Probabilistic Models in IR and Their Relationships*

Robin Aly · Thomas Demeester · Stephen Robertson

Delivery Date

Abstract A solid research path towards new information retrieval (IR) models is to further develop the theory behind existing models. A profound understanding of these models is therefore essential. In this paper, we revisit Probability Ranking Principle (PRP)-based models, Probability of Relevance (PR) models, and language models, finding conceptual differences in their definition and interrelationships. The probabilistic model of the PRP has not been explicitly defined previously, but doing so leads to the formulation of two actual principles with different objectives. First, the belief probability ranking principle (BPRP), which considers uncertain relevance between known documents and the current query, and second, the popularity probability ranking principle (PPRP), which considers the probability of relevance of documents among multiple queries with the same features. Our analysis shows how some of the discussed PR models implement the BPRP or the PPRP while others do not. However, for some models the parameter estimation is challenging. Finally, language models are often presented as related to PR models. However, we find that language models differ from PR models in every aspect of a probabilistic model and the effectiveness of language models cannot be explained by the PRP.

* This paper is an extended version of Aly and Demeester [2011], with the following additional contributions: (1) the finding that there are actually two distinct PRPs, (2) the investigation of three additional models from the unified framework of PR models, and (3) a more elaborate discussion of related work. **The final publication is available at http://link.springer.com**

Robin Aly
University of Twente

Thomas Demeester
Ghent University - iMinds

Stephen Robertson
University College London

1 Introduction

One of the main goals of today's research in Information Retrieval (IR) systems is to invent ranking functions that order the documents of a collection by their likelihood of answering a user's information need. A solid way to define ranking functions is to propose a ranking model that gives an intuition why a corresponding ranking function would answer the users' information needs effectively. For example, the vector space model proposes to rank by the angle between the vector representation of a considered document and the considered query [Salton et al., 1975]. However, the ranking models, like the vector space model, do not give any guarantees on whether or not, and why, they would lead to strong performance, for example a high precision. To overcome such limitations, the research community introduces ranking principles, which show explicitly that ranking by a certain criterion optimizes specific effectiveness measures. Hence, increasing the accuracy of a ranking model that follows a ranking principle also improves its effectiveness.

For around four decades, ranking models have been formulated in a probabilistic way. One of the main reasons for this trend is the Probability Ranking Principle (PRP) by Robertson [1977] that provides a theoretical connection between ranking by the probability of relevance and several evaluation measures. However, the derivation of effective ranking functions from the PRP has proven to be difficult. Some researchers refer to this as the theory effectiveness gap, see for example Lv [2012]. A recent trend is to abandon formal ranking models and to argue about ranking functions in an axiomatic way, without explicitly relating them to a ranking model, see for example [Fang and Zhai, 2005]. These axioms, however, do not give any performance guarantees. In this paper, we take an alternative approach and investigate whether the connections between several popular ranking models are fully understood in the literature.[1] We find that the understanding is not always complete. We clarify a number of issues during the derivation of these ranking models. The improved understanding that comes with this clarification can help researchers to address the theory effectiveness gap in the future.

As the PRP is one of the most frequently used ranking principles, it is important to understand its definition and properties. We show that the understanding of the PRP is currently incomplete by finding two distinct ranking principles based on different probabilities of relevance that optimize different effectiveness measures: one is the principle for different beliefs of a system about relevance of documents to a particular query and the other principle is based on the popularity of documents to different queries with the same representation. We clarify the differences between these principles and discuss the influence of these findings on well-known probabilistic ranking models.

In a following step we investigate in how far two popular types of probability-based ranking models are connected to the PRP. First, we revisit the four classes of ranking models described in the unified framework of Probability of Relevance (PR) models by Robertson et al. [1982], commonly assumed to follow the PRP because they all calculate a specific probability of relevance. But now that there are two principles, we need to examine which follows which. We find that not all PR models, even popular ones, can be mapped onto one of the PRP's. As the second type of models, we consider four variations of language models: the

[1] We provide a formal definition of the term *connection* in Appendix B.

query likelihood model [Ponte and Croft, 1998], the language model by Hiemstra [2001] (referred to as *Hiemstra's model*), the *risk-minimization model* [Zhai and Lafferty, 2006], and the *relevance model* [Lavrenko and Croft, 2003]. These models are commonly thought to implement the PRP by being comparable to PR models. However, a careful analysis of the PR models and language models reveals that they are fundamentally different. Therefore, although we cannot prove the absence of a connection between the two models, we propose on the basis of these differences that a connection does not exist.

This paper adds to the series of works that discuss the connection between PR models and language models. The conclusions of these works differ significantly: the works by Lafferty and Zhai [2003], Luk [2008] and Zhai [2008] propose a connection between PR models and language models exists, while the work by Spärck-Jones et al. [2003] and Robertson [2005] state the opposite. We believe the difference in these conclusions originates from the fact that these works make slightly different assumptions about the discussed models. One possible reason why these differences have gone unnoticed so far is that existing literature focuses on the events and their probabilities and other aspects of probability theory are assumed implicitly. In order to make progress in this discussion, this paper considers all elements of the investigated probabilistic ranking models, i.e., the underlying *process*, the *sample space*, *event spaces*, and the *probability measure*.

In summary, this paper makes the following contributions.

1. We find that the original PRP should be seen as two distinct ranking principles.
2. We identify connections between the PR models and these principles.
3. We find that language models are too different from the probabilistic models considered by these principles or PR models to be connected with them.

We would like to point out to the reader that we do not invent new models in this paper but investigate the connection of the existing models mentioned above. We assume that these models are IR applications of the notion of probabilistic models in probability theory. As this paper makes heavy use of the basic elements of probabilistic models, which are seldom used to this extent in IR literature, we provide their definitions in Appendix A for the reader's reference.

This paper is structured as follows: Section 2 clarifies basic assumptions about the modeled ad-hoc retrieval task, and introduces the notations used in this paper. Section 3 discusses possible probabilistic models for the PRP, Section 4 defines the basic probabilistic aspects of PR models, and Section 5 discusses language models and their differences to PR models. Section 6 puts this paper in context with related work, and finally, Section 7 concludes the paper.

2 The Ad-hoc Retrieval Process: Assumptions and Notation

When comparing models it is important to clarify the real-world process they consider. In this paper, we consider ad-hoc retrieval, which is also considered, for example, by many tasks of the TREC evaluation workshop [Voorhees et al., 2005]. In ad-hoc retrieval, a user formulates each information need (a topic in the TREC terminology) in a single query and submits this query to a retrieval system. The retrieval system returns a ranked list of documents that the user is assumed to read starting from the top. Documents are either relevant or non-relevant to the

Symbol	*Meaning*	Mathematical definition
	Basic IR Objects Sect. 2	
$q/\hat{q}$	Query ($\hat{q}$ is the current query)	
$\mathcal{Q}$	Considered queries	$\mathcal{Q} := \{q_1, ..., q_m\}$
$d/\hat{d}$	Document ($\hat{d}$ is the current document)	
$\mathcal{D}$	Considered documents (collection)	$\mathcal{D} := \{d_1, ..., d_{\lvert\mathcal{D}\rvert}\}$
$\mathbf{d}$	A specific ranking of documents	
t	Term	
$\mathcal{T}$	Term space (vocabulary)	$\mathcal{T} := \{t_1, ..., t_{\lvert\mathcal{T}\rvert}\}$
L	Query length	$L : \mathcal{Q} \to \mathbb{N}$
Tx_i	ith query term	$Tx_i : \mathcal{Q} \to \mathcal{T}$
R	Relevance	$R : \mathcal{Q} \times \mathcal{D} \to \{0, 1\}$
	BPRP and PPRP Sect. 3	
$\Phi_{\hat{q}}$	Sample space for the BPRP for query $\hat{q}$	$\Phi_{\hat{q}} := \{0, 1\} \times ... \times \{0, 1\}$ ($\lvert\mathcal{D}\rvert$ times)
$\hat{R}_{\hat{q},d_i}$	Relevance of ith component in $\Phi_{\hat{q}}$	$\hat{R}_{\hat{q},d_i} : \Phi \to \{0, 1\}$
$U^n_{\hat{q},\mathbf{d}}$	Utility of reading $\mathbf{d}$ for $\hat{q}$ until rank n	$U^n_{\hat{q},\mathbf{d}} : \Phi \to \mathbb{R}$
$Prec^n_{\hat{q},\mathbf{d}}$	Precision at n of ranking $\mathbf{d}$ for $\hat{q}$	$Prec^n_{\hat{q},\mathbf{d}}(\phi \in \Phi) \to [0, 1]$
$Rec^n_{\hat{q},\mathbf{d}}$	Recall at n of ranking $\mathbf{d}$ for $\hat{q}$	$Rec^n_{\hat{q},\mathbf{d}} : \Phi \to [0, 1]$
$\hat{\mathcal{Q}}$	Sample space for the PPRP	$\hat{\mathcal{Q}} := \{q \in \mathcal{Q} \lvert \mathbf{Tx}(q) = \mathbf{Tx}(\hat{q})\}$
U_d	Utility of document d in the PPRP	$U_d : \hat{\mathcal{Q}} \to \mathbb{R}$
$U^n_{\mathbf{d}}$	Utility of ranking $\mathbf{d}$ until rank n	$U^n_{\mathbf{d}} : \hat{\mathcal{Q}} \to \mathbb{R}$
	PR Models Sect. 4	
Ω	Query-document pair sample space	$\Omega = \mathcal{Q} \times \mathcal{D}^+$
QF	Query feature	depends on model instance
Q	Trivial query feature	$Q : \Omega \to \mathcal{Q}$
DF	Document feature	depends on model instance
D	Trivial document feature	$D : \mathcal{D} \to \mathcal{D}$
	Language Models Sect. 5	
$\mathcal{T}_n$	Drawn term sequence (sample space)	$\mathcal{T}_n = \mathcal{T} \times ... \times \mathcal{T}$ (n times)
T_i	ith term in a sequence	$T : \mathcal{T}_n \to \mathcal{T}$
$\mathcal{H}$	Sample space of Hiemstra's model	$\mathcal{H} := \mathcal{T}_{L(\hat{q})} \times \mathcal{D}$
D'	Document a user has in mind	$D' : \mathcal{H} \to \mathcal{D}$
$\mathcal{RM}$	Sample space of the relevance model	$\mathcal{RM} := \{(d, t) \in \mathcal{D} \times \mathcal{T}_1 \lvert R(\hat{q}, d){=}1\}$
D''	Drawn, relevant document	$D'' : \mathcal{RM} \to \mathcal{D}$
R'	Relevant document	$R' : \mathcal{RM} \to \{0, 1\}$
θ	Distribution parameters	$\theta : \mathcal{D} \to [0, 1]$
	Probabilistic Models Appendix A	
$P_{\mathcal{E}}$	Probability measure of model $\mathcal{E}$.	

Table 1: Overview of the notation used in this paper.

user's information need. Additionally, queries and documents have properties. In this paper, we focus on textual properties, although there are also other properties, for example the query submission time or the document genre. Note that some related work defines the term 'query' differently. In this paper, a query should not be confounded with its properties, such as the query's terms. Furthermore, a query, which we defined as a single submission to a search engine, is different to the set of submissions with the same text. The reader may think of a query in our definition as an entry in a query log. The query's terms are part of the log entry. Similar relationships exist with documents and their text.

Before turning to the notation for the ad-hoc retrieval process, we state the principles used for the notation throughout this paper. We denote sets in upper-

case calligraphic letters, set elements and values in lower case letters, vectors in boldface, and functions and random variables in upper case letters.

Table 1 gives an overview of most of the symbols used in this paper, some of which are only introduced in the indicated sections. We denote queries and documents by lower case q's and d's respectively. The considered set of queries is denoted by $\mathcal{Q}$ and the considered set of documents (the collection) by $\mathcal{D}$. Lower case t's are used for terms, and $\mathcal{T}$ indicates the considered set of terms (the vocabulary). The terms of a query are modeled as a vector, denoted as $\mathbf{Tx}(q) = (Tx_1(q), ..., Tx_{L(\hat{q})}(q))$ where $L(\hat{q})$ is the query length. Finally, we define the relevance random variable between a query q and document d as:

$$R(q,d) := \begin{cases} 1 & \text{if document } d \text{ is relevant to query } q, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Note that it would be clearer to define relevance based on information needs rather than on queries. However, because information needs and queries have a one-to-one mapping in the ad-hoc retrieval scenario (there is exactly one need per query), we adapt to the common practice and define relevance based on queries. Note that the ad-hoc retrieval scenario always considers a single user per query, even if multiple TREC assessors have to agree on the definition of the relevance variable R.

3 Probability of Relevance Ranking Principles for IR

A ranking principle states a criterion and shows that ranking by this criterion achieves an objective, usually the maximization of an objective function. Robertson [1977] proposes the probability ranking principle of IR (PRP) that states documents should be ranked by their probability of relevance. He provides a mathematical proof that ranking documents by their probability of relevance maximizes several objective functions that are defined further on. However, the paper also gives in the appendix an example where ranking by the probability of relevance does not maximize the user's utility of a ranking, which was one of the objective functions mentioned in the main text. Therefore, if the example would apply to the assumption made in the PRP, the proof would be contradicted, hence jeopardizing the mathematical justification of many existing ranking models that are declared to follow the PRP. To our knowledge, whether or not the example contradicts the assumptions of the PRP, still needs to be sorted out, see Cooper [1994]. The investigation of this matter requires a complete definition of the probabilistic model assumed by the PRP, which is only partially provided in the original work of Robertson. In fact, we propose for the main text and the example in the appendix of the publication by Robertson consider two different probabilistic models, which correspond to the maximization of different objective functions. The example therefore appears to be not contradictory but rather makes use of another ranking principle.

3.1 Degrees of Belief in the PRP

In the original PRP paper, Robertson [1977] shows that ranking documents by their probability of relevance maximizes three objective functions for the issuer of

the *current* query: the expected recall, the expected precision, and the expected utility. However, the original PRP does not explicitly state on which model the probability of relevance for each document is defined. In this section, we define a probabilistic model based on Bayesian beliefs on which the PRP could be based, and refer to the corresponding principle as the belief probability ranking principle (BPRP). Note that Thomas Bayes made several contributions to probability theory, which are sometimes used ambiguously. In Appendix A we contrast the contribution of Bayesian belief with his other contributions to clarify how we use this term.

In the following, we show how the Bayesian beliefs are used to maximize the objective functions mentioned in the original PRP paper. Note that, although the mathematical development here is similar to the one of the original PRP paper, we provide the necessary proofs using a probabilistic model over all documents, whereas the original paper only uses a comparison between any two documents.

The probabilistic model of the BPRP considers for each document two states: relevant and non-relevant to the current query. Therefore, the sample space of the BPRP consists of all the possible relevance configurations, the set of all possible relevance states of the documents in the collection to the query $\hat{q}$:

$$\Phi_{\hat{q}} := \underbrace{\{0,1\} \times ... \times \{0,1\}}_{|\mathcal{D}| \text{ times}} \tag{2}$$

where each component of $\Phi_{\hat{q}}$ corresponds to an arbitrary but fixed document. For a particular relevance configuration $\phi \in \Phi_{\hat{q}}$, we define the relevance state of document d as $\phi_d \in \{0,1\}$ (using the fixed position of d in $\Phi_{\hat{q}}$), and we define the (trivial) relevance random variable of d as $\hat{R}_{\hat{q},d}(\phi \in \Phi_{\hat{q}}) := \phi_d$. Note that the random variable $\hat{R}_{\hat{q},d}$ differs from the relevance random variable R defined in Eq. (1): $\hat{R}_{\hat{q},d}$ states the relevance of a *given* query $\hat{q}$ and document d in a (unknown) relevance configuration $\phi \in \Phi_{\hat{q}}$ while R states the relevance of *any* query and document in the collection. The probability $P_{\Phi}(\hat{R}_{\hat{q},d}{=}1)$ is the probabilistic relevance, our degree of belief that document d is relevant to query $\hat{q}$.

In the following, we explicitly show how the probabilities of relevance are used to maximize the objective functions of the BPRP, using the example of the expected utility. For the current query $\hat{q}$ and a ranking $\mathbf{d}$, we define the utility at rank n as a function of the relevance random variables:

$$U^n_{\hat{q},\mathbf{d}}(\phi \in \Phi) := \sum_{j=1}^{n} U(\hat{R}_{\hat{q},d_j}(\phi)) \tag{3}$$

where n is the rank at which the user stops reading, $\mathbf{d}$ is a ranking of the collection $\mathcal{D}$, d_j is the jth document in the ranking $\mathbf{d}$ (note that d_j is usually not the jth component in Φ), and $U(r \in \{0,1\})$ is a utility function that assumes that the user issuing $\hat{q}$ has utility u_r from a relevant document ($r = 1$) and a utility u_n from a non-relevant document ($r = 0$). Using the basic laws of expectations, the expected

utility for a user who reads the top-n documents of a particular ranking $\mathbf{d}$ is:

$$E[U^n_{\hat{q},\mathbf{d}}] = \sum_{j=1}^{n} E[U(\hat{R}_{\hat{q},d_j})]$$
$$= \sum_{j=1}^{n} \Big(u_r \; P_{\Phi}(\hat{R}_{\hat{q},d_j}{=}1) + u_n P_{\Phi}(\hat{R}_{\hat{q},d_j}{=}0) \Big) \tag{4}$$

where all variables are defined as above.

Based on the probabilistic model above, it can be seen that the BPRP maximizes the expected utility for the current query because the ranking

$$(d_1, ..., d_{|\mathcal{D}|}) \text{ with } P_{\Phi}(\hat{R}_{\hat{q},d_1}{=}1) \geq ... \geq P_{\Phi}(\hat{R}_{\hat{q},d_{|\mathcal{D}|}}{=}1)$$

satisfies

$$(d_1, ..., d_{|\mathcal{D}|}) = \operatorname*{argmax}_{\mathbf{d}} E[U^n_{\hat{q},\mathbf{d}}]$$

where $\mathbf{d}$ iterates over all possible rankings of the documents in the collection. In a similar manner it can be shown that the BPRP maximizes the expectations of the precision and recall of the user issuing the current query reading until a rank n, which can be defined as follows:

$$Prec^n_{\hat{q},\mathbf{d}}(\phi \in \Phi) := \frac{1}{n} \sum_{j=1}^{n} \hat{R}_{\hat{q},d_j}(\phi) \tag{5}$$

$$Rec^n_{\hat{q},\mathbf{d}}(\phi \in \Phi) := \frac{1}{|\mathcal{R}|} \sum_{j=1}^{n} \hat{R}_{\hat{q},d_j}(\phi) \tag{6}$$

Therefore, the BPRP states that documents should be ranked by $P_{\Phi}(\hat{R}_{\hat{q},d}{=}1)$, and ranking models that implement the BPRP have to define this probability for each document $d \in \mathcal{D}$.

3.2 Popularity in the PRP

We propose that the example in the appendix of the original PRP paper uses a different probabilistic model than the BPRP. The model is related to the model used by Maron and Kuhns [1960] that ranks documents by the probability of a document being relevant among *multiple* queries with the same query terms. Note that this probability is different from the one of the BPRP, which considers only a single query. Because the used probabilities of relevance can be seen as popularity measures of documents for queries with the same query terms, we refer to this ranking principle as the Popularity-based Probability Ranking Principle (PPRP). In the following, we show that the PPRP maximizes the expected utility of a search engine serving a random query.

In the PPRP, we consider the sample space to be the set of queries that share a number of properties with the current query. For the purpose of this definition, we consider the set of queries that have the same query terms as the current query:

$$\hat{\mathcal{Q}} := \{q \in \mathcal{Q} | \mathbf{Tx}(q) = \mathbf{Tx}(\hat{q})\},$$

where $\mathbf{Tx}(\hat{q})$ are the query terms of the current query. Note that this definition can be extended to other properties than the equality of query terms, as done in Sect. 4. It is also important to see that *every* query, issued by a user, is a separate element of $\mathcal{Q}$, even for different queries that have exactly the same intent. Based on the defined sample space, we define the relevance random variable of a document $d \in \mathcal{D}$ for a query q:

$$R_d(q \in \hat{\mathcal{Q}}) := R(q, d) \tag{7}$$

where R is defined in Eq. (1). The probability of relevance, which is the probability that document d is relevant to *a random* query in $\hat{\mathcal{Q}}$, is defined as:

$$P_{\hat{\mathcal{Q}}}(R_d = 1) := |\{q \in \hat{\mathcal{Q}} | R_d(q) = 1\}| / |\hat{\mathcal{Q}}|$$

Under the assumption that all users have the same constant utility for reading a relevant document, respectively, a non-relevant document, we can define the utility random variable for a document $d \in \mathcal{D}$ with respect to a query q based on its relevance:

$$U_d(q \in \hat{\mathcal{Q}}) := \begin{cases} u^+ & \text{if } R_d(q) = 1, \\ u^- & \text{otherwise.} \end{cases} \tag{8}$$

where u^+ is the utility for reading a relevant document and u^- is the utility of reading a non-relevant document, with $u^+ > u^-$. Based on the utility of a single document, we define the utility of reading the first n documents of a ranking $\mathbf{d}$:

$$U^n_{\mathbf{d}}(q \in \hat{\mathcal{Q}}) := \sum_{j=1}^{n} U_{d_j}(q) \tag{9}$$

where U_{d_j} is the utility of the jth document in ranking $\mathbf{d}$. It is important to note that the utility $U^n_{\mathbf{d}}$ is different from the utility $U^n_{\hat{q},\mathbf{d}}$ considered in the BPRP defined in Eq. (3). The PPRP utility $U^n_{\mathbf{d}}$ considers a fixed ranking $\mathbf{d}$ and yields the utility for *any* query q, which is defined on the fixed relevance states of the documents in $\mathbf{d}$ to q, while the BPRP utility $U^n_{\hat{q},\mathbf{d}}$ considers a fixed ranking $\mathbf{d}$ *and* query $\hat{q}$ and states the utility for *any* relevance configuration between the two, with the goal to model the uncertainty which of the configurations is reality (in particular, the relevance of a given document is uncertain). Using the basic laws of expectations, the expected utility for a random query $q \in \hat{\mathcal{Q}}$ whose issuer reads n documents of the ranking $\mathbf{d}$, becomes:

$$\begin{aligned} E[U^n_{\mathbf{d}}] &= \sum_{j=1}^{n} E[U_{d_j}] \\ &= \sum_{j=1}^{n} u^+ \; P_{\hat{\mathcal{Q}}}(R_{d_j}{=}1) + u^- \; P_{\hat{\mathcal{Q}}}(R_{d_j}{=}0) \end{aligned} \tag{10}$$

Based on the probabilistic model above, the PPRP maximizes the expected utility of a random query with the same query terms, because the ranking

$$(d_1, ..., d_{|\mathcal{D}|}) \text{ with } P_{\hat{\mathcal{Q}}}(R_{d_1}{=}1) \geq ... \geq P_{\hat{\mathcal{Q}}}(R_{d_{|\mathcal{D}|}}{=}1))$$

satisfies

$$(d_1, ..., d_{|\mathcal{D}|}) = \operatorname*{argmax}_{\mathbf{d}} E[U^n_{\mathbf{d}}]$$

where $\mathbf{d}$ iterates over all possible rankings of the documents in the collection. Therefore, the PPRP states that documents should be ranked by the probability $P(R_d{=}1)$, which refers to the event that document d is relevant of an *unknown* query in Q. Note that this probability is different from the probability $PP(RB_{\hat{q},d}{=}1)$ used in the BPRP, which refers to the uncertain relevance of a document d to the *known* query $\hat{q}$. Ranking models that want to implement this the PPRP and maximize the expected utility for a search engine serving a user with results for a random query from $\hat{\mathcal{Q}}$, have to define the probabilities $P_{\hat{\mathcal{Q}}}(R_d{=}1)$ for each document $d \in \mathcal{D}$.

3.3 Discussion

In this section, we investigated probabilistic models on which the PRP could be based. We found that there are actually two distinct ranking principles, depending on the considered probabilistic model: the BPRP that ranks a document according to our belief of relevance for a *single* query, and the PPRP that ranks a document according to the probability that it is relevant among *multiple* queries with the same query terms. This new perspective on the PRP has the following impact on IR theory.

1. The rankings produced by models that implement the BPRP or the PPRP can be substantially different. To clarify these differences, Figure 1 depicts an example query-document matrix, see also Robertson [2005], of five queries and six documents. Let us assume that the queries have the same representation (e.g., the same query terms), but apart from that, they are different. For example, they were issued for distinct information needs. The shading of each cell denotes the relevance between a query and a document. Based on their relevance pattern, we divide the documents into two groups: d_1, d_2, d_3 and d_4, d_5, d_6. Note that we intentionally chose this extreme relevance pattern to demonstrate the main differences between the two principles. A ranking model following the PPRP ranks the documents d_1, d_2, d_3 above the documents d_4, d_5, d_6 because they are relevant to more queries, in this case three out of five. A ranking according to the BPRP, on the other hand, depends on the degree of belief that the search engine has about the relevance of each document to each individual query. For example, a search engine could use a different document representation for each query, which leads to different degrees of belief according to a BPRP-based model. Figure 1 shows two possible degrees of belief settings of the six documents for the two queries q_1 and q_5.[2] Therefore, the similarity of the results according to the BPRP and the PRPR depend on the query representation used for the PPRP and the relevance pattern for each query, and the model that generates the degrees of belief used for the BPRP.
2. The probabilities of relevance $P_{\Phi}(\hat{R}_{\hat{q},d}{=}1)$ and $P_{\hat{\mathcal{Q}}}(R_d{=}1)$ of the respective principle have to be estimated differently. However, this has not been accounted for in the literature. In the next section, we will investigate models that consider random draws of query-document pairs to estimate these probabilities.

[2] We discuss possible ways to determine the degrees of belief of relevance in the following Sect. 4.

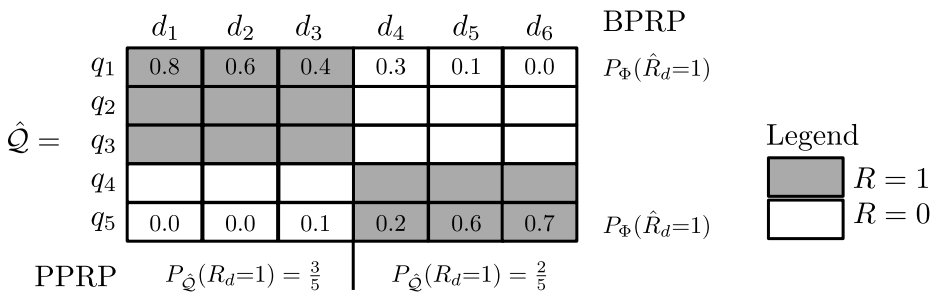

Fig. 1: Comparison of the BPRP and the PPRP based on an example of five queries with the same representation $\hat{\mathcal{Q}} = \{q_1, ..., q_5\}$ and a collection of six documents $\mathcal{D} = \{d_1, ..., d_6\}$. The probabilities in the rows for q_1 and q_5 show two possible sets of beliefs in the relevance of the individual documents for the respective query.

3. Principles stated in recent work build upon the PRP by including the relevance dependencies between documents, see for example [Wang and Zhu, 2009, Chen and Karger, 2006]. However, these principles do not explicitly state on which PRP they are based, although this clearly affects their interpretation and estimation methods.

As a consequence of the discovery that there are two ranking principles, the relationship between each of the ranking models that was originally motivated by the PRP and the two alternative principles have to be analyzed. We provide this analysis for probability of relevance models and language models in the following section and in Sect. 5 respectively.

4 Probability of Relevance Models

Robertson et al. [1982] propose a unified framework of probability of relevance (PR) models, which are generally believed to implement the original PRP. However, Robertson et al. consider draws of *random* query-document pairs in their framework, while the two PRPs consider *given* documents, see Sect. 3. The argumentation of how the differences of those models can be formally overcome is missing in literature. In this section, we investigate under which conditions PR models can be used to define the probabilities used by the respective PRP.

4.1 The Unified Framework of PR Models

Before investigating the relation of the PRP and PR models, we define the four basic probabilistic aspects underlying the unified framework of PR models using a notation based on random variables. We do not use the event-based notation by Robertson et al. [1982], which considers events such as "the document is similar to the current document", because we believe this notation has led to confusion in the comparison of PR models to language models. The first aspect is the considered process, which we already identified to be a drawing of random query-document pairs as stated by Robertson et al.. In the following, we define the remaining three basic probabilistic aspects of PR models.

Sample Space Robertson et al. [1982] do not mention the considered sample space explicitly and refer to the Cartesian product of queries and all documents[3], $\Omega := \mathcal{Q} \times \mathcal{D}^+$, as the considered event space. However, these events are "elementary" events, which we call samples in this paper. This makes Ω the sample space of the unified framework. Note that because Ω is a set of pairs, it cannot be an event space, which is a set of sets.

Event Space The unified framework consists of four models (Models 0−3) that differ in the way that they partition the event space. The partitioning is achieved by features, which are sometimes also referred to as representations or descriptors. Strictly speaking, Model 0−3 are meta models because the unified framework does not explicitly define the considered features. For the discussions below, we give the following abstract definition of features:

$$\mathbf{QF} := (QF_1, ..., QF_m) \tag{11}$$

$$\mathbf{DF} := (DF_1, ..., DF_n) \tag{12}$$

where QF_i is the ith query feature (a function of the query q of a query-document pair $(q, d) \in \Omega$), and $\mathbf{QF}$ is a vector of m query features. DF_i is a document feature (a function of the document d of a query-document pair $(q, d) \in \Omega$), and $\mathbf{DF}$ is the vector of n document features[4]. We refer to $QF(q)$ as the query feature *value* of feature QF for query q, and $DF(d)$ as the document feature *value* of DF for document d. Note that there are also features that are defined on queries *and* documents, for example, the fact that a document was clicked in response to a query. However, following the unified framework, we do not consider such query-document features. For later use, we define two concrete features: let $Q((q, d) \in \Omega) := q$ be the query of a query document pair, and let $D((q, d) \in \Omega) := d$ be the document of the query-document pair. We refer to these features as the trivial query feature and the trivial document feature, respectively. Note that vectors are only one out of multiple mathematical structures to denote features, which we chose to conform to current works in IR.

Additionally to the query and document features, PR models consider the relevance of query-document pairs as a random variable defined in Eq. (1). The combination of query and document feature values and relevance values, induces the event space of PR models. For example, the set $\{(q, d) \in \Omega | R(q, d) = 1\}$ is the relevance event, and the set $\{(q, d) \in \Omega | \mathbf{DF}(d) = \mathbf{DF}(\hat{d})\}$ is the event that a query-document pair has the same document features as the current document.

Probability Measure The unified framework considers a query-document pair $(\hat{q}, \hat{d})$ and uses the conditional probability that any (q, d) pair with the same query

[3] The notion of "all documents" has not been explicitly defined in the unified framework, but could, e.g., be interpreted as "containing the current collection, but extended with other documents that could have belonged to it". An example will be given further on.

[4] In order to keep the notation lean, we denote query features and document features as depending on queries and documents respectively, which is also their intuitive meaning. However, we define them on query-document pairs, to accommodate for the mathematical formalism of the unified framework.

features and document features, is relevant. We define this probability measure from a Frequentist's perspective, similar to Robertson et al. [1982]:

$$P_\Omega(R\,|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), \mathbf{DF}{=}\mathbf{DF}(\hat{d})) := \frac{|\{(q,d) \in \Omega \mid R(q,d){=}1, \mathbf{QF}(q){=}\mathbf{QF}(\hat{q}), \mathbf{DF}(d){=}\mathbf{DF}(\hat{d})\}|}{|\{(q,d) \in \Omega \mid \mathbf{QF}(q){=}\mathbf{QF}(\hat{q}), \mathbf{DF}(d){=}\mathbf{DF}(\hat{d})\}|} \quad (13)$$

where $\hat{q}$ is the current query, and $\hat{d}$ is the current document. Note that Eq. (13) is a definition of a probability measure, which in reality might be estimated using sophisticated machine learning techniques. Equation 13 makes the difference between the BPRP and PRPR on the one hand, and PR models on the other hand apparent: while the BPRP and PPRP consider the probabilities $P_\Phi(\hat{R}_{\hat{q},d} = 1)$ and $P_{\hat{\mathcal{Q}}}(R_d = 1)$ for a particular document d, PR models consider the probability of relevance of random query-document pairs given certain feature values, see Eq. (13).

4.2 PR Models and Their Connection to the PRP

Based on the definition of the basic probabilistic aspects of the unified framework, this section investigates in how far the probability calculated by each of the models can be used in the PRPs, introduced in Sect. 3. For instructive reasons, we consider the models not in their numerical order.

4.2.1 Model 2

Model 2 ranks the document $\hat{d}$ for the query $\hat{q}$ by the probability $P_\Omega(R|Q{=}\hat{q}, \mathbf{DF}{=}\mathbf{DF}(\hat{d}))$. Therefore, Model 2 considers the relevance between the current query and all documents with the same feature values as the current document. If we assume that the only knowledge we have about documents are the features $\mathbf{DF}$, documents with the same feature values are indistinguishable. Under this assumption, it is reasonable to define the probabilistic relevance for document d of the BPRP as the probability of relevance calculated by Model 2:

$$P_\Phi(\hat{R}_{\hat{q},\hat{d}}{=}1) := P_\Omega(R|Q{=}\hat{q}, \mathbf{DF}{=}\mathbf{DF}(\hat{d})) \quad (14)$$

As a result, instances of Model 2 produce a ranking motivated by the BPRP. This connects the BPRP with Model 2. Note that Fuhr [1992] discusses the influence of the chosen document features $\mathbf{DF}$ on the probability of relevance, $P_\Omega(R|Q{=}\hat{q}, \mathbf{DF}{=}\mathbf{DF}(\hat{d}))$. However, the choice of $\mathbf{DF}$ only influences our certainty about the relevance of query-document pairs – the more discriminative $\mathbf{DF}$, the more certain we are about the relevance of a pair – but did not lead to the discovery of the difference between the BPRP and Model 2.

As an illustration that the assumption on which Eq. (14) is based does not always hold, consider the following issue: The probability measure P_Ω is defined on a sample space involving the notion of all documents $\mathcal{D}^+$. The more the feature distribution in $\mathcal{D}^+$ differs from the distribution in collection $\mathcal{D}$, the more unrealistic the assumption in Eq. (14) becomes. In other words, the considered documents

$\mathcal{D}^+$ should be created in such a way that the current collection $\mathcal{D}$ is a representative sample. For example, if we add to a considered collection of web pages $\mathcal{D}$ a collection of news articles to form $\mathcal{D}^+$, the appearance of query terms (the features) better differentiates between relevant and non -relevant documents because journalists have a clearer language usage. However, the probability measure P_Ω in Eq. (14), based on $\mathcal{D}^+$, no longer necessarily reflects our belief of the relevance of documents in $\mathcal{D}$. Therefore, maximizing the expected utility, which is based on these beliefs, is in this case not a good objective.

Furthermore, because Model 2 considers only the current query, it is unsuitable for the PPPR, which considers multiple queries.

4.2.2 Model 1

Model 1 ranks the document $\hat{d}$ to query $\hat{q}$ by the probability $P_\Omega(R\,|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), D{=}\hat{d})$. In other words, Model 1 considers for each document the probability of relevance of query-document pairs where the queries have the same query feature values as the current query, and the document is the current document. Therefore, on the one hand, the probability of relevance calculated by Model 1 is not necessarily suitable to express the probabilistic relevance in the BPRP, which only considers the current query.[5] On the other hand, the probability of relevance calculated by Model 1 can be used in the PPRP by assuming the following equality:

$$P_{\hat{\mathcal{Q}}}(R_{\hat{d}} = 1) = P_\Omega(R\,|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), D{=}\hat{d})$$

where $P_{\hat{\mathcal{Q}}}(R_d = 1)$ is the probability of relevance of document d considered by the PPRP considering the query set $\hat{\mathcal{Q}} := \{q \in \mathcal{Q} | \mathbf{QF}{=}\mathbf{QF}(\hat{q})\}$. This definition effectively connects Model 1 and the PPRP.

In Model 2, the choice of documents considered as "all documents" $\mathcal{D}^+$ limited the adequacy of the connection between probabilities calculated in the model and the ones of the BPRP. The situation for Model 1 is comparable, but now the choice of queries considered as "all queries" $\mathcal{Q}$ limits the adequacy of the connection between the model and the PPRP. If the queries in $\mathcal{Q}$ do not reflect the current distribution of information needs, the maximization of the expected utility of the PPRP, defined by the probabilities $P_\Omega(R\,|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), D{=}\hat{d})$ is not a good ranking objective.

Note that apart from the interpretation of the probability measure of Model 1 for the PPRP, it can also be used for the BPRP, by defining the following new document feature for the current query

$$PO(d \in \mathcal{D}) := P_\Omega(R\,|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), D{=}d)$$

where PO is a document feature expressing the popularity of a document among queries with the same query feature values. We can use this document feature in the probability of relevance measure from Model 2, $P_\Omega(R|Q{=}\hat{q}, PO{=}PO(\hat{d}))$, to implement the BPRP. If we consider this measure as a function of $PO(d)$, its shape will depend on the considered query. For example, for many queries the probability of relevance of Model 2 will increase with the popularity PO.

[5] Although we are free to choose our Bayesian degree of belief, it depends on the considered query if Model 1's probability of relevance is a good measure in the BPRP.

However, for other queries popular documents with a high PO might have a lower probability of relevance in Model 2. For example, this might hold for queries posted by researchers, who are sometimes not interested in popular documents.

4.2.3 Model 3

Model 3 ranks the document $\hat{d}$ for the query $\hat{q}$ by the probability $P_\Omega(R|Q{=}\hat{q}, D{=}\hat{d})$, where Q and D are the previously defined trivial query and document features. Model 3 is a special case of Model 2 that uses the trivial document feature instead of the general document features $\mathbf{DF}$, and analogously it is a special case of Model 1. Therefore, in principle Model 3 can be used to implement both the BPRP and the PPRP. However, we find that the consideration of Model 3 and hence its use in the BPRP or PPRP is only of academic nature. To see this, we expand the Model's conditional probability of relevance by its definition:

$$\begin{aligned}
& P_\Omega(R=1|Q=\hat{d}, D=\hat{d}) \\
= & \frac{P_\Omega(\{(q,d) \in \Omega | R(q,d)=1\} \cap \{(q,d) \in \Omega | q=\hat{q}, d=\hat{d}\})}{P_\Omega(\{(q,d) \in \Omega | q=\hat{q}, d=\hat{d}\})} \\
= & \begin{cases} \frac{P_\Omega(\{(\hat{q},\hat{d})\})}{P_\Omega(\{(\hat{q},\hat{d})\})} & \text{if } R(\hat{q},\hat{d})=1, \\ \frac{P_\Omega(\{\})}{P_\Omega(\{(\hat{q},\hat{d})\})} & \text{otherwise.} \end{cases}
\end{aligned}$$

We can see that, for any probability measure P_Ω that maps the empty event $\{\}$ to zero probability, this probability can only take two values: one, if document $\hat{d}$ is relevant to query $\hat{q}$, and zero otherwise. Therefore, ranking by the probability of relevance of Model 3 would solve the ad-hoc retrieval task (we can tell the relevance of each document to each query). However, we propose that it seems unlikely that one can ever find a method to accurately estimate a probability measure for the mentioned events.

4.2.4 Model 0

Model 0 ranks the document $\hat{d}$ to query $\hat{q}$ by the probability $P_\Omega(R|\ \mathbf{QF}{=}\mathbf{QF}(\hat{q}), \mathbf{DF}{=}\mathbf{DF}(\hat{d}))$. Therefore, Model 0 considers for each document the probability of relevance of multiple query-document pairs with equal feature values. As a result, Model 0 considers multiple queries in contrast to Model 2, which only considers the current query. Furthermore, Model 0 considers multiple documents in contrast to Model 1, which considers only a single document for multiple queries. Therefore, Model 0 cannot be used in the BPRP, which considers a single query, or the PPRP, which considers each document in multiple queries.

4.3 Discussion

In this section, we investigated the four basic probabilistic aspects of the unified framework of PR models (Models 0–3). In the following, we discuss the possible connections of PR models and the BPRP or the PPRP:

1. We found that the probabilities calculated by Model 2 and Model 3 can be used for the BPRP. However, we found that Model 3 is only of academic interest because it requires knowledge of the relevance of the currently considered query-document pair. Furthermore, because Model 2 is only defined on the current query, there is often no, or only limited, training data available to estimate the model's parameters.
2. Model 1 considers multiple queries with the same query feature values for one particular document, and the calculated probability of relevance can be used for the PPRP.
3. Current search approaches use relevance examples from seen query-document pairs and therefore rank similar to Model 0. These approaches often produces strong performance, see for example the literature about learning to rank Liu [2009]. However, because Model 0 cannot be used to implement the BPRP or the PPRP, these principles cannot explain the strong performance of these approaches. Therefore, if the development of these approaches should be guided by a ranking principle there are the following two alternatives: first, the underlying Model 0 must be shown to implement another, possibly new, ranking principle, or second, search approaches have to find ways to estimate parameters of different models using past queries.
4. The features of documents are in practice often unique in the collection. If we consider only the current collection ($\mathcal{D}^+ = \mathcal{D}$), Model 2 is equivalent to Model 3. Note, however, that instances of Model 2 usually consider a larger set of documents that have a similar distribution. If we consider Model 2 as a classifier, see for example Lewis [1998], this assumption is the same as in many works in machine learning [Bishop, 2006].

5 Language Models

In this section, we compare PR models presented in Sect. 4 to the following four popular language models: the *query likelihood model* by Ponte and Croft [1998], the language model by Hiemstra [2001], which we refer to as *Hiemstra's model*, the *risk minimization model* by Zhai and Lafferty [2006], and the *relevance model* by Lavrenko and Croft [2003]. Note that we focus here on the probabilistic aspects of the mentioned models because their more conceptual aspects are discussed in other work, for example the one mentioned above. Before analyzing the connection between PR models and these mentioned above models, we define the basic probabilistic aspects, which are common to all of them.

5.1 Common Elements in Language Models

The four language models discussed in this paper have in common that they consider term draws. For the definition of the individual models, we define the (in some cases partial) sample space of drawing terms and the random variables expressing

the outcome of this process as follows:

$$\mathcal{T}_n := \overbrace{\mathcal{T} \times ... \times \mathcal{T}}^{n \text{ times}} \tag{15}$$

$$T_i(\mathbf{t} \in \mathcal{T}_n) := \text{the } i\text{th drawn term in } \mathbf{t} \tag{16}$$

$$\mathbf{T}(\mathbf{t} \in \mathcal{T}_n) := \mathbf{t} \tag{17}$$

where $\mathcal{T}_n$ is the sample space of drawing n terms (the set of all possible term combinations resulting from n term draws), the random variable T_i states the ith drawn term, and $\mathbf{T}$ denotes a sequence of drawn terms (a vector of random variables).

Because it will be used in the comparison between PR models and language models, please note that there is a difference between the random variable for the ith query term, Tx_i, see Sect. 2, which is defined on queries, and the ith drawn term, T_i, which is defined on the drawn text. For example, given the current query $\hat{q}$, its ith term $Tx_i(\hat{q})$ is a fixed value, whereas T_i denotes a random term.

Note that Roelleke and Wang [2006] consider a slightly different probabilistic model for language models, which is based on a sample space of text locations, where locations contain terms. We use term sequences instead of locations as the sample space of language models, because the simpler notation suffices for our needs. Nevertheless, it can be shown that using text locations as the sample space of language models does not change the findings in this paper.

For the probability measure in language models, we limit our discussion to unigram models, which are most frequently used in IR. In unigram models, we assume terms are independently drawn from a multinomial distribution. The probability measure of drawing a sequence of terms is hence:

$$P_d(\mathbf{T}{=}\mathbf{t}) := \prod_{i=1}^{L(\hat{q})} P_d(T_i{=}t_i) = \prod_{i=1}^{L(\hat{q})} \theta_i(d) \tag{18}$$

where $\mathbf{t}$ is the considered term sequence, $L(\hat{q})$ is the length of the sequence, t_i is the ith term, $P_d(T_i{=}t_i)$ is the probability of drawing the ith term from document d, and $\theta_i(d)$ is the parameter of the multinomial distribution for term t_i in the language model of document d.

Note that the language model parameters $\boldsymbol{\theta}(d)$ of document d are usually unknown and estimated from the document text. For this estimation, some literature, see for example Zhai and Lafferty [2004], uses Bayesian estimators that are also based on a probabilistic model. Here, the model parameters are usually included in the notation: $P_d(T{=}t|\boldsymbol{\theta}(d))$. In this paper, we focus on probabilistic models for ranking, and assume that we can determine the language model parameters with sufficient precision. Therefore, we exclude the parameter estimation from our discussion, and use the parameters in the probability notation.

5.2 Individual Language Models

In order to be able to compare language models to PR models, we define the basic probabilistic aspects of the four language models mentioned above, using the common definitions from Sect. 5.1.

5.2.1 Query Likelihood Model

Ponte and Croft [1998] propose the *query likelihood model* that considers for each document a hypothetical process in which $L(\hat{q})$ terms are drawn. It ranks the documents by the likelihood, $P_d(\mathbf{T}{=}\mathbf{Tx}(\hat{q}))$, of the event that the query terms were drawn from their language model.[6] The *event space* hence consists of all possible term sequences.

5.2.2 Hiemstra's Model

Hiemstra [2001] proposes a language model that considers a process of generating the document that the user has in mind, and the terms the user draws using the document's language model. Using the common definitions of language models in Sect. 5.1, we define the following random variables:

$$\begin{aligned} \mathcal{H} &:= \mathcal{T}_{L(\hat{q})} \times \mathcal{D} \\ D'((\mathbf{t}, d) \in \mathcal{H}) &:= \text{the document } d\text{, which the user has in mind} \end{aligned}$$

where $\mathcal{H}$ is the model's *sample space*, and D' states the document the user has in mind. The *event space* is defined by the values of the random variables D' and $\mathbf{T}$, see Eq. (17). Hiemstra's model ranks a document $\hat{d}$ by the probability that the user had this document in mind, given that the query terms were observed: $P_{\mathcal{H}}(D'{=}\hat{d}|\mathbf{T}{=}\mathbf{Tx}(\hat{q}))$. Note that in practice this probability is "reversed" using Bayes' law, leaving out components that do not influence the ranking.

5.2.3 Risk-minimization Model

Zhai and Lafferty [2006] propose the *risk-minimization model* that considers drawing a single term (the *sample space* is $\mathcal{T}_1$) from a query language model and from the language model of each document. The model ranks a document d by the Kullback-Leibner (KL) divergence between the two distributions:

$$KL(P_q||P_d) := \sum_{t \in \mathcal{T}} P_q(T{=}t)\ \log\left(\frac{P_q(T{=}t)}{P_d(T{=}t)}\right) \tag{19}$$

where P_q is the probability measure of the query language model, P_d is the probability measure of the current document's language model, and T is the random variable expressing the drawn term.[7]

Note that the literature rarely mentions that the risk-minimization framework considers only a single term draw, which is different from considering $L(\hat{q})$ term draws in the query likelihood model or Hiemstra's model. However, that Eq. (19) considers a single term draw can be seen from the original definition of the KL divergence, which measures the difference between a true distribution and a proposed distribution of sending a *single* message, see Kullback and Leibler [1951].

[6] Ponte and Croft [1998] use binary random variables expressing the event that a certain term was drawn or not. We follow more common notation and consider events of drawing query terms, which leads to equivalent results.

[7] The KL divergence is defined over the domain $\mathcal{T}$ of the random variable T, not to be confused with the sample space $\mathcal{T}_1$ of the probabilistic model (although in this case equivalent)

5.2.4 Relevance Model

Lavrenko and Croft [2003] propose the *relevance model* that considers drawing a single term (the *sample space* is $\mathcal{T}_1$) from each document's language model. The relevance model ranks a document $\hat{d}$ by the negative cross entropy (CE) between the term distribution of the relevance language model[8] and the document's language model:

$$-CE(P_r||P_d) := -\sum_{t \in \mathcal{T}} P_r(T{=}t)\ \log\left(P_d(T{=}t)\right)$$

where the term distribution of the document model $P_d(T{=}t)$ is defined by the probabilistic model in Sect. 5.1, and hereunder we define the term distribution of the relevance language model.

The relevance language model first considers drawing a relevant document and then a term from this document [Lavrenko and Croft, 2003, p. 24]. Therefore, the *sample space*, the random variable for the drawn document, and the relevance of the relevance language model are defined as follows:

$$\begin{aligned} \mathcal{RM} &:= \{(d,t) \in \mathcal{D} \times \mathcal{T}_1 | R(\hat{q},d){=}1\} \\ D''((d,t) \in \mathcal{RM}) &:= d \text{ was drawn} \\ R'((d,t) \in \mathcal{RM}) &:= R(\hat{q},d) \end{aligned}$$

where $\mathcal{RM}$ is the sample space of the relevance language model (the set of relevant documents with the corresponding drawn terms), $\hat{q}$ is the current query, D'' states the drawn relevant document, and R' states the relevance of the drawn document to the current query, which is always one because only relevant documents are considered. The probability of drawing a term t from the relevance language model is the marginalization over documents:

$$P_r(T{=}t) := \sum_{\{d \in \mathcal{D} | R(\hat{q},d){=}1\}} P_r(T{=}t|D''{=}d) P_r(D''{=}d)$$

Note that the set $\{d \in \mathcal{D} | R(\hat{q},d){=}1\}$ is unknown in practice, and Lavrenko and Croft [2003] and others propose estimation methods for this probability.

5.3 PR Models vs. Language Models

Given the definitions of PR models and language models in Sect. 4 and above, we now investigate whether language models can be used in the definition of PR models. Table 2 summarizes the models' definitions.

We find that PR models and language models exhibit fundamental differences on the level of the underlying process, the sample space, event space, and probability measure. These differences are discussed in the following paragraphs. Note that there is related work that proposes that PR models and language models are related. We discuss the differences between these findings and our work in Sect. 6.

[8] We use "relevance model" to refer to the ranking model and relevance "language model" for the probabilistic model against which document models are compared.

Basic probabilistic aspect	PR Models	Query likelihood model	Hiemstra's model	Risk minimization model	Relevance model
Processes	A random query-document pair is drawn.	Each document produces $L(\hat{q})$ terms.	User finds a random document relevant and draws $L(\hat{q})$ terms from it.	Each document produces one term and the query (user) produces one term	Each document produces one term, and a random, relevant doc. produces one term
Sample space	$\Omega = \mathcal{Q} \times \mathcal{D}^+$	$\mathcal{T}_{L(\hat{q})} = \underbrace{\mathcal{T} \times ... \times \mathcal{T}}_{L(\hat{q}) \text{ times}}$	$\mathcal{H} = \mathcal{T}_{L(\hat{q})} \times \mathcal{D}$	Document model $\mathcal{T}$ Query lang. model $\mathcal{T}$	Document model $\mathcal{T}$ Rel. lang. model $\mathcal{RM} = \mathcal{T} \times \mathcal{D}$
Event space (induced by the considered random variables)	Unified framework $R : \Omega \to \{0, 1\}$ $DF_i : \Omega \to \mathcal{F}$ $QF_i : \Omega \to \mathcal{QF}$ $i \in \{1, ..., L(\hat{q})\}$ where DF_i and QF_i correspond to query term $Tx_i(\hat{q})$	$T_i : \mathcal{T}_{L(\hat{q})} \to \mathcal{T}$ $i \in \{1, ..., L(\hat{q})\}$	$T_i : \mathcal{H} \to \mathcal{T}$ $i \in \{1, ..., L(\hat{q})\}$ $D' : \mathcal{H} \to \mathcal{D}$	Document model $T : \mathcal{T} \to \mathcal{T}$ Query lang. model $T : \mathcal{T} \to \mathcal{T}$	Document model $T : \mathcal{T} \to \mathcal{T}$ Rel. lang. model $T : \mathcal{RM} \to \mathcal{T}$ $D'' : \mathcal{RM} \to \mathcal{D}$ $R' : \mathcal{RM} \to \{1\}$
Considered probabilities and distributions	Unified framework $P_\Omega(R{=}1\|\mathbf{QF}{=}\mathbf{QF}(\hat{q}), \mathbf{DF}{=}\mathbf{DF}(\hat{d}))$	$P_d(\mathbf{T} = \mathbf{Tx}(\hat{q}))$	$P_\mathcal{H}(D'{=}d\|\mathbf{T}{=}\mathbf{Tx}(\hat{q}))$	Document model $P_d(T)$ Query lang. model $P_q(T)$	Document model $P_d(T)$ Rel. lang. model $P_r(T\|R{=}1)$

Table 2: Comparison between PR models and language models. Notation: q is a query, d is a document, t is a term, $L(\hat{q})$ is the query length, $Tx_i(\hat{q})$ are the query terms, $\mathcal{Q}$ is the set of queries, $\mathcal{D}$ is the set of documents, DF are document features with hypothetical range $\mathcal{F}$, QF are query features with hypothetical range $\mathcal{QF}$, D' is the document the user found relevant, D'' is the drawn relevant document, and R' is the (constant) relevance.

Process PR models and language models differ in the process they describe. Although not often mentioned in the literature, we believe this is worth mentioning because it clarifies the correspondence between the process described by the model and the real-world ranking process. On the one hand, PR models envision a process of uncertain relevance of a documents. On the other hand, the mentioned language models consider different processes. In the query likelihood model, a document seems to perform the process, which can be deduced from the common jargon "a term is produced by a document". In Hiemstra's model, the user draws documents and terms. In the risk-minimization model, a single term is produced by a document and the query language model is produced by the language of the user posing the query. Finally, in the relevance model, a single term is produced by the document, but it is unclear who performs the process of the relevance language model.

Sample Space PR models consider query-document pairs, whereas from the four discussed language models, only Hiemstra's model considers drawing documents in connection with the current document[9]. Additionally, while PR models consider queries (objects) in their sample space, language models mainly consider terms in their sample space.

Event Space PR models consider the event of a query-document pair having certain query feature values, document feature values and relevance status. The feature values and the relevance status are fixed for a given query-document pair, although unobservable in the case of relevance. In contrast, language models consider mainly events that we cannot observe as, for example, a term t being produced. For the difference of query features, and events of drawing query terms from language models, see the discussion in Sect. 5.1. Furthermore, the use of a relevance event in language models is different from PR models. The query likelihood model and the risk-minimization model do not mention relevance. Hiemstra's model assumes a single relevant document (random variable D'), which has been mentioned by Spärck-Jones et al. [2003]. What has not been mentioned is that Hiemstra's model also assumes that the relevance of a document is random, which can be seen from the fact that the value of the random variable D' is functionally dependent on the drawn sample. Finally, although the relevance variable in the relevance language models is used in a similar way as the relevance variable of PR models, they only consider relevant documents, such that the role of the relevance variable R' is mainly for reasons of clarity.

Probability Measure PR models and language models also differ in the quantities, mainly probabilities of events, they consider for ranking. On the one hand, PR models consider for each document the probability of relevance, with one probabilistic model for all queries and documents. Language models, on the other hand, consider a variety of events. The query likelihood model considers for each document a separate probabilistic model, which describes the drawing of terms from the respective document. Hiemstra's model considers a single probabilistic model

[9] In the relevance model, document draws are only considered in the relevance language model.

per query, similar to PR models. However, instead of varying features in the probability measure, the model varies the documents the user could have had in mind. The risk-minimization model and the relevance model do not consider single probabilities but compare distributions of drawing single terms from a document with a query language model or a relevance language model, respectively.

5.4 Discussion

In this section, we investigated whether the differences between PR Models and language models can be overcome from a probabilistic perspective. From the comparison in Sect. 5.3, we can see that language models and PR models differ in every basic probabilistic aspect. Therefore, we propose that it is unlikely that one can connect the PR models and language models. One could raise the question whether language models could also be directly connected to the BPRP and/or the PPRP. This would require a formal motivation as to why the probabilities calculated by the individual language models represent a suitable degree of belief of relevance for the BPRP or the probability of being relevant among similar queries in the PPRP. However, given the fundamental differences between all aspects of both types of the respective probabilistic models, we argue that such a connection is equally unlikely as the connection between PR models and language models. In summary, the above finding has the following impact on IR theory: language models cannot be motivated by the BPRP or the PPRP because the respective probabilistic models are not comparable to those models or to PR models.

Additionally, the careful mutual comparison of the four discussed language models on the level of basic probabilistic aspects revealed that these language models also substantially differ among themselves. This fact has not been stressed in the literature so far, and we propose a further investigation of these differences and their consequences as future work.

6 Related Work

This paper is not the first to investigate the relationship between probabilistic models in IR. In the following, we will discuss previous contributions and point out their relationship to this paper.

Cooper [1994] proposes that one should refer to the PRP as a hypothesis, because the example that he contributed to the original publication by Robertson [1977] would contradict the principle's proof. In this work, we show that the example does not contradict the main text but the main text and the example refer to two different principles. Crestani et al. [1998] present an overview of estimation methods for the probability of relevance in PR models, therefore focusing on modeling the probability measure of PR models. In this paper, we focus on the comparison of probabilistic models Furthermore, Chen and Karger [2006] propose to rank documents according to the expected value of other metrics than the one proposed in the PRP. Chen and Karger's work is orthogonal to the content of this paper because it proposes new objective functions, whereas we consider the differences between the probabilistic models and principles.

The following works have compared PR models and language models. The proponents of a connection between PR models and language models derive the probabilities calculated by PR models and language models from the probability of relevance given a particular document and a particular query, see Lafferty and Zhai [2003], Luk [2008] and Zhai [2008]. Their contributions are difficult to compare to our work because the basic assumptions differ in at least the following aspects.

1. On the one hand, the proponents assume an event space of the crossproduct of queries, documents, and the possible relevance status of the two, $\mathcal{Q} \times \mathcal{D} \times \{0, 1\}$. On the other hand, we consider a sample space of query-document pairs, $\mathcal{Q} \times \mathcal{D}^+$, and an event space of relevance status and feature values, as originally proposed by the unified framework of PR models by Robertson et al. [1982].
2. On the one hand, the proponents derive language models and the binary independence model (BIM) by Robertson and Spärck-Jones [1976] from the probability of relevance given the current query and document $P(R|q, d)$ defined on the event space $\mathcal{Q} \times \mathcal{D} \times \{0, 1\}$. The proponents consider this probability similar to the probability of relevance used in Model 3 of the unified framework. In the derivation, they assume that the probability of a query given a document can be approximated by the language model based probability of the query terms given the document, which is $P(q|d) \simeq P(\mathbf{Tx}(q)|d)$ in our notation. Furthermore, they assume that the BIM uses an approximation of the probability of the document given relevance $P(d|r) \simeq P(\mathbf{A}(d)|r)$, where $\mathbf{A}(d)$ are binary attributes of the document d. On the other hand, we consider language models as explicitly defined in this paper, and the unified framework of PR models, as originally proposed. We find that the respective sample spaces, event spaces, and probability measures are fundamentally different. Additionally, Robertson et al. [1982] present the BIM as an instance of Model 2, where the attributes $\mathbf{A}$ are used as the documents features used in the model, and not as an approximation of Model 3, as suggested by the proponents.

In summary, the proponents take a different point of view on the connection of PR models and language models. From our point of view, as we argued in Sect. 5.3, we have to conclude that the differences between the PR models and language models cannot be overcome on the level of probabilistic models. Note that Spärck-Jones et al. [2003] and Robertson [2005] already pointed out the differences between PR models and language models in terms of event spaces. The current paper goes even further: we consider all four basic aspects of probabilistic models, and we find additional differences in the PRP and PR models.

Roelleke and Wang [2006] establish a link between the BIM and language models on the level of ranking functions. They focus on documents with the same term occurrences (see their Theorem 2), which correspond to a single point in the domain of the ranking function of the BIM (an existing PR model). This approach is complementary to our paper: we investigate the connection between probabilistic models, whereas Roelleke and Wang investigate connection between ranking functions that are derived from these models. Note that although we focus in this paper on the probabilistic models of PR models and ranking principles, we showed in Aly and Demeester [2011] an alternative connection between the mentioned ranking functions compared to the connection proposed by Roelleke and Wang.

7 Conclusions

In this paper, we revisited the definition of the following probabilistic IR models and their connection with each other: first, the probabilistic model considered by the probability ranking principle (PRP), second, the probability of relevance (PR) models, and finally, language models.

The first issue treated in this paper concerned the probabilistic model of the PRP as well as the objectives followed by that principle, which had not been explicitly defined in the literature. We proposed two probabilistic models that maximize different objective functions. First, the belief probability ranking principle (BPRP) ranks documents based on the belief that a document is relevant to the current query, which is expressed by the probability of relevance. We showed that the BPRP maximizes the expected utility for the current query, which can also be shown for the expected precision and expected recall. Second, the popularity probability ranking principle (PPRP) ranks documents based on the probability that a document is relevant to a query from a set of queries with the same query terms (or feature values in the more general case). We showed that the PPRP maximizes the expected utility of a search engine serving a random query from the *set of* queries with the same features. We found that the differences between the principles, which for example influences the goals of parameter estimation methods, is not always reflected in the literature that is based on the PRP. We identified the BPRP as the more desirable principle than the PRPR, because the BPRP optimizes the effectiveness for each individual query while the PPRP focuses on queries with the same representation.

Furthermore, in Sect. 4.2 we investigated for each of the four models of the unified framework of PR models by Robertson et al. [1982] whether the calculated probabilities can be used in the BPRP or the PPRP. We found that Model 2 and Model 3, which both consider only the current query, can be used to define the probabilistic relevance of the BPRP, under the assumption that we cannot differentiate between distinct documents with the same feature values. Model 1 considers for each document the probability that this document is relevant among queries with the same query features. We showed that the probability calculated by Model 1, but also the Model 3 probability, can be used in the PPRP. We also found that Model 3 is mainly of academic interest because its definition only allows a probabilistic relevance of 0 or 1, depending on the relevance of the only considered query-document pair. Therefore, Model 2 was the only model of the unified framework that can be realistically used to implement the BPRP. A major weakness of Model 2 is that it partitions the sample space of the unified framework by individual queries. Therefore, example-based learning methods cannot use examples from past queries for parameter estimation. Model 0, which considers query-document pairs with the same query features and document features, cannot be used in the BPRP or the PPRP because it considers multiple queries and documents at the same time.

Additionally, we investigated the difference between PR models, which consider random query-document pairs, and language models, which consider term draws. Previous work proposed that there is a connection between PR models and language models, see for example Lafferty and Zhai [2003], Luk [2008], Zhai [2008]. However, we found that those works used a slightly different definition of PR models, compared to the original publication by Robertson et al. [1982]. From

the definition of the probabilistic model of PR models and language models as given in this paper, we found that the two types of models differ in every basic probabilistic aspect.

According to the authors, the main merit of this paper is to bring insights and to open new perspectives, which can be used as research directions in the future. We propose some of these research directions as follows:

1. Recently, the research community has been considering ranking principles that address diversity and relate them to the PRP. However, we found that there are actually two distinct PRPs. Therefore, we believe that an important research direction is to investigate in how far this distinction affects ranking principles for diversity.
2. Model 0, which depends on query and document features, is one of the most widely used ranking models in practice but we found that it does not follow the BPRP or the PPRP. Therefore, finding out which principle Model 0 follows, if there is one, is an important research direction.
3. We identified Model 2 as the most promising of the unified framework because it optimizes effectiveness measures for individual queries. However, Model 2 cannot use example relevance judgments of past queries for parameter estimation. On the other hand, there are also other learning methods than example-based methods, which have not received much attention so far. We propose a more thorough investigation of such methods as a promising research direction.
4. Language models would benefit from a connection to a ranking principle, which can guide their development orthogonally to the improvement of their scoring functions axiomatically. Therefore, we believe a promising research direction is to define new ranking principles that language models do follow. An alternative direction is to investigate the similarity of language model ranking functions with score functions from models that do follow an existing ranking principle, akin to but more general than our approach in Aly and Demeester [2011] (Sec. 5) or the one by Roelleke and Wang [2006].

Acknowledgements The work reported in this paper was funded partly by the EU Project AXES (FP7-269980) and carried out at the University of Twente, in the Netherlands, and partly by Ghent University – iMinds in Flanders. We would like to thank Djoerd Hiemstra, Arjen de Vries and the anonymous reviewers for their comments that greatly helped to improve this work and especially forced us to reflect critically on our motivations and results.

A Extract from Probability Theory

Similar to Feller [1968] and Manning and Schuetze [1999, chap. 2], we use the following definitions of probability theory. The definition of a probabilistic model uses four basic aspects: a *sample* is a possible outcome of a process[10]. The corresponding *sample space* is the set of all possible samples. An *event* is a subset of the sample space. An *event space* is a set of events. A *probability measure* is a function that maps events to probabilities. We use a subscript to indicate the sample space on which the measure is defined. For example $P_X : \mathcal{E} \to [0:1]$ is a probability measure defined on the event space $\mathcal{E}$ for the process connected with event space X.[11] A *conditional probability* is the probability of an event e_1 given a conditioning event e_2,

[10] Samples are also referred to as basic outcomes, sample points or elementary events.

[11] Other literature assumes that each sample has an elementary probability, say $\mu(s \in X) \in [0:1]$, where X is the sample space, and the probability of an event is then defined as the sum

which is defined as the probability of the intersection of events divided by the probability of the conditioning event: $P(e_1|e_2) = P(e_1 \cap e_2)/P(e_2)$. A *random variable* is a function of a sample.[12]

The literature often refers to Thomas Bayes in the context of probability theory. However, there are at least three concepts in probability theory that are attributed to Thomas Bayes, which make such references ambiguous. In this paper we differentiate three contributions of Thomas Bayes: 1) *Bayes rule*, which establishes the equality between a conditional probability and its inverse together with two priors regardless of probability measure and event space, 2) the *Bayesian estimation framework* where estimated parameters are assumed to have a prior distribution and one chooses, for example, the parameters with the maximum a posteriori probability, and 3) *Bayesian beliefs*, as opposed to Frequentist probabilities, where the random process can only be executed once, see Bishop [2006]. A typical example for Bayesian beliefs is "the probability that the polar caps melt in 10 years". Here, it is clear that the polar cap can only melt or not and this process cannot be repeated. The reason for establishing a Bayesian belief is to be able to reason about consequences, for example, by means of a utility function. For the interested reader, Cox [1946] uses a similar to Bayesian beliefs.

B Connections between Probabilistic Models and Their Objectives

This paper treats the connection between probabilistic ranking principles and their objectives as well as several probabilistic models. This section formalizes our notion of a connection. We use the term connection in two different senses, which we define as follows:

1. A connection between a probabilistic ranking principle X and an objective, represented by the maximization of a function Y, exists if following that principle implies that the objective will be met. In mathematical terms, a connection exists if a ranking

$$(d_1, ..., d_{|\mathcal{D}|}) \text{ follows } X$$

   implies

$$(d_1, ..., d_{|\mathcal{D}|}) = \underset{\mathbf{d}}{\operatorname{argmax}}\, Y(\mathbf{d}).$$

   where $\mathbf{d}$ iterates over all possible rankings of the documents in the collection $\mathcal{D}$.
2. Let $M_1 =< S_1, E_1, P_1 >$ and $M_2 =< S_2, E_2, P_2 >$ be two probabilistic models, where the components are the sample space, the event space, and the probability measure respecitively. We say, a connection between M_1 and M_2 exists, if there is a justifiable correspondence between any event $e_1 \in E_1$ and an event $e_2 \in E_2$, such that we can assume the equality of the events' probabilities $P_1(e_1) = P_2(e_2)$. Note that such a correspondence between events is often subjective and proposing its existence requires careful argumentation.

References

R. Aly and T. Demeester. Towards a better understanding of the relationship between probabilistic models in ir. In G. Amati and F. Crestani, editors, *ICTIR '11: Proceedings of the 3nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, volume 6931, pages 164–175, 2011. doi: 10.1007/978-3-642-23318-0_16.

C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006.

of these elementary probabilities over those samples that constitute the event, $P(e \in \mathcal{E}) = \sum_{s \in e} \mu(s)$. Note that these assumptions are compatible, and we will attach probabilities to events in this paper.

[12] Feller [1968] refers to random variables as the biggest misnomer in probability theory, because they are denoted equivalent to variables in the P notation, although they are actually functions. Therefore, random functions would have been more suitable.

H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR'06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 429–436. ACM, 2006. doi: 10.1145/1148170.1148245.

W. S. Cooper. The formalism of probability theory in ir: A foundation for an encumbrance? In *SIGIR'94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 242–247, 1994. ISBN 3-540-19889-X.

R. Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14 (1):1–13, 1946. doi: 10.1119/1.1990764.

F. Crestani, M. Lalmas, C. J. V. Rijsbergen, and I. Campbell. "Is this document relevant?. . .probably": a survey of probabilistic models in information retrieval. *ACM Comput. Surv.*, 30(4):528–552, 1998.

H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 480–487. ACM, 2005. doi: 10.1145/1076034.1076116.

W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Edition*. Wiley, 3 edition, 1968. ISBN 0471257087.

N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, 2001.

S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951. ISSN 0003-4851.

J. Lafferty and C. Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, volume 13, chapter 1, pages 1–10. Kluwer Academic Publishers, 2003.

V. Lavrenko and W. B. Croft. *Language Modeling for Information Retrieval*, chapter Relevance models in information retrieval, pages 11–56. Kluwer Academic Publishers, 2003.

D. D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *ECML-98: Machine Learning*, volume 1398/1998 of *Lecture Notes in Computer Science*, pages 4–15. Springer Berlin / Heidelberg, 1998. doi: 10.1007/BFb0026666.

T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3:225–331, 2009. doi: 10.1561/1500000016.

R. W. P. Luk. On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval*, 11(6):539–561, 2008.

Y. Lv. *Improving the Effectiveness of Language Modeling Approaches to Information Retrieval: Bridging the Theory-Effectiveness Gap*. PhD thesis, University of Illinois at Urbana-Champaign, 2012. URL `http://hdl.handle.net/2142/34306`.

C. D. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition, 1999. ISBN 0-26213-360-1.

M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998. doi: 10.1145/290941.291008.

S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33: 294–304, 1977.

S. E. Robertson. On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2):319–329, 2005. ISSN 1386-4564 (Print) 1573-7659 (Online).

S. E. Robertson and K. Spärck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. doi: 10.1002/asi.4630270302.

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

T. Roelleke and J. Wang. A parallel derivation of probabilistic information retrieval models. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 107–114. ACM, 2006. doi: 10.

1145/1148170.1148192.
G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communication of the ACM*, 18(11):613–620, 1975. doi: 10.1145/361219.361220.
K. Spärck-Jones, S. E. Robertson, H. Zaragoza, and D. Hiemstra. *Language modelling for information retrieval*, chapter Language modelling and relevance, pages 57–71. Kluwer, 2003.
E. Voorhees, D. Harman, N. I. of Standards, and T. (US). *TREC: Experiment and evaluation in information retrieval*. MIT press USA, 2005.
J. Wang and J. Zhu. Portfolio theory of information retrieval. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 115–122. ACM, 2009. doi: 10.1145/1571941.1571963.
C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2(3):137–213, 2008.
C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.
C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006.