

Language Models, Smoothing, and IDF Weighting

Najeeb Abdulmutalib, Norbert Fuhr

University of Duisburg-Essen, Germany

{najeeb|fuhr}@is.inf.uni-due.de

Abstract

In this paper, we investigate the relationship between smoothing in language models and idf weights. Language models regard the relative within-document-frequency and the relative collection frequency; idf weights are very similar to the latter, but yield higher weights for rare terms. Regarding the correlation between the language model parameters and relevance for two test collections, we find that the idf type of weighting seems to be more appropriate. Based on the observed correlation, we devise empirical smoothing as a new type of term weighting for language models, and retrieval experiments confirm the general applicability of our method. Finally, we show that the most appropriate form of describing the relationship between the language model parameters and relevance seems to be a product form, which confirms a language model proposed before.

1 Introduction

Since several years, language models are the preferred type of IR models [Hiemstra, 1998; Ponte and Croft, 1998; Berger and Lafferty, 1999]. In contrast to other models, they explicitly include a document indexing model that relates the within-document frequency of a term to its indexing weight. On the other hand, there is no explicit notion of probability of relevance. Closely related to this statement, there is the somewhat unclear relation between tf*idf weighting (like e.g. in the classic vector space model or in BM25) and the probabilistic parameters of language models.

In this paper, we present some empiric results that relate language models to tf*idf weights, which leads us to a new smoothing method giving us good retrieval results.

2 Language Models

Language models regard a text in form of a sequence of words as a stochastic process. Thus, for a given vocabulary (set of terms) T , a language model θ is defined as a probability distribution

$$\theta = \{(t_i, P(t_i|\theta)|t_i \in T)\} \quad \text{with} \quad \sum_{t_i \in T} P(t_i|\theta) = 1$$

In the most simple form, one assumes independence of term occurrences, and thus the probability of a document text $d = t_1 t_2 t_3 \dots t_m$ wrt. to language model θ can be computed as $P(d|\theta) = \prod_{j=1}^m P(t_j|\theta)$.

The basic idea for defining a retrieval function is to compare the document's d language models to that of the query q . One way for doing this is to compute the probability that the query was generated by the document's language model:

$$\begin{aligned} P(q|d) &\approx \prod_{t_i \subseteq q^T} P(t_i|d) \\ &= \prod_{t_i \in q^T \cap d^T} P_s(t_i|d) \prod_{t_i \in q^T - d^T} P_u(t_i|d) \\ &= \prod_{t_i \in q^T \cap d^T} \frac{P_s(t_i|d)}{P_u(t_i|d)} \prod_{t_i \in q^T} P_u(t_i|d) \quad (1) \end{aligned}$$

Here d^T denotes the set of terms occurring in the document, and q^T refers to the set of query terms. $P_s(t_i|d)$ denotes the probability that the document is about t_i , given that t_i occurs (is seen) in the document. On the other hand, $P_u(t_i|d)$ denotes the same probability for those terms t_i not occurring (is unseen) in the document.

The estimation of these parameters suffers from the problem of sparse data. Thus, a number of smoothing methods have been developed. Let F denote the total number of tokens in the collection and $cf(t)$ the collection frequency of term t , $l(d)$ the number of tokens in document d and $tf(t, d)$ the corresponding within-document frequency of term t . Then we estimate

$$P_{avg}(t) = \frac{cf(t)}{F} \quad \text{and} \quad P_{ml}(t|d) = \frac{tf(t, d)}{l(d)}$$

where $P_{avg}(t)$ is the average relative frequency of t in the collection, and $P_{ml}(t|d)$ is the maximum likelihood estimate for the probability of observing t at an arbitrary position in d .

Various smoothing methods have been developed in the past. In this paper, we only regard the most popular one, namely Jelinek-Mercer (JM) smoothing [Jelinek and Mercer, 1980]:

$$P_s(t_i|d) = (1 - \lambda)P_{ml}(t|d) + \lambda P_{avg}(t) \quad (2)$$

Here λ (with $0 \leq \lambda \leq 1$) is a global smoothing parameter that allows for collection-specific tuning. For the unseen terms, [Zhai and Lafferty, 2001] propose the following estimate:

$$\begin{aligned} P_u(t_i|d) &= \alpha_d P_{avg}(t) \\ \text{with } \alpha_d &= \frac{1 - \sum_{t_i \in q^T \cap d^T} P_{avg}(t)}{1 - \sum_{t_i \in q^T \cap d^T} P_{ml}(t|d)} \end{aligned}$$

As we can see from eqn 2, the term weight is a weighted sum of its relative within-document frequency P_{ml} and its relative frequency in the whole collection, P_{avg} . Classic tf*idf weighting formulas are based on the same parameters. However, whereas the tf part of these types of weights usually is some monotonic transformation of P_{ml} , the idf part is the negative logarithm of the document frequency, i.e. similar to $-\log(P_{avg})$. (Note, however, that P_{avg} refers to tokens, whereas the idf weight regards the number of documents in which a term occurs. A theoretic treatment of this aspect can be found in [Roelleke and Wang, 2008]. Here we assume that this difference is negligible.) The theoretic justification of idf weights goes back to [Croft and Harper, 1979], who showed that the relative document frequency is an estimate for the probability of the term occurring in a nonrelevant document, thus linking this parameter to relevance. Thus, we have the language model interpretation of P_{avg} on one hand, and the relevance-oriented interpretation of the idf weight on the other hand. In the former, the weight of a term grows monotonically with P_{avg} , whereas the opposite is true for idf weights. Although [Roelleke and Wang, 2008] shows how the two kinds of weightings relate to each other, this paper does not resolve the apparent contradiction.

3 Language model parameters vs. probability of relevance

As an alternative to a theoretic treatment, we performed an empirical study on the distribution of the language model parameters P_{ml} and P_{avg} in relevant and nonrelevant documents. For that, we regarded two test collections:

1. The INEX 2005 collection consisting of 16819 journal articles (764 MB), where we regard each of the 21.6 million XML elements as a document¹. Due to this document definition, we have a great variation in document lengths, as is illustrated in figure 1. As query set, we use the corresponding 29 content-only queries.
2. The AP part of the TREC collection containing 240,000 documents, along with TREC queries 51-100 and 101-150.

For computing our statistics for the given query sets, we considered all query-document pairs where the document

¹Retrieval of XML elements can be used as a first step in a two-stage process for focused XML retrieval, where the second step picks the most specific elements from each XML document that answer the query in the most exhaustive way.

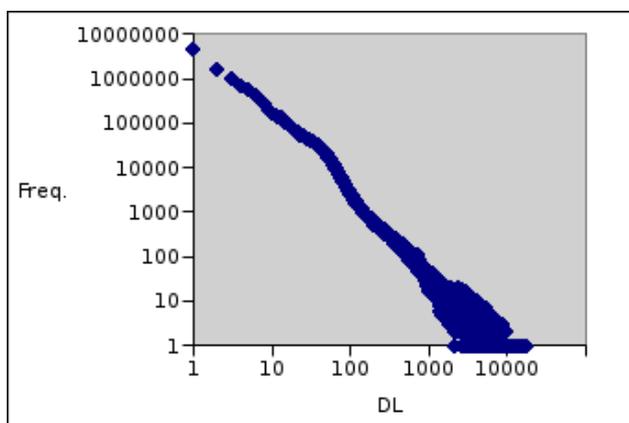


Figure 1: Distribution of document lengths in INEX

contains at least one query term. In case the document is relevant, all query terms are regarded as being relevant for this document otherwise all terms are irrelevant. Now we aim at relating the (P_{ml}, P_{avg}) pairs of terms to their probability of relevance $P(R|t)$ that a document containing t will be judged relevant to a random query containing t as query term. For that, we perform an appropriate binning of (P_{ml}, P_{avg}) pairs into two-dimensional intervals, and then we compute the ratio of relevant pairs among all pairs in an interval.

Figure 2 shows the corresponding statistics for the INEX collection². At first glance, we already see that this statistics confirms the tf*idf heuristics: the higher P_{ml} and the smaller P_{avg} , the higher $P(R|t)$. Moreover, P_{avg} is dominating and P_{ml} has only a minor effect: For any given P_{avg} interval, the $P(R|t)$ values are roughly all in the same order of magnitude (ignoring the case where $P_{ml} = 0$), whereas for any P_{ml} interval the $P(R|t)$ values vary by several orders of magnitude. This observation contrasts with the standard justification of smoothing methods in language models, where it is said that P_{ml} is the dominating factor and P_{avg} is used only for dealing with data sparsity. The results also show that for $P_{ml} = 0$ (terms not occurring in the document), $P(R|t)$ is much smaller than for $P_{ml} > 0$. For higher values of P_{avg} , $P(R|t)$ seems to be zero. However, using a logarithmic scale, we can see that $P(R|t)$ decreases monotonically when P_{avg} increases.

The corresponding results for the TREC collection are shown in figure 3. The major difference to the TREC collection is that in TREC, the slope in the P_{avg} direction is not as high as in INEX. One possible explanation could be the fact that the relevance definition used in TREC is less strict than the INEX one. Furthermore, for terms not occurring in the document, there is only a minor $P(R|t)$ difference in comparison to those having low P_{ml} values.

Overall, these empirical observations confirm the dominant role of P_{avg} wrt. retrieval quality. This is in stark contrast to the standard language model justification, saying that P_{ml} is more important and P_{avg} only helps in smoothing.

4 Implementing empirical smoothing

Based on the observations described above, we now want to propose a new approach for smoothing, which we call empirical smoothing. The basic idea is already illustrated in figures 2-3: For each possible combination of (P_{ml}, P_{avg}) values of a term, these plots show the corresponding probability $P(R|t)$. So it seems straightforward to use these values as result of the smoothing process.

In principle, there are three different ways for implementing this idea:

Direct use of interval values: As outlined above, we can directly use the probability estimates of $P(R|t)$ from figures 2-3. Thus, given a (P_{ml}, P_{avg}) pair, we determine the corresponding 2-dimensional interval, and then look up its $P(R|t)$ value from the training set. However, this method needs large amounts of training data to avoid overfitting. Moreover, it does not give us any insights into the relationship between (P_{ml}, P_{avg}) and $P(R|t)$.

Application of probabilistic classification methods:

This approach has been investigated already in [Fuhr

²In order to derive a meaningful statistics, elements with less than 100 words were not considered here.

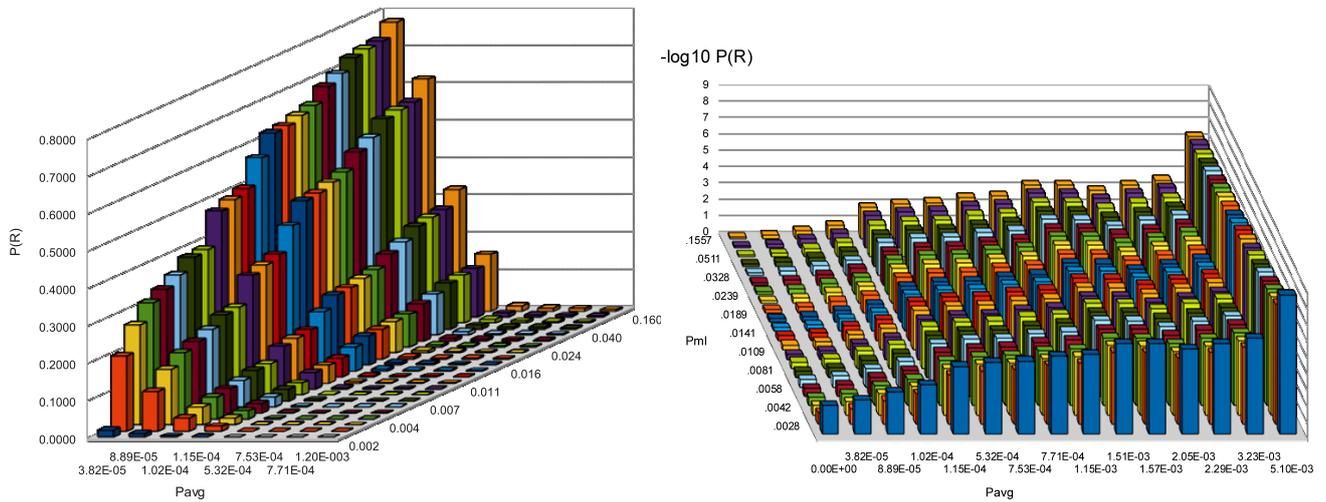


Figure 2: $P(R|t)$ for different (P_{ml}, P_{avg}) values (INEX), linear/log scale

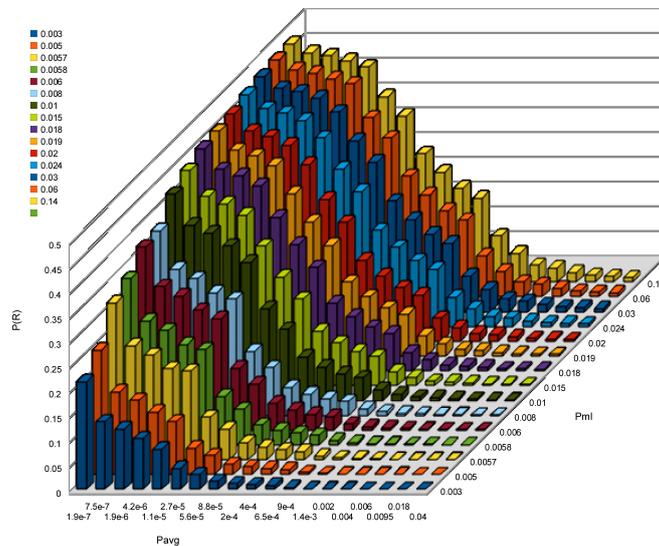


Figure 3: $P(R|t)$ for different (P_{ml}, P_{avg}) values (TREC)

and Buckley, 1991]. As input, the machine learning method would use the raw data underlying the figures from above, i.e., for each term in each query-document pair considered, we have a training instance consisting of the P_{ml} and P_{avg} values as features and the relevance decision as class variable. In recent years, this kind of approach has also become very popular for developing retrieval functions in the so called 'learning to rank' approaches (see e.g. [Fuhr, 1989; Liu, 2009]). Like the previous method, however, this approach operates like a black box, giving us no further insights.

Application of numeric prediction: Here we start with the data shown in figures 2 - 3, and now seek for a function that describes the relationship between P_{ml} , P_{avg} and $P(R|t)$. As classic smoothing functions perform the same kind of task, we can compare the outcome of the machine learning method with these functions.

From these three possibilities, we only consider the last one in the following. Furthermore, we only regard the most

simple variant of numeric prediction, namely linear regression.

5 Linear regression

First, we use a purely linear function of the form:

$$P_s(t_i|d) = \alpha P_{ml} + \beta P_{avg} + \gamma \quad (3)$$

As a second variant, we start from the observation in figure 2 that a linear function of P_{avg} may not be very appropriate. Therefore we use $\log(P_{avg})$ instead:

$$P_s(t_i|d) = \alpha P_{ml} + \beta \log(P_{avg}) + \gamma \quad (4)$$

Table 1 shows the actual coefficients which have been predicted using linear regression, along with the average squared error. As we can see, replacing P_{avg} by its logarithm (LR linear vs. LR log) reduces the error substantially for both collections.

For further analysis, we regard the difference between the linear predictions of equation 3 and the actual $P(R|t)$ values, as illustrated in figure 4 for the INEX collection (for TREC, the figure looks very similar). In the ideal case, there would be random errors; instead, these figures show

Table 1: Coefficients derived using linear regression

Method	Collection	α	β	δ	γ	Error
LR linear	INEX	0.97	-60.43		0.12	0.053
LR log	INEX	-9.12	-2		9.7	0.011
LR quadratic	INEX	0.97	-209.58	41064.69	0.18	0.022
LR linear cnst.=0	INEX	2.59	-23.4		0	0.060
LR linear	TREC	1.07	-6.93		0.13	0.091
LR log	TREC	-6.23	-0.5		3.43	0.012
LR quadratic	TREC	1.07	-28.03	660.81	0.16	0.041
LR linear cnst.=0	TREC	2.65	-2.69		0	0.094

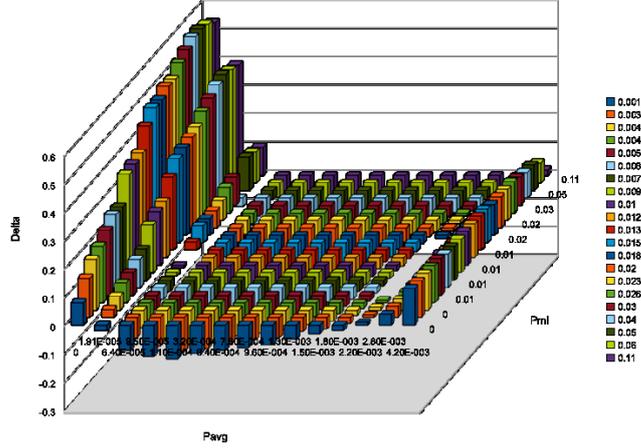


Figure 4: Residuals for linear regression (INEX)

us systematic deviations from the predicted values. The distribution of these errors suggests that a quadratic function of P_{avg} would be more appropriate:

$$P_s(t_i|d) = \alpha P_{ml} + \beta P_{avg} + \delta P_{avg}^2 + \gamma \quad (5)$$

Looking at the corresponding quadratic errors in table 1 (LR quadratic), we see that the quadratic form is better than the linear one, and rates as good as the variant with $\log(P_{avg})$.

Since JM smoothing also uses a linear function with P_{ml} and P_{avg} as inputs, we want to compare its outcome with that of our linear regression. For that, we used the equation 2 with $\lambda = 0.7$ which gave the best results for this method. For better comparison, we also tried a variant of the regression function 3, where we dropped the constant γ , so in this case it has the same structure as the JM smoothing function. However, looking at the corresponding regression coefficients listed in table 1 (LR linear cnst.=0), we see that P_{avg} has a negative coefficient, whereas JM smoothing assumes both coefficients to be positive. In JM smoothing, P_{ml} is the dominating factor (although P_{avg} has a higher weight with $\lambda = 0.7$, it is at least an order of magnitude smaller than P_{ml}), whereas the empirical data as well as the result of our regression put major emphasis on P_{avg} , and P_{ml} just serves as a minor correction factor.

6 Retrieval experiments

Finally, we performed retrieval experiments with the retrieval function 1 and the various regression functions and

compared them with standard retrieval functions. The results are depicted in table 2 and figure 5.

For the three variants of linear regression, we did not separate between training and test sample, so their results are a bit optimistic. Only for the purely linear form, we performed experiments with 2-fold cross validation (LR linear (cv)), showing that the choice of the training sample has little effect on the quality of results.

Comparing the results of the three variants of linear regression, we can see that for both collections, already the linear form gives good results, which can be improved by using one of the variants. For INEX, $\log(P_{avg})$ gives the best quality overall, whereas the quadratic form yields improvements for the top ranking elements only. With TREC, both the logarithmic and the quadratic form are much better than the linear one. In both cases, the quality of JM smoothing is comparable to that of the linear form. BM25 performs poorly for INEX, but very good for TREC.

Furthermore, we also present results for our odds-like language model presented in [Abdulmutalib and Fuhr, 2008], where the retrieval function is shown in eqn. (6); as estimate of $P(d)/P(\bar{d})$, we use the ratio of the length of d and the average document length, and ω and γ are tuning parameters for smoothing.

$$\rho_{o,e}(q, d) = \prod_{t_i \in q^T \cap d^T} \left(\frac{P_{ml}(t_i|d)}{P_{avg}(t_i|C)} \right)^\omega \cdot \prod_{t_i \in q^T - d^T} P_{avg}(t_i|C)^\gamma \cdot \frac{P(d)}{P(\bar{d})} \quad (6)$$

Table 2: Retrieval results: empirical smoothing vs. standard retrieval methods (INEX / TREC)

Method	MAP	P@5	P@10	P@20
LR linear	0.0729	0.355	0.339	0.334
LR log	0.1004	0.397	0.366	0.315
LR quadratic	0.0668	0.389	0.389	0.359
JM	0.0667	0.303	0.245	0.216
LR linear (cv)	0.0862	0.331	0.324	0.299
Odds	0.0800	0.348	0.348	0.323
ZL	0.0780	0.338	0.324	0.307
BM25	0.0063	0.096	0.087	0.070

Method	MAP	P@5	P@10	P@20
LR linear	0.0286	0.283	0.253	0.213
LR log	0.0633	0.359	0.312	0.273
LR quadratic	0.0654	0.304	0.247	0.222
JM	0.0307	0.214	0.238	0.231
LR linear (cv)	0.0355	0.345	0.339	0.333
Odds	0.0572	0.232	0.211	0.191
ZL	0.0611	0.279	0.233	0.228
BM25	0.0844	0.445	0.432	0.352

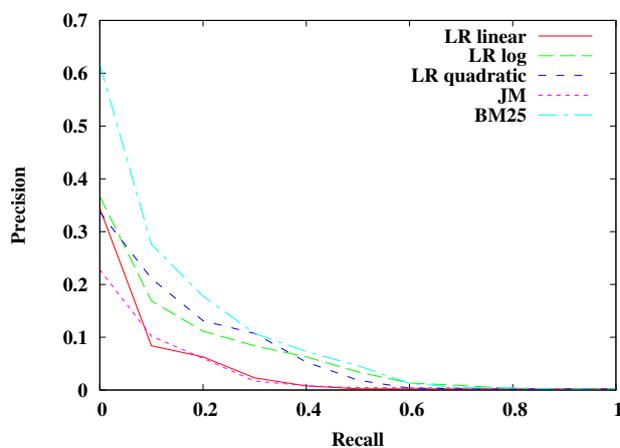
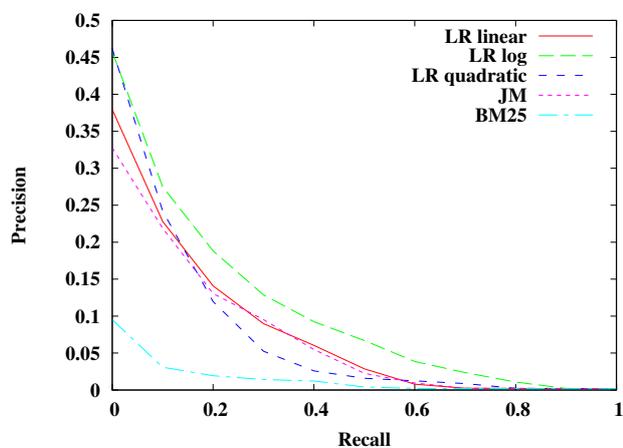


Figure 5: Recall-precision graphs for various smoothing methods and BM25 (INEX / TREC)

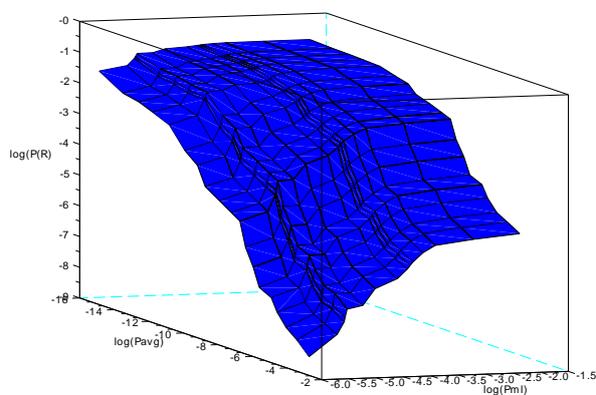
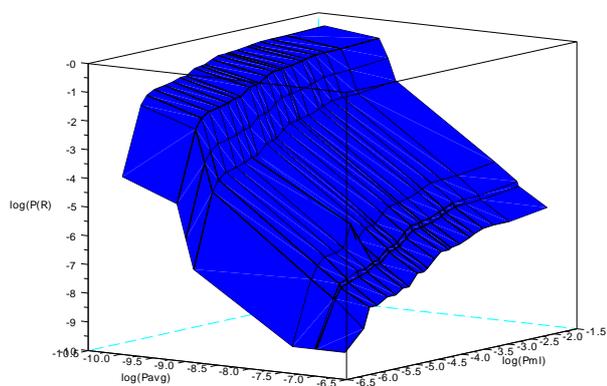


Figure 6: $\log(P(R|t))$ as function of $\log(P_{ml})$ and $\log(P_{avg})$ (INEX / TREC)

Overall, these results show that empirical smoothing in combination with nonlinear regression functions is superior to classic smoothing methods.

7 Further analysis

Figures 2–3 illustrating the relationship between P_{ml} , P_{avg} and $P(R|t)$ do not plot the first two dimensions in proportion to their size. In contrast, figure 6 uses a logarithmic scale for all three dimensions and also plots them proportionally. These figures indicate that the relationship could be fairly well described by a plane in the log-log-log space. In fact, looking at the odds-like-retrieval function (6), this is exactly the form that would result from such a plane (modulo the document length component). Based on this function, we also performed a few experiments with logistic regression, but the results were inferior to that of a grid search for the parameters ω and γ [Abdulmutalib, 2010, sec. 7.8].

8 Conclusion and Outlook

In this paper, we have investigated the relationship between smoothing in language models and idf weights. Although the relative collection frequency P_{avg} and idf weights are very similar, there is a contradiction in the weighting strategy. Regarding the correlation between the language model parameters and relevance, we find that the idf type of weighting seems to be more appropriate. Based on the observed correlation, we have devised empirical smoothing as a new type of term weighting for language models, and retrieval experiments confirm the general applicability of our method. Finally, we showed that the most appropriate form of describing the relationship between the language model parameters and relevance seems to be a product form, which confirms a language model proposed by us before.

In this paper, we have not considered the influence of document length. In fact, other smoothing methods like Dirichlet smoothing or absolute discount (see e.g. [Lafferty and Zhai, 2001]) consider this parameter. Thus, empirical smoothing could also be extended to document length as third parameter.

The comparison between theoretic models and empirical data in this paper has brought us interesting observations. However, this comparison does not answer the question why JM smoothing gives fairly reasonable retrieval results, its structure contradicts our empirical findings. A reasonable explanation for this effect remains the subject of further research.

References

- [Abdulmutalib and Fuhr, 2008] Najeeb Abdulmutalib and Norbert Fuhr. Language models and smoothing methods for collections with large variation in document length. In A. M. Tjoa and R. R. Wagner, editors, *DEXA Workshops*, pages 9–14. IEEE Computer Society, 2008.
- [Abdulmutalib, 2010] Najeeb Abdulmutalib. Language models and smoothing methods for information retrieval. PhD thesis (submitted), 2010.
- [Berger and Lafferty, 1999] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pages 222–229, New York, 1999. ACM.
- [Croft and Harper, 1979] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295, 1979.
- [Fuhr and Buckley, 1991] Norbert Fuhr and Chris Buckley. A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems*, 9(3):223–248, 1991.
- [Fuhr, 1989] N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.
- [Hiemstra, 1998] Djoerd Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Lecture Notes In Computer Science - Research and Advanced Technology for Digital Libraries - Proceedings of the second European Conference on Research and Advanced Technology for Digital Libraries: ECDL'98*, pages 569–584. Springer Verlag, 1998.
- [Jelinek and Mercer, 1980] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, 1980.
- [Lafferty and Zhai, 2001] J. Lafferty and C. Zhai. Probabilistic ir models based on document and query generation. In B. Croft J. Callan and J. Lafferty, editors, *Proceedings of workshop on Language Modeling and Information Retrieval*, 2001.
- [Liu, 2009] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, 2009.
- [Ponte and Croft, 1998] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR'98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM Press.
- [Roelleke and Wang, 2008] Thomas Roelleke and Jun Wang. Tf-idf uncovered: a study of theories and probabilities. In *SIGIR*, pages 435–442, 2008.
- [Zhai and Lafferty, 2001] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In W. B. Croft, D. Harper, D. H. Kraft, and J. Zobel, editors, *Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval*, New York, 2001. ACM.