Finding an Exponential Model for Text Retrieval through Textual Analysis

Jaime Teevan & David Karger

Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Text retrieval is a difficult problem that has become both more difficult and more important in recent years. The problem has grown due to the increased amount of electronic information available and the greater demand for text search as a result of the World Wide Web. People are surrounded with large quantities of digital information, but are sometimes unable to use that information effectively because of its overabundance. By improving information retrieval, we can improve people's access to the available information.

One popular approach to information retrieval (IR) is to develop a model for a user's information need and the corpus, and retrieve documents that satisfy the user's need using the model. There is a trade off in building a model between how closely the model matches reality versus the complexity of the model. We are in the process of using textual analysis to learn the best model for IR that is also computationally feasible.

Motivation: There are several benefits to working with a model for information retrieval. First, the assumptions are explicit. For example, we might assume that word order is not important, or that whether or not a term occurs in a document is independent of all the other terms that occur in that document. Given these clearly stated assumptions, we can understand how closely our retrieval method matches reality and recognize areas that we might improve. Additionally, the trade-offs when changes are made are clear. When we change an assumption, we can clearly understand what level of complexity that will add to our model, and predict the effects of such changes on performance.

Now, while we may have a good idea of whether any of the assumptions made in a model are true or not (as most are not), we do not necessarily know which assumptions are the least realistic, and therefor the most worth investigating further. In our research, we hope to break away from relying on our preconceived beliefs by using textual analysis. This allows us to go directly to the data and improve upon the current IR models by understanding which assumptions are most inaccurate, and therefore most need to be relaxed. While we do not revisit every assumption IR models make, by grounding at least some of the assumptions in reality, we are able to build a model with a stronger foundation.

Some information retrieval methods are similarly based on textual analysis and a good understanding of the written language. For example, Sparck Jones [2] suggests using the inverse document frequency for term weighting based on textual analysis, and likewise Warren Greiff suggests improvements to tf.idf in more recent work come from an understanding of the statistical nature of text [1]. However, these methods are not model based.

Approach: We restrict ourselves throughout the course of our analysis to naïve Bayesian models. Lewis gives a good overview of the use of naïve Bayesian models for text retrieval and classification [4]. Although doing this greatly restricts the course of our exploration, many information retrieval models are naïve Bayesian. Thus, by restricting ourselves, we are able to encompass much of what is currently used and understand better the significant problems with them. We also chose to do this because naïve Bayesian models allow for efficient IR implementations.

We use the very simple naïve Bayesian model, the multinomial model, as a starting point for our textual analysis. This model is common in text classification, but not for text retrieval [4], although Kalt does explore it briefly [3]. In its simplest form, we find this model performs retrieval well, with retrieval performance comparable to tf.idf, a popular retrieval benchmark. However, when we tried to take advantage of the model and use more likely estimates of the model parameters, we found our performance worsened. This lead us to conclude that the multinomial model is not a good model for text documents, but rather is so bad that it actually hinders retrieval.

We find that one major mismatch of the multinomial model with actual text data is that there are many fewer unique terms in a document of a given length than the multinomial model expects. This is because in reality once a term is used in a document, it is likely to be used again. We can confirm this by looking at the probability distribution for a term. If, once a term has occurred in a document, it is much more likely to occur again, we should see the probability of that term occurring a large number of times in a document to be much higher than the multinomial model predicts, as we do indeed see in Figure 1.

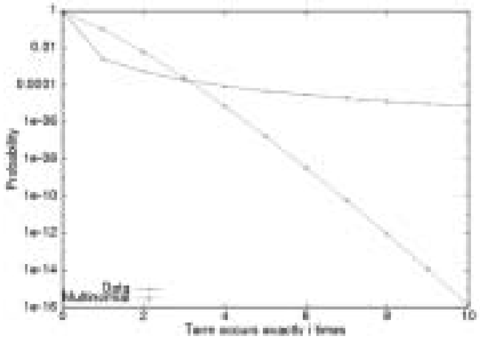

Figure 1: The empirical term occurrence probability distribution for terms compared with what is expected given a multinomial model. Term occurrence appears to be more clustered than expected.

Impact: Using what we have learned, we improve upon the multinomial model, remaining all the while within the bag-of-words, naïve Bayesian framework. We would like to match the empirical term distribution more closely. We propose a new textual model that allows for greater flexibility in the term distributions. Instead of restricting ourselves to one particular family of distributions, we want to learn the best distribution family within all possible families.

However, we do further restrict the set of all possible families in two ways. The first is because we would like to maintain the simplicity of the original multinomial model. To avoid increasing the complexity of the model, we continue to describe each specific term distribution with a single parameter. This restriction rules out many plausible new models, such as one that describes a particular term's distribution as a mixture of multinomial distributions. However, it makes for a simpler and more efficient model. It also helps us avoid the danger of over-fitting our model to the data.

Additionally, we decide to restrict ourselves to one-parameter exponential family of distributions. We choose to do this for many reasons. The exponential family can be used to retrieve documents efficiently, but remains general enough to include the multinomial model and many other models currently used. What's more, the exponential family is flexible and well studied, and learning within the family is straight forward. By using the corpus to learn the best exponential family to model the term occurrence probability distribution, we have found that we can build a model that more closely matches the data.

Future Work: We continue to work understand how to learn the best possible exponential distribution. Additionally, while the new exponential model more closely matches the data, that does not necessarily mean that it can be used to perform better retrieval. We look forward to testing the new model's search performance, and plan to do this shortly.

There is also a problem with naïve Bayesian models in general that does not occur with the multinomial model that we would like to explore. Most naïve Bayesian models expect all documents to be the same length. The document length is determined by how many times the various terms are selected to appear and is not an external parameter determined by the document writer. We are in the process of investigating whether or not this poses a significant problem.

Research Support: This research is supported by NTT, the Packard Foundation, Project Oxygen, the Authur P. Sloan Foundation and the National Science Foundation.

References:

[1] Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proceedings of SIGIR-98*, Melbourn, Australia, August 1998.

[2] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[3] T. Kalt. A new probabilistic model of text classification and retrieval. Technical Report IR-78, University of Massachusetts Center for Intelligent Information Retrieval, 1996.

[4] David D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *EMCL*, pages 4–15, 1998.