# Is Information Retrieval Anything More Than Smoothing?

Jay M. Ponte
The MITRE Corporation
202 Burlington Rd.
Bedford MA, 01703
ponte@mitre.org

**Abstract**

This paper looks at a specific problem in the area of document retrieval from the standpoint of language modeling. From this perspective, the following question is posed: What utility, if any, is there in additional sources of information beyond collection statistics to improve retrieval effectiveness?

## 1 A Simple Question

Is information retrieval anything more than smoothing? If one is talking about the whole field of "Information Retrieval," the answer is clearly yes, as there is a lot more to the field than the act of retrieving documents. Researchers have investigated user interaction, time and space efficiency issues, manual indexing schema, interaction with intermediaries and numerous other topics. The topic of this article, however, is more limited. We will consider information retrieval (in lower case) to be the act of retrieving relevant documents in response to a user, a task exemplified by the TREC *ad hoc* track and tasks very much like it such as web retrieval, query by example, and small sample relevance feedback. These problems are similar in that they require the retrieval of a small number of relevant documents out of a very large collection of mostly non-relevant documents and the information about the user is limited to either a query consisting of a few words or perhaps by a small number of example documents. Contrast this to document routing where a larger number of examples is available, potentially enough to construct a reasonable user model. For our purposes, this problem can be described as small user sample information retrieval.

Clearly, in this restricted case, one has only a small amount of information regarding the user's intentions. A seemingly necessary step in solving this problem is the reduction of small sample variance error, smoothing being a convenient mechanism to accomplish this in the language modeling formalism. The question in the title can now be better rephrased as: Is retrieval with a small user sample a matter of variance reduction, or can we do better?

## 2 Background: Language Modeling and Information Retrieval

The language modeling approach to retrieval investigated by Ponte and Croft [PC98] treats the problem of document retrieval as the problem of estimating the probability that the user's query was

drawn from a language model estimated for each document in the collection. The intuition behind this approach is that:

- Users are formulating their queries by choosing words that they believe to be good discriminators (as opposed to writing a description of the information they would like).
- The words the users are choosing are, in fact, good discriminators.

If one accepts this model, or to the extent that one accepts it, it becomes clear that estimating an entire distribution for a document, a distribution over the vocabulary of the collection, is where the difficulty of the problem lies. Clearly, a document, by itself, is an extremely small sample from which to estimate a distribution using a maximum likelihood criterion. In order to produce a reasonable distribution, one has to take the estimation problem seriously. The approach taken by Ponte and Croft [PC98] and in additional work by Ponte [Pon98] used a variety of smoothing techniques, a shrinkage style estimator using the average probability of a term in documents containing it, a backoff estimator using the collection probabilities, and a histogram estimator for low frequency terms. Stated briefly, smoother was better.

Miller et al [MLS99] use a two state hidden Markov model with one state for general English words and the other state for on topic words. This two state model, can be viewed as a two part mixture model and though the motivations are different, the effect is to smooth the estimates of the query terms by, essentially, backing off to the collection probabilities.

Berger and Lafferty [BL99], use a retrieval model based on statistical machine translation. This model replaces *ad hoc* query expansion techniques with statistical machine translation by explicitly modeling the difference between the user's query and the ideal query as the translation of one to the other. As part of the training of this model, a smoothing technique using backoff to the collection model was employed to prevent excessive variance error. It is important not to overlook the contribution of the translation model and the similar contribution of query expansion techniques employed by more traditional retrieval algorithms. However, first, we will consider the role of variance reduction in *tf.idf* retrieval.

## 3 Variance Reduction in Traditional Approaches to IR

In [RW94], the authors describe what has become known as Robertson's TF, one of the most important practical developments in IR. In its simplest form, Robertson's TF is defined to be $\frac{count_{t,d}}{count_{t,d}+k}$, where $count_{t,d}$ is the count of term $t$ in document $d$ and $k$ is a (usually small) constant. This function rapidly approaches unity as $count_{t,d}$ increases. The authors' motivation comes from the 2-Poisson model, but one way to understand the effectiveness of this formula was provided by Greiff [Gre99]. A plot of the log-odds of relevance vs. increasing query term counts (from the AP newswire) is shown in Figure 1. Note that this is a residual plot that takes into account term occurrence as well as *idf*. This accounts for the negative values. Notice that the smoothed curve has a similar asympotic behavior to Robertson's TF, which though not intended to be a variance reducing estimator per se, does have the effect of reducing variance just as the smoothing based estimator does.

In [XC96], the technique of Local Context Analysis (LCA) was developed to address the vocabulary mismatch between queries and documents. One key feature of LCA is bias towards terms and phrases that co-occur with more of the query terms. LCA can be viewed as analogous to Berger and Lafferty's [BL99] translation based approach, but it functions by adding co-occuring terms from the

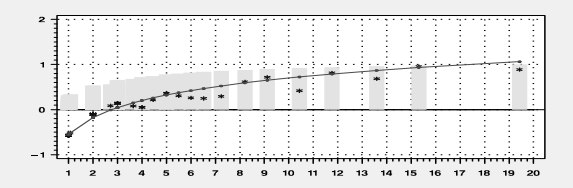


Figure 1: **Residual log-odds or relevance as function of tf smoothed with regression.**

top retrieved documents to the query. The advantage of this approach over the technique of Latent Semantic Indexing (LSI) [DDL$^+$90], is that while the latter loses information in the processing of reduing the feature space, Local Context Analysis is able to achieve increased recall without harming precision. This is an important property for an algorithm to have and we will revisit it in the final section.

## 4 A Simple Question Restated

The success stories in the TREC evaluations, to date, have been largely statistical. A necessary component has been smoothing techniques or, at least, techniques that work rather like smoothing. This makes perfect sense. The nature of the problem is to pull a needle out of a haystack in response to an impoverished description of the needle. Matching documents to queries has necessitated the use of smoothing because smoothing reflects the lack of available information. With few exceptions, attempts to add external knowledge sources, more elaborate language processing techniques, or extended feature sets has met with limited success.

One possible reason for this is simply that word based features reduce the entropy of relevance to a degree and it is possible that when more elaborate features are included, they may not reduce the conditional entropy of relevance any further. I.e., it may be that these other features just don't provide useful information. In this case, we can stop looking at such techniques.

On the other hand, if these techniques can reduce the conditional entropy of relevance, even when used in addition to word based features, it is possible that their ineffectiveness to date has been the result of something else.

In order to achieve state of the art performance on retrieval tasks, one needs to smooth. One needs to smooth a lot. When a distribution is smoothed to a great degree, it is possible to smooth out more than the noise. In fact, it may be the case that the current retrieval techniques are both a blessing and a curse. A blessing because they provide a high baseline performance and a curse because the excessive smoothness precludes finer grained distinctions from having a positive effect on retrieval.

In addition, when, for example, one goes from single word features to phrasal features, the small sample variance problem exhibited by single words is going to be significantly more of an issue for phrases simply because they are rarer. The effect will only become more pronounced as feature sets get more elaborate.

This leads us to the question being posed by this paper. Can we make information retrieval more accurate by adding in new sources of information or by enriching the feature set? Have these

techniques failed to work because the necessarily smooth retrieval functions have not allowed them to? Have they failed to work because the information they provide is overcome by estimation error? Or, have they failed to work because there is no new information to be gained from them?

These questions are currently a matter for debate but should be a matter for science. Language modeling can provide us with a new set of tools to examine these problems more closely. In conjunction with the data now available to us as a result of TREC [Har93], we have a real opportunity to resolve these and similar issues.

## 5 For Example

A simple example of variance reduction is case folding. Partially due to empirical evidence and partially as an engineering convenience, most IR systems do not make case distinctions during retrieval. Each term in the collection is typically converted to lower case at indexing time and likewise for the query terms at retrieval time. One can view this as k-nearest neighbor smoothing with an appropriate distance function. The neighbors are the case variants of the term and all are weighted equally when any of them occurs. This has the effect of altering the feature space in a variance reducing (and bias increasing) manner. Case-folding is a knowledge-based operation, but it can be viewed as a variance reduction technique. As an aside, one can view more interesting operations, such as stemming, in the same way. When one case folds, it is an 'all or nothing' operation and whether to do it or not is likely to be determined by empirical study. Viewed as a smoothing technique, however, one can alter the kernel function in order to trade off bias for variance to the optimal level. If one does this for case folding, stemming and additional techniques, a general conflation mechanism emerges. More importantly, adding a knowledge source, such as knowledge of English morphology in the case of stemming, one does not have to make a yes or no decision as to whether it is better to stem and lose information or not stem and keep noise, but instead a principled tradeoff can be made.

In addition to the greater space of possibilities, the language modeling formalism is amenable to analysis. Adding prior knowledge increases the bias of the model; ignoring that prior knowledge increases the variance. The tradeoff of the two is a measurement of the usefulness of the combination of the knowledge source and of processing done to exploit it. Without this view, one only has a binary choice, use the knowledge source or not. In this case, the information gain of a knowledge source in the form of reduced variance error may be offset by increased bias error. Viewed as smoothing, one can also learn about the space in between and make informed tradeoffs.

## References


[BL99] Adam Berger and John D. Lafferty. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 222–229. ACM, 1999.

[DDL+90] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[Gre99] Warren R. Greiff. *Maximum Entropy, Weight of Evidence and Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, Massachusetts, September 1999.

[Har93] Donna Harman. Overview of the first Text REtrieval Conference (TREC-1). In D. K. Harman, editor, *The First Text REtrieval Conference (TREC1)*, pages 1–20, Gaithersburg, Md., February 1993. NIST Special Publication 500-207.

[MLS99] David R. H. Miller, Tim Leek, and Richard M. Schwartz. A hidden markov model information retrieval system. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*, pages 214–221. ACM, 1999.

[PC98] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, Melbourne, Australia, August 1998. ACM Press.

[Pon98] Jay Michael Ponte. *Probabilistic Language Models for Topic Segmentation and Information Retrieval*. PhD thesis, University of Massachusetts, Amherst, Massachusetts, May 1998.

[RW94] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland, July 1994.

[XC96] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, aug 1996.