Modeling the Score Distributions of Relevant and Non-relevant Documents

Evangelos Kanoulas, Virgil Pavlu, Keshi Dai, and Javed A. Aslam*

College of Computer and Information Science
Northeastern University, Boston, USA
ekanou,vip,daikeshi,jaa@ccs.neu.edu

Abstract. Empirical modeling of the score distributions associated with retrieved documents is an essential task for many retrieval applications. In this work, we propose modeling the relevant documents' scores by a mixture of Gaussians and modeling the non-relevant scores by a Gamma distribution. Applying variational inference we automatically trade-off the goodness-of-fit with the complexity of the model. We test our model on traditional retrieval functions and actual search engines submitted to TREC. We demonstrate the utility of our model in inferring precision-recall curves. In all experiments our model outperforms the dominant exponential-Gaussian model.

1 Introduction

Information retrieval systems assign scores to documents according to their relevance to a user's request and return documents in a descending order of their scores. In reality, however, a ranked list of documents is a mixture of both relevant and non-relevant documents. For a wide range of retrieval applications (e.g. information filtering, topic detection, meta-search, distributed IR), *modeling* and *inferring* the distribution of relevant and non-relevant documents over scores in a reasonable way could be highly beneficial. For instance, in information filtering and topic detection modeling the score distributions of relevant and non-relevant documents can be utilized to find the appropriate threshold between relevant and non-relevant documents [16, 17, 2, 19, 7, 15], in distributed IR it can be used for collection fusion [3], and in meta-search to combine the outputs of several search engines [10].

Inferring the score distribution for relevant and non-relevant documents in the absence of any relevance information is an extremely difficult task, if at all possible. *Modeling* score distributions in the right way is the basis of any possible inferences. Due to this, numerous combinations of statistical distributions have been proposed in the literature to model score distributions of relevant and non-relevant documents. In 60's and 70's Swets attempted to model the score distributions of non-relevant and relevant documents with two Gaussians

* We gratefully acknowledge the support provided by NSF grants IIS-0533625 and IIS-0534482

of equal variance [16], two Gaussians of unequal variance and two exponentials [17]. Bookstein instead proposed a two Poisson model [6] and Baumgarten a two Gamma model [3]. A negative exponential and a Gamma distribution [10] has also been proposed in the literature. The dominant model, however, has been an exponential for the non-relevant documents and a Gaussian for the relevant ones [2, 10, 19].

As mentioned before the right choice of distributions (that is distributions that reflect the underline process that produces the scores of relevant and non-relevant documents) can enhance the ability to infer these distributions, while a bad choice may make this task practically impossible. Clearly a strong argument for choosing any particular combination of distributions is the goodness-of-fit to a set of empirical data. However, the complexity of the underline process that generates documents' scores makes the selection of the appropriate distributions a hard problem. Hence, even though the exponential - Gaussian model is the dominant one, there is no real consensus on the choice of the distributions. For instance, recently, Bennett [4], by utilizing the two Gaussians model for text classification and based on the observation that documents' scores outside the modes of the two Gaussians (corresponding to "extremely irrelevant" and "obviously relevant" documents) demonstrate different empirical behavior than the scores between the two modes (corresponding to "hard to discriminate" documents) introduced several asymmetric distributions to capture these differences.

Even though the goodness-of-fit can be a reasonable indicator of whether a choice of statistical distributions is the right one, from an IR perspective, these distributions should also possess a number of IR theoretical properties. Robertson considered various combinations of distributions and examined whether these combinations exhibit anomalous behavior with respect to theoretical properties of precision and recall [13].

In this work, we revisit the choice of distributions used to model documents' scores. Similarly to Bennett [4] we observed that the scores of relevant documents demonstrate different behavior in different score ranges. In order to study what is the appropriate choice of distributions for relevant and non-relevant documents we assume that the relevance information for all documents is available. We utilize a richer class of density functions for modeling the score distributions. In particular, we empirically fit a Gamma distribution in the scores of the non-relevant documents and a mixture of Gaussians in the scores of the relevant documents. Note that, the Gamma distribution represents the sum of M independent exponentially distributed random variables. In order to balance between the flexibility and the generalization power of the model we take a Bayesian treatment on the model that automatically trades-off the goodness-of-fit with the complexity of the model. We show that the data alone suggest that a mixture of two Gaussians for the relevant documents and a Gamma distribution with $M > 1$ is often times the right choice to model documents' scores. Further, we examine the IR utility of our model by testing how well one can infer precision-recall curves from the fit probability distributions. We show that our model outperforms the dominant exponential - Gaussian model.

2 Modeling Score Distributions

In this work, we empirically fit a Gamma distribution in the scores of the non-relevant documents and a mixture of Gaussians in the scores of the relevant documents (*GkG* model) and compare it to the dominant exponential-Gaussian model (*EF* model).

To avoid the effects of arbitrary query manipulations and score transformations that systems submitted to TREC (Text REtrieval Conference) often apply, in the sections that follow, we instead use scores produced by traditional IR models. Later, in Section 4, we validate our model on TREC systems.

The document collections used are the ones contained in TREC Disk 4 and 5, excluding the *Congressional Record* sub-collection, that is the exact same document collection used in TREC 8. The topics used are the TREC topics $401-450$ (the topics in TREC 8) [18]. Indexing and search was performed using the Terrier search engine [11]. Porter stemming and stop-wording was applied. The document scores obtained are the outputs of (a) Robertson's and Spärck Jones' TF-IDF [14], (b) BM25 [12], (c) Hiemstra's Language Model (LM) [9], and (d) PL2 divergence from randomness [1] (with Poisson estimation for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization). Further, three different topic formulations were used, (a) topic titles only, (b) topic titles and descriptions, and (c)topic titles, descriptions and narratives.

2.1 Methodology

The Gamma distribution was used to model the scores of the non-relevant documents. The Gamma density function with scale θ and shape M is given by,

$$P(x|M,\theta) = x^{M-1}\frac{\exp^{-M/\theta}}{\theta^M \Gamma(M)}$$

where, $\Gamma(M) = (M-1)!$. The mean of the distribution is $M\theta$, while the variance is $M\theta^2$. The maximum likelihood estimation (MLE) was used to estimate the Gamma parameters. When $M = 1$, the Gamma distribution degrades to an exponential distribution with rate parameter $1/\theta$.

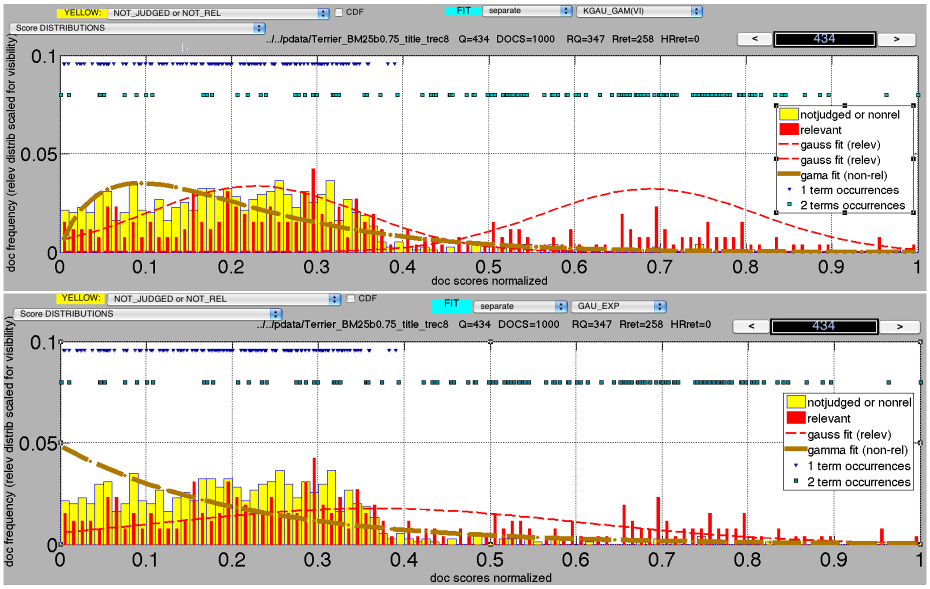
The scores of relevant documents are modeled by a mixture of K Gaussians. Fitting the mixture of Gaussians into the scores could be easily done by employing the EM algorithm if the number of Gaussian components K was known. However, we considered as known only an upper bound on K. Given the fact that the larger the number of components is the better the fit will be and that EM finds the maximum likelihood mixture of Gaussians regardless of the model complexity, the EM algorithm is not appropriate for our problem. Instead, to avoid over-fitting, we employ a Bayesian treatment on the model by utilizing Variational Bayesian model selection for the mixture of Gaussians [5, 8].

The mixture distribution of K Gaussian components is given by,

$$P(x|\pi,\mu,\Lambda) = \sum_{i=1}^{K} \pi_i \mathcal{N}(x|\mu_i, \Lambda_i^{-1})$$

where π_i are the mixing coefficients, and satisfy $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^{K} \pi_i = 1$, μ_i and Λ_i the mean and the precision of the i^{th} Gaussian component.

The mixture coefficients π essentially give the contribution of each gaussian to the model. A fully Bayesian treatment of the mixture modeling problem involves the introduction of prior distributions over all the parameters, that is including π. Given a fixed number of potential components (an upper bound on K) the variational inference approach causes the mixing coefficients of unwanted components to go to zero and essentially leads to an automatic trade-off between the goodness-of-fit and the complexity of the model. The approach used in this paper to determine the number of components is to treat the mixing coefficients π as parameters and make point estimates of their value instead of maintaining a probability distribution over them [8].



YELLOW: NOT_JUDGED or NOT_REL □ CDF FIT separate KGAU_GAM(VI)

Score DISTRIBUTIONS

../../pdata/Terrier_BM25b0.75_title_trec8 Q=434 DOCS=1000 RQ=347 Rret=258 HRret=0

YELLOW: NOT_JUDGED or NOT_REL □ CDF FIT separate GAU_EXP

Score DISTRIBUTIONS

../../pdata/Terrier_BM25b0.75_title_trec8 Q=434 DOCS=1000 RQ=347 Rret=258 HRret=0

Fig. 1. The histogram over the scores of non-relevant and relevant documents and the Gamma and k Gaussians distribution (top) along with the negative exponential and single Gaussian distributions (bottom) fit into these scores separately.

2.2 Results and Analysis

We separately fit the Gamma distribution and the mixture of Gaussians into the scores of the non-relevant and relevant documents, respectively, per topic. There are 50 topics available and 3 query formulations (title, title and description and title, description and narrative), along with the relevance information for the top 1000 documents returned by 4 IR systems (TF-IDF, BM25, LM and PL2). Thus, there are in total 600 ranked lists of documents. The scores of the documents were first normalized into a 0 to 1 range.

An example of fitting an exponential-Gaussian model and a Gamma and a mixture of two Gaussians into scores of non-relevant and relevant documents (separately) for query 434 ("Estonia economy") is shown in Figure 1. The wide yellow-bar and the thin red-bar histograms in both plots correspond to the non-relevant and relevant documents scores, respectively (scaled). Further, the top plot shows a negative exponential and a single Gaussian density functions fit into the scores, while the bottom plot shows a Gamma density function and a mixture of two Gaussians fit into the scores. As it can be observed, the Gamma and the mixture of two Gaussians can better fit the data than the choice of the exponential and the single Gaussian. To summarize our results we report the parameter M of the Gamma distribution, which as mentioned earlier corresponds to the number of independent exponential density functions averaged, and the number K of Gaussian components in the mixture, for all four systems, all 150 topics (50 topics and 3 query formulations). Figure 2 shows the histograms over M and K. As it can be observed, both M and K, in most of the cases, are different from 1, which shows that, taken into account the complexity of the model, the data suggest that a Gamma distribution and a mixture of Gaussians is a better fit than a negative exponential and a single Gaussian. In particular, the mean number of Gaussian components is 1.7, while the mean value of the parameter M is 1.3. In order to quantify and compare the goodness-of-fit for the different statistical distributions fit into the scores of relevant and non-relevant documents we employ hypothesis testing. The null hypothesis tested is that the scores of relevant (non-relevant) documents come from a certain distribution. The Kolmogorov-Smirnov test (using the maximum distance between the empirical and the theoretical cumulative distributions as a statistic) was utilized. The histogram of the p-values for all systems and all queries is shown in Figure 3. The top row corresponds to the p-values of testing the relevant documents scores against the single Gaussian distribution and mixture of K Gaussians, while the

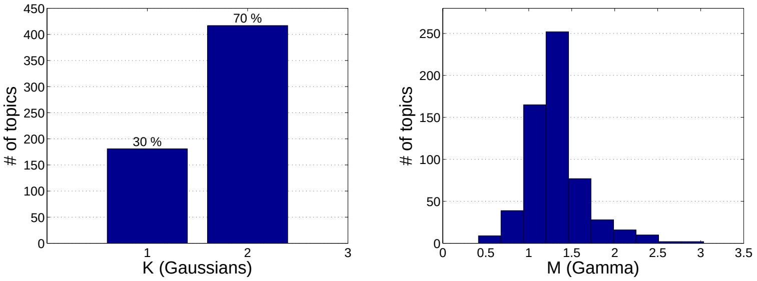


Fig. 2. The histograms over the number K of Gaussian components and the parameter M of the Gamma distribution, over all IR models, topics and topic formulations.

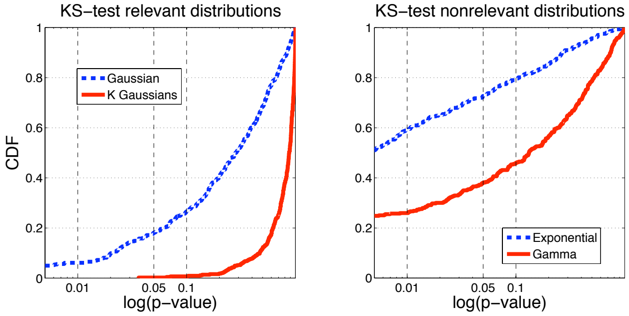


Fig. 3. The histogram of p-values of the Kolmogorov-Smirnov test on all systems, topics and topic formulations for relevant and non-relevant documents score distribution.

bottom row corresponds to the p-values of testing the non-relevant documents scores against the negative exponential and the Gamma distributions. As it can be observed, in the case of the relevant documents' scores distribution the single Gaussian distribution yields the worst results (as expected), with most of the p-values being less than the significance level of 0.05 and thus rejecting the null hypothesis, while the mixture of two Gaussian distributions yields clearly much higher p-values. In particular, for 82% of the system-query pairs the null hypothesis that the score distribution is a single Gaussian distribution could not be rejected, while the corresponding percentage for the mixture of two Gaussians is **100%**. For the case of non-relevant documents the corresponding percentages for the exponential and Gamma distributions are 27% and **62%**, respectively.

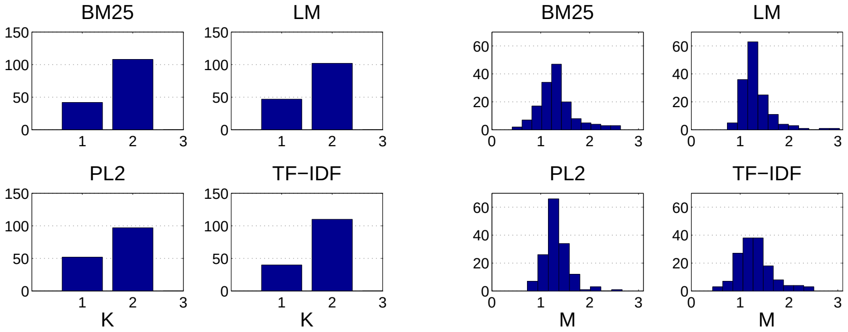


Fig. 4. The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and topic formulations for each IR model.

Finally, we tested how the different IR systems and topic formulations affect the parameter M and the number K of Gaussian components. In Figures 4 and 5, we report the histograms over K for each system separately (50 topics with 3 topic formulations) and the histograms over K for each query formulation (all 50 topics and 4 IR systems). As it can be observed, the distribution of K appears to be independent both with respect to the IR model and with respect to query formulation. To validate our observations we run an n-way ANOVA testing whether the mean values of K per IR model - query formulation are equal and we could not reject the hypothesis.

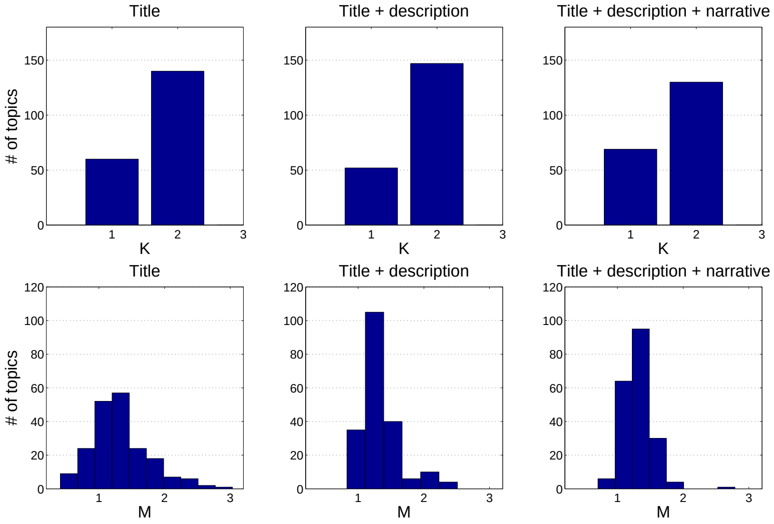


Fig. 5. The histogram over the number K of Gaussian components and the parameter M of Gamma distribution, over all topics and IR models for each topic formulation.

2.3 On the Choice of Score Distributions

So far the optimal distributions to model the scores of relevant and non-relevant documents have been dictated by the data. In this section, we give an intuitive explanation of choice of a Gamma distribution to model non-relevant documents' scores and a mixture of Gaussians to model relevant documents' scores from an IR point of view.

An intuition behind the shape of the distribution that models the scores of relevant documents is given by Manmatha et al. [10]. Assuming that a query consists of a single term, Manmatha shows that the scores of relevant documents can be modeled as a Poisson distribution with a large λ parameter, which approaches a Gaussian distribution. Now, let's consider queries that consist of multiple terms and let's revisit the top plot in Figure 1. The query used in the example is: “Estonia economy”. Each relevant document in the plot corresponds either to a triangular or to a rectangular marker at the top of the plot. The triangular markers denote the relevant documents for which only one out of the two query terms occur in the document, while the rectangular ones denote the relevant documents for which both terms occur in the document. By visual inspection, the relevant documents containing a single term clearly correspond to the low-scores' Gaussian, while the relevant documents containing both terms clearly correspond to the high-scores' Gaussian. Essentially, the former documents get a low score due to the fact that only one terms appear in them but they happen to be relevant to the query, while the latter correspond to documents that are obviously relevant. We observed the same phenomenon for many different queries independently of the IR model used for retrieval and independent of the query formulation. In the case of queries with multiple terms (e.g. queries that consists of both the title and the description), even though the possible number of query terms that may co-occur in a document is greater than 2 (e.g. for a query with 3 terms, all terms may occur in a document or only two of them or only a single one of them), we observed that there is a threshold on the number of terms occurring in the document; relevant documents containing a number of terms that is less than this threshold are clustered towards low scores (first Gaussian), while relevant documents containing a number of terms that is greater than the threshold are clustered towards high scores (second Gaussian).

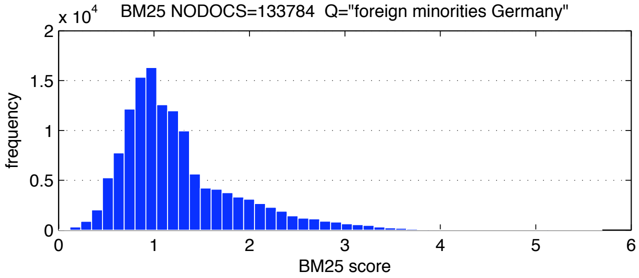


Fig. 6. The distribution of BM25 scores for all 133, 784 documents (containing at least one query term) on query "foreign minorities Germany". Note the different slopes at the left and at the right of the mean. Truncating the list at rank 1, 000 would cause the scores' distribution to look like an exponential one.

Regarding the non-relevant documents, given that the number of them is orders of magnitude larger than the number of the relevant ones, a modeling distribution over non-relevant documents' scores is essentially a modeling distribution over all scores. Previous work [10, 13] argues that this distribution is a negative exponential but often times a more flexible distribution is necessary. The Gamma distribution, which can range (in skewness) from an exponential to a Gaussian distribution is flexible enough. In order to explain why a Gamma distribution is a better choice, several factors should be considered.

- Truncation cut-off: If a list is arbitrarily truncated very early (say at rank 1, 000) the distribution of the top scores may indeed look as an exponential. However looking deep down in the list (say up to rank 200, 000), the scores' distribution shape changes (Figure 6).
- Query complexity: Arguments for the scores' distribution for single term queries have been given in the literature [10]. For a query with two or more terms, most non-trivial documents (i.e. the ones that contain at least two

query terms) will have the following property; the contribution of the two or more terms to the final score of a document would often times be very different for the two or more terms, with some terms having a low contribution while others having a higher contribution. Averaging such effects is likely to produce a "hill" of score frequencies, perhaps with different slopes at the left and the right side of the mean; the Gamma distribution is known to be an average of exponential distributions.
- Retrieval function: We mostly look at scoring functions that are decomposable into a sum of scores per query terms, like TF-IDF or Language Models (after taking logs); such scores also induce averaging effects.

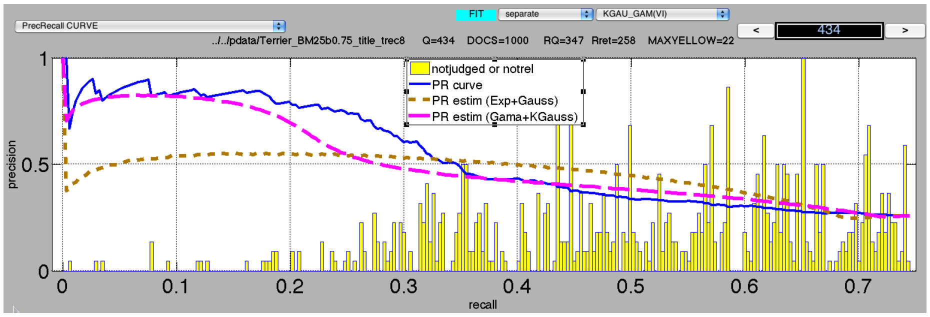


Fig. 7. Precision-Recall curve (blue) for query 434 and the BM25 retrieval function implemented by Terrier. It is easy to see that the PR curve estimated from the GkG model (magenta) is much better than the PR estimated from the EG model (brown). Yellow bars indicate the number of non-relevant documents in each recall interval.

3 Precision-Recall Curves

As a utility of our model for IR purposes, we estimate the precision-recall (PR) curve separately from both the *EG* and *GkG* model. Similarly to Robertson [13], let f_r and f_n denote the model densities of relevant and non-relevant scores, respectively; $F_r(x) = \int_x^1 f_r(x)dx$ and $F_n(x) = \int_x^1 f_n(x)dx$ are the cumulative density functions *from the right*. While the density models might have support outside the range [0,1], we use integrals up to 1 because our scores are normalized. For each recall level r we estimate the retrieval score at which r happens, from the relevant cumulative density: $score(r) = F_r^{-1}(r)$, which we compute numerically. Then we have $n(r) = F_n(score(r))$ as the percentage of non-relevant documents found up to recall r in the ranked list. Finally, the precision at recall r can be computed as in [13], $prec(r) = \frac{r}{r+n(r)*G}$, where G is the ratio of non-relevant to relevant documents in the collection searched. Computing precision at all recall levels from the score distribution models f_r and f_n gives an estimated PR curve.

In the reminder of this section we show that estimating PR curves from the *GkG* model clearly outperforms PR curves estimated from the dominant *EG* model.

To measure the quality of the estimated PR curves we report the RMS error between the actual and the predicted precisions at all recall levels for both models. The results are summarized in Table 1, separately for each model. Language model LM and Divergence from randomness PL2 seem to produce slightly better PR estimates, independent of the query formulation. The over-all RMSE of *GkG* vs. *EG* is .094 vs .117, or about 20% improvement.

	title		title+desc		title+desc+narrative	
	EG	GkG	EG	GkG	EG	GkG
BM25	.135	.106	.122	.093	.117	.099
LM	.117	.098	.101	.085	.091	.076
PL2	.113	.092	.116	.094	.113	.092
TFIDF	.137	.106	.122	.095	.120	.100

Table 1. RMS error between the actual and the inferred precision-recall curves.

Further, we report the mean absolute error between the actual and predicted precisions at all recall levels. This is the area difference between the estimated and the actual curve, which immediately gives a bound for the difference in Average Precision of the two curves (because the AP metric is approximated by the area under the PR curve). The results are reported in Table 2. Note that the best fit with respect to MAE are given for the full query formulation (title, description and narrative); the overall MAE for *GkG* is .055 vs *EG* with .074, or an improvement of about 25%.

	title		title+desc		title+desc+narrative	
	EG	GkG	EG	GkG	EG	GkG
BM25	.091	.067	.076	.052	.071	.056
LM	.078	.063	.064	.052	.055	.043
PL2	.072	.056	.070	.052	.065	.049
TFIDF	.092	.067	.076	.053	.072	.055

Table 2. Mean Absolute Error between actual and inferred precision-recall curves.

.

4 TREC Search Engines

To avoid the effects of arbitrary query manipulations and score transformations that systems submitted to TREC (Text REtrieval Conference) often applied, we used in our experiments scores produced by traditional IR models. In this section we apply our methodology over the score distributions returned by search engines submitted to TREC 8. Out of the 129 manual and automatic systems submitted

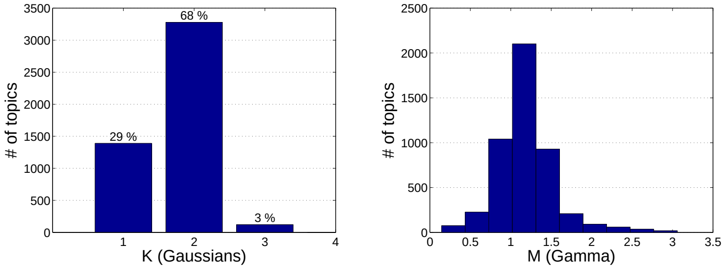


Fig. 8. The histograms over the number K of Gaussian components and the parameter M of the Gamma distribution, over all IR models, topics and topic formulations.

to TREC 8 30 of them were excluded from our experiments since they transform document scores into ranks. No other quality control was performed. As earlier, we report the parameter M of the Gamma distribution, and the number K of Gaussian components in the mixture, for all systems and all queries as histograms in Figure 8. As it can be observed, similarly to the case of the traditional IR models, both M and K, in most cases, are different from 1, confirming that a Gamma distribution and a mixture of Gaussians is a better fit than a negative exponential and a single Gaussian.

5 Conclusions

In this work, we proposed modeling the relevant documents' scores by a mixture of Gaussians and modeling the non-relevant scores by a Gamma distribution. In all experiments conducted our model outperformed the dominant exponential-Gaussian model. Further, we demonstrated the utility of our model in inferring precision-recall curves. Some intuition about the choice of the particular model from an IR perspective was also given.

References

1. G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.
2. A. Arampatzis and A. van Hameran. The score-distributional threshold optimization for adaptive binary classification tasks. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 285–293, New York, NY, USA, 2001. ACM.
3. C. Baumgarten. A probabilistic solution to the selection and fusion problem in distributed information retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 246–253, New York, NY, USA, 1999. ACM.

4. P. N. Bennett. Using asymmetric distributions to improve text classifier probability estimates. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 111–118, New York, NY, USA, 2003. ACM.
5. C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006.
6. A. Bookstein. When the most "pertinent" document should not be retrieved—an analysis of the swets model. *Information Processing & Management*, 13(6):377–383, 1977.
7. K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *In Proceedings of the 11th Text Retrieval Conference*, 2003.
8. A. Corduneanu and C. M. Bishop. Variational bayesian model selection for mixture distributions. In *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, pages 27–34. Morgan Kaufmann, 2001.
9. D. Hiemstra. *Using language models for information retrieval.* PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.
10. R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–275, New York, NY, USA, 2001. ACM.
11. I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Next Generation Web Search, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 8(1):49–56, 2007.
12. E. Robertson, S. and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
13. S. Robertson. On score distributions and relevance. In G. Amati, C. Carpineto, and G. Romano, editors, *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007*, volume 4425/2007 of *Lecture Notes in Computer Science*, pages 40–51. Springer, June 2007.
14. S. E. Robertson and S. K. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.
15. M. Spitters and W. Kraaij. A language modeling approach to tracking news events. In *In Proceedings of TDT workshop 2000*, pages 101–106, 2000.
16. J. A. Swets. Information retrieval systems. *Science*, 141(3577):245–250, July 1963.
17. J. A. Swets. Effectiveness of information retrieval methods. *American Documentation*, 20:72–89, 1969.
18. E. M. Voorhees and D. K. Harman. *TREC: Experiment and Evaluation in Information Retrieval.* Digital Libraries and Electronic Publishing. MIT Press, September 2005.
19. Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 294–302, New York, NY, USA, 2001. ACM.