) argues that the event space  $Y$  of  $p_Y(t)$  is not the same as the event space  $X \times Y$  of  $p_{X \times Y}(s, t)$  or the event space  $Y$  of  $p_{Y \times X}(t|s)$  where  $s$  is in  $X$  and  $t$  is in  $Y$ . Therefore, he cautioned those who worked in developing language models for information retrieval. However, it should be noted that both the event space  $Y$  for  $p_Y(t)$  and the event space  $Y$  of  $p_{Y \times X}(t|s)$  are always subsets of the event space  $X \times Y$  (or  $Y \times X$ ) for  $p_{X \times Y}(s, t)$ . This observation enables us to obtain the marginal probability  $p_Y(t)$  from the more detailed tuple events in both  $X$  and  $Y$  as follows.

We provide an interpretation of the marginal probability based on random sampling similar to that of Robertson (2005), but we do not require the marginal probabilities to be evenly distributed (although sometimes they can be). Using Robertson's example, let an event be a tuple  $(s, t)$  in the set  $X \times Y$  of events where  $s$  is a star,  $t$  is a planet,  $X$  is a set of stars and  $Y$  is a set of planets. In this paper, we ignore the magnetic field in (Robertson 2005) because if we know the particular star/planet, then we know whether the star/planet has or has not a magnetic field. Let us randomly pick event tuples from an urn  $U$  that contains the following three tuples:  $(s_1, t_{11})$ ,  $(s_1, t_{12})$  and  $(s_2, t_{21})$ . Assume that all these tuples are equally likely to be randomly selected. The marginal probabilities for events in  $Y$  are  $p_{X \times Y}(t_{11}) = p_{X \times Y}(t_{12}) = p_{X \times Y}(t_{21}) = 1/3$ . We can obtain this probability by sampling the tuples from  $U$  and ignoring their first elements (i.e.,  $X$ , or  $s_1$  and  $s_2$ ) of the selected tuples. Likewise, we can obtain  $p_{X \times Y}(s_1) = 2/3$  and  $p_{X \times Y}(s_2) = 1/3$  by randomly sampling the tuples from  $U$  and ignoring their second elements (i.e.,  $Y$ , or  $t_{11}$ ,  $t_{12}$  and  $t_{21}$ ) of the selected tuples. Therefore, we may obtain the marginal probabilities based on randomly sampling tuples over the cross product space of  $X$  and  $Y$ , without the need to sample the subspace  $X$  first and then  $Y$ , or vice versa (although we can do so).

As in Robertson (2005), the previous example shows that the probability  $p_{X \times Y}(t) = 1/3$  is based on randomly picking a tuple of star and planet with equal likelihood. However, Robertson (2005) indicated that this probability calculated using the marginal probability will not be  $1/3$  if we assume that randomly picking a star is equally likely, and if randomly picking a planet of a given star is also equally likely. Since the marginal probability calculation is based on Kolmogorov's axioms, and we assume that necessary condition 1 holds, the calculation must be accepted as ground truth. The only possibility to reconcile  $p_{X \times Y}(t)$  and its marginal probability calculation using Eq. 1 is *not to assume* that randomly picking a star is equally likely nor randomly picking a planet given a star is necessarily equally likely. In fact, there is no logical necessity that probability distributions are uniform. Language models typically leave some of the probability distributions unspecified and later estimate the probability distributions rather than assuming the distributions are uniform. While many language models are Bayesian probabilistic models, the Bayesian prior probabilities (of these language models) are not necessarily uniformly distributed.

We summarize our discussion using probability generation trees as in Fig. 1 that shows three different examples. Figure 1a is Robertson's example where  $p_{X \times Y}(t)$  is not the same as calculating it based on the marginal probability using Eq. 1. Figure 1b is our example that is consistent with the marginal probability calculation using Eq. 1, but the probability



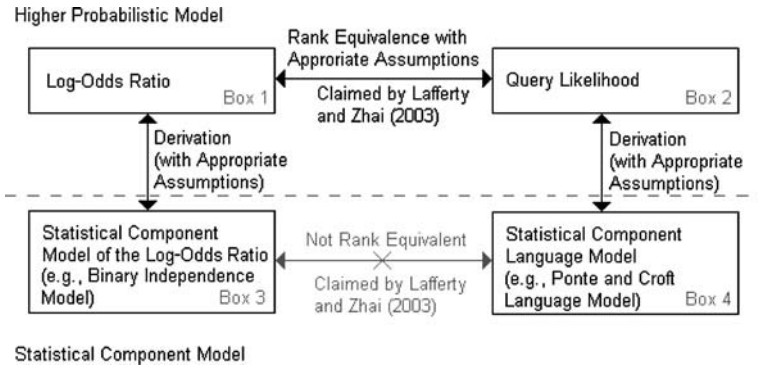
**Fig. 1** (a) Robertson’s example with uniform probability distributions for randomly picking stars with probabilities  $\{p_{X \times Y}(s)\}$ , planets with probabilities  $\{p_{X \times Y}(t)\}$  and planets given a chosen star with probabilities  $\{p_{X \times Y}(t|s)\}$ . The problem is that the product of  $p_{X \times Y}(s)$  and  $p_{X \times Y}(t|s)$  is inconsistent with the marginal probability of  $p_{X \times Y}(t) = 1/3$ . (b) Our example that is consistent with Eq. 1 has equal likelihood of picking a planet at random, but the probability (i.e.,  $p_{X \times Y}(s)$ ) of picking a star is not uniformly distributed. (c) Our example that is consistent with Eq. 1 has equal likelihood of picking a star at random, but the probability (i.e.,  $p_{X \times Y}(t)$ ) of picking a planet is not uniformly distributed

of picking a star randomly is not uniformly distributed (as explained earlier). Figure 1c is our other example that is consistent with the marginal probability calculation using Eq. 1, but the probability of picking a planet randomly is not uniformly distributed. This shows that by not requiring that all probability distributions be uniformly distributed, the problem raised by Robertson may not occur. In fact, there is no logical necessity that the probability distributions are uniformly distributed.

Apart from assuming probabilities are uniformly distributed, there is a possibility that asserting assumptions may lead to data inconsistencies for some probability distributions (but not necessarily for all possible probability distributions). If we take Robertson’s example (i.e., planets and stars), it is possible that estimating the probabilities resulted in uniform probability distributions and data inconsistencies can arise. The underlying problem seems to be the assumptions made as indicated earlier by Cooper (1995), instead of the event space as Robertson suggested, because the corollary, lemma and theorem derived from the probability axioms are true once the axioms are accepted while the assumptions can be invalid. For example, if our model assumes that  $p_{X \times Y}(s) = 1/2$  and  $p_{X \times Y}(t|s) = 1/2$ , it will predict that  $p_{X \times Y}(t) = 1/4$  or  $1/2$  depending on  $t$ . If our relative frequency estimate of  $p_{X \times Y}(t)$  is  $1/3$ , then there are data inconsistencies between the predicted and actual values. We believe that the problem of data inconsistencies is our assumption about certain values that  $p_{X \times Y}(s)$  or  $p_{X \times Y}(t|s)$  should take, instead of the different event spaces. In summary, it seems more appealing to question whether the assumptions asserted lead to inconsistencies for some probability distributions. Let us start by revisiting the rank equivalence claimed by Lafferty and Zhai (2003).

### 3.2 Rank equivalence of Lafferty and Zhai

Figure 2 depicts the claim by Lafferty and Zhai (2003) on the rank equivalence between the query likelihood and the log-odds ratio (i.e., between Box 1 and 2 in Fig. 2). Separately, these probabilistic ranking functions derive the corresponding operational ranking functions of their statistical component models, so they belong to the probabilistic models at a level higher (or more general) than the statistical component model. These probabilistic models are more general in the sense that they have fewer details (e.g., assumptions and component specifications) than the statistical component models. Also, they may be related to the probability ranking principle (Robertson 1977) that forms the unified basis



**Fig. 2** Rank equivalence between higher probabilistic models and between their related statistical component models

for ranking documents [even for interactive retrieval (Fuhr 2008)] and for modeling the evaluation of ranked lists (Wu et al. 2007).

Lafferty and Zhai (2003) claimed that the statistical component models of the log-odds ratio are not rank equivalent to the statistical component language model (i.e., between Box 3 and 4) because their probabilities are estimated differently. Apart from the issues about probability estimation for establishing rank equivalence, the derivations and rank equivalence relation assert different sets of assumptions. The impact of asserting multiple sets of assumptions is important because the rank equivalence claimed by Lafferty and Zhai (2003) will be purely theoretical if there does not exist a statistical component model that is rank equivalent to both the log-odds ratio and the query likelihood. In other words, if the log-odds ratio is not rank equivalent to statistical component language models (i.e., Box 1, 2 and 4 are not rank equivalent), or vice versa (i.e., Box 2, 1 and 3 are not rank equivalent), then the rank equivalence for the higher probabilistic models has no direct practical significance. This practical issue about rank equivalence is discussed in Sect. 6.

#### 4 Rank equivalence

Showing two retrieval models are rank equivalence is an important theoretical task as it may avoid the use of Eq. 1 without the need to assert necessary condition 1. Its practical significance may imply that (a) there is no need to compare such models for better retrieval effectiveness, and that (b) the model that may be carried out by a more time-space efficient algorithm is preferred over slower ones. In this section, we discuss a motivating example of rank equivalence, identify different types of rank equivalence relations, and discuss the impact of different probability estimation methods on them. This discussion differs from the equivalence of BIM and the novel language model by Roelleke and Wang (2006) because (1) the BIM and their language model use the same probability estimation method (i.e., relative frequency), (2) their language model ranks documents by a ratio of two query likelihoods instead of a single query likelihood, (3) their equivalence relation does not assert assumptions of Lafferty and Zhai, and (4) their equivalence relation makes an assumption that the contribution to the retrieval status value from relevant documents is constant.

### 4.1 Motivating example

Lafferty and Zhai (2003) use rank equivalence to reconcile different probabilistic models. For example, the two sides of the probabilities in the following equation are related by the definition of conditional probability,

$$p_{\Omega}(d|q, r) = \frac{p_{\Omega}(q|d, r)p_{\Omega}(d, r)}{p_{\Omega}(q, r)}, \tag{2}$$

where  $\Omega = D \times Q \times R$  (as in Wu et al. 2007),  $D$  is a set of documents,  $Q$  is a set of queries,  $r$  is the relevance value,  $\bar{r}$  is the non-relevance value,  $R = \{r, \bar{r}\}$ ,  $d$  is a document in  $D$ , and  $q$  is a query in  $Q$ . We are interested in Eq. 2, because (1)  $p_{\Omega}(d|q, r)$  is related to the evaluation model (Wu et al. 2007) for the probability ranking principle (PRP as in Robertson 1977), (2)  $p_{\Omega}(q|d, r)$  is the query likelihood of the language model (e.g., Miller et al., 1999), and (3) this rank equivalence illustrates that assumptions 1 and 2 of Lafferty and Zhai are sufficient conditions in Sect. 6. Consequently, Eq. 2 can be considered as reconciling two different probabilistic retrieval models.

Implicitly, the conditional probability,  $p_{\Omega}(d|q, r)$ , is valid for  $p_{\Omega}(q, r) > 0$ . Interestingly,  $p_{\Omega}(d|q, r)$  is rank equivalent (Wu et al. 2007) to the log-odds ratio of the RSJ model by  $p_{\Omega}(r|d, q)$  assuming that  $p_{\Omega}(d, q)$  is a constant. However, this assumption is not true for some language models because  $p_{\Omega}(d, q)$  of these language models is not a constant. Otherwise,  $p_{\Omega}(q|d)$  becomes problematic for ranking as it is independent of  $q$ . This motivates us to find another alternative to reconcile different probabilistic models using rank equivalence instead of the equality as follows.

Suppose that the probability  $p_{\Omega}(d|q, r)$  on the left hand side (LHS) of Eq. 2 is estimated by one model with a set  $S_1$  of assumptions and the probability on the right hand side (RHS) of Eq. 2 is estimated by another model with another set  $S_2$  of assumptions. While the two probabilistic models may not assign uniform probability distributions, it is possible that, for some data set, the probability distributions estimated by the two retrieval models can lead to inconsistencies as suggested by Robertson (2005). In general, assuming a certain probabilistic relationship holds can lead to the exclusion of certain data distributions. In other words, some data distributions contradict with the assumptions asserted. This problem does not appear in the derivation by Lafferty and Zhai (2003) because the log-odds ratio is equivalent to query likelihood of the language model on the basis of producing the same ranking output, i.e., the orderings of the probabilities/similarities on both sides of the equation are the same. Instead of Eq. 2 in the previous instructive example, we consider an alternative type of rank equivalence between the evaluation model (Wu et al. 2007) for PRP and the language model

$$p_{\Omega}(d|q, r) \stackrel{rank}{=} \frac{p_{\Theta}(q|d, r)p_{\Theta}(d, r)}{p_{\Theta}(q, r)},$$

where  $\Omega$  is the event space as in (Wu et al. 2007),  $\Theta$  is the event space of a language model,  $\stackrel{rank}{=}$  is the binary relation operator for rank equivalence (as in Lafferty and Zhai 2003). Because these are two different event spaces, the RHS of the previous relation may relate to the PRP because only the evaluation model assumes that  $p_{\Omega}(d, q)$  is a constant, and not the language model on the RHS. We call such rank equivalence between two different probabilistic models defined over two different event spaces weak rank equivalence. Weak rank equivalence can be identified by observing that the event spaces on its two sides are different.

**Table 1** Hypothetical ranking results of three documents using two different models based on the probabilities on the different sides of the conditional probability in Eq. 2

Rank	Document	Model A	Model B			
		$p_{\Omega}(d q, r)$	$\frac{p_{\Theta}(q d, r)p_{\Theta}(d, r)}{p_{\Theta}(q, r)}$	$p_{\Theta}(q d, r)$	$p_{\Theta}(q, r)$	$p_{\Theta}(d, r)$
1	$d_1$	0.70	0.40	0.10	0.10	0.40
2	$d_2$	0.20	0.35	0.10	0.10	0.35
3	$d_3$	0.10	0.25	0.10	0.10	0.25

For illustration, Table 1 shows an example of rank equivalence between Model A and Model B. Here, we suppose that Model A and Model B use different methods to estimate the probabilities. The probabilities in bold and underlined font are used for ranking the documents. In this example, the probabilities of Model A for ranking are different from the corresponding probabilities of Model B for ranking, but the assigned ranks of the documents are the same for both Model A and B. For Model B,  $p_{\Theta}(q, r)$  is not necessary for ranking, but it is included here for clarity of presentation. For claiming that Model A is rank equivalent to Model B, the ranking using Model A should be the same as that using Model B for all queries and for all collections. To make such a claim, therefore, there is a practical problem of finding such probabilities or such probability estimation methods that lead to rank equivalence between Model A and B, for all queries and for all collections.

One possible approach that may guarantee two probabilistic models are rank equivalent for all queries and for all collections assumes that the probabilities on both sides of a rank equivalence relation are derived from a single event space of an underlying (random) experiment as suggested by Robertson (2005). We call this type of rank equivalence strict rank equivalence which can be identified easily because the probabilistic models participating in the rank equivalence relation have the same event space. Even when the probabilistic models are strict rank equivalent, there may be an additional problem that these models estimate probabilities differently from the same random experiment as indicated by Lafferty and Zhai (2003). If the probabilities are measured differently for these different probabilistic models operating with the same event space, will these probabilistic models not necessarily be rank equivalent?

#### 4.2 Problem with different probability estimation methods

To answer the previous question, one method to show that these models are strict rank equivalent is to prove such equivalence mathematically by finding how the probability estimates of both sides of the rank equivalence are related. For illustration, suppose a probability that is estimated by relative frequency is denoted by the superscript  $f$ . For example,  $p_{\Omega}^f(d|q, r)$  is the relative frequency estimate of  $p_{\Omega}(d|q, r)$ . Likewise, suppose a probability that is estimated by additive smoothing (e.g., Lidstone 1920) is denoted by the superscript  $a$ . In Eq. 2, suppose that the probability on the LHS is estimated by relative frequency and the probabilities on the RHS are estimated by additive smoothing. Since these probabilities estimated differently are not expected to be the same, they are not expected to be strict rank equivalent. However, when we want to claim that the language model is rank equivalent to the evaluation model (Wu et al. 2007), they are required to be strict rank equivalent by demanding



$$p_{\Omega}^f(d|q, r) \stackrel{\text{rank}}{=} \frac{p_{\Omega}^a(q|d, r) \cdot p_{\Omega}^a(d, r)}{p_{\Omega}^a(q, r)} \stackrel{\text{rank}}{=} p_{\Omega}^a(q|d, r) \cdot p_{\Omega}^a(d, r). \tag{3}$$

The probabilities estimated by additive smoothing (e.g., Lidstone 1920) are linearly related to their relative frequency estimates:

$$p_{\Omega}^a(x) \equiv \frac{f(x) + \delta}{N + \delta \cdot V} = k \cdot p_{\Omega}^f(x) + \eta$$

where  $f(x)$  is the occurrence frequency of  $x$ ,  $N$  is the total frequency,  $V$  is the vocabulary size,  $\delta$  is a small additive constant for smoothing,  $p_{\Omega}^f(x) = f(x)/N$ ,  $k = N/(N + \delta \cdot V)$  and  $\eta = \delta/(N + \delta \cdot V)$ . The RHS of relation 3 can be rewritten as

$$p_{\Omega}^a(q|d, r) \cdot p_{\Omega}^a(d, r) = k^2 \cdot p_{\Omega}^f(q|d, r) \cdot p_{\Omega}^f(d, r) + \eta \cdot k \cdot (p_{\Omega}^f(q|d, r) + p_{\Omega}^f(d, r)) + \eta^2.$$

Substituting the above equation into relation 3 shows that the ranking by  $p_{\Omega}^a(q|d, r) \cdot p_{\Omega}^a(d, r)$  is not rank equivalent to  $p_{\Omega}^f(d|q, r)$  in general. However, for some special case, relation 3 holds. If  $p_{\Omega}^f(d, r)$  is a constant, then

$$p_{\Omega}^a(q|d, r) \cdot p_{\Omega}^a(d, r) = \left[ k^2 \cdot p_{\Omega}^f(d, r) + \eta \cdot k \right] \cdot p_{\Omega}^f(q|d, r) + \eta \cdot k \cdot p_{\Omega}^f(d, r) + \eta^2 \\ \stackrel{\text{rank}}{=} p_{\Omega}^f(q|d, r) \cdot p_{\Omega}^f(d, r) \stackrel{\text{rank}}{=} p_{\Omega}^f(d|q, r).$$

Here, we want to show that even different estimation methods may lead to strict rank equivalence under specific conditions, albeit the current example condition (i.e.,  $p_{\Omega}^f(d, r)$  is a constant) does not invoke the assumptions of Lafferty and Zhai, and it may not be realistic in practice. While it is obvious that two different estimation methods may lead to different ranking, it might be possible for some (unknown) probabilistic models estimated using different estimation methods to be strict rank equivalent. Given this possibility, there may be a reason other than the use of different probability estimation methods, which explains why the statistical component language model is not strict rank equivalent to the statistical component model of the log-odds ratio.

### 4.3 Equivalence between probabilistic and statistical component models

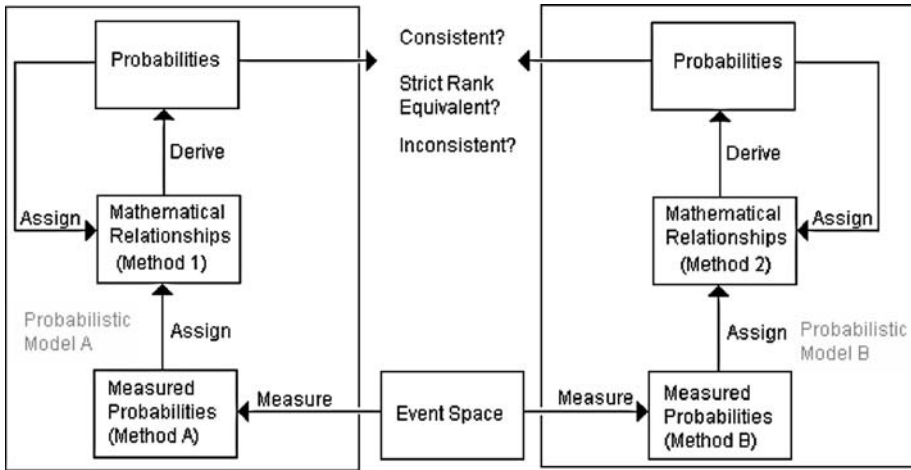
When a statistical component model is derived from a probabilistic model at a higher level, such a statistical component model may have data inconsistencies (e.g., Cooper 1995). For example,  $p_{\Omega}(d|q, r)$  can be rewritten in terms of its components

$$p_{\Omega}(d|q, r) \equiv p_{\Omega}(A_{d1}, \dots, A_{dn}|q, r), \tag{4}$$

where a document is rewritten as a set of  $n$  attributes from  $A_{d1}$  to  $A_{dn}$ . If we assume that attributes are conditionally independent, then

$$p_{\Omega}(A_{d1}, \dots, A_{dn}|q, r) = \prod_{i=1}^n p_{\Omega}(A_{di}|q, r). \tag{5}$$

While we assumed identity 4 holds, it may not hold in practice. We clarify this problem using a schematic diagram as shown in Fig. 3. More specifically, probabilistic model A and its statistical component version B in Fig. 3 correspond to the LHS and the RHS of identity 4, respectively. Suppose model A directly measures  $p_{\Omega}(d|q, r)$  by relative frequency counting, and suppose method 1 of model A assigns this measured probability as the



**Fig. 3** A schematic diagram of strict rank equivalence between two probabilistic/statistical models

probability on the LHS of identity 4. Suppose model B measures  $p_{\Omega}(A_{di}|q, r)$  by relative frequency counting, and  $p_{\Omega}(A_{d1}, \dots, A_{dn}|q, r)$  is calculated by Eq. 5. In most practical cases, the following does not hold:

$$p_{\Omega}^f(d|q, r) = \prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r).$$

This is partly because

$$\sum_{d \in D} p_{\Omega}^f(d|q, r) = 1 \neq \sum_{d \in D} \prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r).$$

A common remedy is to use Bayes’ theorem (as done for language models, e.g., Hiemstra 2002):

$$p_{\Omega}^f(d|q, r) = \frac{\prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r)}{\sum_{d \in D} \prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r)}.$$

However, this assumes that necessary condition 1 holds for different documents. An alternative without assuming that such a condition holds uses strict rank equivalence to link these probabilities:

$$p_{\Omega}^f(d|q, r) \stackrel{rank}{=} \prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r).$$

Despite that we use the same probability estimation method, and we use a probabilistic model and its derived statistical component version, there will still be no guarantee that the equivalence relation will hold for all queries and for all collections if we measure the probabilities on its LHS and calculate the derived probability ratio using measured probabilities on its RHS. To avoid the inconsistency between the probabilistic model and its statistical component model, the probability of the former probabilistic model is assigned with the probability derived from its latter statistical component model. For our previous example, the probability on the LHS of identity 4 is not estimated from the event space

even though it can be done. Instead, this probability is assigned with the calculated probabilities that are derived from the measured probabilities on the RHS of identity 4. In this case, the probability on the LHS of identity 4 represents the model’s belief that a certain event will happen. This option to avoid inconsistencies by probability assignment is not available when two statistical component models are strict rank equivalent. This is because the two statistical component models need to estimate their probabilities from the same event space. This suggests that even if the same probability estimation method is used for both statistical component models, these models may not be strict rank equivalent.

#### 4.4 Equivalence between statistical component models

Let us discuss an instructive example to appreciate the difficulty of making two statistical component models rank equivalent. Suppose we have two such models for the LHS and RHS of Eq. 2:

$$p_{\Omega}(d|q, r) \stackrel{rank}{=} \prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r)$$

$$p_{\Omega}(q|d, r) \cdot p_{\Omega}(d, r) \stackrel{rank}{=} p_{\Omega}^f(d, r) \cdot \prod_{i=1}^n p_{\Omega}^f(A_{qi}|d, r),$$

where  $A_{qi}$  is the  $i$ -th attribute of the query  $q$ . The probabilities on the LHS of the above strict rank equivalence relations are the probabilities of the probabilistic models without any specified probability estimation methods. The probabilities on the RHS of the above strict rank equivalence relations are the measured probabilities that are estimated by the relative frequency method  $f$ . When these probabilities are substituted into equation 2, and the equality is changed to strict rank equivalence, i.e.,

$$\prod_{i=1}^n p_{\Omega}^f(A_{di}|q, r) \stackrel{rank}{=} p_{\Omega}^f(d, r) \cdot \prod_{i=1}^n p_{\Omega}^f(A_{qi}|d, r),$$

it becomes difficult to prove the above relation is true for all queries and for all collections. The reason is not because different estimation methods are used. Actually, these estimation methods are identical on both sides of the above relation. The reason is that the  $i$ -th attribute of the document,  $A_{di}$ , and the  $i$ -th attribute of the query  $q$ ,  $A_{qi}$ , are the same attribute but of different objects. Even if we make the same attributes of different objects to have the same value for strict rank equivalence, i.e.,  $A_{di} = A_{qi}$ , the conditioning part of these two probabilities will be different. For the conditional probabilities on the LHS, we count the frequencies for a single query  $q$ , where as we count the frequencies of a single document  $d$  for the conditional probabilities on the RHS. Consequently, it is difficult to prove that the previous strict rank equivalence relation holds for all queries and for all collections. This discussion will be inconsistent with the remark by Laffery and Zhai (2003) about “the component models (that are) estimated quite differently” (added our bracketed text in the quotation) if this remark solely refers to the different probability estimation methods used (e.g., additive smoothing (Lidstone 1920) or absolute discounting (Ney et al. 1994)). However, it will be consistent with their remark if “estimated quite differently” means getting different probability values because of the use of different measured probabilities and/or different probability estimation methods.

Roelleke and Wang (2006) show that it is possible to create a novel statistical component language model using query likelihood ratio that is rank equivalent to the BIM without using the rank equivalence of Laffery and Zhai (2003). This is done using models that employ the

same probability estimation methods. Is it possible that two different statistical component models will be strict rank equivalent to each other if they use different probability estimation methods? To illustrate this possibility, suppose we are given that

$$\prod_{i=1}^n p_{\Omega}^F(A_{di}|(A_{q1}, \dots, A_{qn}), r) \stackrel{\text{rank}}{=} p_{\Omega}^F(d, r) \prod_{i=1}^n p_{\Omega}^F(A_{qi}|(A_{d1}, \dots, A_{dn}), r),$$

is there a probability estimation method  $b$  such that

$$\prod_{i=1}^n p_{\Omega}^F(A_{di}|(A_{q1}, \dots, A_{qn}), r) \stackrel{\text{rank}}{=} p_{\Omega}^b(d, r) \prod_{i=1}^n p_{\Omega}^b(A_{qi}|(A_{d1}, \dots, A_{dn}), r)?$$

Suppose the probability estimation method,  $b$ , is related to its corresponding relative frequency estimation method  $F$ :

$$p_{\Omega}^b(x) \equiv c \cdot p_{\Omega}^F(x)^k$$

where  $c$  is a normalization constant and  $k$  is a parameter of this estimation method. If  $k > 1$ , then smaller relative frequency probability estimates will be even smaller. This reduces the effect of over-estimation by the relative frequency estimation method (i.e., the maximum likelihood estimator) when the frequency count is small. When the frequency count is zero, this method maintains that the probability is zero. To avoid zero probabilities, 0.5 is added to the frequency counts and the total frequency as in the binary independence model (BIM of Robertson and Spärck 1976), i.e.,

$$p_{\Omega}^F(x) \equiv \frac{f(x) + 0.5}{N + 0.5}.$$

Taking the logarithm, the probabilities estimated by these two methods are related:

$$\log p_{\Omega}^b(x) = \log c + k \log p_{\Omega}^F(x).$$

Using this equality, the query likelihood estimated by method  $b$  is rank equivalent to the query likelihood estimated by relative frequency, i.e.,

$$\begin{aligned} & \log p_{\Omega}^b(d, r) + \sum_{i=1}^n \log p_{\Omega}^b(A_{qi}|(A_{d1}, \dots, A_{dn}), r) \\ &= k \log p_{\Omega}^F(d, r) + \log c_1 + n \log c_2 + k \sum_{i=1}^n \log p_{\Omega}^F(A_{qi}|(A_{d1}, \dots, A_{dn}), r) \\ &\stackrel{\text{rank}}{=} \log p_{\Omega}^F(d, r) + \sum_{i=1}^n \log p_{\Omega}^F(A_{qi}|(A_{d1}, \dots, A_{dn}), r), \end{aligned}$$

where  $c_1$  is the normalization constant for estimating  $p_{\Omega}^b(d, r)$ ,  $c_2$  is the normalization constant for estimating  $p_{\Omega}^b(A_{qi}|A_{d1}, \dots, A_{dn}, r)$  and  $k$  is a parameter of estimation method  $b$ . In general, we derive

$$\begin{aligned} p_{\Omega}(r|d, q) &\stackrel{\text{rank}}{=} \frac{p_{\Omega}(d|q, r)}{p_{\Omega}(d, q)} \stackrel{\text{rank}}{=} \frac{\prod_{i=1}^n p_{\Omega}^F(A_{di}|q, r)}{p_{\Omega}^F(d, q)} \\ &\stackrel{\text{rank}}{=} \frac{p_{\Omega}^b(d, r) \cdot p_{\Omega}^b(q|d, r)}{p_{\Omega}^b(d, q)} \stackrel{\text{rank}}{=} \frac{p_{\Omega}^b(d, r) \times \prod_{i=1}^n p_{\Omega}^b(A_{qi}|d, r)}{p_{\Omega}^b(d, q)}. \end{aligned}$$

This illustrates that it is possible that two statistical component models are rank equivalent even though the probability estimation methods are different. Therefore, different probability estimations do not necessarily lead to different ranking.

Strict rank equivalence is related to weak rank equivalence under some special circumstances. Suppose we have two retrieval models,  $A$  and  $B$ . They rank documents by the probabilities,  $p_{\Omega,A}(x)$  and  $p_{\Omega,B}(y)$ , respectively, where  $x$  and  $y$  are some expression of (conditional) events, and  $p_{\Omega,C}(\cdot)$  is model  $C$ 's probability that is defined over the event space  $\Omega$ . The following obvious remark can be observed:

**Remark 1** If  $p_{\Omega,A}(x) \stackrel{\text{rank}}{=} p_{\Omega,B}(y)$  and the event space  $\Theta$  is a superset of the event space  $\Omega$ , and  $p_{\Omega,B}(y) \stackrel{\text{rank}}{=} p_{\Theta,B}(y)$ , then  $p_{\Omega,A}(x) \stackrel{\text{rank}}{=} p_{\Theta,B}(y)$ .

For  $\Theta \supset \Omega$ ,  $p_{\Theta,B}(y)$  is a marginal probability since  $y$  is in  $\Omega$  and so it specifies only a subset of attributes in  $\Theta$ . This remark illustrates that weak rank equivalence is possible under specific conditions.

### 5 Combining other models

In this section, we show that the language model by Ponte and Croft (1998) is rank equivalent to the language model in Lafferty and Zhai (2003) using their assumptions 1 and 2. This enables us to deal with both types of language models together in subsequent sections. In addition, we show that the log-odds ratio is rank equivalent to minus the probability of non-relevance (Wu et al. in press) that relates to the TF-IDF term weights.

As Lafferty and Zhai (2003) have already shown, both the log-odds of the RSJ model and the query likelihood of language models can be thought of as estimating and/or ranking on the basis of the log-odds ratio,  $O_{\Omega}(r|q, d)$  (Fuhr 1992), that is defined over the event space  $\Omega$ :

$$O_{\Omega}(r|d, q) \equiv \log \left[ \frac{p_{\Omega}(r|q, d)}{p_{\Omega}(\bar{r}|q, d)} \right].$$

Lafferty and Zhai (2003) further show that

$$O_{\Omega}(r|q, d) \stackrel{\text{rank}}{=} \log p_{\Omega}(q|d, r), \tag{6}$$

where  $\stackrel{\text{rank}}{=}$  denotes the rank equivalence relation. Equation 6 assumes that the following two statistical (conditional) independence assumptions (Lafferty and Zhai 2003) hold without asserting any assumption of uniform probability distributions:

**Assumption 1**  $p_{\Omega}(d, q|\bar{r}) = p_{\Omega}(d|\bar{r}) \times p_{\Omega}(q|\bar{r})$ .

**Assumption 2**  $p_{\Omega}(d, r) = p_{\Omega}(d) \times p_{\Omega}(r)$ .

For binary relevance, the following holds:

$$p_{\Omega}(d) = p_{\Omega}(d, r) + p_{\Omega}(d, \bar{r}). \tag{7}$$

Together with assumption 2, the previous condition logically implies  $p_{\Omega}(d, \bar{r}) = p_{\Omega}(d) \times p_{\Omega}(\bar{r})$ , because, from Eq. 7, we have:

$$\begin{aligned} p_{\Omega}(d, \bar{r}) &= p_{\Omega}(d) - p_{\Omega}(d, r) \text{ (applying equation 2)} \\ &= p_{\Omega}(d) - p_{\Omega}(d) \cdot p_{\Omega}(r) \\ &= p_{\Omega}(d) \cdot (1 - p_{\Omega}(r)) \\ &= p_{\Omega}(d) \times p_{\Omega}(\bar{r}). \end{aligned} \tag{8}$$

Hence, Eq. 8 is a consequence of assumption 2 in Lafferty and Zhai (2003).

The query likelihood function (Lafferty and Zhai 2003) is defined as  $p_{\Omega}(q|d, r)$ . This has been used as the basis of ranking by a number of language models (e.g., Miller et al. 1999). Alternatively, another ranking function,

$$p_{\Omega}(q|d)$$

is also commonly found in a variety of language models (e.g., Ponte and Croft 1998; Hiemstra 1998; Berger and Lafferty 1999; Song and Croft 1999; Hiemstra 2000; Spärck et al. 2003; Zhai and Lafferty 2004; Gao et al. 2004) for information retrieval. We will show that these two ranking functions of language models are rank equivalent to each other if assumptions 1 and 2 are true. Given that:

$$p_{\Omega}(q, d) = p_{\Omega}(q, d, r) + p_{\Omega}(q, d, \bar{r}), \tag{9}$$

we rewrite the query likelihood:

$$\begin{aligned} p_{\Omega}(q|d, r) &= \frac{p_{\Omega}(q, d) - p_{\Omega}(q, d, \bar{r})}{p_{\Omega}(d, r)} \\ &= \frac{p_{\Omega}(q, d) - p_{\Omega}(q, d|\bar{r}) \cdot p_{\Omega}(\bar{r})}{p_{\Omega}(d, r)} \text{ (applying assumption 1)} \\ &= \frac{p_{\Omega}(q, d) - p_{\Omega}(d|\bar{r}) \cdot p_{\Omega}(q|\bar{r}) \cdot p_{\Omega}(\bar{r})}{p_{\Omega}(d, r)} \text{ (applying assumption 2)} \\ &= \frac{p_{\Omega}(q|d)}{p_{\Omega}(r)} - \frac{p_{\Omega}(d, \bar{r}) \cdot p_{\Omega}(q|\bar{r})}{p_{\Omega}(d) \cdot p_{\Omega}(r)}. \end{aligned}$$

Applying Eq. 8, the previous equation is simplified to

$$p_{\Omega}(q|d, r) = \frac{p_{\Omega}(q|d)}{p_{\Omega}(r)} - \frac{p_{\Omega}(d) \cdot p_{\Omega}(\bar{r}) \cdot p_{\Omega}(q|\bar{r})}{p_{\Omega}(d) \cdot p_{\Omega}(r)} = \frac{p_{\Omega}(q|d)}{p_{\Omega}(r)} - \frac{p_{\Omega}(\bar{r}) \cdot p_{\Omega}(q|\bar{r})}{p_{\Omega}(r)}.$$

Since the second term on the RHS of the previous equation only depends on the query  $q$ , we simplify the previous equation further using strict rank equivalence,

$$p_{\Omega}(q|d, r) = \frac{p_{\Omega}(q|d)}{p_{\Omega}(r)} - \frac{p_{\Omega}(\bar{r})p_{\Omega}(q|\bar{r})}{p_{\Omega}(r)} \stackrel{rank}{=} \frac{p_{\Omega}(q|d)}{p_{\Omega}(r)} \stackrel{rank}{=} p_{\Omega}(q|d).$$

Since logarithm is an order-preserving transformation, it follows that the Lafferty and Zhai model (2003) and Ponte and Croft model (1998) are strict rank equivalent with the log-odds ratio:

$$O_{\Omega}(r|q, d) \stackrel{rank}{=} \log p_{\Omega}(q|d, r) \stackrel{rank}{=} \log p_{\Omega}(q|d). \tag{10}$$

provided that assumptions 1 and 2 hold (an alternative derivation is shown in Appendix I), and Eq. 9 holds. Even though the relevance random variable does not appear in the Ponte and Croft (1988) model, the information in the relevance random variable might be embedded in the implicit assumptions of the model.

Recently, Wu et al. (in press) showed that the TF-IDF term weight is derived from the probabilistic retrieval model that ranks on the basis of  $-\log p_{\Omega}(\bar{r}|d, q)$ . This ranking is strict rank equivalent to the log-odds ratio:

$$O_{\Omega}(r|d, q) \stackrel{rank}{=} \log p_{\Omega}(r|d, q) \stackrel{rank}{=} -\log p_{\Omega}(\bar{r}|d, q), \tag{11}$$

given that  $p_{\Omega}(r|d, q) + p_{\Omega}(\bar{r}|d, q) = 1$  (Eq. 1.1. in Lafferty and Zhai 2003).

## 6 Tackling the assumptions

Assumptions 1 and 2 bring the work by Lafferty and Zhai (2003) and by Ponte and Croft (1998) on the same ground of the log-odds ratio and it is time to question whether these assumptions are reasonable. Intuitively, assumption 2 is reasonable since one does not expect that documents and relevance are statistically dependent without specifying the query. That is, the marginal probability,  $p_{\Omega}(d, r)$ , is determined by the product of two marginal probabilities,  $p_{\Omega}(d)$  and  $p_{\Omega}(r)$ . Note that assumption 2 does not state the likelihood of retrieving a document is uniformly distributed. For example, documents that have more terms are more likely to be matched and therefore retrieved (but whether these documents are relevant or not are unknown). Although assumption 2 implies that  $p_{\Omega}(r|d_i) = p_{\Omega}(r|d_j) = p_{\Omega}(r)$ ,  $p_{\Omega}(r)$  may be different from  $p_{\Omega}(\bar{r})$ , and  $p_{\Omega}(r, d_i)$  may be different from  $p_{\Omega}(r, d_j)$  because  $p_{\Omega}(d_i)$  may be different from  $p_{\Omega}(d_j)$ . Note that  $p_{\Omega}(r|d)$  is a marginal probability that is aggregated for all queries given a document  $d$ , so this probability has no relationship with a particular query  $q$ . It should not be confused with  $p(L|d)$  of the BIM (Spärck et al. 2000) where  $L$  maps to relevance. This is because BIM is defined for a single query, so  $p(L|d)$  of the BIM is actually  $p_{\Omega}(r|d, q)$  (and not the marginal probability  $p_{\Omega}(r|d)$ ).

Assumption 1 is specified by conditional independence that is more complex to justify. First, we note that many probabilistic retrieval models have asserted assumption 1 to hold (e.g., Yu and Salton 1976; Robertson et al. 1982a, b), including the binary independence model of Robertson and Spärck (1976), according to Cooper (1995). The assertion of assumptions 1 and 2 together does not fall into the data inconsistency problem discussed by Cooper (1995) and discovered by Robertson (1974) because the conditional probability in assumption 2 here only assumed the independence between documents and relevance, instead of documents and queries (as in Cooper, 1995). Second, if the following assumption holds:

**Assumption 3**  $p_{\Omega}(d, q, \bar{r}) \times p_{\Omega}(\bar{r}) = p_{\Omega}(d, \bar{r}) \times p_{\Omega}(q, \bar{r})$

then we can derive assumption 1 instead of asserting it. This is obtained by dividing  $p_{\Omega}(\bar{r})^2$  on both sides of assumption 3, which does not need to assume that  $p_R(\bar{r}) > 0$ .

For the probabilistic models at a higher level (Fig. 2), we show that

**Sufficient condition 1** Assumptions 1 and 2 of Lafferty and Zhai are sufficient conditions for the rank equivalence between the log-odds ratio and the query likelihood.

Since the rank equivalence between the log-odds ratio and the query likelihood is already established by Lafferty and Zhai the remaining task is to show that these assumptions are not necessary conditions. This is done by establishing the strict rank equivalence using a different set of assumptions (that are mentioned as uniform distribution by Robertson 2005):

**Assumption 4**  $p_{\Omega}(d, q)$  is a constant (i.e., uniformly distributed).

**Assumption 5**  $p_{\Omega}(d, r)$  is a constant (i.e., uniformly distributed).

Using these assumptions, we deduce that  $p_{\Omega}(r|d, q) \stackrel{\text{rank}}{=} p_{\Omega}(d|q, r) \stackrel{\text{rank}}{=} p_{\Omega}(q|d, r)$ . This relation is applicable to a special form of the hidden Markov language model (HMLM) of Miller et al. (1999) without asserting assumptions 1 and 2 of Lafferty and Zhai (2003). However, these two assumptions are not applicable to language models (e.g., Ponte and Croft 1998) which rank documents by  $p_{\Omega}(q|d)$ . This is because assumption 4 implies that

Eq. 10 does not hold as the document ranking is problematic (i.e., it assigns the same value for document ranking across different queries) for such language models. For BIM, this assumption does not cause any problems, because  $p_{\Omega}(d, q)$  is a marginal probability combining the probability of relevance and the probability of non-relevance.

Lafferty and Zhai (2003) have claimed that while probabilistically the log-odds is rank equivalent to the query likelihood, statistical component models of the log-odds ratio and the query likelihood are not rank equivalent. In their words,

... the two approaches (based on the log-odds or the query likelihood) are not equivalent from a statistical point of view, since the component models are estimated quite differently. (page 2, Lafferty and Zhai 2003)

(added our bracketed text in the above quotation). Lafferty and Zhai (2003) do not mention whether the statistical component language model (e.g., Berger and Lafferty 1999) can be derived from the log-odds ratio at a higher probabilistic modeling level (i.e., Box 1, 2 and 4 in Fig. 2). Such derivations are important for claiming that the statistical component language models are strict rank equivalent to the log-odds ratio. Similarly, they do not mention whether the statistical component model (e.g., BIM [Robertson and Spärck 1976]) of the log-odds ratio can be derived from the query likelihood at a higher probabilistic modeling level (i.e., Box 2, 1 and 3 in Fig. 2). Such derivations are important for claiming that the statistical component models of the log-odds ratio are strict rank equivalent to the query likelihood. These derivations require the assumptions by Lafferty and Zhai (2003) to be asserted with the assumptions of the related statistical component models.

Let us focus on the strict rank equivalence between the log-odds ratio at the higher probabilistic level and the statistical component language models. We identify at least three possible claims about such a strict rank equivalence relation:

**Claim 1** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *no* statistical component language models are strict rank equivalent to the log-odds ratio that is at the higher probabilistic level.

**Claim 2** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *all* statistical component language models are strict rank equivalent to the log-odds ratio that is at the higher probabilistic level.

**Claim 3** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *some* statistical component language model is strict rank equivalent to the log-odds ratio that is at the higher probabilistic level.

We argue for claim 3 by showing that claims 1 and 2 are not tenable. First, we assert that claim 1 is true and disprove it by a counter example. For ranking using the language model by Lafferty and Zhai (2003), we estimate the probability  $p_{\Omega}(q|d, r)$  and assume their assumptions 1 and 2 hold so that the language model corresponds to the log-odds ratio. Since

$$\log p_{\Omega}(q|d, r) \stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d, r)}{p_{\Omega}(d, r)} \stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d, r)}{p_{\Omega}(d)},$$

we can directly estimate  $p_{\Omega}(q, d, r)$  and derive  $p_{\Omega}(d)$  based on the marginal probability calculation using Eq. 1, i.e.,

$$p_{\Omega}(d) \equiv \sum_{q \in Q} \sum_{x \in R} p_{\Omega}(d, q, x).$$



Since only  $p_{\Omega}(q, d, r)$  is estimated from data and the others are derived from it based on the assumptions, there are no problematic rankings (unlike the planet and star example of Robertson). If we further simplify the above (see Appendix I) and replace the above ranking function by the one in (Ponte and Croft 1998) (i.e.,  $\log p_{\Omega}(q|d)$ ), then we need to simply estimate  $p_{\Omega}(q, d)$  from the data and derive the marginal probability  $p_{\Omega}(d)$  using Eq. 1 based on the estimated joint probabilities  $\{p_{\Omega}(q, d)\}$ , i.e.,  $p_{\Omega}(d) = \sum_{q \in Q} p_{\Omega}(q, d)$  assuming that necessary condition 1 holds in this case.

Similarly, to show that claim 2 is not tenable, we assert that it is true, and find a counter example as follows. A special form of the HMLM by Miller et al. (1999) is strict rank equivalent to the query likelihood given assumptions 1, 2 and 5 hold as follows. By Eqs. 10, 11 and assumption 5, we deduce that

$$p_{\Omega}(r|d, q) \stackrel{\text{rank}}{=} p_{\Omega}(q|d, r) \stackrel{\text{rank}}{=} p_{\Omega}(d, r|q).$$

The previous relation is rewritten as

$$\frac{p_{\Omega}(r, d, q)}{p_{\Omega}(d, q)} \stackrel{\text{rank}}{=} \frac{p_{\Omega}(d, r, q)}{p_{\Omega}(q)}.$$

Since  $p_{\Omega}(r, d, q) = p_{\Omega}(d, r, q)$ , we deduce that

$$p_{\Omega}(d, q) \stackrel{\text{rank}}{=} p_{\Omega}(q)$$

where  $p_{\Omega}(q)$  is a marginal probability. By the above relation and by Eq. 10 that asserts assumptions 1 and 2 of Lafferty and Zhai hold, we obtain

$$p_{\Omega}(q|d, r) \stackrel{\text{rank}}{=} p_{\Omega}(q|d) \stackrel{\text{rank}}{=} p_{\Omega}(q, d)/p_{\Omega}(d) \stackrel{\text{rank}}{=} p_{\Omega}(q)/p_{\Omega}(d) \stackrel{\text{rank}}{=} 1/p_{\Omega}(d).$$

This ranking is problematic because it is only dependent on the document and independent of the query (i.e., assigning the same rank to a document for all queries). If this special form of the HMLM is strict rank equivalent to the log-odds ratio via assumptions 1 and 2 of Lafferty and Zhai, then the resultant ranking will be problematic as the ranking is query independent. This shows that claim 2 does not hold. Since both claim 1 and 2 do not hold, claim 3 holds.

When deriving the strict rank equivalence between the statistical component models of the log-odds ratio and the query likelihood, assumptions 1 and 2 by Lafferty and Zhai are asserted with the model-specific assumptions of these statistical component models. Again, there are three possible claims:

**Claim 4** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *no* statistical component models of the log-odds ratio are strict rank equivalent to the query likelihood that is at the higher probabilistic level.

**Claim 5** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *all* statistical component models of the log-odds ratio are strict rank equivalent to the query likelihood that is at the higher probabilistic level.

**Claim 6** If assumptions 1 and 2 by Lafferty and Zhai (2003) are true, then *some* statistical component model of the log-odds ratio is strict rank equivalent to the query likelihood that is at the higher probabilistic level.

When all statistical component models (like the BIM) of the log-odds ratio are defined for single queries, such models are making the following singleton query assumption in the event space,  $\Omega$ :

**Assumption 6**  $Q = \{q\}$ .

This assumption implies that  $p_{\Omega}(d|r) = p_{\Omega}(d|q, r)$  and  $p_{\Omega}(d|\bar{r}) = p_{\Omega}(d|q, \bar{r})$ . As a result, the log-odds ratio becomes:

$$\begin{aligned} O_{\Omega}(r|q, d) &\stackrel{\text{rank}}{=} \log \left[ \frac{p_{\Omega}(d|q, r)}{p_{\Omega}(d|q, \bar{r})} \right] \text{ (applying assumption 6)} \\ &\stackrel{\text{rank}}{=} \log \left[ \frac{p_{\Omega}(d|r)}{p_{\Omega}(d|\bar{r})} \right] \text{ (applying assumption 2)} \\ &\stackrel{\text{rank}}{=} \log[1]. \end{aligned}$$

Therefore, claim 5 does not hold. Claim 4 will hold if the statistical component models of the log-odds ratio are defined for single queries. However, claim 4 will not hold if we extend the statistical component models (e.g., the BIM) of the log-odds ratio for multiple queries as in Lafferty and Zhai (2003). In the case of the BIM, we use a group of such BIMs for the set of queries such that each query is assigned with a BIM. Since both claim 4 and claim 5 do not hold, claim 6 holds. In summary, we conclude that the rank equivalence claimed by Lafferty and Zhai (2003) has immediate practical significance because some statistical component language model is rank equivalent to the log-odds ratio, and some statistical component model of the log-odds ratio is rank equivalent to the query likelihood.

## 7 Topic variable

In Sect. 2, we have shown that it is possible to enlarge the event space by performing a cross-product with other event spaces. This enlargement of the event space is the same as the problem of missing (hidden) variables. For probabilistic information retrieval, one variable that seemed to be missing is the “conceptual” user information need. When we consider a query  $q$ ,  $q$  is a (textual) description or (computer) representation of the user information need. It is possible that many different queries are referring to the same user information need (e.g., Robertson 2005; Zhai and Lafferty 2006). For instance, the TREC user information need (called a topic) in ad hoc retrieval is specified by multiple types of queries (e.g., title queries, concept queries and narrative queries). We believe that the user information need is an important hidden variable because it can clarify the confusion between query text and the identification of the user information need.

Many of the probabilistic retrieval models deal with a single user information need (e.g., BIM) and it does not seem to be necessary to model it explicitly because the query representing it is already unique. This becomes important when reconciling different retrieval models when one model deals with a single request and another model deals with multiple requests. From the perspective of probabilistic modeling, if the probabilistic retrieval model needs the marginal probabilities without query dependence (e.g.,  $p_{\Omega}(d, r)$ ), these marginal probabilities are calculated using Eq. 1 that requires necessary condition 1 to hold. This condition is satisfied by making each query text to be distinct from each other. This is done by adding an identifier for the user information need to the query. Apart from modeling, a web query may contain different kinds of information (like session identifiers, user profiles, etc.). In this case, the distinction between a topic and a query may be helpful

because many different queries may be generated for the same user information need, and because different user information need may generate the same query (i.e., when the query text is ambiguous). Furthermore, documents are considered to be relevant to the user information need and not on the basis of being relevant to a piece of query text. Based on this, we show later that the introduction of the user information need into the probabilistic model enables language models to perform (pseudo) relevance feedback which has been identified to be difficult before (Robertson 2005).

For presentation clarity, we treat the user information needs as topics and we propose to add a topic set  $T$  so that the event space is:

$$\Psi = Q \times T \times D \times R.$$

Equation 10 is considered as a notation convenience of Eq. 12:

$$\log \frac{p_{\Psi}(r|q, d, t)}{p_{\Psi}(\bar{r}|q, d, t)} \stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d, r) \stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d) \tag{12}$$

by taking away the particular topic  $t$  and the topic set  $T$  in Eq. 12. For Eq. 12 to hold, we need to assert assumption 2 and we need to modify assumption 3 to assumption 7:

**Assumption 7**  $p_{\Psi}(d, q, \bar{r}, t) \times p_{\Psi}(\bar{r}) = p_{\Psi}(d, \bar{r}) \times p_{\Psi}(q, \bar{r}, t).$

The above assumption does not imply that any of the above probability distributions are uniform. Details of the derivations to obtain Eq. 12 are in Appendix II.

The topic variable bridges the difference in representing the query by the RSJ model and the language models as follows. According to Robertson (2005), the query in the RSJ model is supposed to be a topic identifier that uniquely identifies the topic (although not explicitly written down) where as the query in the language model is a feature vector or a text string. We reconcile these differences by adding the topic variable so that the query  $q$  in Eq. 12 is interpreted as a feature vector/text string as in the language models and the tuple  $(q, t)$ , which is a combination of the query  $q$  and the topic  $t$  in Eq. 12. This combination provides a unique value to identify the particular topic as required by the RSJ model. This unique value for each query ensures that the necessary condition 1 for calculating marginal probabilities holds. For example, this is needed for estimating the marginal probability,  $p_{\Psi}(d) = \sum_{q \in Q} p_{\Psi}(q, d)$ , from the estimated joint probabilities,  $\{p_{\Psi}(q, d)\}$  for the language model by Ponte and Croft (1998).

Based on Eq. 12, (pseudo) relevance feedback for language models is the same as the usual (pseudo) relevance feedback based on the vector space model. The language model ranks on the basis of the expanded query  $q'$  by reformulating the initial query  $q$  using the feedback information, i.e.,  $\log p_{\Psi}(qt, t|d, r)$ . Since relevance is defined in terms of the topic  $t$ , the same set of relevance is used for both query  $q$  and query  $q'$ . This facilitates a single topic to associate with any number of pieces of query text. Conversely, a piece of query text can associate with any number of topics since a piece of query text can be ambiguous.

In practice, the topic variable is already used in open IR evaluation workshops (e.g., TREC, INEX and NTCIR). A query text string is the concatenation of the topic identifier with the sequence of query terms. When this query text string is processed into a vector, the query vector may contain the unique topic identifier. Since the topic identifier does not appear in any documents, it does not match with any documents and it does not affect document ranking. In practice, most query processing modules in IR systems filter the topic identifiers.

For notation convenience, it is possible to drop  $T$  in the probabilistic retrieval models as in the past because it is clear that the probabilities are defined for a particular topic, and that relevance is defined in terms of the topic and not in terms of the query. With the advent of more sophisticated web query processing, the topic random variable appears to be helpful to distinguish information needs and query text, as well as reconciling different probabilistic retrieval models.

## 8 Summary

It seems that if we accept Kolmogorov's probability axioms and if necessary condition 1 holds, then the marginal probability will follow immediately. However, Robertson (2005) highlights the significance of the event space when reconciling different probabilistic models by rank equivalence (Lafferty and Zhai 2003). In this paper, we highlight that, perhaps, the assumptions made in the retrieval models or in the unification of different models need to be questioned as well. Apart from looking at event spaces, it seems that it is appealing to question whether multiple probability distributions can be assumed to be uniformly distributed simultaneously without causing the probabilistic models to produce problematic ranking. We have also shown that the probability of relevance (Robertson 1977) can be considered as the common ground (as in Eq. 10) of the earlier language model by Lafferty and Zhai (2003), as well as the earlier language model by Ponte and Croft (1998) that requires assumptions 1 and 2 to hold (provided that suitable probabilities can always be estimated). This implies that under specific conditions (i.e., when assumptions of Lafferty and Zhai are true) the relevance random variable may be implicit in the ranking formula. Assumption 1 is mentioned in previous retrieval model design (e.g., Robertson et al. 1982b) and it can be replaced by assumption 3 that does not require any additional condition to hold. Assumption 2 seems intuitively plausible when its probabilities are marginal probabilities. We show that the rank equivalence claimed by Lafferty and Zhai (2003) has practical significance because some statistical component language model is strict rank equivalent to the log-odds ratio, and some statistical component model of the log-odds ratio is strict rank equivalent to the query likelihood. In this case, assumptions 1 and 2 are the sufficient conditions for the rank equivalence between the log-odds ratio and the query likelihood. While some of the earlier probabilistic models do not include the user information need (called the topic) as a random variable for notation convenience, we suggest adding the topic random variable in the probabilistic retrieval models for notation clarity when these models deal with multiple requests.

**Acknowledgements** I am grateful to Prof. Stephen Robertson and two anonymous reviewers for their valuable comments about the paper. This work is supported by the Project # PolyU 5226/05E.

## Appendix I: Alternative derivation

According to the Lafferty and Zhai (2003) model, the ranking formula (i.e., Eq. 1.18 in Lafferty and Zhai 2003) is:

$$\begin{aligned}
 O_{\Omega}(r|q, d) &\stackrel{\text{rank}}{=} \log p_{\Omega}(q|d, r) \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d, r)}{p_{\Omega}(d, r)} \text{ (applying assumption 2)} \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d, r)}{p_{\Omega}(d) \cdot p_{\Omega}(r)} \text{ (applying equation 9)} \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d) - p_{\Omega}(q, d, \bar{r})}{p_{\Omega}(d)} \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d) - p_{\Omega}(q, d|\bar{r}) \cdot p_R(\bar{r})}{p_{\Omega}(d)} \text{ (applying assumption 1)} \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d) - p_{\Omega}(q|\bar{r}) \cdot p_{\Omega}(d|\bar{r}) \cdot p_{\Omega}(\bar{r})}{p_{\Omega}(d)} \text{ (applying assumption 2)} \\
 &\stackrel{\text{rank}}{=} \log \frac{p_{\Omega}(q, d) - p_{\Omega}(q, \bar{r}) \cdot p_{\Omega}(d)}{p_{\Omega}(d)} \\
 &\stackrel{\text{rank}}{=} \log [p_{\Omega}(q|d) - p_{\Omega}(q, \bar{r})] \\
 &\stackrel{\text{rank}}{=} \log [p_{\Omega}(q|d)].
 \end{aligned}$$

**Appendix II: Derivations with the topic variable**

Starting with the log-odds ratio, we have:

$$\log \left[ \frac{p_{\Psi}(r|q, d, t)}{p_{\Psi}(\bar{r}|q, d, t)} \right].$$

We assume that assumption 2 and 4 holds. Next, we derive similar to Lafferty and Zhai (2003):

$$\begin{aligned}
 O_{\Psi}(r|q, d, t) &= \log \left[ \frac{p_{\Psi}(q, d, t, r)}{p_{\Psi}(q, d, t, \bar{r})} \right] \\
 &= \log \left[ \frac{p_{\Psi}(q, t|d, r) \cdot p_{\Psi}(d, r)}{p_{\Psi}(q, t|d, \bar{r}) \cdot p_{\Psi}(d, \bar{r})} \right] \\
 &= \log \left[ \frac{p_{\Psi}(q, t|d, r)}{p_{\Psi}(q, t|d, \bar{r})} \right] + \log \left[ \frac{p_{\Psi}(d, r)}{p_{\Psi}(d, \bar{r})} \right] \\
 &= \log \left[ \frac{p_{\Psi}(q, t|d, r)}{p_{\Psi}(q, t, d|\bar{r})/p_{\Psi}(d|\bar{r})} \right] + \log \left[ \frac{p_{\Psi}(d, r)}{p_{\Psi}(d, \bar{r})} \right] \text{ (applying assumption 7)} \\
 &= \log \left[ \frac{p_{\Psi}(q, t|d, r)}{p_{\Psi}(q, t|\bar{r})} \right] + \log \left[ \frac{p_{\Psi}(d, r)}{p_{\Psi}(d, \bar{r})} \right] \\
 &\stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d, r) + \log \left[ \frac{p_{\Psi}(d, r)}{p_{\Psi}(d, \bar{r})} \right] \text{ (applying assumption 2)} \\
 &\stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d, r) + \log \left[ \frac{p_{\Psi}(r) \cdot p_{\Psi}(d)}{p_{\Psi}(\bar{r}) \cdot p_{\Psi}(d)} \right] \\
 &\stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d, r) + \log \left[ \frac{p_{\Psi}(r)}{p_{\Psi}(\bar{r})} \right] \\
 &\stackrel{\text{rank}}{=} \log p_{\Psi}(q, t|d, r).
 \end{aligned}$$

To derive the Ponte and Croft model (1998), we have:

$$\begin{aligned}
 p_{\Psi}(q, t|d, r) &= \frac{p_{\Psi}(q, t, d) - p_{\Psi}(q, t, d, \bar{r})}{p_{\Psi}(d, r)} \text{ (applying equation 9)} \\
 &= \frac{p_{\Psi}(q, t, d) - p_{\Psi}(q, t, d|\bar{r}) \cdot p_{\Psi}(\bar{r})}{p_{\Psi}(d, r)} \text{ (applying assumption 7)} \\
 &= \frac{p_{\Psi}(q, t, d) - p_{\Psi}(d|\bar{r}) \cdot p_{\Psi}(q, t|\bar{r}) \cdot p_{\Psi}(\bar{r})}{p_{\Psi}(d, r)} \text{ (applying assumption 2)} \\
 &= \frac{p_{\Psi}(q, t, d) - p_{\Psi}(d) \cdot p_{\Psi}(\bar{r}) \cdot p_{\Psi}(q, t|\bar{r})}{p_{\Psi}(d)p_{\Psi}(r)} \\
 &= \frac{p_{\Psi}(q, t, d)}{p_{\Psi}(d) \cdot p_{\Psi}(r)} - \frac{p_{\Psi}(\bar{r})}{p_{\Psi}(r)} \cdot p_{\Psi}(q, t|\bar{r}) \\
 &\stackrel{\text{rank}}{=} p_{\Psi}(q, t|d).
 \end{aligned}$$

## References

- Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In F. Gey, M. Hearst, & R. Tong, (Eds.), *Proceedings of SIGIR '99: The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 222–229). Berkeley, CA, USA, New York: ACM Press.
- Cooper, W. S. (1995). Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. *ACM Transactions on Information Systems*, 13(1), 100–111.
- Croft, W. B., & Lafferty, J. (Eds.). (2003). *Language modeling for information retrieval*. Kluwer Academic Publishers.
- Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3), 243–255.
- Fuhr, N. (2008). A probability ranking principle for interactive information retrieval. *Information Retrieval*, 11(3), 251–265.
- Gao, J., Nie, J.-Y., Wu, G., & Cao, G. (2004). Dependence language model for information retrieval. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza, (Eds.), *Proceedings of SIGIR '04: The 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 170–177). Sheffield, UK, New York: ACM Press.
- Hiemstra, D. (1998). A linguistically motivated probabilistic model of information retrieval. In C. Nicolaou & C. Stephanidis, (Eds.), *Lecture notes in computer science: Research and advanced technology for digital libraries* (Vol. 1513, pp. 569–584). Greece: Springer-Verlag.
- Hiemstra, D. (2000). A probabilistic justification for using tf.idf term weighting in information retrieval. *International Journal on Digital Libraries*, 3(2), 131–139.
- Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term. In K. Järvelin, M. Beaulieu, R. Baeza-Yates, & S. H. Myaeng, (Eds.), *Proceedings of SIGIR '02: The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 35–41). Tampere, Finland, New York: ACM Press.
- Lafferty, J., & Zhai, C. (2003). Probabilistic relevance models based on document and query generation. In W. B. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 1–10). Kluwer Academic Publishers.
- Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In D. H. Kraft, W. B. Croft, D. J. Harper, & J. Zobel, (Eds.), *Proceedings of SIGIR '01: The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 120–127). New Orleans, Louisiana, US, New York: ACM Press.
- Lidstone, G. J. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8, 182–192.
- Miller, D. H., Leek, T., & Schwartz, R. (1999). A hidden Markov model information retrieval system. In F. Gey, M. Hearst, & R. Tong, (Eds.), *Proceedings of SIGIR '99: The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 214–221). Berkeley, CA, USA, New York: ACM Press.

- Ney, H., Essen, U., & Kenser, R. (1994). On structuring probabilistic dependencies in stochastic language modeling. *Information Technology: Research and Development*, 3, 33–42.
- Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilinson, & J. Zobel, (Eds.), *Proceedings of SIGIR '98: The 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281). Melbourne, Australia, New York: ACM Press.
- Robertson, S. E. (1974). Specificity and weighted retrieval. *Journal of Documentation*, 30(1), 41–46.
- Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4), 294–304.
- Robertson, S. E. (2003). The unified model revisited. In S. Dominich, K. van Rijsbergen, & M. Lalmas, (Eds.), *ACM SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*. Toronto, Canada.
- Robertson, S. E. (2005). On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2), 319–329.
- Robertson, S. E., & Spärck, J. K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3), 129–146.
- Robertson, S. E., Maron, M., & Cooper, W. S. (1982a). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology-Research and Development*, 1, 1–21.
- Robertson, S. E., Maron, M., & Cooper, W. S. (1982b). The Unified Probabilistic Model for IR. In H.-J. Schneider & G. Salton (Eds.), *Proceedings of SIGIR '82: Proceedings of the 5th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 108–117). West Berlin, Germany, New York: ACM Press.
- Roelleke, T., & Wang, J. (2006). A parallel derivation of probabilistic information retrieval models. In S. Dumais, E. N. Efthimiadis, D. Hawking, & K. Järvelin (Eds.), *Proceedings of SIGIR '06: The 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 107–114). Seattle, Washington, USA, New York: ACM Press.
- Song, F., & Croft, W. B. (1999). A general language model for information retrieval. In F. Gey, M. Hearst, & R. Tong (Eds.), *Proceedings of SIGIR '99: The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 279–280). Berkeley, CA, US, New York: ACM Press.
- Spärck, J. K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments, part 1. *Information Processing and Management*, 36(6), 779–808.
- Spärck, J. K., Robertson, S. E., Hiemstra, D., & Zaragoza, H. (2003). Language modelling and relevance. In W. B. Croft & J. Lafferty (Eds.), *Language modeling for information retrieval* (pp. 57–73). Kluwer Academic Publishers.
- Wu, H. C., Luk, R. W. P., & Wong, K. F. (2007). Probability ranking principle via optimal expected rank. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando, (Eds.), *Proceedings of SIGIR '99: The 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 713–714). Amsterdam, The Netherlands, New York: ACM Press.
- Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (in press). Interpreting TF-IDF term weights as making relevance decisions. To appear in *ACM Transactions on Information Systems*.
- Yu, C., & Salton, G. (1976). Precision weighting—an effective automatic indexing method. *Journal of the ACM*, 23(1), 76–88.
- Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 179–214.
- Zhai, C., & Lafferty, J. (2006). A risk minimization framework for information retrieval. *Information Processing and Management*, 42(1), 31–55.