Estimation, Sensitivity, and Generalization in Parameterized Retrieval Models

Donald Metzler
metzler@cs.umass.edu
University of Massachusetts
Amherst, MA 01003

ABSTRACT

In this work we investigate three important aspects of parameterized retrieval models: estimation, sensitivity, and generalization. Since all parameterized models, even those based on heuristics, have inherent uncertainty, we study these issues using statistical tools.

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

General Terms: Algorithms, Experimentation, Theory

Keywords: Estimation, sensitivity, generalization

1. INTRODUCTION

In information retrieval, most of the commonly used retrieval models are *parameterized models*. A parameterized model is one whose scoring function has one or more tunable parameters. In these kinds of models, there are several important issues that have not received a great deal of attention, but are critical for developing a deeper understanding of such models.

In this paper[1] we analyze parameter estimation, sensitivity, and generalization in parameterized retrieval models from a statistical point of view. Our work formalizes and extends the current understanding of the relationship between parameters and models. In addition to providing deeper insight into existing models, the analytical tools proposed here can also be used to develop better models in the future.

There has been relatively little work in the information retrieval literature that has looked at these issues in any substantial detail. Previous work [1, 3] provides some sensitivity analysis via the use of plots, but little has been done to address the issue formally.

2. PARAMETER ESTIMATION

In information retrieval, the goal of parameter estimation is to choose a parameter setting that yields the most effective model possible. Since uncertainty is inherent in all parameterized retrieval models, even those based on heuristics, the problem is best motivated and studied in a statistical framework.

[1]See [2] for an extended version of this paper.

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.
ACM 1-59593-433-2/06/0011.

Given a model, parameterized by θ, let $\mathcal{M}_\theta$ be the parameter space and $m(\theta', \mathcal{T})$ be the value of the effectiveness metric evaluated at θ' with regard to training data $\mathcal{T}$. Furthermore, let $P(\theta'|\mathcal{T})$ be the likelihood that parameter setting θ' is the *optimal* parameter setting given the training data. This is the *posterior distribution* over optimal parameter settings after observing the training data.

The most common approach to parameter selection involves choosing a single parameter setting for use on all queries. Such estimates are *point estimates*. Two point estimation techniques based on the metric surface and the posterior distribution may be used. The first technique is based on the metric surface. In this approach, the parameter that maximizes the metric over the parameter space is selected. That is, $\hat{\theta} = \arg\max_{\theta'} m(\theta', \mathcal{T})$, where $\hat{\theta}$ is our estimate. The other technique is based on the posterior distribution. Here, the parameter that maximizes the posterior is selected. That is, $\hat{\theta} = \arg\max_{\theta'} P(\theta'|\mathcal{T})$. This is the maximum *a posteriori* estimate. Both approaches assume that the maximum/mode of the training set metric/posterior and the maximum/mode of the test set metric/posterior will be similar.

Alternatively, we can take a Bayesian approach. Rather than choosing a single estimate to use across all queries, a new parameter is selected for each query. This is done by repeatedly sampling parameters from some underlying distribution. The most straightforward choice is to sample parameters from the posterior. Such sampling can overcome the problems involved when the posterior may be multimodal. In such cases, sampling can be 'safer' and ensure that parameters around each mode are used.

3. PARAMETER SENSITIVITY

Given a retrieval model and effectiveness metric, how sensitive is the effectiveness to perturbations of the parameter? This is the question that parameter sensitivity analysis tries to answer. If a slight perturbation causes the effectiveness to drastically change then the model is sensitive, but if the effectiveness is unchanged even with large shifts in the parameter then the model is insensitive. The reason for studying parameter sensitivity is because models that are sensitive are less prone to drastic changes in effectiveness if a poor parameter setting is chosen.

We propose two statistically motivated measures of sensitivity. Our first sensitivity measure is the entropy of the posterior distribution. The entropy of a distribution can be thought of as the uncertainty inherent in it. The entropy

alone is not a valid indicator of sensitivity. A posterior distribution with a large entropy is not necessarily sensitive, because the metric surface may still be flat. Similarly, if the posterior has low entropy, but the metric surface varies widely over the high confidence parameter values then there exists parameter sensitivity. Therefore, we must also include some notion of the flatness of the metric surface.

In order to measure how flat a distribution is over a set of parameter values, we compute the *spread* of the effectiveness metric, which is computed as:

$$S = \max_{\theta \in \{\theta' : P(\theta'|\mathcal{T}) > 0\}} m(\theta, \mathcal{T}) - \min_{\theta \in \{\theta' : P(\theta'|\mathcal{T}) > 0\}} m(\theta, \mathcal{T})$$

where $\{\theta : P(\theta|\mathcal{T}) > 0\}$ is the support of θ.

The spread, when combined with the entropy, provides a novel, robust way of looking at parameter sensitivity. For example, a model with high entropy, but low spread is more stable than a model with low entropy, but large spread. An ideal model will have low entropy and low spread, which indicates very high confidence over a small, *flat* range of the parameter space.

4. PARAMETER GENERALIZATION

We are particularly interested in intracollection and intercollection generalization, which are two different ways of measuring the generalization properties of a model.

Intracollection generalization deals with how well a model trained on a set of topics from some collection generalizes to another set of topics on that same collection. This is a common setting in TREC evaluations, where collections are often reused from year to year, and systems are typically trained on the topics from the previous year(s).

Intercollection generalization measures how well a model trained on a topic set from one collection generalizes to a different topic set on a different collection. This is a practical scenario for 'off the shelf' retrieval systems that may be used across a wide range of different collections.

We measure generalization properties of a model by computing the *effectiveness ratio*, which is the ratio of the observed effectiveness of a (trained) model to the optimal effectiveness. Thus, an effectiveness ratio of 100% represents a model that generalizes optimally.

5. RESULTS

We use the tools developed here to analyze the properties of BM25, F2EXP, language modeling (Dirichlet smoothing), and Metzler's dependence models using two newswire collections (AP, Robust 2004) and two web collections (wt10g, GOV2). Due to space constraints, we present an overview of our analysis and results. See [2] for more details.

In terms of parameter estimation, we investigated whether one of the point estimates or the Bayesian sampling technique resulted in a clearly optimal parameter selection strategy. Our results showed that no single method clearly dominates. Each method, in fact, works well for all model/collection pairs, with negligible differences between the methods in most cases. There are several interesting trends in the data, however. For the wt10g collection, it is almost always better to use the sampling method. The posterior distribution estimated using the Dirichlet model has several modes, which indicates that there may actually be multiple optimal parameter values that are appropriate, rather than a single, fixed value.

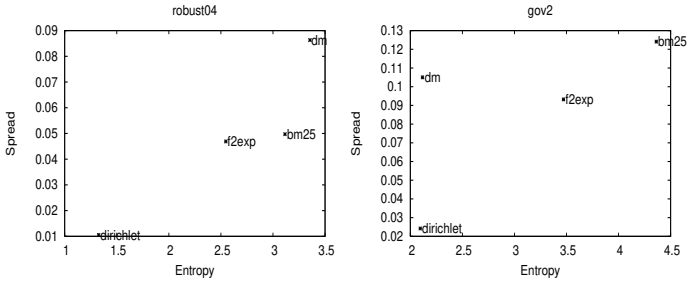

Figure 1: Sensitivity plots for robust04 and gov2.

When looking at parameter sensitivity, both the entropy and spread of the posterior distribution are considered together. Figure 1 plots the models with respect to these measures for two collections.

Our results indicate that, in terms of sensitivity, the F2EXP and Dirichlet models are the least sensitive models, and that slight variations in their parameter settings are less likely to produce drastic changes in effectiveness. In addition, the results indicate that both of these models have rather focused (low entropy) posterior distributions.

The intracollection generalization results indicate that all of the models do a relatively good job of generalizing across topic sets within the same collection, with average effectiveness ratios well above 98%. We note that the F2EXP model tends to generalize better within newswire collections, while the dependence model generalizes better for web collections. The BM25 model, however, has the best average effectiveness ratio, which indicates its parameters do a particularly good job of capturing collection-dependent characteristics, rather than topic set-specific ones.

For intercollection generalization, the dependence model, on average, comes within 1% of the optimal setting regardless of which collection is what trained on, whereas the F2EXP model only comes within 4% of the optimal on average. The Dirichlet and BM25 models have average effectiveness ratios of 98.9% and 96.9%, respectively. Therefore, the dependence model and Dirichlet models are more robust when it comes to cross-collection generalization which make them good "out of the box" algorithms.

Finally, we note that there is a certain disconnect between sensitivity and generalization. Models that are less sensitive are not necessarily those that generalize the best. This is mainly caused by the characteristics of the posterior distribution. If the distribution changes across collections, then the model is unlikely to generalize.

Acknowledgments

This work was supported in part by the CIIR, in part by NSF grant #CNS-0454018, in part by ARDA and NSF grant #CCF-0205575, and in part by NSF grant #IIS-0527159. Any opinions, findings and conclusions or recommendations expressed in this material are the author's and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proc. 27th SIGIR*, pages 49–56, 2004.

[2] D. Metzler. Estimation, sensitivity, and generalization in parameterized retrieval models (extended version). Technical report, University of Massachusetts, 2006.

[3] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proc. 25th SIGIR*, pages 49–56. ACM Press, 2002.