Unsupervised Estimation of Dirichlet Smoothing Parameters

Jangwon Seo
jangwon@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

ABSTRACT

A standard approach for determining a Dirichlet smoothing parameter is to choose a value which maximizes a retrieval performance metric using training data consisting of queries and relevance judgments. There are, however, situations where training data does not exist or the queries and relevance judgments do not reflect typical user information needs for the application. We propose an unsupervised approach for estimating a Dirichlet smoothing parameter based on collection statistics. We show empirically that this approach can suggest a plausible Dirichlet smoothing parameter value in cases where relevance judgments cannot be used.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval Models

General Terms: Algorithms, Measurement, Experimentation

Keywords: Dirichlet smoothing, unsupervised approach, parameter estimation

1. INTRODUCTION

Dirichlet smoothing is known to be one of the most effective smoothing techniques for the language modeling-based retrieval framework [5]. This smoothing technique has a free parameter, i.e. the Dirichlet smoothing parameter. A standard approach for determining this parameter is to choose a value which maximizes a retrieval performance metric using relevance judgments. We call this supervised approach metric-based estimation of Dirichlet smoothing parameters.

We do not, however, always have relevance judgments as given by TREC standard test collections. For example, we may use new document collections where there are no relevance judgments. Even when we have relevance judgments for a collection, we may be addressing different search tasks from those for which relevance judgments are made. Furthermore, the characteristics of actual user queries can be different from the queries associated with relevance judgments used for training the smoothing parameter. For example, if most queries used in relevance judgments are long, while real queries are short, then the trained value may not work well because the smoothing parameter is sensitive to query lengths as well as document lengths [2]. In such cases, we cannot use metric-based estimation.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

To tackle these situations, we propose an unsupervised estimation approach. This method estimates a Dirichlet smoothing parameter from collection statistics, specifically, a variance of multinomial parameters associated with each term. Therefore, this estimation is independent of specific queries or relevance judgments. Note that if a test collection with relevance judgments is available, we cannot say that our unsupervised approach can produce a better smoothing parameter than the supervised approach. In this work, we intend to introduce an estimation technique which can be used when the supervised approach cannot be used.

There are few formal studies for determining Dirichlet smoothing parameters for retrieval models in an unsupervised manner. However, the average document length of a collection is sometimes used as the parameter value [1, 6, 4]. Also, in the Machine Learning literature, Minka [3] has presented maximum likelihood estimation for Dirichlet distributions.

2. UNSUPERVISED ESTIMATION

Dirichlet smoothing assumes that a document can be represented by a multinomial distribution, Multi$(\theta_1, \theta_2, \cdots, \theta_N)$, where N is the size of vocabulary of collection C. Introducing a Dirichlet prior, Dir$(\alpha_1, \cdots, \alpha_N)$, we choose the mean of the posterior distribution as a smoothed document representation given by $p(i|D) = (tf_{i,D} + \alpha_i)/(|D| + \alpha_0)$, where D is a document, i is an index corresponding to a unique term, and $\alpha_0 = \sum_j \alpha_j$. A typical choice for α's is $\alpha_i = \mu \cdot m_i$, where $m_i = cf_i/|C|$. Then, the mean of the Dirichlet prior, $\mathrm{E}[\theta_i] = \alpha_i / \sum_j \alpha_j = m_i$, is independent of μ. On the other hand, the variance of the Dirichlet prior, $\mathrm{Var}[\theta_i] = [\alpha_i(\alpha_0 - \alpha_i)]/[\alpha_0^2(\alpha_0 + 1)] = m_i(1 - m_i)/(\mu + 1)$, is closely related to the choice of μ. Therefore, the variance can be parameterized by μ.

Assuming that a smoothing parameter should reflect collection statistics well, we choose μ which minimizes the following squared error of variances.

$$e(\mu) = \sum_i \left(\frac{\bar{V}_i - \mathrm{Var}[\theta_i]}{\mathrm{Var}[\theta_i]} \right)^2 = \sum_i \left(\frac{\bar{V}_i(\mu + 1)}{m_i(1 - m_i)} - 1 \right)^2$$

where $\bar{V}_i$ is the sample variance.

Via $\frac{de(\mu)}{d\mu} = 0$, a closed form solution is obtained by

$$\mu = \left(\sum_i \frac{\bar{V}_i}{m_i(1 - m_i)} \right) / \left(\sum_i \frac{\bar{V}_i^2}{m_i^2(1 - m_i)^2} \right) \qquad (1)$$

$\bar{V}_i$ can be computed by $\sum_{D \in C} (p_{ML}(i|D) - m_i)^2$, where

	AP	WSJ	GOV2
Avg.#terms of short queries	2.5	2.5	2.4
Avg.#terms of long queries	5.1	5.1	3.8
μ_{short}	4000	2300	3700
μ_{long}	1900	1200	800
μ_{avgdl}	464	449	949
μ_{est}	2560	1563	1011

Table 1: Average query lengths of split topic sets and four Dirichlet smoothing parameters. μ_{short} and μ_{long} are parameters trained for short queries and long queries, respectively. μ_{avgdl} is the average document length. μ_{est} is estimated by our proposed method.

	AP		WSJ		GOV2	
	Short	Long	Short	Long	Short	Long
μ_{short}	0.1359	0.1097	0.2255	0.1840	0.1532	0.1367
μ_{long}	0.1344	0.1114	0.2206	0.1853	0.1456	0.1479
μ_{avdl}	0.1304	0.1030	0.2107	0.1769	0.1466	0.1479
μ_{est}	0.1344	0.1109	0.2235	0.1847	0.1477	0.1477

Table 2: Retrieval results for short queries and long queries according to different Dirichlet smoothing parameters. A number is a MAP score.

$p_{ML}(i|D)$ is the maximum likelihood estimator of a language model, i.e. $tf_{i,D}/|D|$. However, since computations crossing all terms and all documents are required, this is practically infeasible in case of large collections. Therefore, we use a sampling and approximation approach. First, we randomly sample T terms from a collection and consider only these terms instead of all terms in vocabulary. Then, we exploit the fact that each term occurs very sparsely in documents. That is, in many cases, $tf_{i,D} = 0$. Accordingly, we consider an approximation, $\bar{V}_i \approx m_i^2$. Using this approach, Equation (1) can be easily computed. We call this unsupervised approach variance-based estimation of Dirichlet smoothing parameters.

3. EXPERIMENTS

We conducted experiments to evaluate our unsupervised estimation method. We used three standard TREC collections: AP (topic 51-150), WSJ (topic 51-150) and GOV2 (topic 701-800). Each document is stemmed by the Krovetz stemmer and stopped by a standard stopword list. To simulate situations where the characteristics of training queries are different from those of test queries, we split the topics into two subsets with the same size according to the number of terms in the topic titles, i.e. short queries and long queries.

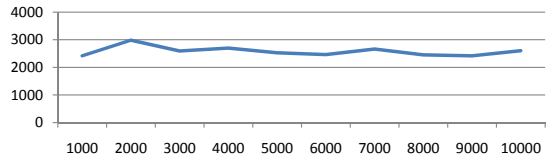

Figure 1: Estimated Dirichlet smoothing parameters (y-axis) according to the numbers of sample terms (x-axis) on the AP collection.

For each collection, we considered four Dirichlet smoothing parameters. Two of them are values which maximize mean average precision (MAP) for short queries and long queries, respectively. To find the values, we swept [500, 4000] with stepsize 100. Another is the average document length of each collection that is often used as an unsupervised heuristic for Dirichlet smoothing parameters. The last one is a value computed by our proposed method (with $T = 3000$). Table 1 shows these values. As you see, even though relevance judgments are built on the same collection, there is a substantial divergence between the Dirichlet smoothing parameters trained for different types of queries. While the average document length does not appear close to the trained values, a parameter estimated by our unsupervised approach appears between two trained values. That is, this method seems to produce a plausible value.

We evaluated retrieval performance of these smoothing parameters for short queries and long queries. Table 2 shows the results. The average document length produces consistently poor performance. Also, parameters trained with a specific type of query (μ_{short} and μ_{long}) do not generalize well to different types of queries. This shows that when making relevance judgments, accurate prediction of the characteristics of actual user queries is necessary so that the supervised approach is effective. On the other hand, parameters estimated by our unsupervised method, while not the best, do produce reasonable (i.e., the second best) performance regardless of the type of query for all collections.

To see how our method depends on the number of sample terms T, we tried various T's as shown in Figure 1. This shows that the Dirichlet smoothing parameter value appears stable after $T = 3000$. That is, the dependence on T is not substantial when a sufficient number of terms are used.

4. CONCLUSIONS

We proposed an unsupervised estimation approach for determining Dirichlet smoothing parameters. This method was shown empirically to be able to produce a plausible parameter. Furthermore, this method is relatively stable and robust in that it is independent of the characteristics of queries and relevance judgments. Therefore, it can be applied to cases that relevance judgments cannot be used or are not applicable.

Acknowledgments: This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-0534383. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] H. Fang and C. Zhai. An exploration of axiomatic approaches to information retrieval. In *SIGIR '05*, 2005.
[2] D. E. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11(2), 2008.
[3] T. Minka. Estimating a Dirichlet distribution.
[4] D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *ICTAI '06*, 2006.
[5] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *SIGIR '01*, 2001.
[6] J. Zheng and Z. Nie. Language models for web object retrieval. In *NISS '09*, 2009.