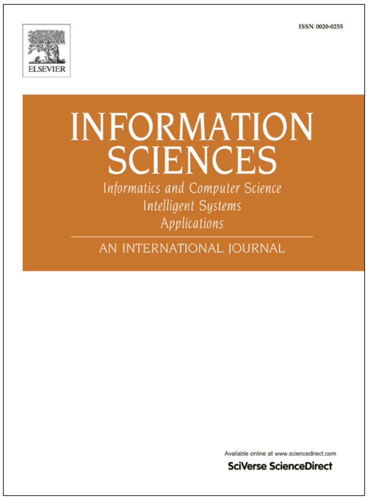

Score Distributions for Pseudo Relevance Feedback

Javier Parapar, Manuel A. Presedo-Quindimil, Álvaro Barreiro

PII: S0020-0255(14)00318-1
DOI: http://dx.doi.org/10.1016/j.ins.2014.03.034
Reference: INS 10712

To appear in: *Information Sciences*

Received Date: 24 July 2013
Revised Date: 18 February 2014
Accepted Date: 10 March 2014

Please cite this article as: J. Parapar, M.A. Presedo-Quindimil, Á. Barreiro, Score Distributions for Pseudo Relevance Feedback, *Information Sciences* (2014), doi: http://dx.doi.org/10.1016/j.ins.2014.03.034

Score Distributions for Pseudo Relevance Feedback

Javier Parapar*, Manuel A. Presedo-Quindimil, Álvaro Barreiro

Information Retrieval Lab, Department of Computer Science, University of A Coruña,
Campus de Elviña, 15071 A Coruña, Spain

Abstract

Relevance-Based Language Models, commonly known as Relevance Models, are successful approaches to explicitly introduce the concept of relevance in the statistical Language Modelling framework of Information Retrieval. These models achieve state-of-the-art retrieval performance in the Pseudo Relevance Feedback task. It is known that one of the factors that more affect to the Pseudo Relevance Feedback robustness is the selection for some queries of harmful expansion terms. In order to minimise this effect in these methods a crucial point is to reduce the number of non-relevant documents in the pseudo relevant set. In this paper, we propose an original approach to tackle this problem. We try to automatically determine for each query how many documents we should select as pseudo-relevant set. For achieving this objective we will study the score distributions of the initial retrieval and trying to discern in base of their distribution between relevant and non-relevant documents. Evaluation of our proposal showed important improvements in terms of robustness.

Keywords: Information Retrieval, Pseudo Relevance Feedback, Score Distributions, Pseudo Relevance Feedback Set, Relevance Models

1. Introduction and Motivation

In the history of the Information Retrieval research, efforts to improve retrieval effectiveness have been centred in both developing better retrieval

[*] Corresponding author. Tel.: +34 981 167 000 x 1276; Fax: +34 981 167 160.

Email addresses: javierparapar@udc.es (Javier Parapar), mpresedo@udc.es (Manuel A. Presedo-Quindimil), barreiro@udc.es (Álvaro Barreiro) models by including new features or using different theoretical frameworks; and in designing new techniques to be incorporated on top of existing models to improve their performance. Particularly on the later, Query Expansion (QE) has proven to be effective from very early research stages. QE approaches can be classified between global techniques which produce a query rewriting without considering the original rank produced by the query, and local techniques in which the expanded query is generated using the information of the initial retrieval list.

In [33] Salton presented the initial efforts on exploiting the local information to improve the query formulation introducing, among others, Rocchio approach [29] working on the Vector Space Model framework. This family of local techniques is called Relevance Feedback (RF) [30] and it is based on using the relevant documents in the initial retrieval set in order to reformulate the query based on their content. Nevertheless, in a real retrieval scenario it is not realistic to assume that relevance judgements are available. Because of this, Pseudo Relevance Feedback (PRF) algorithms have been investigated [9, 39]. PRF methods are based on assuming relevance of a set of documents retrieved by the original query. The set of documents which are assumed to be relevant and the way in which their information is exploited to improve the original query varies from one PRF method to another.

One crucial aspect of the pseudo-relevance feedback methods is robustness. In this context, robustness is defined as the quality of not hurting the effectiveness values achieved by the retrieval model in the initial rank for every query. Most of existing pseudo-relevance feedback methods outperform the effectiveness of the initial retrieval in average but they tend to harm some of the queries. This is an important point for solving in order to popularise the use of these methods in the commercial search engines. The most common phenomenon causing the decrease of effectiveness for a query is the *topic drift*. Topic drift refers to the situation where the expansion of the query produced that the topic of the original user need has moved (drifted) away to a different one. For instance, for the TREC topic 101: *Design of the "Star Wars" Anti-missile Defense System*, a very clear example of topic drift would be the returning of documents about the film. The topic drift can be naturally produced by the addition of terms, but this problem can be greatly intensified when the pseudo-relevant set (RS) has plenty of irrelevant documents.

This problem has been exposed very early in the literature [24] and caused lots of works on areas such as query performance prediction [11, 8] which investigates how to predict the performance of a query anticipating those queries that will be negatively affected by the expansion, selective pseudo-relevance feedback[32, 2] which tries to decide for which queries PRF should or not be applied, and adaptive pseudo-relevance feedback [21] that is centred on adjust the weight of the expansion terms over the original query automatically depending on the nature of the given query.

The different approaches to decide when of how much apply PRF have considered pre-query processing indicators and initial ranking examination.

Several evidences have been considered such as the number of query terms in the pseudo-relevant documents, the similarity between query and the relevant set, term proximity measures, etc. But it was only recently when some works started to consider the scores of the initial retrieval [34]. Shtok et al. argue that query-drift can potentially be estimated by measuring the diversity (e.g., standard deviation) of the retrieval scores of the documents in the ranking.

In this paper we also exploit the scoring information but in a different way, we use the scores of the initial retrieval for determining the pseudo-relevant set itself, trying to minimise the amount of non-relevant documents in it. For achieving this objective we used a framework for modelling the score distributions of a retrieval model [23] and adapt the threshold optimization solution for recall-oriented retrieval [4] for our particular problem, where we want to stop selecting documents from the top of the initial retrieval when non-relevant documents appear. Score distributions research investigates the idea of using the documents' scores for separating relevant and non-relevant documents. For doing this, different statistical modelling choices over both groups of documents are taken and the parameters of the statistical distributions are inferred from the observed scores. Although it has been already used for other task such as meta-search and high recall oriented task such as legal retrieval, this is a novel and especially adequate use of the score distributions analysis. We are really pursuing a high precision for our task in such a way that ideally if no relevant documents are present on the top of the initial retrieval we want to return an empty RS producing a way of selective PRF. Furthermore, and not less important, our approach reduces the number of parameters to tune in the training phase of PRF methods by suppressing the necessity of tune r, the number of documents on the RS.

For assessing our proposal we will use one of the most successful PRF methods in the state-of-the art: Relevance-Based Language Models (RM) [18]. In particular, we will use the best performing estimation for RM the so called RM3 estimation [1]. Although, when averaged over a query set the differences in performance in terms of Average Precision when selecting different top sizes for RS in a particular collection may not differ too much for RM3, it varies a lot at query level (see Figure 1). Meanwhile some queries present a stable behaviour (as query 81), most of them have either an increasing behaviour (as queries 60 and 62) or a decreasing behaviour (as queries 54 and 63). Thus, it is clear that it is import to be able to automatically adjust the RS at query level, which motivates the work in this paper.

We performed evaluation to assess how our proposal affects to the effectiveness and robustness of RM on standard settings. Results showed that both characteristics are improved with the extra advantage of the reduction of the number of parameters involved in the training phase. The rest of the paper is as follow: next Section (2) starts with some specific background on score distributions, in Section 3 we present our proposal for modelling the score distributions and automatically limit the RS size, Section 4 shows the evaluation results, related work is briefly reviewed in Section 5 and finally

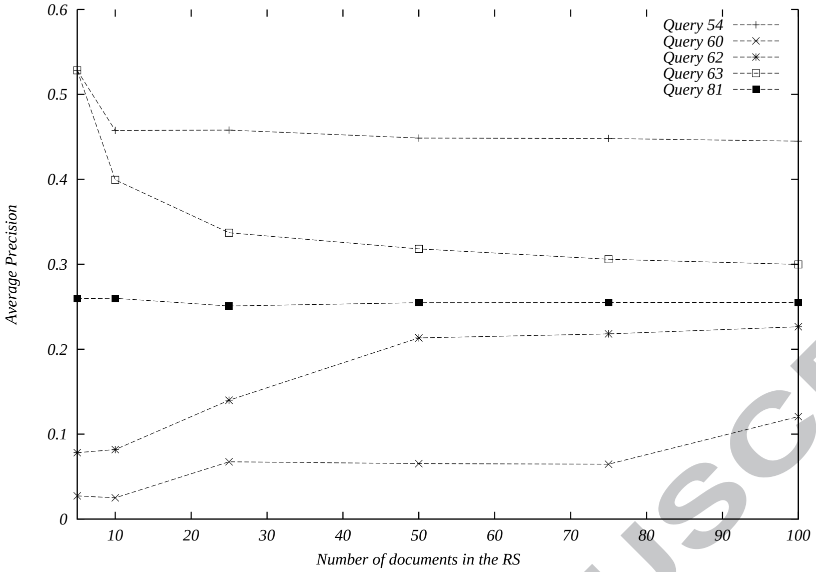

Figure 1: RM3 behaviour in terms of Average Precision for different queries from the training query set of the AP88-89 collection with $t = 100$ and $\lambda = 0.8$ and $\mu = 1000$

we conclude with our main findings in Section 6.

2. Background

2.1. Relevance-Based Language Models

The RM for PRF was presented within the Language Modelling (LM) theoretical framework. In Language Modelling the probability of a document given a query, $P(d|q)$, is estimated using the Bayes' rule as presented in Eq. 1.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \overset{rank}{=} \log P(q|d) + \log P(d) \tag{1}$$

In practice $P(q)$ is dropped for document ranking purposes. The prior $P(d)$ encodes a-priori information on documents and the query likelihood,

$P(q|d)$, incorporates some form of smoothing. In this paper we consider uniform priors and uni-gram language models with Dirichlet smoothing [40].

After obtaining the initial ranking using the original query, the PRF methods assume relevance over a subset of retrieved documents. This set is usually called *relevance set*. The information of those documents is then used to improve the initial retrieval. The most common way of achieving this objective is expanding the original query and producing a second retrieval with the reformulated query. Next, different models to produce expanded queries are analysed.

The RM approach builds better query models using the information given by the pseudo relevant documents. Two estimations were originally presented in [18]. RM1 assumes that the words in the relevant documents and the query words are sampled identically and independently from the relevance model. The result is an estimation where the query likelihood for every document is used as the weight for the document and the probability of a word is averaged over every document language model. In contrast, RM2 assumes that the query words are independent of each other, but they are dependent of the words of the relevant documents (conditional sampling). The result is that relevant documents containing query words can be used for computing the association of the their words with the query terms. A quite detailed explanation of the RM for PRF is given in the Chapter 7 of the book by Croft et al. [10].

In RM the original query is considered a very short sample of words obtained from the relevance model (R). If more words from R are desired then it is reasonable to choose those words with highest estimated probability when considering the words for the distribution already seen. So the terms in the lexicon of the collection are sorted according to that estimated probability, which after doing the assumptions using the RM1 method, is estimated as in Eq. 2.

$$P(w|R) \propto \sum_{d \in C} P(d) \cdot P(w|d) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{2}$$

Usually $P(d)$ is assumed to be uniform. $\prod_{i=1}^{n} P(q_i|d)$ is the query likelihood given the document model, which is traditionally computed using Dirichlet smoothing. Then for assigning a probability to the terms in the relevance model we have to estimate $P(w|d)$; in order to do so it is also common to use Dirichlet smoothing. The final retrieval is obtained by four steps:

1. Initially the documents in the collection C are ranked using their query likelihood using Dirichlet smoothing.
2. A certain top r documents from the initial retrieval are taken for the estimation instead of the whole collection C, let us call this pseudo relevance set RS.
3. The relevance model probabilities $P(w|R)$ are calculated using the estimate presented in Eq. 2, with RS instead of C.
4. To build the expanded query the e terms with highest estimated $P(w|R)$ are selected. The expanded query is used to produce a second document

ranking using negative cross entropy as in Eq. 3.

$$\sum_{i=1}^{e} P(w_i|R) \cdot \log P(w_i|d) \tag{3}$$

RM3 is a later extension of RM that performs better than RM1 in terms of effectiveness. RM3 interpolates the terms selected by RM1 with the original query as in Eq. 4 instead of using them directly. The final query is used in the same way as in RM1 to produce a second ranking using negative cross entropy.

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot P(w|R) \tag{4}$$

2.2. Score Distributions

The Probability Ranking Principle (PRP, [28]) states that the ranking of the documents should be according to their probability of relevance. However, retrieval models, in the ideal case where the document ranking strictly honours the PRP, do not provide with a method for delimiting when the non-relevant documents start to appear. In this context, score distributions have been studied and modelled since the early days of IR. Initial works date from the sixties [35], when the idea of using the scores for separating relevant and non-relevant documents was originally formulated. However, it was only recently when the benefit of these approaches was demonstrated for the retrieval task [23]. Score distribution modelling techniques try to infer statistical properties from the seen data (the scores of the ranking documents) and take advantages of such inferred properties, and not directly from the observed data, for classifying documents between relevant and not relevant.

Score distribution models generally assume that the scores of the relevant documents were generated by a different distribution from the distribution of the non-relevant documents. The research efforts have been centred on two aspects: which family of statistical distributions corresponds with each group of documents and how the parameters of the distributions can be learned or estimated from the observed documents' scores. Different combinations of statistical distributions were proposed for modelling the score distributions. Swets [35] originally proposed to model the relevant and non-relevant groups as two Gaussian distribution with different parameter values (see Fig. 2 as an example), although later on, Swets considered two negative exponential distributions [36]. Bookstein [7] tested with two Poisson and Baumgarten

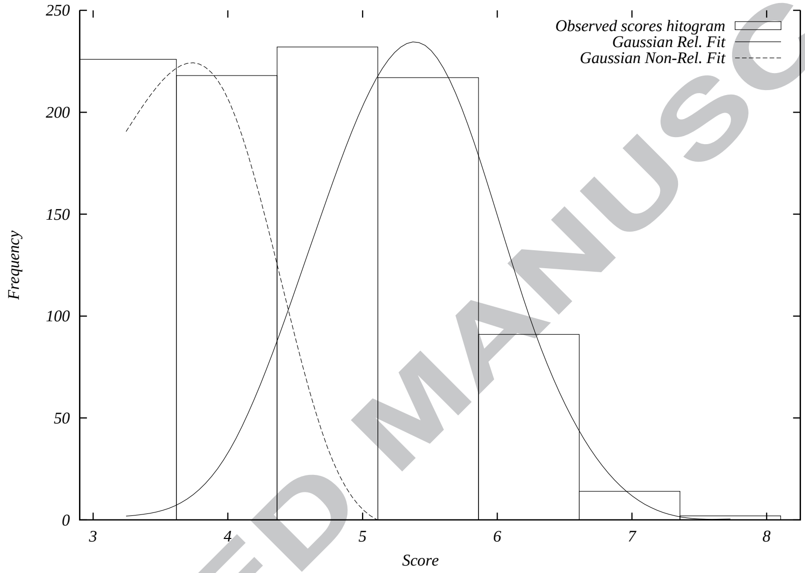

Figure 2: Mixture of Gaussians fit to relevant and non-relevant data obtained processing the scores of TREC query 154 over the AP88-89 collection produced with the LM retrieval function with Dirichlet smoothing ($\mu = 1000$)

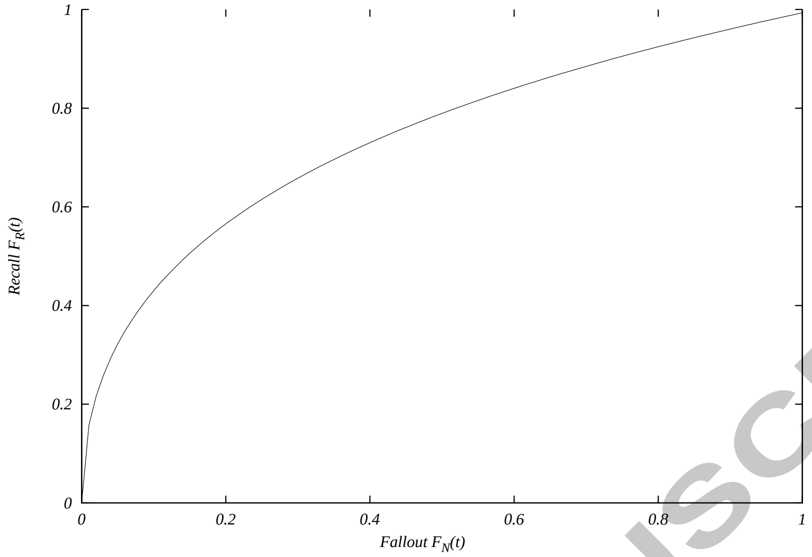

Figure 3: Example of idealised Receiver Operating Characteristic (ROC) for a cut-off or threshold t

proposed [6] a two Gamma choice. It was only lately when the mixture model of a Gaussian distribution for the relevant and a negative exponential distribution for the non-relevant documents was proposed [3]. Also recently, when Kanoulas et al. [17] proposed a mixture of Gaussian distributions for relevant documents and a Gamma for non-relevant documents.

In this context, Robertson [27] presented the convexity hypothesis which stated that for all good systems, the recall-fallout curve (when viewed from the top left (0,1), see Fig. 3) is convex. In this case, recall should be interpreted as the proportion of the relevant distribution exceeding a given threshold t and fallout the proportion of the non-relevant distribution exceeding that point. In the graph, the point (0,0) corresponds with a very high threshold that is (nothing retrieved), while the point (1,1) corresponds with a very low threshold (everything retrieved). So, if this graph presents concave parts it means that the proportion of the relevant distribution over the non-relevant decreases when the scores increase for some segment of values. This is related, but somewhat stronger than, the inverse recall-precision relationship and it means that the higher the score of a document the higher the probability of relevance. Over the graph, a random ordering of the collection of documents (identical relevant and non-relevant score distributions) would produce a straight line from (0,0) to (1,1). Any other straight segment may also be interpreted of random ordering of sub-sets of the documents. We can easily improve the performance eliminating the concavity segments of the curves by simply randomising the sub-list of scores corresponding with those segments and thus, replacing the concavity parts by straight segments. Indeed, we can just reversing the scores in the sub-list and converting the concavity segments in their convex mirror reflections. In this way, if we depart from a convex curve, we can easily improve the initial performance of our model, so convexity seems to be a desirable property.

In this work, Robertson probes that although the most of the previously presented distributions choice honour the convexity principle, some of them, do not. In particular the Gaussian-negative exponential mixture model [23, 4], one of the most popular choices, does not accomplish this property. In particular, this model presents concavity problems both in the top right end (low threshold values) and the bottom left end (high threshold values) for any parameters' values.

The model presented in [4], besides not honouring the recall-fallout curve convexity (about the 60% of the queries in the experiments suffer from this anomaly), presents good practical results for a high recall retrieval task such as legal retrieval. One of the most popular effectiveness measures on legal retrieval is the $F_1@K$ where K is the cut-off selected by the system to stop providing with results. The objective pursued with score distributions is to automatically determine the value of K for each query. So for achieving that objective Arampatzis et al. presented a threshold optimisation method over the learned distributions which we adapted for our problem in the next section.

Most of the existing works on score distributions use relevance information and so the learning of the different distributions' parameters is an easy task (the groups of relevant and non-relevant documents are already defined). When there are no relevance judgements, the learning of the distributions' properties from the observed scores also includes the learning of the weights of the mixture. The Expectation Maximisation (EM) algorithm [13] has been the standard approach to finding the mixing and the distribution's properties in this area. Recently, extended versions of this method have been developed for this specific task [12]. EM is an iterative algorithm which is used for finding maximum likelihood estimates of the parameters in probabilistic models, when dependency exists on unobserved hidden variables

3. Modelling Score Distributions for Pseudo-Relevance Feedback

Our objective is the use of score distributions models to automatically determine the size of the pseudo-relevant set, i.e., we want to select for each query the optimal top of documents which will feed the PRF process. Ideally these top documents will be only relevant ones. We formulate this problem as a threshold optimisation task. In order to adapt the score distribution models to work under this paradigm we have to (i) select an appropriate distribution modelling choice, (ii) select a learning strategy for inferring the distributions' parameters and (iii) formulate the corresponding cut-off conditions.

Referring to the first decision, the straightforward choice should be to use the popular Negative exponential-Gaussian mixture [3] or its truncated version [4]. However, as stated before, this model clearly violates the convexity hypothesis [27]. Moreover, our experiments using these models showed results consistently worse than with our final choice. The model which resulted to perform better than those alternatives was the Gaussian-Gaussian mixture [35] which honours the convexity hypothesis for fixed variances and for almost every situation of different variances (it only presents anomalies in the ends of the intervals). Particularly, apart from the honouring of the convexity hypothesis, we chose to use the later because it presented more robust results across collections (this fact can be observed in Section 4) moreover, Madigan et al [22] produced an analytical study observing the expected precision and contamination (the number of non-relevant documents in a given top of documents) values depending on the election of different distributions, showing again the better behaviour of the Guassian-Gaussian mixture over other alternatives as the aforementioned Gaussian-Exponential.

Regarding to the second point, EM is an efficient and popularly used method to estimate model parameters from a set of observed values by maximising the likelihood. In this case, we decided to use a generalisation of the EM algorithm known a Bregman soft clustering [5]. Bregman soft clustering allows estimating the parameters of a mixture of exponential families [14], given a set of observations. This Bregman soft clustering algorithm shares with the EM the initialisation, expectation and maximisation steps. The main advantage of using this method instead of the EM algorithm is that it allows to estimate the parameters of *any* mixture of exponential family distributions. The Statistical Exponential Family [25] is a set of probability distributions admitting the following canonical decomposition:

$$P(x, \Theta) = \exp(\langle t(x), \Theta \rangle - F(\Theta) + k(x)) \tag{5}$$

where

- $t(x)$ is the sufficient statistic, a function of the data that fully summarizes the data.
- Θ are the natural parameters,
- $\langle ., . \rangle$ is the inner product,
- $F(.)$ is called the log-normalizer because it is the logarithm of a nor-

malization factor,

- $k(x)$ the carrier measure.

In particular, this family includes the following well-known distributions: Gaussian, Poisson, Bernoulli, binomial, multinomial, Laplacian, Gamma, Beta, negative exponential, Wishart, Dirichlet, Rayleigh, probability simplex, negative binomial, Weibull, von Mises, Pareto distributions, skew logistic, etc. In our case, we use a mixture of Gaussian distributions, in this case the mapping for the canonical decomposition is:

- $t(x) = (x, x^2)$
- $\Theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$
- $F(\Theta) = -\frac{\Theta_1^2}{4\Theta_2} + \frac{1}{2}\log(-\frac{\pi}{\Theta_2})$
- $k(x) = 0.$

where μ is, in this case, the mean of the Gaussian distribution and σ its standard deviation. More details of the canonical decomposition can be found in [25].

For estimating the parameters of a mixture of exponential families with Bregman soft clustering over the observed scores a general expectation-maximisation procedure is used. As result of this process, the natural parameters of the distributions involved in the mixture are obtained as well as the weights of the distributions in the mixture. In our case, those natural parameters correspond with the means and variances of the Gaussian distribution. Details of the initialisation, expectation and maximisation steps of the process are reported in section 1.5.4 of [25]. In the initialisation step, the scores are grouped in so many clusters as distributions in the mixture with the K-Means algorithm estimating the weight for each component as the proportion of scores in each cluster. The initial values for the parameters of each distribution are estimated in the corresponding clusters. In the expectation step the probabilities of the observed scores of belonging to each distribution are recomputed. Finally, the maximisation step recomputes the values the parameters of the probability distributions given the new belonging probabilities of the observed scores.

The only remaining aspect to be defined is the cut-off strategy. Given the following definitions, Arampatzis et al. [3] state the following threshold optimisation problem.

$$\begin{aligned} &R = nG_n \\ &R_+(s) = R(1 - F(s|1)) \\ &N_+(s) = (n - R)(1 - F(s|0)) \\ &R_-(s) = R - R_+(s) \\ &N_-(s) = (n - R) - N_+(s) \end{aligned} \tag{6}$$

where R is the number of relevant documents for the query, $R_+(s)$ and $R_-(s)$ the number of relevant documents over and below the given score respectively, $N_+(s)$ and $N_-(s)$ the number of non-relevant documents over and below the given score respectively, G_n is the fraction of relevant documents in the collection, n is the number of documents in the collection and $F(s|1)$ and $F(s|0)$ are values of the cumulative distribution functions at the score s for the relevant and non-relevant distributions respectively.

Then, the optimal score where to perform the cut-off (s_{opt}) is that one such maximise a given effectiveness measure M of the form of a linear combination of the document count of the categories defined in Eq. 6:

$$s_{opt} = \arg\max_s \{M(R_+(s), N_+(s), R_-(s), N_-(s))\} \tag{7}$$

In our case, we ideally want to obtain a RS for RM where every document is relevant. This is a quite strict condition and for many queries the apparition of a non-relevant document as the highest scored document would produce an empty RS, discarding a lot of useful information. For this reason we decided to relax this constraint and formulate the effectiveness measure for cut-off problem as:

$$M(R_+(s), N_+(s), R_-(s), N_-(s)) = \frac{R_+(s)}{N_+(s)} \tag{8}$$

That is, we will cut the top for building the RS in the point of maximum relevance density.

This is our approach to automatically estimate the size of the pseudo-relevance feedback set for RM. Only some final estimation details remain to be explained. As commented before, we chose to model the relevant and non-relevant distributions as a mixture of two Gaussian distributions. From the two Gaussian distributions, learned with the Bregman soft clustering method, the one corresponding with the relevant documents will be assumed that one with highest mean. The G_n and n parameters will be replaced by their estimated values, corresponding with the fraction of relevant documents in the top and the size of the top, respectively. The fraction of relevant documents in the top will be estimated as the weight of the Gaussian distribution corresponding with the relevant documents in the mixture.

4. Experiments and Results

The evaluation of our approach was performed over four TREC collections, using Terrier [26] and comparing with a baseline retrieval model and the baseline feedback model (training the size of the pseudo-relevant set) In this section we describe the evaluation methodology, including collection and metric election, and we carefully analyse the results comparing the behaviour of our proposals with respect to the baselines.

4.1. Collections

To evaluate the different approaches we chose the same collections used in previous works about RM estimations [20]: a subset of the Associated Press collection corresponding to the 1988 and 1989 years (AP88-89) [37], the Small Web Collection WT2G and the disk 4 and 5 from TREC (TREC-678). Additionally, given the fact that TREC Conference only provided with a set

Table 1: Collections and topics (short queries: title only, average length in words) for training and test used in the document retrieval evaluation

Col.	*# of Docs*	*avg. words per doc*	*Topics (avg. length)*	
			Training	Test
AP88-89	164,597	284.7	51-100 (3.8)	151-200 (6.5)
WT2G	247,491	645.3	401-450 (2.4)	–
TREC-678	528,155	297.1	301-350 (2.7)	351-400 (2.5)
WT10G	1,692,096	399.3	451-500 (3.46)	501-550 (4.62)
GOV2	25,205,179	647.9	701-750 (3.14)	751-800 (3.08)

of topics for the WT2G collection, we decide to use the WT10G collection, which was not used in [20], to report test values in a web collection and GOV2 dataset (a crawl of the `.gov` domain from 2004) for examining the results in huge collection. In AP88-89, TREC-678, WT10G, and GOV2 we used training and test evaluation: we performed training for Mean Average Precision in a set of topics and testing over another set. For WT2G we report well-tuned values over the trained topics, as it was done in [20]. Short queries (title only) were used because they are the most suitable to be expanded. All the collections were preprocessed with standard stop-word removal and Porter stemmer, as it has been demonstrated the best performing scenario for this task [20]. In Table 1 the evaluation settings are summarized. stop-word removal and Porter stemmer.

4.2. Compared Methods

We compared four methods:

- **LM**: the baseline Language Modelling retrieval model with Dirichlet smoothing. This approach was also used by the other methods for producing the initial retrieval.

- **RM3**: the standard formulation of RM3, as explained in Section 2.1 training the size of the RS.
- **SDRM3-GE**: the standard formulation of RM3 but automatically determining for each query the size of the RS in this case using the previously proposed model of Gaussian-Exponential mixture [4], we report these values to highlight the importance of the model selection.
- **SDRM3**: the standard formulation of RM3 but automatically determining for each query the size of the RS as described in Section 3

4.3. Training and Evaluation

The two basic metrics in IR evaluation are Precision and Recall. The precision P_r of a ranking produced by a retrieval method at some cut-off point r is the fraction of the top r documents that are relevant to the query. On the other side, the recall R_r of a method at a value r is the proportion of the total number of known relevant documents retrieved at that point. Average Precision (AP) was designed to provide a fair comparison across multiple precision levels and is considered as a standard evaluation metric in IR. AP is defined as the arithmetic mean of the precision at all the levels where a relevant document occurs. When averaging AP across a set of topics the resulting evaluation metric is what it is called Mean Average Precision

Table 2: Trained values for every method and collection

Col.	LM	*RM3*				*SDRM3-GE*			*SDRM3*		
	μ	μ	e	r	λ	μ	e	λ	μ	e	λ
AP88-89	1000	500	50	5	0.2	1000	100	0.1	1000	100	0.1
WT2G	2000	2000	50	5	0.4	1000	100	0.2	2000	100	0.3
TREC-678	2000	500	100	10	0.2	500	75	0.1	500	75	0.1
WT10G	1000	500	10	5	0.6	500	10	0.6	500	10	0.7
GOV2	1500	1500	50	10	0.6	1500	50	0.6	1500	50	0.6

(MAP). In order to follow with the traditional evaluation procedure for this task and report effectiveness results for MAP.

As commented we performed a training and test strategy (training for MAP). There are several parameters to train. Namely, the smoothing parameter μ was tuned for LM, RM3, SDRM3-GE and SDRM3 ($\mu \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$). The parameter e, the number of expansion terms, and λ, the interpolation factor, for the pseudo feedback based query expansion were trained in the RM3, SDRM3-GE and SDRM3 methods ($e \in \{5, 10, 25, 50, 75, 100\}$ and $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$). Furthermore, for RM3 the parameter $r = |RS|$, the size of the pseudo-relevant set, was also trained ($r \in \{5, 10, 25, 50, 75, 100\}$) (See Table 2).

Finally Robustness Index (RI) over the initial retrieval (LM) are also reported. We decided to use the RI measure specially designed to evaluate the behaviour of PRF methods to assess the robustness of our proposal. The Robustness Index ($-1 \leq RI(Q) \leq 1$), also called Reliability of Improvement Index, of a model with respect to a baseline was formulated in [31] as in Eq. 9:

Table 3: Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon $p < 0.1$, and Wilcoxon $p < 0.05$ underlined) with respect to LM, RM3, SDRM3-GE and SDRM3 are superscripted with l, r, g and d respectively. Best values are bolded.

	MAP			
Col.	*LM*	*RM3*	*SDRM3-GE*	*SDRM3*
AP88-89	.2775	$.3408^{\underline{l}}$ (+22%)	$.3714^{\underline{l}r}$ (+34%)	$\mathbf{.3794}^{\underline{lr}}$(+37%)
WT2G	.3115	$\mathbf{.3376}^{\underline{lg}}$(+08%)	.3239 (+04%)	$.3345^{\underline{lg}}$ (+08%)
TREC-678	.1915	$.2194^{\underline{l}}$ (+15%)	.2144 (+12%)	$\mathbf{.2245}^{\underline{lg}}$(+17%)
WT10G	.2182	$\mathbf{.2402}^{l}$ (+10%)	.2307 (+06%)	$.2322^{l}$ (+06%)
GOV2	.3295	$.3529^{lg}$ (+07%)	$.3490^{\underline{l}}$ (+06%)	$\mathbf{.3570}^{\underline{lg}}$(+08%)

$$RI(Q) = \frac{n_+ - n_-}{|Q|} \tag{9}$$

where Q is the set of queries over the RI has to be calculated, n_+ is the number of improved queries, n_- the number of degraded queries and $|Q|$ the total number of queries in Q.

4.4. Results

The first comment is that, as expected both RM3 and SDRM3 outperform the initial retrieval with statistical significant differences. Analysing the MAP values for the query expansion methods for the test topics (see Table 3) the best values are obtained by our proposal in three collections and by traditional RM3 in the other two. However, it has to be notice that the differences in favour of RM3 are never statistically significant and in one case the improvements occur with optimal trained values (WT2G). Meanwhile, our method achieves statistical significant improvements in the AP88-89 collection surpassing RM3 in more than 11%. This fact is even more remarkable

Table 4: Values for Robustness Index (RI) with respect to the LM baseline model for every collection. Best values are bolded.

	RI		
Col.	*RM3*	*SDRM3-GE*	*SDRM3*
AP88-89	.23	.64	**.70**
WT2G	.35	.18	**.36**
TREC-678	.22	.14	**.29**
WT10G	**.12**	.05	**.12**
GOV2	**.28**	.26	**.28**

when considering that our proposal has one parameter less that RM3 with its implications in efficiency and method stability across collections. Particularly, it is also clarifying that our election for the mixture of distributions performs constantly better than the Gaussian-Exponential mixture, producing in the majority of the cases statistical significant improvements.

An interesting fact, is that our proposal seems to be more adequate on the text collection meanwhile it is not able to outperform traditional RM3 (in terms of MAP) in two of the three web collections. This can be partially explained by the fact that the fitting of the chosen score distribution model (mixture of two Gaussian distributions) over the web documents is not as good as it is in textual documents. We explain this, because the retrieval model will produce more separated scores for relevant and non-relevant documents if the documents are more focused and shorter (which is the case of the textual documents), where the risk of spurious signals of relevance is lower, that is not the case of the web pages of the Web Collections.

Another important point to analyse is the robustness of the methods, and how this is maintained across collections. Considering the values presented in Table 4 we can see that our method obtains the best values in terms of RI in every collection. Again the differences between the RI values of RM3 and SDRM3 are higher in the text collections than in the web collections. In fact, for the AP88-89 text collection the RI for our method is 0.70 which is the highest RI value reported in this paper, it improves the RM3 method in more than 38% (please, note that RI spans from -1 to 1) and it is close to the maximum RI.

5. Related Work

Related with score distributions *per se*, several works have addressed the finding of best distributions models. In Section 2 we already reviewed the most important works about this topic. Recently, some efforts have been presented in the direction of modelling the score distributions in a systematic way [16], producing an analytical process based on the form of the scoring formulas of the retrieval models.

Score distributions modelling has been applied to tasks such as information filtering or distributed IR, but, in particular, we shall remark the work of Manmatha et al. [23] where score distribution modelling was applied in order to combine the outputs of different search engines for the meta search task, and the works of Arampatzis et al. [3, 4] which formulated the threshold optimization problem over the score distributions models for locating a good cut-off point in the legal search task. The objectives in both cases are quite different from ours, for instance, the legal search task is a high recall task, meanwhile in our case we desire the opposite: a high precision cut for determining the RS.

Very few works have been presented in the direction of refining the RS. Winaver et al. [38] presented a language modelling approach for improving the robustness of the PRF methods. This approach, given a query, computes a set of different language models corresponding with different parameter settings, then, the best computed language model (two different strategies for deciding which one is the best are presented) is selected as initial retrieval. Secondly, different language models are computed using different configurations of r and e over the chosen initial retrieval, selecting that one with the minimum KLD with the query model for processing a second retrieval. Evaluation is not conclusive and no comparison with train and test approach is presented. Moreover, this method requires of a high number of computations of language models for each query, which is quite expensive in terms of computational costs. This last fact is more evident if we compare with our proposal which does not require any extra relevance or language model computation but a very efficient expectation-maximisation process over a limited set of scores.

Huang et al. [15] remarked the importance of selecting the adequate number of feedback documents for the PRF methods. This work explores two different approaches for query-specific feedback document selection. The first approach determines the size of the RS for a given query using either clarity score or cumulative gain. The second one instead of locating the optimal number of documents in the RS uses a mixture model by combining all the query language models rather than only selecting one with the hope of smoothing the effects of the different models. Neither the clarity score base method, nor the cumulative gain strategy, nor the mixture model are able to achieve significant improvements in any collection over the training-test strategy.

In [41] a different view to the problem of the presence of irrelevant documents in the RS is presented. This paper proposes a distribution separation model than taking as input a seed of non-relevant documents and the mixed distributions of the RS will try to estimate an approximation to the true relevance distribution. Evaluation results are interesting but they depend of the existence of relevance judgements to determine the irrelevant seeds (up to the 30% of the known non-relevant documents in the RS are used by the algorithm).

Another open research line is to produce models less sensitive to the composition of the RS. Li [19] presented a new estimation for the relevance models which combines three different aspects: common word discounting, non-uniform document priors and the modification of the traditional pseudo feedback paradigm by considering the original query as a pseudo feedback document rather than combining it with the expanded query. With the introduction of three additional parameters in the model, the method seems to be more robust to the variation in the number of feedback documents than RM3, the effectiveness, once reached the optimal size of the RS, drops slower than for RM3 when increasing the number of pseudo-relevant documents.

6. Conclusions

In this paper we showed how the size of the RS greatly affects to the performance of the RM methods. Motivated by that fact, we presented a method which introduces the use of the threshold optimisation problem over score distribution modelling for automatically selecting the size of the RS. Particularly our method assumes a mixture of two Gaussian distributions and based on this assumption computes the threshold point as the score over which the highest density of relevant documents is obtained.

We have used Bregman soft clustering in order to learn the distributions' parameters from the observed scores. The results of the evaluation showed that in terms of MAP our method is equivalent of better to standard RM3. Important improvements in terms of robustness are obtained with respect to RM3, achieving more than a 38% in the case of the AP88-89 collection. Analyses of the results suggest that our modelling decisions perform better in textual collection than in web collections. Overall, the general objective of improving the robustness of the RM estimations is achieved and moreover, we present the extra advantage of reducing the number of parameters involved in the estimation of the Relevance Models.

Score distribution modelling is not an easy field because it depends in weak assumptions on distributions choice. More work on selecting good mixtures of appropriate statistical distributions has to be carried out. In particular we have observed that the distribution fitting depends not only on the queries but also in the nature of the collection of documents. We envisage future work on automatically selecting for each situation the distribution combination that best fit with the observed data in an attempt of improving the performance of these methods.

Vitae

Javier Parapar (`http://www.dc.fi.udc.es/~parapar/`) holds a Ph.D.in Computer Science. He is an Assistant Professor at the Department of Computer Science of the University of A Coruña. His research interests comprise: pseudo-relevance feedback, clustering and cluster based retrieval, blog and news search, document processing and engineering, text summarisation and retrieval over degraded information.

Manuel A. Presedo-Quindimil holds a Ph.D. from University of Santiago de Compostela. He is an Associated Professor at the Department of Mathematics of the University of A Coruña where he works on the area of Statistics. His research focus is in applied statistics and in particular bootstrapping methods.

Álvaro Barreiro (`http://www.dc.fi.udc.es/~barreiro/`) holds a Ph.D. from University of Santiago de Compostela. He is a Professor at the Department of Computer Science of the University of A Coruña where he leads the Information Retrieval Lab (`http://www.irlab.org`). He has been the main researcher of several IR research projects funded by the Spanish Government.

References

[1] N. Abdul-jaleel, J. Allan, W.B. Croft, O. Diaz, L. Larkey, X. Li, M.D. Smucker, C. Wade, UMass at TREC 2004: Novelty and HARD, in: Proceedings of TREC-13, NIST Special Publication, National Institute for Science and Technology, 2004.

[2] G. Amati, C. Carpineto, G. Romano, Query Difficulty, Robustness and Selective Application of Query Expansion, in: S. McDonald, J. Tait (Eds.), Proceedings of the 26th European conference on Advances in Information Retrieval, volume 2997 of *ECIR'04*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 127–137.

[3] A. Arampatzis, J. Beney, C.H.A. Koster, T.P. van der Weide, Incrementality, Half-life, and Threshold Optimization for Adaptive Document Filtering., in: Proceedings of TREC-9, NIST Special Publication, National Institute for Science and Technology, 2000.

[4] A. Arampatzis, J. Kamps, S.E. Robertson, Where to stop reading a ranked list?, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR'09, ACM Press, New York, New York, USA, 2009, pp. 524–531.

[5] A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh, Clustering with Bregman Divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.

[6] C. Baumgarten, A probabilistic solution to the selection and fusion problem in distributed information retrieval, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99, ACM, New York, NY, USA, 1999, pp. 246–253.

[7] A. Bookstein, When the most pertinent document should not be retrieved - An analysis of the Swets model, Information Processing & Management 13 (1977) 377–383.

[8] D. Carmel, E. Yom-Tov, A. Darlow, D. Pelleg, What makes a query difficult?, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '06, ACM Press, New York, New York, USA, 2006, p. 390.

[9] W.B. Croft, D.J. Harper, Using Probabilistic Models of Document Retrieval without Relevance Information, Journal of Documentation 35 (1979) 285–295.

[10] W.B. Croft, D. Metzler, T. Strohman, Search Engines: Information Retrieval in Practice, 1st ed., Addison-Wesley Publishing Company, USA, 2009.

[11] S. Cronen-Townsend, Y. Zhou, W.B. Croft, Predicting query performance, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02, ACM Press, New York, New York, USA, 2002, p. 299.

[12] K. Dai, V. Pavlu, E. Kanoulas, J.A. Aslam, Extended expectation maximization for inferring score distributions, in: Proceedings of the 34th European conference on Advances in Information Retrieval, ECIR'12, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 293–304.

[13] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society 39 (1977) 1–38.

[14] V. Garcia, F. Nielsen, Simplification and hierarchical representations of mixtures of exponential families, Signal Processing 90 (2010) 3197–3212.

[15] Q. Huang, D. Song, S. Rüger, Robust query-specific pseudo feedback document selection for query expansion, in: Proceedings of the 30th European conference on Advances in Information Retrieval, ECIR'08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 547–554.

[16] E. Kanoulas, K. Dai, V. Pavlu, J.A. Aslam, Score distribution models: assumptions, intuition, and robustness to score manipulation, in: Proceedings of the 33rd international ACM SIGIR conference on Research

and development in information retrieval, SIGIR '10, ACM, New York, NY, USA, 2010, pp. 242–249.

[17] E. Kanoulas, V. Pavlu, K. Dai, J.A. Aslam, Modeling the score distributions of relevant and non-relevant documents, in: Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory, ICTIR '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 152–163.

[18] V. Lavrenko, W.B. Croft, Relevance based language models, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'01, ACM, New York, NY, USA, 2001, pp. 120–127.

[19] X. Li, A new robust relevance model in the language model framework, Information Processing & Management 44 (2008) 991–1007.

[20] Y. Lv, C. Zhai, A comparative study of methods for estimating query language models with pseudo feedback, in: Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, ACM, New York, NY, USA, 2009, pp. 1895–1898.

[21] Y. Lv, C. Zhai, Adaptive relevance feedback in information retrieval, in: Proceeding of the 18th ACM conference on Information and knowledge management, CIKM '09, ACM, New York, NY, USA, 2009, pp. 255–264.

[22] D. Madigan, Y. Vardi, I. Weissman, Extreme value theory applied to document retrieval from large collections, Inf. Retr. 9 (2006) 273–294. URL: http://dx.doi.org/10.1007/s10791-006-0882-4. doi:10.1007/s10791-006-0882-4.

[23] R. Manmatha, T. Rath, F. Feng, Modeling score distributions for combining the outputs of search engines, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'01, ACM Press, New York, New York, USA, 2001, pp. 267–275.

[24] M. Mitra, A. Singhal, C. Buckley, Improving automatic query expansion, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, ACM Press, New York, New York, USA, 1998, pp. 206–214.

[25] F. Nielsen, V. Garcia, Statistical exponential families: A digest with flash cards, CoRR abs/0911.4863 (2009).

[26] C.C.V. Ounis, I.;Lioma, Research directions in terrier, Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper (2007).

[27] S. Robertson, On score distributions and relevance, in: Proceedings of the 29th European conference on Advances in Information Retrieval, ECIR'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 40–51.

[28] S.E. Robertson, The probability ranking principle in IR, in: K. Sparck Jones, P. Willett (Eds.), Readings in information retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 281–286.

[29] J. Rocchio, Relevance feedback in information retrieval, in: G. Salton (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, Inc., 1971, pp. 313–323.

[30] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, Knowl. Eng. Rev. 18 (2003) 95–145.

[31] T. Sakai, T. Manabe, M. Koyama, Flexible pseudo-relevance feedback via selective sampling, ACM Transactions on Asian Language Information Processing (TALIP) 4 (2005) 111–135.

[32] T. Sakai, S.E. Robertson, Flexible pseudo-relevance feedback using optimization tables, in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, ACM Press, New York, New York, USA, 2001, pp. 396–397.

[33] G. Salton, The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[34] A. Shtok, O. Kurland, D. Carmel, Predicting Query Performance by Query-Drift Estiamation, in: Proceedings of the 2nd International Con-

ference on Theory of Information Retrieval: Advances in Information Retrieval Theory, volume 5766 of *ICTIR '09*, Springer, Berlin, Heidelberg, 2009, pp. 305–312.

[35] J.A. Swets, Information retrieval systems, Science 141 (1963) 245–250.

[36] J.A. Swets, Effectiveness of information retrieval methods, American Documentation 20 (1969) 72–89.

[37] E. Voorhees, D. Harman, N.I. of Standards, T. (US), TREC: Experiment and evaluation in information retrieval, volume 63, MIT press Cambridge, 2005.

[38] M. Winaver, O. Kurland, C. Domshlak, Towards robust query expansion: model selection in the language modeling framework, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, ACM, New York, NY, USA, 2007, pp. 729–730.

[39] J. Xu, W.B. Croft, Query expansion using local and global document analysis, in: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'96, ACM, New York, NY, USA, 1996, pp. 4–11.

[40] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to information retrieval, ACM Trans. Inf. Syst. 22 (2004) 179–214.

[41] P. Zhang, Y. Hou, D. Song, Approximating true relevance distribution from a mixture model based on irrelevance data, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR'09, ACM Press, New York, New York, USA, 2009, p. 107.