Towards a Better Understanding of the Relationship Between Probabilistic Models in IR

Robin Aly[1] and Thomas Demeester[2]

[1] University of Twente r.aly@ewi.utwente.nl
[2] Ghent University thomas.demeester@intec.ugent.be

Abstract. Probability of relevance (PR) models are generally assumed to implement the Probability Ranking Principle (PRP) of IR, and recent publications claim that PR models and language models are similar. However, a careful analysis reveals two gaps in the chain of reasoning behind this statement. First, the PRP considers the relevance of particular documents, whereas PR models consider the relevance of any query-document pair. Second, unlike PR models, language models consider draws of terms and documents. We bridge the first gap by showing how the probability measure of PR models can be used to define the probabilistic model of the PRP. Furthermore, we argue that given the differences between PR models and language models, the second gap cannot be bridged at the probabilistic model level. We instead define a new PR model based on logistic regression, which has a similar score function to the one of the query likelihood model. The performance of both models is strongly correlated, hence providing a bridge for the second gap at the functional and ranking level. Understanding language models in relation with logistic regression models opens ample new research directions which we propose as future work.

1 Introduction

The Probability Ranking Principle (PRP) of IR [10] is one of the widest acknowledged ranking principles in IR, and the fact that probability of relevance (PR) models [13] implement the PRP is commonly accepted without arguing [1]. Furthermore, to explain the empirically strong performance of language models, recent publications reason that language models are similar to PR models and therefore also implement the PRP [5, 14]. We identify two gaps in this chain of reasoning: (Gap1) The PRP considers the relevance of particular documents, which cannot be directly related to the relevance of query-document pairs considered by the PR models, and (Gap2) the relevance of query-document pairs cannot be directly related to the term and document draws considered by language models. In this paper, we investigate the above mentioned gaps and examine how they can be bridged. Figure 1 shows an overview of the content of this paper.

The PRP shows that ranking a document d by the probability of its relevance, for example, maximizes the expected precision of a ranking. On the other hand,

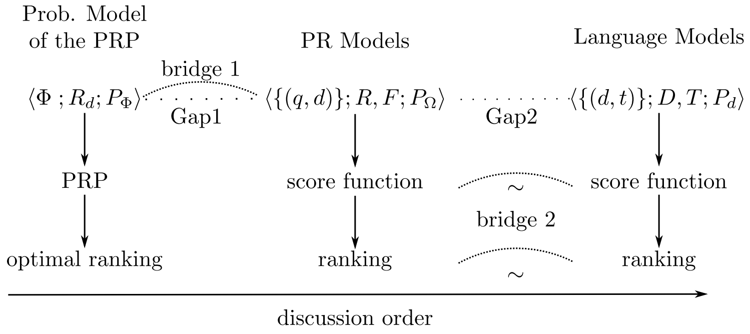


Fig. 1: Graphical overview over this paper's contents. Gap1's bridge translates models. Gap2's bridge relates score functions and rankings. The notation $\langle X;Y;Z\rangle$ denotes a probabilistic model where X are samples, Y are events, and Z is a probability measure. The detailed definition of the symbols used in this figure will be given in further sections.

PR models rank by the probability of any query-document pair (q,d) being relevant given the pair has certain features F, see Sect. 3.2. Therefore, Gap1 is the difference among the considered relevance events and among their probabilities. We argue that Gap1 has so far gone unnoticed because the probabilistic model considered by the PRP has not been defined on a mathematical basis yet. To bridge Gap1, we define the PRP's *probabilistic model*, and show how PR models can be related to this definition.

Language models consider variations of drawing terms and documents as samples. First, the *query likelihood model* [9] considers drawing query terms, second, *Hiemstra's model* [3] additionally considers drawing documents, and finally, the *risk-minimization model* [20] as well as the *relevance language model* [6] consider drawing a single term. The difference between the drawing of query-document *pairs* in PR models and the drawing of terms and documents in language models forms Gap2, whose existence is controversially discussed in literature [16, 11, 7, 18]. Similar to [11], we argue that this controversy originates from the fact that the concept of sample spaces in language models has received little attention so far. Therefore, we first define the sample spaces of the above language models on a mathematical basis. Given these definitions, we claim that PR models and the above language models are too dissimilar for Gap2 to be bridged at the probabilistic model level.

If Gap2 cannot be bridged at the probabilistic model level, it is interesting to investigate to what extent language models are related to PR model in terms of score functions and rankings. Roelleke and Wang [14] are the first to find a relation on an analytical level between the score functions of the Binary Independence Model (BIM) [12] and Hiemstra's Language model. However, this relation only holds for documents with the same term occurrences (apparent from Theorem 2 in [14]). To overcome this limitation, we define a new PR model based on logistic regression, the score function of which is similar to the score functions of the query likelihood model in terms of structure, weights, and ranking results. Although we are not able to bridge Gap2 at the probabilistic model level, we can therefore bridge Gap2 at the functional and ranking level.

This paper is structured as follows: Section 2 introduces the notation and basic definitions. Section 3 describes Gap1 and the probabilistic model we propose for the PRP to bridge it. Section 4 discusses Gap2, and why we cannot bridge it at the probabilistic model level. Section 5 defines a new PR model which ranks similarly to language model score functions and bridges Gap2 at the functional and ranking level. Finally, Section 6 concludes the paper.

2 Notation and Definitions

In this section, we introduce basic notations and central concepts from information retrieval and probability theory.

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 query terms of a query are modeled as the vector $\boldsymbol{qt} = (qt_1, ..., qt_{ql})$ where ql is the query length. Furthermore, the random variable R, relevance, is defined 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 a query on its own should not be confounded with its properties. For example, the reader may think of a query as an object in an object-oriented programming language, the symbol q as a reference, and the query terms $\boldsymbol{qt}$ as some of the object's properties. The same holds for documents.

Following Manning and Schuetze [8, chap. 2], we define the basic concepts of probability theory as follows: a *sample* is a possible outcome of a process. 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 which maps events to probabilities. We use a subscript to the probability measure P to indicate the process on which the measure is defined, for example $P_X : \mathcal{E} \rightarrow [0 : 1]$ is a probability measure defined on the event space $\mathcal{E}$ for process X. A *random variable* is a function mapping samples to the function's range. Note that a value of a random variable defines an event: the subset of samples in the sample space for which the random variable yields this value.

3 Gap1 - Between the PRP and PR models

In this section, we bridge Gap1, the difference between the PRP and PR models. First, we describe the PRP and the unified framework of PR models. After that, we show a way to relate the two probabilistic models.

3.1 The PRP

In the following we sketch the PRP and propose a definition for the underlying sample space and events, which has not yet, on a mathematical basis, been proposed in literature. Note that the proposed sample space is not necessarily the one Robertson [10] had in mind, however we consider it likely that this is indeed the case.

For a given query, the PRP considers the expected precision (together with the expected recall and reading costs) for a reader who stops reading at any rank n. This expected precision can be defined as follows:

$$E[Prec^n_{\boldsymbol{d}}] = \frac{1}{n} \sum_{j=1}^{n} P_{\varPhi}(R_{d_j}{=}1)$$

Here, $\boldsymbol{d}$ is a ranking of documents which is read until rank n. The PRP then shows that a ranking of documents

$$(d_1, ..., d_{|\mathcal{D}|}) \text{ for which } P_{\varPhi}\left(R_{d_1}\right) \geq ... \geq P_{\varPhi}\left(R_{d_{|\mathcal{D}|}}\right), \tag{2}$$

maximizes the expected precision for *any* rank n:

$$(d_1, ..., d_{|\mathcal{D}|}) = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, E[Prec^n_{\boldsymbol{d}}] \tag{3}$$

Here, $\boldsymbol{d}$ varies over all possible rankings of the documents in the collection $\mathcal{D}$ and each document d can either be labeled relevant $R_d = 1$, or irrelevant $R_d = 0$. Therefore, we propose that the PRP's sample space $\varPhi$ consists of all possible relevance labeling combinations of the documents in the collection, for the current query:

$$\varPhi = \underbrace{\{0,1\} \times ... \times \{0,1\}}_{|\mathcal{D}| \text{ times}}$$

Here, each component corresponds to a document in the collection. For a particular sample (a specific relevance labeling of all documents) $\phi \in \varPhi$, we denote the relevance label of document d as ϕ_d, and we define a (trivial) relevance random variable for each document $d \in \mathcal{D}$ as the relevance of that document within the sample, shortly, $R_d(\phi \in \varPhi) = \phi_d$. The event $R_d{=}1$ is the set of all samples ϕ with $R_d(\phi){=}1$. The sample space $\varPhi$ requires a Bayesian perspective on probabilities, because in a Frequentist's perspective a document can never be relevant or irrelevant to the same query, according to our assumptions in Sect. 2. As a result, the probability measure $P_{\varPhi}(R_d{=}1)$ expresses the degree of belief that document d is relevant. A retrieval model has to define these probabilities for each document $d \in \mathcal{D}$ in order to implement the PRP.

3.2 PR Models

Robertson et al. [13] propose a unified framework of PR models which rank by the probability that any query-document pair from the sample space $\varOmega = \mathcal{Q} \times \mathcal{D}$

is relevant[3]. The unified framework of PR models comprises four (meta-) models (Model 1−4), which consider variations to partition the sample space Ω by abstract query features and document features (or random variables)[4].

$$\boldsymbol{QF} = (QF_1, ..., QF_m) \tag{4}$$
$$\boldsymbol{QF}(q) = (QF_1(q), ..., QF_m(q)) \tag{5}$$
$$\boldsymbol{F} = (F_1, ..., F_n) \tag{6}$$
$$\boldsymbol{F}(d) = (F_1(d), ..., F_n(d)) \tag{7}$$

Here, QF_i is a query feature (a function of $q \in \mathcal{Q}$), $\boldsymbol{QF}$ is a vector of m considered query features, and $\boldsymbol{QF}(q)$ are the query features of query q. Furthermore, F_i is a document feature (a function of $d \in \mathcal{D}$), $\boldsymbol{F}$ is the vector of n document features, and $\boldsymbol{F}(d)$ are the document features of document d. For example, a query feature could be "query q contains term t", defined as $W_t : \mathcal{Q} \to \{0, 1\}$. The sets of considered features $\boldsymbol{QF}$ and $\boldsymbol{F}$ are usually selected by considering the query terms $\boldsymbol{qt}$ or terms from query expansion [2]. For later use, we introduce two trivial features: let $Q(q) = q$ be the query of a query document pair, and let $D(d) = d$ be the document of the query-document pair.

Because of space limitations, we focus our discussion to the BIM, an instance of Model 2. The BIM considers ql indexing document features, $I_i : \mathcal{D} \to \{0, 1\}$, indicating whether or not a document is indexed with query term qt_i. Documents are then ranked by the probability that any query-document pair is relevant, which we display for instructive reasons from a Frequentist's perspective, similar to [13]:

$$P_\Omega(R\,|Q(q){=}q^*, \boldsymbol{F}{=}\boldsymbol{F}(d^*)) = \frac{|\{(q,d) \in \Omega \mid R(q,d){=}1, Q(q){=}q^*, \boldsymbol{F}(d){=}\boldsymbol{F}(d^*)\}|}{|\{(q,d) \in \Omega \mid Q(q){=}q^*, \boldsymbol{F}(d){=}\boldsymbol{F}(d^*)\}|} \tag{8}$$

Here, q^* is the current query, and d^* is the current document. Now, Gap1 exists between the probabilistic model of the PRP, which considers relevance of particular documents to particular queries, and PR models which consider the relevance of any query-document pairs.

3.3 A Bridge for Gap1

In this section, we bridge Gap1 by showing how PR models can be used in the definition of the probability measure used by the PRP. Considering Model 2, if we assume that the only knowledge we have about documents are their features $\boldsymbol{F}$,

[3] Note that Robertson [11] refers to Ω as an event space. However, Ω is a set of pairs whereas an event space is a set of sets according to our definitions in Sect. 2. We assume Ω to be a sample space.

[4] In PR model literature, document features are also referred to as document representations, and descriptors, and they are often denoted by D. We denote features by $\boldsymbol{F}$ to avoid confusion with a document d.

we can decide to treat documents with the same features as indistinguishable. Under this assumption, it is reasonable to define the degree of belief $P_{\Phi}(R_d)$ that document d is relevant, as the probability that a document of any random query-document pair is relevant, given that the query is the current query and the document has the same features $\boldsymbol{F}(d)$ as the current document d:

$$P_{\Phi}(R_d) = P_{\Omega}(R|Q{=}q^*, \boldsymbol{F}{=}\boldsymbol{F}(d)) \tag{9}$$

Because of this equality of the two probability measures, PR models which rank by the probability $P_{\Omega}(R|Q{=}q^*, \boldsymbol{F}{=}\boldsymbol{F}(d))$ produce the same ranking as the PRP, see Equation 2. Therefore, Equation 9 bridges Gap1 between the PRP and PR models derived from Model 2. Note that Fuhr [2] discusses the influence of the chosen features $\boldsymbol{F}$ on the probability $P_{\Omega}(R|Q{=}q^*, \boldsymbol{F}{=}\boldsymbol{F}(d))$. However, although the choice of $\boldsymbol{F}$ influences the strength of the bridge (the more selective the features, the more realistic the assumption in Equation 9), this did not lead to the discovery of or answer to Gap1.

Furthermore, for example, Model 1 of the unified framework of PR models ranks a document d by the probability $P_{\Omega}(R|\boldsymbol{QF}{=}\boldsymbol{QF}(q^*), D{=}d)$, where $\boldsymbol{QF}$ are query features. Therefore, for each document d, Model 1 considers different queries with the same features. It is however less intuitive, why this probability would express our degree of belief $P_{\Phi}(R_d)$ that document d would be relevant to the *current* query. We postpone the investigation of this issue to future work.

4 Gap2 - Between PR Models and Language Models

In this section, we analyze Gap2, the difference between PR models and language models. First, we define the corresponding probabilistic model for four popular language models and then point out the differences to PR models described in Sect. 3.2.

4.1 Language Models

Language models have in common that they consider draws of terms, for which we define the (partial) sample space and the considered random variables:

$$\mathcal{T}_n = \overbrace{\mathcal{T} \times ... \times \mathcal{T}}^{n \text{ times}} \tag{10}$$

$$T_i(\boldsymbol{t} \in \mathcal{T}_n) = \text{the } i\text{th term in } \boldsymbol{t} \tag{11}$$

$$\boldsymbol{T}(\boldsymbol{t} \in \mathcal{T}_n) = \boldsymbol{t} \tag{12}$$

Here, $\mathcal{T}_n$ is the (partial) 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 term, and $\boldsymbol{T}$ does the same for sequences of term draws. Furthermore, in a uni-gram model, to which we limit the discussion, the random variable T_i represents the results of the ith independent trial from a multinomial distribution, and we have $P_d(\boldsymbol{T}{=}\boldsymbol{qt}) = \prod_{i=1}^{ql} P_d(T_i{=}qt_i) = \prod_{i=1}^{ql} \theta_{d,qt_i}$. Here, $P_d(T{=}t)$ is the probability of drawing the term t, and θ_{d,qt_i} is the distribution parameter of the term qt_i in language model of document d. To show that the language model parameters $\boldsymbol{\theta}_d$ are estimations, they are sometimes included in the notation of Bayesian probabilities, $P_d(T{=}t) = P_d(T{=}t|\boldsymbol{\theta}_d)$. Here, we focus on the probabilistic model used for ranking and consider the language model parameters as fixed.

For a given query, the *query likelihood model* [9] considers for each document in the collection the drawing of ql random terms[5]. Documents are ranked by the probability that the query terms are drawn, $P_d(\boldsymbol{T}{=}\boldsymbol{qt})$.

Hiemstra's model [3] considers documents, which the user has in mind for a query, and terms which the user drew from the language model of this document:

$$\begin{aligned} \mathcal{H} &= \mathcal{D} \times \mathcal{T}_{ql} \\ D'((d, \boldsymbol{t}) \in \mathcal{H}) &= \text{the document } d \text{ which the user had in mind} \end{aligned}$$

Here, $\mathcal{H}$ is the sample space of Hiemstra's model, D' is the random variable stating which document the user had in mind, and $\boldsymbol{t}$ are the drawn terms, see Equation 11. Hiemstra's model ranks a document by the probability that the user had document d in mind given the observed query terms, $P_{\mathcal{H}}(D'{=}d|\boldsymbol{T}{=}\boldsymbol{qt})$.

The *risk-minimization model* [20] considers the process of drawing a single term (sample space $\mathcal{T}_1$) from a query language model and from the language model of each document. Documents are ranked by the Kullback-Leibner divergence between the distribution of the query language model and the document's language model.

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

Here, P_q is the probability measure of the query language model. Note that it is rarely mentioned in literature that the risk-minimization framework only considers a single term draw. However, this must be the case because if the Kullback-Leibner divergence were considered for, say, ql term draws, the above summation would run over $|\mathcal{T}|^{ql}$ possible outcomes of the draws.

The *relevance language model* [6] considers for each document the process of drawing a single term from this document. The distribution is compared with a relevance model of the current query which considers the sample space of first drawing a relevant document and subsequently a term from this document. The sample space and the random variable for the drawn document of the relevance model are defined as follows:

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

[5] Following common usage, we interpret the query likelihood model as multinomial trials; it leads to the same ranking as the multi-Bernoulli interpretation considered in [9].

Here, $\mathcal{RM}$ is the sample space of the relevance model (a document-term pair), q^* is the current query, D' states the drawn relevant document. The relevance language model ranks by the negative cross entropy between drawing a term from the relevance model and from the document language model $-CE(P_r||P_d)$ of drawing terms. Here, the probability of drawing a term from the relevance model is determined by marginalization over D''.

4.2 Differences between PR Models and Language Models

Based on the definitions of PR models in Sect. 3.2 and the presented language models in the previous Sect. 4.1, we investigate whether we can bridge Gap2 at the probabilistic model level. To compare PR models and the presented language models, they are usually presented as different derivations of the probability $P(R|Q,D)$ [5, 7, 15]. However, the definition of each of these symbols differs among PR models and language models. In PR models, Q are query features, denoted in this paper as $\boldsymbol{QF}$, which are functions of the considered query. Therefore, given a query, its feature values are not random. On the other hand in the presented language models, the random variable Q, which is in our notation $\boldsymbol{T}_{ql}$, represents the outcome of randomly drawing ql terms and this does not depend on a query.

Furthermore, in PR models, D stands for document features, denoted in this paper as $\boldsymbol{F}$, which are functions of the considered document. Therefore, given a particular document the feature value is not random. On the other hand, in Hiemstra's model, D stands for the document the user had in mind and which is modeled as the outcome of a random process.

Also, the notion of relevance differs between its use in PR models, where it is a function of query-document pairs, and its use in the four presented language models. First, the query likelihood model and the risk-minimization model do not use the notion of relevance. Second, Hiemstra's model assumes only a single relevant document [16]. Finally, the notion of relevance in the relevance language model can be seen to be the same as in PR models. However, in the relevance language model, single, particular documents are drawn from the relevance model while PR models consider drawing any relevant query-document pair with certain features $P_{\Omega}(\boldsymbol{F}|R)$.

Therefore, we propose that the reasoning for the similarity between PR models and the presented language models is mainly guided by similar notation, and that PR models and the presented language models are too different to bridge Gap2 at the probabilistic model level.

5 Bridging Gap2 at the Functional and Ranking Level

In this section, we propose a new PR model which ranks similarly as the score function of the query likelihood model. Instead of considering probabilities of drawing a term, $P_d(T=t)$, we consider language scores as document features (functions of a document), a particular feature F in Equation 6:

$$LS_i(d) = \log\left(\frac{\theta_{d,qt_i}}{\alpha_d\ \theta_{\mathcal{D},qt_i}}\right) \tag{13}$$

Here, $LS_i(d)$ is the language score of document d for query term qt_i, θ_{d,qt_i} is the language model parameter for query term qt_i in document d, see Sect. 4.1, α_d ensures that $LS_i(d)$ is zero if query term qt_i is not in the document [19], and $\theta_{\mathcal{D},qt_i}$ is the constant collection prior. We denote the vector of language score feature functions for the current query as $\boldsymbol{LS} = (LS_1, ..., LS_{ql})$ and, evaluated for a document d, as $\boldsymbol{LS}(d) = (LS_1(d), ..., LS_{ql}(d))$. Based on these features, we define a PR model in which the probability of any query-document pair being relevant is represented by a discriminative logistic regression model [4]:

$$P_\Omega(R\,|Q{=}q^*, \boldsymbol{LS}{=}\boldsymbol{LS}(d^*)) = \frac{1}{1 + exp(-w_0 - \sum_{i=1}^{ql} w_i\ LS_i(d^*))} \propto \sum_{i=1}^{ql} w_i\ LS_i(d^*) \tag{14}$$

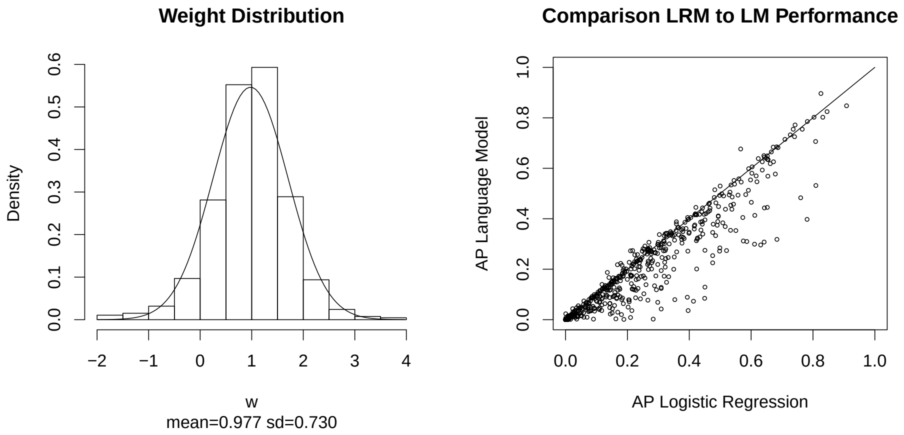
Here, q^* is the current query, $\boldsymbol{LS}(d^*)$ are the language scores of the current document d^*, w_0 is the intercept of the logistic regression model representing the relevance prior, and w_i is the language score weight of query term qt_i. From the calculated probability $P_\Omega(R\,|Q{=}q^*, \boldsymbol{LS}{=}\boldsymbol{LS}(d^*))$ we see that the PR model implements Model 2 of the unified framework of PR models, for which we have shown that it bridges Gap1. The middle term of Equation 14 is the definition of the logistic regression model, and the rightmost term is a rank equivalent score function see [17] for a derivation.

Now, we compare the logistic regression model in Equation 14 with the score function of the query likelihood model, which is defined as follows [19]:

$$P_d(\boldsymbol{T}{=}\boldsymbol{qt}) \propto \sum_{i=1}^{ql} \log\left(\frac{P_d(T{=}qt_i)}{\alpha_d\ \theta_{\mathcal{D},qt_i}}\right) + |\mathcal{T}|\ \alpha_d + const$$

Here, α_d has the same function as in Equation 13. From expanding the rightmost term of Equation 14 by the definition of language scores in Equation 13 and using the relationship $P_d(T{=}qt_i) = \theta_{d,qt_i}$, we see that the logistic regression model score function has a similar structure to the score function of the query likelihood model, except for the missing expression $|\mathcal{T}|\ \alpha_d$ and the non-uniform language score weights w_i.

In order to quantify their similarity in practice, we compare the performance of the score functions of the logistic regression models and the query likelihood model using 550 queries from the ROBUST 04+05, TREC 09, TERABYTE 04-06 data sets. If we assume uniform language score weights w_i in the logistic regression model, the model practically performs identically to the query likelihood model in terms of mean average precision (MAP). Therefore, the term $|\mathcal{T}|\ \alpha_d$ has no significant influence on the ranking. Furthermore, we consider the hypothetical case of using the language score weights w_i which we trained on all



(a) (b)

Fig. 2: (a) Weight distribution over the query terms of 550 queries of the proposed logistic regression model (LRM) trained on all relevance data. (b) Performance comparison to the query likelihood model (LM).

relevance data for each query separately. Figure 2 (a) shows that the trained language score weights w_i are Gaussian distributed with an expected weight around one. Therefore, for a random query term we can *expect* the logistic regression weight w_i of the corresponding language score to be one. This expected weight also coincides with the uniform weight of the query likelihood model. Figure 2 (b) compares the performance of the hypothetical logistic regression model against the performance of the query likelihood model in terms of average precision. The models have a high performance correlation (Pearson correlation coefficient 0.92). We suggest that the uniform weights of the query likelihood score function can also be seen as a first approximation of the ideal language score weights w_i from Equation 8. As a result, the newly proposed logistic regression PR model bridges the Gap2 to the query likelihood model at a functional and ranking level.

Additionally, the score functions of the risk-minimization framework and the relevance model can be seen as methods to improve upon the uniform weights of the query likelihood model for an expanded set of query terms [18]. Hence, these weights potentially could also be approximations to the weights w_i of newly selected features. We postpone these investigations to future work.

Note that the similarity of the *score functions* of the described logistic regression model and the query likelihood model does not imply that ranking by the query likelihood *model* could not be justified otherwise.

6 Conclusions

In this paper, we bridged two gaps in the chain of reasoning used for two popular probabilistic IR models, PR models and language models. (Gap1) The PRP considers the relevance of particular documents which cannot directly be related to the relevance of query-document pairs considered by the PR models, and (Gap2) the relevance of query-document pairs cannot directly be related to the term draws considered by language models.

In order to bridge Gap1, we defined a probabilistic model underlying the PRP, which considers all possible combinations of relevance labels of the documents in the collection. Probabilistic models which implement the PRP need to define the degree of belief that document d is relevant $P_{\Phi}(R_d)$. Furthermore, the (meta) Model 2 of the unified framework of PR models [13] considers the probability of relevance of any query-document pair with the query being the current query q^* and the document having the same features $\boldsymbol{F}(d)$ as the current document d^*, $P_{\Omega}(R|Q{=}q, \boldsymbol{F}{=}\boldsymbol{F}(d^*)$. We argued that, under the assumption that we can only distinguish documents by the features $\boldsymbol{F}$, we can take $P_{\Omega}(R|Q{=}q, \boldsymbol{F}{=}\boldsymbol{F}(d^*))$ as the degree of belief of relevance $P_{\Phi}(R_d)$. With this assumption, Gap1 was bridged. Similar assumptions of the other models of the unified framework require further investigations, which we will discuss in future work.

Furthermore, from the definition of the probabilistic model of PR models and language models, we found that the two models are different and we observed that previous comparisons were mainly based on similar notation with different meaning. Therefore, Gap2 could not be bridged at the probabilistic model level. Additionally, we proposed a new PR model derived from Model 2 of the unified framework of PR models, based on logistic regression. For 550 queries in six collections, we showed that the score functions of the logistic regression model and the query likelihood model were similar, and the performance of the two score functions was strongly correlated. Comparing the weights of both score functions showed that the uniform weights of the query likelihood model score function can be seen as the expected logistic regression weights for a random query. Therefore, we bridged Gap2 at the functional and ranking level, leading to an alternative explanation for the strong performance of language models.

Understanding and further exploring the apparent connection between language models and logistic regression models (or possibly other discriminative models) opens ample new research directions which we propose for future work. The proposed logistic regression model could for instance be used for score normalization, and existing research on feature selection for logistic regression models could be used for query expansion.

References

[1] 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. ISSN 0360-0300.

[2] N. Fuhr. Probabilistic models in information retrieval. *Comput. J.*, 35(3): 243–255, 1992.

[3] D. Hiemstra. *Using Language Models for Information Retrieval.* PhD thesis, University of Twente, Enschede, January 2001.

[4] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*. Wiley-Interscience Publication, September 2000. ISBN 0471356328.

[5] J. Lafferty and C. Zhai. *Probabilistic Relevance Models Based on Document and Query Generation*, chapter 1, pages 1–10. Kluwer Academic Pub., 2003.

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

[7] R. W. P. Luk. On event space and rank equivalence between probabilistic retrieval models. *Information Retrieval*, 11(6):539–561, December 2008. ISSN 1386-4564 (Print) 1573-7659 (Online). doi: 10.1007/s10791-008-9062-z.

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

[9] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98*, pages 275–281, New York, NY, USA, 1998. ACM. ISBN 1-58113-015-5. doi: 10.1145/290941.291008.

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

[11] S. E. Robertson. On event spaces and probabilistic models in information retrieval. *Information Retrieval*, 8(2):319–329, April 2005. ISSN 1386-4564 (Print) 1573-7659 (Online). doi: 10.1007/s10791-005-5665-9.

[12] 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.

[13] 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, January 1982.

[14] T. Roelleke and J. Wang. A parallel derivation of probabilistic information retrieval models. In *SIGIR '06*, pages 107–114, New York, NY, USA, 2006. ACM. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148192.

[15] T. Roelleke and J. Wang. Tf-idf uncovered: a study of theories and probabilities. In *SIGIR '08*, pages 435–442, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390409.

[16] 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, Dordrecht, 2003.

[17] R. Yan. *Probabilistic Models for Combining Diverse Knowledge Sources in Multimedia Retrieval*. PhD thesis, Canegie Mellon University, 2006.

[18] C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2(3):137–213, 2008. ISSN 1554-0669. doi: 10.1561/1500000008.

[19] 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. ISSN 1046-8188. doi: 10.1145/984321.984322.

[20] C. Zhai and J. Lafferty. A risk minimization framework for information retrieval. *Inf. Process. Manage.*, 42(1):31–55, 2006. ISSN 0306-4573. doi: 10.1016/j.ipm.2004.11.003.