Formal Multiple-Bernoulli Models for Language Modeling

Donald Metzler
metzler@cs.umass.edu

Victor Lavrenko
lavrenko@cs.umass.edu

W. Bruce Croft
croft@cs.umass.edu

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

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Theory

Keywords: Information retrieval, language modeling.

1. INTRODUCTION

Statistical language modeling allows formal methods to be applied to information retrieval. As a result, such methods are preferred over their heuristic *tf.idf*-based counterparts. In language modeling, a statistical model is estimated for each document in the corpus. Documents are then scored by the likelihood the query was generated by the document's model. Typically, the underlying model is assumed to be of a specific parametric form. In the past, a number of different assumptions have been made about this distribution. In [1], documents were modeled by a multiple-Bernoulli distribution. However, the estimation and smoothing techniques used to estimate the model were non-standard and somewhat heuristic. The predominant modeling assumption used today, as described in [2], is to model documents by a multinomial distribution. Such models may be smoothed in a number of ways [4]. Among these is Bayesian (Dirichlet) smoothing that takes a formal, Bayesian approach to smoothing by assuming a Dirichlet prior over the document model. Unlike Ponte and Croft's multiple-Bernoulli estimation techniques, the multinomial assumption combined with Bayesian smoothing results in a completely formal statistical model. In this paper, we revisit the multiple-Bernoulli assumption and formalize it by taking a Bayesian approach to estimating smoothed document models.

2. MULTINOMIAL MODEL

Here we provide a quick review of multinomial language models. Assume that the underlying document model takes the form of a multinomial distribution over words. Documents and queries are then modeled as finite sequences of words. Given a document $D = w_1 w_2 \ldots w_{|D|}$, where each w_i is a word, we wish to compute the *maximum a priori* (MAP) distribution and use it as the document's language model ($\hat{\theta}$). Here,

$$\hat{\theta}_D = \arg\max_{\theta_D} P(\theta_D|D) = \arg\max_{\theta_D} P(D|\theta_D)P(\theta_D)$$

where $P(D|\theta)$ is the likelihood of the document given model θ and $P(\theta)$ is the prior on the model. When θ parameterizes a multinomial and the model prior is Dirichlet, the conjugate prior for the multinomial, we get:

$$\hat{\theta}_D = \arg\max_{\theta_D} \frac{\Gamma(|D| + \sum_{i=1}^{|V|} \alpha_i)}{\prod_{i=1}^{|V|} \Gamma(tf_{w_i,D} + \alpha_i)} \prod_{i=1}^{|V|} \theta_i^{tf_{w_i,D}+\alpha_i-1} \quad (1)$$

where Γ is the Gamma function, $\theta_i = P(w_i|\theta_D)$, $tf_{w_i,D}$ is the number of times word w occurs in document D, $|V|$ is the size of the vocabulary (number of unique words occurring in the corpus), and the α_i's are the parameters of the Dirichlet prior.

The solution to Equation 1 yields the following form of probability estimates:

$$\hat{\theta}_i = P(w_i|\hat{\theta}_D) = \frac{tf_{w_i,D} + \alpha_i - 1}{|D| + \sum_{i=1}^{|V|} \alpha_i - |V|}$$

As was shown in [3], setting $a_i = 1$ results in the maximum likelihood estimate, $a_i = 2$ gives Laplace smoothing, and $a_i = \mu \frac{cf_i}{|C|} + 1$ yields the popular Dirichlet smoothing. Here, μ is the smoothing parameter, cf_i is the number of times word w_i appears in the collection, and $|C|$ is the total number of words in the collection.

3. MULTIPLE-BERNOULLI MODEL

We now examine two formal methods for statistically modeling documents and queries based on the multiple-Bernoulli distribution.

3.1 Model A

Let us assume that a document is a single sample from a multiple-Bernoulli distribution, where each binary trial corresponds to the event that some word appears in the document or not. Therefore, a document can be represented by a vector $r \in \{0,1\}^{|V|}$, where $r_k = 1$ if and only if word w_k occurs in the document. From this single sample, we wish to estimate a smoothed language model for the document. As in the multinomial case, we will assume a prior over the model. In the case of the multiple-Bernoulli distribution, we

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
Copyright 2004 ACM 1-58113-881-4/04/0007 ...$5.00.

choose a multiple-Beta distribution, which is the conjugate prior. We get:

$$\hat{\theta}_D = \arg\max_{\theta_D} \prod_{i=1}^{|V|} \frac{1}{B_i} \theta_i^{r_i+\alpha_i-1}(1-\theta_i)^{\beta_i-r_i} \qquad (2)$$

where $\theta_i = P(w_i|\theta_D)$ parameterizes the multiple-Bernoulli distribution, $B_i = \frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_i+\beta_i)}$ is the Beta function, and the α_i's and β_i's are parameters of the multiple-Beta prior. The solution to Equation 2 results in probabilities of the form:

$$\hat{\theta}_i = P(w_i|\hat{\theta}_D) = \frac{r_i + \alpha_i - 1}{\alpha_i + \beta_i - 1}$$

We show how the α_i and β_i parameters can be chosen in Section 3.3.

3.2 Model B

The term probability estimates in Model A are based on whether or not a term appears in a document. The model fails to take into account the number of times a term appears, which is an important factor in information retrieval. Model B deals with this issue by modeling the document as a collection of samples from a multiple-Bernoulli distribution. Here, we assume that we sample from a multiple-Bernoulli distribution once for each word in the document. Each sample $r \in \{0,1\}^{|V|}$ contains a single element set to 1 corresponding to the word that appears in that location. That is, the sample associated with the occurrence of word w_i is the vector r such that $r_i = 1$ and $r_{j\neq i} = 0$. Thus, each word in the document is associated with an indicator vector. Modeling documents in this manner leads to the following MAP distribution:

$$\hat{\theta}_D = \arg\max_{\theta_D} \prod_{i=1}^{|V|} \frac{1}{B_i} \theta_i^{tf_{i,D}+\alpha_i-1}(1-\theta_i)^{|D|-tf_{i,D}+\beta_i-1} \qquad (3)$$

where $tf_{i,D}$ is the number of times word w_i occurs in document D. The solution to Equation 3 gives:

$$\hat{\theta}_i = P(w_i|\hat{\theta}_D) = \frac{tf_{i,D} + \alpha_i - 1}{|D| + \alpha_i + \beta_i - 2}$$

As we see from the term probability estimates, this model makes use of term frequencies, and should result in better performance over Model A.

3.3 Smoothing

The probability estimates derived for the two models contain the free parameters α_i and β_i. Notice that setting $\alpha_i = 1$ and $\beta_i = 1$ results in the maximum likelihood estimate in both cases. However, such estimates suffer from the "zero frequency" problem and need to be smoothed. For Model B we choose to set $\alpha_i = \mu\frac{cf_i}{|C|} + 1$ and $\beta_i = \frac{|C|}{cf_i} + \mu(1 - \frac{cf_i}{|C|}) - 1$ where μ again is the smoothing parameter. Note that this choice results in the mean of the multiple-Beta prior equaling $\frac{cf_i}{|C|}$ for each word. This is exactly analogous to what is done in Dirichlet smoothing. This results in the following smoothed term estimates:

$$P(w_i|\hat{\theta}_D) = \frac{tf_{i,D} + \mu\frac{cf_i}{|C|}}{|D| + \frac{|C|}{cf_i} + \mu - 2}$$

A similar form of smoothing results for Model A, but details are omitted due to space constraints. We note that our choice of smoothing is meant to mimic the form of Dirichlet smoothing and that many other choices for α_i and β_i are possible.

	Model A		Model B		Multinomial	
	$\hat{\mu}$	AvgP	$\hat{\mu}$	AvgP	$\hat{\mu}$	AvgP
DOE	10	0.1616	100	0.1966	200	0.1968
WSJ	500	0.1050	2000	0.2540	1500	0.2592

Table 1: Comparison of results

4. RESULTS AND CONCLUSION

Some preliminary experiments were done with the two proposed models. The two data sets used were the DOE abstracts (title queries, TREC topics 51-200) and WSJ news articles (title queries, TREC topics 1-200). Results, in terms of mean average precision (AvgP) are given in Table 4 for smoothed version of both multiple-Bernoulli models and the multinomial model with Dirichlet smoothing. Note that $\hat{\mu}$ denotes the smoothing parameter that resulted in the best performance for the given method. Scoring is done by query likelihood. That is, queries are modeled exactly as documents are under each model. Documents are then ranked by the likelihood $\hat{\theta}_D$ generates the query. From the results, we see that Model B performs equivalently to the multinomial model, whereas Model A performs very poorly due to its lack of a tf component. Results also showed that the multiple-Bernoulli and multinomial models were equally sensitive to the choice of smoothing parameter.

In this paper we presented a formal approach to modeling and smoothing documents with a multiple-Bernoulli model. Although the preliminary results obtained are not better than the multinomial model, there may exist applications where a multiple-Bernoulli approach is a more appropriate choice.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, SPAWARSYSCEN-SD grant number N66001-02-1-8903, Advanced Research and Development Activity, and NSF grant #CCF-0205575. 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] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[2] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 316–321. ACM Press, 1999.

[3] H. Zaragoza, D. Hiemstra, M. Tipping, and S. Robertson. Bayesian extension to the language model for ad hoc information retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 4–9. ACM Press, 2003.

[4] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Research and Development in Information Retrieval*, pages 334–342, 2001.