Language Models and Smoothing Methods for Collections with Large Variation in Document Length

Najeeb Abdulmutalib and Norbert Fuhr
University of Duisburg-Essen
Information Systems, Department of Computational
and Cognitive Sciences - Duisburg, Germany
najeeb@is.inf.uni-due.de , norbert.fuhr@uni-due.de

Abstract

In this paper we present a new language model based on an odds formula, which explicitly incorporates document length as a parameter. Furthermore, a new smoothing method called exponential smoothing is introduced, which can be combined with most language models. We present experimental results for various language models and smoothing methods on a collection with large document length variation, and show that our new methods compare favorably with the best approaches known so far.

KEYWORDS

Information retrieval,Smoothing methods

1 Introduction

Since the first language models for information retrieval were presented [8], [3], [5], a large variety of models of this kind have been proposed. However, with the exception of [4], little attention has been paid to the influence of document length, and only a few approaches have considered this parameter explicitly.

In the next section, we present a new language model which includes document length as a genuine parameter. We start along the lines of the classic Zhai/Lafferty [9] model and present a probability and an odds model as variations of this basic model. Section 3 introduces a new smoothing method for combining the relative term frequencies in the current document and the whole collection into a single probability estimate. As alternative to this smoothing method, we also consider the classic smoothing methods regarded by Zhai and Lafferty, and then present experimental results for the INEX collection in Section 5, where we regard each XML element as a document, thus having a collection with a large variation of document lengths.

2 Models

2.1 Basic model

We first present the basic probabilistic language model, which forms the basis for Zhai and Lafferty's model.

Let q denote a query containing the set of terms q^T, and d is a document with the set of terms d^T. Furthermore, let t_i denote a term and C stands for the collection. Then we compute the conditional probability of observing the query q given the document d as follows:

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

Here we have the following probabilities:

$P(d)$ Probability that d implies an arbitrary query

$P_s(t_i|d)$ Probability that d implies term t_i, given that t_i occurs in d

$P_u(t_i|d)$ Probability that d implies term t_i, given that t_i does not occur in d

2.2 Zhai/Lafferty model

Building on the basic model (1) the core idea of Zhai and Lafferty is the estimation of the probability $P_u(t_i|d)$ of terms not occurring in the document, by means of the following formula: $P_u(t_i|d) = a_d P(t_i|C)$. Here a_d is a document-dependent constant estimated in the following way:

$$a_d = \frac{1 - \sum_{t_i \in q^T \cap d^T} P_s(t_i|d)}{1 - \sum_{t_i \in q^T \cap d^T} P(t_i|C)}$$

Regarding the logarithmic form of (1) their retrieval function yields:

$$\log P(q|d) = \sum_{t_i \in q^T \cap d^T} \log \frac{P_s(t_i|d)}{a_d \cdot P(t_i|C)} + n \log a_d + \sum_{t_i \in q^T} \log P(t_i|C)$$

where n is the length of the query.

2.3 An Odds model

In [1], we have shown that language models can be interpreted in terms of uncertain inference, and that Hiemstra's model [3] regards the probability $P(q \rightarrow d) = P(d|q)$ of the query implying the document. In contrast, the basic probability model from above focuses on the implication in the reverse direction, i.e. $P(d \rightarrow q) = P(q|d)$. In order to develop a language model formula with explicit consideration of document length, we start from Hiemstra's approach, but consider odds instead of probabilities. Thus, we regard the probability that the query implies the document, which we divide by the probability $P(q \rightarrow \bar{d}) = P(\bar{d}|q)$ that the query implies an arbitrary document $\bar{d}$ different from d. Applying Bayes' theorem and the standard independence assumptions, we get:

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

In addition to the parameters defined for the probability model, we have the following probabilities here:

$P(\bar{d})$ Probability that an arbitrary document $\neq d$ implies an arbitrary query.

$P_s(t_i|\bar{d})$ Probability that an arbitrary document $\neq d$ implies term t_i, given that t_i occurs in that document.

$P_u(t_i|\bar{d})$ Probability that an arbitrary document $\neq d$ implies term t_i, given that t_i does not occur in that document.

3 Smoothing methods

The major problem in the application of language models is the estimation of the probability $P_s(t_i|d)$. For this purpose, various methods have been proposed in the past. In principle, all methods aim at combing the tf weight of a term (its within-document-frequency) with its idf weight. Thus, as input for the estimation step, all known methods consider (at least) the following two parameters:

$P_{ML}(t_i|d)$ Maximum likelihood-estimate of $P(t_i|d)$, i.e. the relative frequency of t_i in d

$P_{avg}(t_i|C)$ The average probability of observing t_i in the collection C, i.e. the relative frequency of t_i in the text of the whole collection

In the following, we first describe three popular smoothing methods, and then we present our own new method.

3.1 The Jelinek-Mercer method

This method involves a linear interpolation of the maximum likelihood model with the collection model, using a smoothing coefficient λ to control the influence of collection model. The resulting probability estimate is called $P_{s,\lambda}$ here:

$$P_{s,\lambda}(t_i|d) = (1 - \lambda) \cdot P_{ML}(t_i|d) + \lambda \cdot P_{avg}(t_i|C)$$

3.2 Bayesian parameter estimation

Typical smoothing methods in language models are length-independent. On the other hand, it is obvious, that the maximum likelihood estimate is more biased for shorter documents. When the documents in the collection are of almost uniform length (which e.g. is the case for the largest part of the TREC collections), this effect can be compensated by document-independent smoothing parameters. However, in a collection with a big variation in document lengths, a document-dependent smoothing factor may be more adequate. One possible approach following this strategy is Bayesian parameter estimation. Since a language model is a multinomial distribution, the corresponding conjugate prior is the Dirichlet distribution with parameters

$$(\mu P(t_1|C), \mu P(t_2|C), ..., \mu P(t_n|C))$$

and the estimate of $P_u(t_i, d)$ is given as

$$P_{s,\mu}(t_i|d) = \frac{c(t_i; d) + \mu P(t_i|C)}{\sum_{t_i \in d^T} c(t_i; d) + \mu}$$

where $c(t_i; d)$ is the number of occurrences of t_i in d.

3.3 Absolute discount

This method is similar to Jelinek-Mercer, but differs in that it discounts the seen word probability by subtracting a constant instead of multiplying it by $(1 - \lambda)$.

So the estimate of $P_u(t_i, d)$ is given as

$$P_{s,\delta}(t_i|d) = \frac{max(c(t_i;d) - \delta, 0)}{\sum_{t_i \in d^T} c(t_i;d)} + \sigma P(t_i|C)$$

where

δ is a discounting constant,

$\sigma = \frac{\delta \cdot |d^T|}{|d|}$, with

$|d|$ denoting the document length, and

$|d^T|$ is the number of unique terms in document d

3.4 Exponential smoothing

As a new, alternative way of smoothing, we propose an exponential formula for combining $P_{ML}(t_i|d)$ and $P_{avg}(t_i|C)$ as an estimate of $P_s(t_i, d)$. In a similar way, we estimate $P_u(t_i|d)$ as a function of $P_{avg}(t_i|C)$. More precisely, our estimates are:

$$\begin{aligned} P_{s,e}(t_i|d) &= P_{ML}(t_i|d)^{\alpha_d} \cdot P_{avg}(t_i|C)^{1-\alpha_d} \\ P_{u,e}(t_i|d) &= P_{avg}(t_i|C)^{\beta_d} \end{aligned}$$

Here α_d and β_d are (possibly document-dependent) smoothing factors.

In the same way, we estimate $P_{s,e}(t_i|\bar{d})$ and $P_{u,e}(t_i|\bar{d})$ with different smoothing factors.

$$\begin{aligned} P_{s,e}(t_i|\bar{d}) &= P_{ML}(t_i|d)^{\mu_d} \cdot P_{avg}(t_i|C)^{1-\mu_d} \\ P_{u,e}(t_i|\bar{d}) &= P_{avg}(t_i|C)^{\delta_d} \end{aligned}$$

$$\frac{P_{u,e}(t_i|d)}{P_{u,e}(t_i|\bar{d})} = P_{avg}(t_i|C)^{\beta_d - \delta_d} = P_{avg}(t_i|C)^{\gamma_d}$$

$$\begin{aligned} \frac{P_{s,e}(t_i|d)}{P_{s,e}(t_i|\bar{d})} &= P_{ML}(t_i|d)^{\alpha_d - \mu_d} \cdot P_{avg}(t_i|C)^{-\alpha_d + \mu_d} \\ &= P_{ML}(t_i|d)^{\omega_d} \cdot P_{avg}(t_i|C)^{-\omega_d} \end{aligned}$$

Applying exponential smoothing to our odds model, we get the retrieval function

$$\begin{aligned} \rho_{o,e} &= \prod_{t_i \in q^T \cap d^T} P_{ML}(t_i|d)^{\omega_d} \cdot P_{avg}(t_i|C)^{-\omega_d} \\ &\quad \cdot \prod_{t_i \in q^T - d^T} P_{avg}(t_i|C)^{\gamma_d} \cdot \frac{P(d)}{P(\bar{d})} \end{aligned}$$

$$= \prod_{t_i \in q^T \cap d^T} \left(\frac{P_{ML}(t_i|d)}{P_{avg}(t_i|C)} \right)^{\omega_d} \cdot \prod_{t_i \in q^T - d^T} P_{avg}(t_i|C)^{\gamma_d} \cdot \frac{P(d)}{P(\bar{d})}$$

Here we have the additional parameters $P(d)$ and $P(\bar{d})$. The former denotes the probability that document d implies a random query, while the latter denotes the same probability for an arbitrary document different from d. As a first approximation, we assume that these probabilities are proportional to document length, which we use as estimates in the experiments described below. In a similar way, the probability model with exponential smoothing yields

$$\rho_{p,e} = \prod_{t_i \in q^T \cap d^T} \frac{P_{ML}(t_i|d)^{\alpha_d}}{P_{avg}(t_i|C)^{\beta_d + \alpha_d - 1}} \prod_{t_i \in q^T} P_{avg}(t_i|C)^{\beta_d}$$

Since the second factor is independent of the specific document, we can also ignore it when we are only interested in the ranking of the documents.

Losada and Azzopardi [4] studied different Language Modeling smoothing strategies from a document length retrieval perspective and showed that the document length retrieval pattern is of major importance in language modeling for information retrieval. In some initial experiments, we also noticed that document length plays an important role and significantly improves the retrieval quality. For this reason, we decided to regard a variant of the probability model which incorporates document length, thus leading to the retrieval function

$$\rho^d_{p,e} = \prod_{t_i \in q^T \cap d^T} \frac{P_{ML}(t_i|d)^{\alpha_d}}{P_{avg}(t_i|C)^{\beta_d + \alpha_d - 1}} \cdot \frac{p(d)}{p(\bar{d})}$$

4 Test collection and evaluation metrics

For our experiments, we used the INEX 2005 IEEE collection [2], version 1.9. This collection consists originally of 16,819 journal articles in XML format, comprising 764 MB of data. For our experiments, we regarded each XML element as an independent document XML documents, thus leading to a collection of 21.6 million documents with a collection size of more than 253 million words. Figure 1 shows the distribution of document lengths in our test collection.

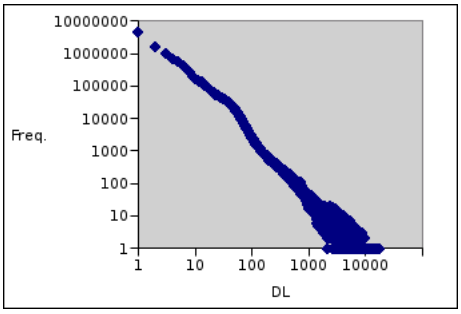

Figure 1: Distribution of document length in our test collection

Here document length ranges from 1 to 17784. We obviously have a linear relationship between the logarithms of document lengths and frequency. This is certainly a kind of document length distribution which can only be found in the special setting we are regarding here, namely retrieval of XML elements. On the other hand, this situation also serves as a good test case for investigating the influence of document length variation on the retrieval quality of language models. In principle, our approach can be regarded as the first step of an XML retrieval engine, where initially the most relevant answer elements are determined, and then the structural relationships between these elements (i.e., two or more elements from the same document, or even one answer element containing another one) are addressed in the second step of the retrieval process. However, since the focus of this paper is on language models and document length, we consider the first step of this process only.

As search requests, we used the so-called CO queries from INEX, which are free text queries. For our experiments, we considered the 29 queries from INEX 2005 (version 003) along with the official adhoc 2005-assessments-v7.0. These assessments judged relevance with respect to two dimensions, namely specificity and exhaustivity [6]. Here we regard the exhaustivity dimension only, since specificity targets at the most specific answer element in a document (which should be addressed in the second step of the retrieval process sketched above). For grading exhaustivity, relevance assessors had to chose from 3 + 1 levels: highly exhaustive (e = 2), somewhat exhaustive (e = 1), not exhaustive (e = 0) and too small (e =?). Given the graded relevance (exhaustivity) scale, we measured retrieval quality with the EPRUM (Expected Precision Recall with User Model) metrics which was developed within INEX [7] [1] . This metric is based on a more realistic user model which encompasses a large variety of user behaviours. It supposes a set of ideal results. Recall is defined as the ratio of the number of retrieved ideal elements to the number of relevant

[1] http://inex.is.informatik.uni-duisburg.de/2005/Metrics.html

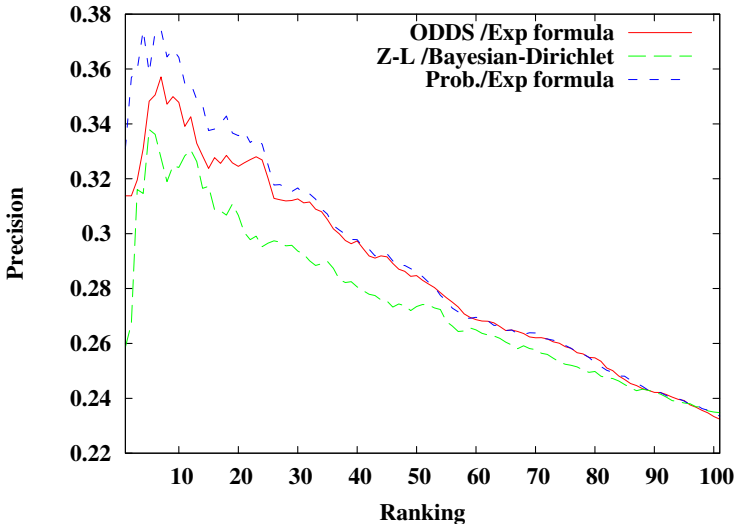

Figure 2: Precision-rank curve for the best runs

elements. The ideal run is defined as the run that maximizes the recall for each rank. Precision is defined as the ratio of the number of the length of an ideal run for achieving the same level of recall to the size of the retrieved list. The two definitions are generalization of precision and recall in the standard case.

5 Experimental results

First, we regarded the effect of considering document length. For the Odds and the Probability model, Tables 1 and 2 show the best results with and without considering document length (using exponential smoothing); for the Odds model, the factor $\frac{p(d)}{p(\bar{d})}$ was omitted from the retrieval formula for $\rho_{o,e}$ when document length was ignored; in the case of the Probability model, the functions for $\rho^d_{p,e}$ and $\rho_{p,e}$ were compared. The experimental results show huge performance differences for both kinds of models. So document length is an important factor for achieving good retrieval results when dealing with collections of varying document size.

In a second series of experiments, we investigated the effect of exponential smoothing on the performance of the Odds and the Probability model. For this purpose, we varied the values of the smoothing parameters between 0 and 1 and performed a large number of runs. The MAP values of these experiments are shown in Tables 3 and 4, Our results indicate that the retrieval performance is very sensitive to the values of the smoothing parameters.
For the Probability model, the best results were achieved when β approaches 1 and α takes values between 0.4 and 0.8.

Table 1: Best results for Odds model ($\gamma = 0.2$) with and without using document length

Omega	Normal model using DL	Ignoring DL
0	0.006	0.016
0.1	0.014	0.002
0.2	0.038	0.002
0.3	0.064	0.002
0.4	0.078	0.002
0.5	0.080	0.002
0.6	0.076	0.002
0.8	0.068	0.002
0.9	0.063	0.002

Table 2: Best results for Prob. model ($\beta = 1$) with and without using document length

Alpha	Normal model using DL	Ignoring DL
0	0.006	0.018
0.1	0.020	0.027
0.3	0.059	0.027
0.4	0.076	0.027
0.5	0.079	0.027
0.6	0.076	0.027
0.8	0.070	0.027
0.9	0.063	0.027

For the Odds model, the retrieval performance was the highest for $\gamma = 0.2$ and ω between 0.4 and 0.6.

Finally, we compared the best results of our new models and smoothing method with those of the Zhai/Lafferty model and known smoothing methods.

The results depicted in Table 5 indicate that the Probability and the Odds model yield their best results when combined with the exponential smoothing, and they even outperform the Zhai/Lafferty model. For the latter, the best results were achieved in combination with Bayesian Dirichlet smoothing. We think that this outcome is due to the fact that Bayesian Dirichlet is the only smoothing method which explicitly considers document length. In contrast, other smoothing methods lead to very poor performance figures for the Zhai/Lafferty model. So this model should only be used in combination with Bayesian Dirichlet smoothing when being applied to collections with varying document size.

The results of the three best combinations (Probability and Odds model with exponential smoothing, Zhai/Lafferty with Bayesian Dirichlet) are also illustrated in the precision-rank curve shown in Figure 2.

Table 3: Influence of Alpha and Beta parameters on MAP when using Prob. model

Alpha	Beta				
	0.1	0.2	0.5	0.9	1
0.1	0.003	0.003	0.003	0.004	0.020
0.3	0.003	0.003	0.003	0.013	0.059
0.4	0.003	0.003	0.003	0.037	**0.076**
0.5	0.003	0.003	0.003	0.057	**0.079**
0.6	0.003	0.003	0.003	0.063	**0.076**
0.8	0.003	0.003	0.007	0.061	0.070
0.9	0.003	0.003	0.014	0.060	0.063

Table 4: Influence of Omega and Gamma parameters on MAP when using Odds model

Omega	Gamma						
	0	0.1	0.2	0.3	0.5	0.8	0.9
0	0.005	0.003	0.006	0.006	0.008	0.012	0.013
0.1	0.011	0.013	0.014				
0.2	0.033	0.036	0.038				
0.3	0.060	0.062	0.064				
0.4	0.076	0.077	**0.078**		0.066		0.052
0.5	0.079	0.079	**0.080**	0.080		0.060	
0.6	0.076	0.076	**0.076**				
0.8	0.068	0.068	0.068		0.061		0.052
0.9	0.063	0.063	0.063	0.062	0.060	0.052	0.059

Table 5: Best results for models and smoothing methods: Prob. model ($\alpha = 0.5, \beta = 1$), odds model ($\omega = 0.5, \gamma = 0.2$) and the ZL model ($\mu = 2,000$)

Smoothing	Model	MAP	Prec at 5	Prec at 10	Prec at 20
Exponential	Odds	**0.080**	**0.348**	**0.348**	**0.323**
	Prob.	**0.079**	**0.359**	**0.364**	**0.336**
	Zhai/Lafferty	0.004	0.015	0.013	0.016
Jelink Mercer	Odds	0.047	0.180	0.180	0.170
	Prob.	0.051	0.180	0.180	0.170
	Zhai/lafferty	0.040	0.180	0.150	0.140
Baysian Dirichlet	Odds	0.041	0.235	0.235	0.235
	Prob.	**0.063**	**0.300**	**0.290**	**0.290**
	Zhai/Lafferty	**0.078**	**0.338**	**0.324**	**0.307**
Absolute-discount	Odds	0.002	0.009	0.006	0.004
	Prob.	0.002	0.012	0.009	0.007
	Zhai/Lafferty	0.004	0.006	0.006	0.006

6 Conclusion and Outlook

In this paper, we presented a new language model based on an odds formula, as well as a new smoothing method called exponential smoothing. Experiments performed on a collection with large variations in document length showed that document length is an important factor for language models, so models ignoring this parameter lead to very poor results. Our new model along with the new smoothing method give very good results. With variants of the document length parameter and (possibly) document-specific smoothing, there are still possibilities for further improvement.

References

[1] N. Fuhr. Models in information retrieval. In M. Agosti, F. Crestani, and G. Pasi, editors, *Lectures in Information Retrieval*, pages 21–50. Springer, Heidelberg et al., 2001.

[2] N. Fuhr, M. Lalmas, S. Malik, and G. Kazai, editors. *Advances in XML Information Retrieval and Evaluation, 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2005, Dagstuhl Castle, Germany, November 28-30, 2005, Revised Selected Papers*, volume 3977 of *LNCS*. Springer, 2006.

[3] D. 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.

[4] D. E. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11(2):109–138, 2008.

[5] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden markov model information retrieval system. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pages 214–221, New York, 1999. ACM.

[6] P. Ogilvie and M. Lalmas. Investigating the exhaustivity dimension in content-oriented xml element retrieval evaluation. In *Proceedings of ACM CIKM*, New York, 2006. ACM.

[7] B. Piwowarski. Eprum metrics and inex 2005. In Fuhr et al. [2], pages 30–42.

[8] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, New York, 1998. ACM.

[9] C. Zhai and J. 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.