Improving Information Retrieval with Textual Analysis: Bayesian Models and Beyond

Jaime B. Teevan
teevan@mit.edu

February 27, 2001

Abstract

Information retrieval (IR) is a difficult problem. While many have attempted to model text documents and improve search results by doing so, the most successful text retrieval to date has been developed in an ad-hoc manner. One possible reason for this is that in developing these models very little focus has been placed on the actual properties of text. In this paper, we discuss a principled Bayesian approach we take to information retrieval, which we base on the standard IR probabilistic model. Not surprisingly, we find this approach to be less successful than traditional ad-hoc retrieval. Using data analysis to highlight the discrepancies between our model and the actual properties of text documents, we hope to arrive at a better model for our corpus, and thus a better information retrieval strategy. Specifically, we believe we will find it is inaccurate to assume that whether a term occurs in a document is independent of whether it has already occurred, and we will suggest a way to improve upon this without adding complexity to the solution.

1 Introduction

Text retrieval is a difficult problem that has become both more difficult and more important in recent years. This is because of 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 information, but unable to use that information effectively because of its overabundance.

The most common solution to this problem are the Web search engines. Here text search is done using a collection of heuristics, each new trick giving a slight improvement over previous methods. While people may develop models to describe these ad-hoc methods, the methods are developed empirically based on what has been shown to work, and what has not. The result has been reasonably successful text retrieval.

However, without a theoretical framework to describe a problem, it is difficult to talk about what is happening. Our goal is to develop a model for information retrieval that performs well because it accurately describes the text documents it is expected to retrieve over, but that is simple enough to allow efficient retrieval. By developing a good model of text documents and their relationships to each other and to a query, it will be easier to understand and improve on retrieval methods. The assumptions inherent in the retrieval are obvious and the trade offs between the complexity of indexing and retrieving and the benefit of improvements are clear.

We begin working with a well established information retrieval framework, namely a Bayesian probabilistic framework. The models developed in this framework contain many simplifying assumptions to make retrieval with them feasible. In this paper we describe how we have worked to understand which assumptions are reasonable, given our textual analysis, and which aren't. From this understanding, we hope to arrive at a better probabilistic model for the text that matches the data as closely as possible while allowing us to retrieve documents in an efficient manner, improving retrieval. In order to develop a system that could reasonably replace the many current ad-hoc information retrieval systems, we impose the goals of one, having retrieval with our model not take significantly more time than they do, and two, not using significantly more information about the corpus in our model.

2 Prior Work in Information Retrieval

There is a large body of work on information retrieval. Here we will first briefly describe the traditional ad-hoc method of text retrieval currently in use by a majority of search engines. Then we will discuss a traditional probabilistic IR model and some variations within that model. We will also discuss the text corpora used by TREC, and how they have been used both for evaluation and to better understand the text retrieval problem.

In discussing retrieval models, it should be noted that any information retrieval model assumes that there exist some features of a document set that are sufficient to determine if each document is relevant or irrelevant to a particular query. Most text retrieval also relies on the assumption that the distribution of index terms throughout a collection is informative about the likely relevance of a document.

2.1 Vector space retrieval

The most common form of retrieval in practice now is vector space retrieval. Vector space models represent a broad class of models that view text documents and queries as feature vectors. This class of models is popular because, by constructing an index which maps the vector coordinates to documents, it is possible to search through large corpora quickly. While the original vector space models were developed empirically, it should be noted that there exist more complex frameworks that in their simple forms can reduce down to a vector space model, and can be used to motivate vector space retrieval, such as the Naive Bayesian model that we are working with.

The general idea behind a vector space model is that each coordinate in the document or query vector represents a term in the corpus. The documents that are "closest" to the the query are considered the most similar, and are returned as search results. This general framework allows for a lot of variation. Closeness can be measured as the angle between the two vectors or as the dot product. The coordinates in the vector can be represented an simple binary values indicating the presence or absence of a term, or can be more complex, attempting to encompass both the emphasis the document or query puts on the term and the rarity of the term.

A common way to represent a term is called tf.idf. The emphasis a document puts on a term is incorporated by factoring the term frequency (TF) of the term in the document into the coordinate's value. The inverse of the number of times a document occurs in the corpus (IDF) is commonly used to represent the rarity of a term. While there are many variations on exactly how these values contribute to a coordinate's value, we will be using the following formula when we use tf.idf:

$$TF(w_i) \cdot \log(\text{number of documents} \cdot IDF(w_i))$$

2.2 Probabilistic retrieval models

A number of principled models have been developed to improve upon the ad-hoc, yet successful, vector space method described above. One of the first of these was the probabilistic retrieval model. vanRijsbergen gives a good introduction to the basics of the probabilistic model in his book [van79].

There are several major assumptions at the base of a probabilistic retrieval model. First, probabilistic models assume that there exists one or more feature distributions that describe each relevant document and one or more feature distributions that describe each irrelevant document. Documents are assumed to have been generated by sampling from these distributions.

Traditionally, probabilistic models also assume that the relevance of a document to a query is independent of the other documents in the collection. From this assumption, it is easy to arrive at the *probability ranking principle*. The probability ranking principle asserts that if you can determine over a corpus the probability that each document is relevant, where this probability is estimated as accurately as possible, ranking the documents by probability of relevance will give the "best" ranking.

Since traditional probabilistic models use this principle to produce results, it becomes important to find the probability that a document is relevant. Bayes rule transforms this problem into one of finding the probability that the document would contain the features it does if it were generated from one of the relevant distributions. For this reason, estimating the feature distributions is at the heart of using a traditional probabilistic model. To further simplify the problem, standard probabilistic models make the additional assumptions that there is only one relevant and one irrelevant distribution, that the important features are terms, and that the terms are independent of each other.

From here there are several major paths that have been explored. A model may be binary, and assume only the presence or absence of a feature, or multinomial, like the model developed by Kalt and Croft [KC96]. A model may try to estimate the feature distributions directly, or it may take a Bayesian approach, like the model proposed by Keim, et al. [KLM97], and place priors on the distributions which are then updated with whatever new information becomes available about the distributions.

In addition to how the model is specified, information about the distributions can be collected in different ways. Some implementations assume that there already exist sufficient labeled documents to estimate the distributions [KC96], others first present a set of results from an ad-hoc retrieval system and rely on relevance feedback [van79], while still others use the present query and previous queries, as well as any feedback information that is available [KLM97]. Nigam, et al., for the purpose of classification, use expectation maximization to estimate the model parameters from the data [NMTM98]. While they still rely on having labeled training data, they are able to use the unlabeled data as well in a feedback loop.

2.3 Working with Text Documents

Understanding the corpus we are retrieving over is essential in building a model for information retrieval. Here we will discuss issues involving the corpus. We will discus work involving exploratory data analysis to improve retrieval. We will also discuss how evaluation is done, and how different retrieval methods are compared.

2.3.1 Data analysis

We all have a lot of experience with text documents. We read the newspaper. We read books, journals and magazines. We know that when we write a document, we're not drawing words at random from some prior probability distribution. Because of this knowledge, it is often tempting when trying to improve upon a text retrieval model to make assumptions about how the data diverges from the model without actually understanding how it diverges. For example, we know that the words we choose influence the other words we choose, so we may try to fix the model by removing the independence assumption. But really this assumption may not be the problem. Without data analysis we cannot really know. Unfortunately, this intuitive manner for improving IR models is how changes are often made.

There has been some recent work, however, that has involved exploratory data analysis to improve retrieval. Warren Greiff uses data analysis of relevant and irrelevant document sets to understand the relationship between term frequency and the mutual information between relevance and term occurrence

[Gre98]. Through this analysis he finds a "flattening" of the typical term rarity indicator, the inverse document frequency, at document frequency extremes, and uses this to improve information retrieval. In our work we will similarly use data analysis in the hopes of improving retrieval, but with the goal of understanding what are the appropriate features for our Bayesian model.

2.3.2 Evaluation

Exactly what to evaluate in an information retrieval system and how to evaluate it are contentious issues. However, we will use the reasonably standard and well established measures of *precision* and *recall* to evaluate our model. Precision is the portion of retrieved documents that are actually relevant and recall is the proportion of the relevant documents that are retrieved. Retrieval effectiveness is measured as the ability of the system to retrieve relevant documents while holding back non-relevant ones. We will use the standard measures for these numbers used within the IR community. A set of standard recall values are specified (10%, 20%, ..., 100%). For each of these values we calculate corresponding precision values for each query. Then the precision values are averaged across queries. vanRijsbergen's book on information retrieval has a good chapter discussing this method and some of the issues involved in further detail [van79].

We will compare our system with the standard tf.idf vector space method of retrieval. Time permitting, we will also compare our model with several other models, such as Robertson and Walker's 2-Poisson model [RW94].

2.3.3 TREC

Testing new information retrieval methods is difficult for a number of reasons beyond the difficulties in deciding what and how to evaluate. One is that the documents in different corpora could conceivably be written in very different manners. This is obvious when considering documents in different languages, but likely true for different topics as well. Retrieving over a collection of law review articles may be different from retrieving over Web pages.

Additionally, which documents are relevant to a person's search may be difficult to determine, especially because searches are often under specified. A child searching for dog documents for a school report needs different results than a veterinarian does. And if testing an individual system seems complicated, just imagine how much more complicated it gets when you want to be able to compare across different systems.

Because of this, the information retrieval community has developed a common framework for testing. Each year at the Text REtrieval Conference (TREC), a new corpus, set of queries, and a result set for testing against, are published. While there is still some room for variation within this framework, in general this greatly facilitates evaluation, both internally to one's experiments, and across different experiments. We use the TREC framework for our data analysis and evaluation.

3 Building a Better Model

We have implemented and tested a Bayesian probabilistic model that treats the query as a document which has been labeled as relevant, and then uses expectation maximization to find a likely relevant term distribution. This relevant term distribution is then used to find each document's probability of relevance, and rank the documents according to the *probability ranking principle*. We have found Heckerman's tutorial to be a good introduction to a Bayesian approach to probability and statistics [Hec95], and have found Bishop's book to provide a good introduction to EM [Bis95].

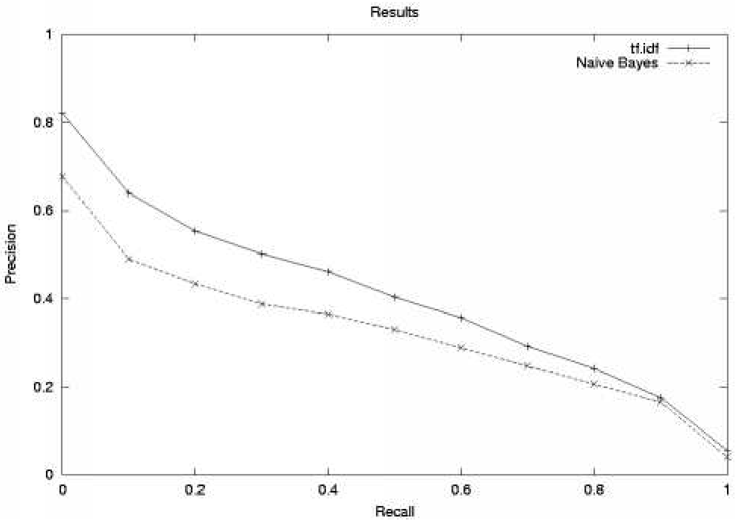

Figure 1: Results for naive Bayesian model.

As can be seen in Figure 1, the results from our initial tests have shown our model to be less successful for retrieval than the industry standard tf.idf. This problem could stem from one of two sources: the method for estimating the underlying distribution could be flawed, or the model could be flawed, and would not be successful even with the correct distribution. When the correct document set is used to generate the initial term distributions, EM will often continue to find more likely distributions that tend away from the correct distribution. Clearly, some assumptions in the model must be wrong. My master's thesis will focus on using data analysis to improve the assumptions made in our model about the relevant and irrelevant term distributions with the end result of improving the retrieval.

From the data analysis that we have done so far, it has become obvious that the naive Bayesian model doesn't match the evidence. The probability of large numbers of occurrences of the same term in a document has been found to be significantly higher than would be predicted by the Bayesian model, as can be seen in Figure 2.

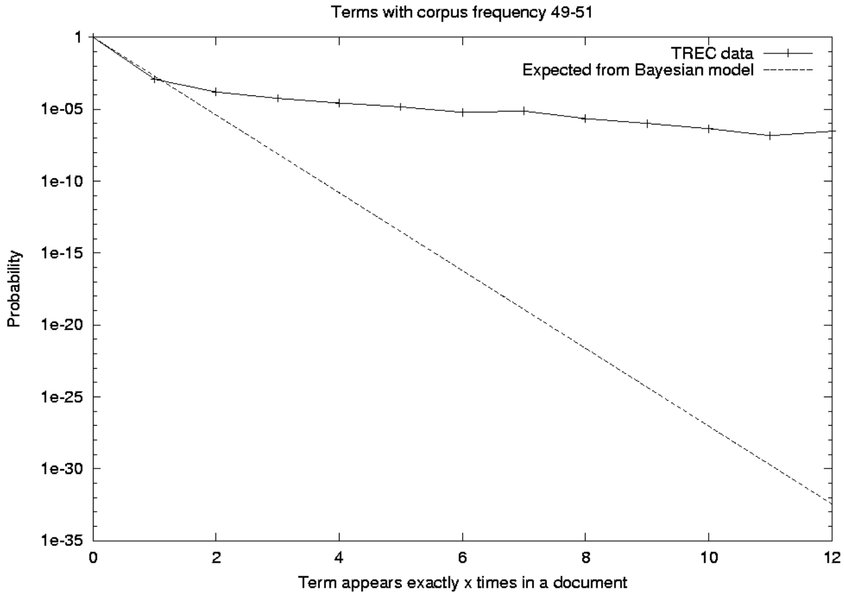

Figure 2: Here we look at the probability that a term will occur exactly x times in a document. We compare what we have actually found in TREC documents with what is predicted by our Bayesian model. We consider only d TREC documents of similar length, and terms with a similar overall corpus frequency of f. The probability distribution for the TREC data is found by counting the number of times the term occurred x times in a document. In our Bayesian model it is approximately proportional to $\left(\frac{f}{d}\right)^x$.

We want to understand more about exactly what affects the term distribution through different documents. To do this we needed to group "similar" terms and look how the different variables we are interested in, such as document length, document frequency, word frequency, etc., affect the distribution. When we have a decent understanding of the properties of the text we will suggest a new model based on what we have found. There are several ways of improving on the current naive Bayesian model that we have begun to think about.

One assumption that clearly is broken is the independence assumption. Accounting for the relationships between each term is very complex, too complex to allow for reasonable retrieval. People have tried to work around the dependence assumption with dependence trees [van79] and Bayes nets [Sah96]. However, we have seen that the independence assumption is broken in a simpler and more fundamental way: once a term occurs in a document, it is more likely to occur again in that document. One possible model for this could involve separating the probability of a term occurring in a document from a term's frequency in a document. This is an option we are likely to pursue.

4 Thesis Proposal

We have already explored the naive Bayesian model and embarked upon the data analysis. By the conclusion of my thesis, I plan to finish up the data analysis that we have begun, and develop and test an improved model. We have already found some interesting results, and the subsequent data analysis should be minimal. Specifically I imagine it will involve:

- Finding a relationship between term frequency, document frequency, and corpus frequency.
- Analyzing the effect of document length on this relationship.
- If there is time, developing some understanding as to where query terms seem to fall and how relevant document sets seem to behave.

From this analysis, we will develop a model that matches what we have observed as closely as possible, while not introducing many new variables, especially not introducing inter-term dependence. I will implement the model we develop within the current framework developed for testing the naive Bayesian model, and test it. I will then write up what we have learned.

4.1 Evaluation of the Model

Results of the new model will initially be evaluated using the Wall Street Journal documents of TREC. Same set which was used in the old model. Comparisons will be done to tf.idf and to the old model, with the hopes of improving both. If our initial theory proves to be successful, and there is time permitting, I will test the model over the whole TREC corpus.

4.2 Schedule

February 22 - March 16 During the end of February and the beginning of March, I plan to finish up the data analysis and form a reasonable model for the data. This will entail having a good understanding of term distribution among similar documents, and how document length and term frequency affect this distribution.

March 19 - April 13 This period, which contains both Spring Break and CHI, will be spent implementing the model and testing. This includes fixing the maximization in the current code to do EM instead of MM, and reproducing previous results.

April 16 - May 11 The end of April will be spent writing up the thesis. I hope to have a strong outline by April 20th, and a rough draft by the

27th. This will leave two weeks for polishing the thesis and tying up loose ends. This is a tight schedule, and will require background writing to take place while during the research and development phase.

References

[Bis95] Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[Gre98] Warren R. Greiff. A theory of term weighting based on exploratory data analysis. In *Proceedings of SIGIR-98*, Melbourn, Australia, August 1998.

[Hec95] D. Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR_TR-95-06, Microsoft Research, Redmond, Washington, 1995. Revised June 1996.

[KC96] T. Kalt and W. B. Croft. A new probabilistic model of text classification and retrieval. Technical Report IR-78, University of Massachusetts Center for Intelligent Information Retrieval, 1996.

[KLM97] Michelle Keim, David D. Lewis, and David Madigan. Bayesian information retrieval: Preliminary evaluation. In *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 303–310, Fort Lauderdale, Florida, January 1997.

[NMTM98] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Using EM to classify text from labeled and unlabeled documents. Technical Report CMU-CS-98-120, School of Computer Science, CMU, Pittsburgh, PA 15213, 1998.

[RW94] S. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, Dublin, Ireland, 1994.

[Sah96] M. Sahami. Learning limited dependence Bayesian classifiers. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[van79] C. J. vanRijsbergen. *Information Retrieval.* Butterworths, London, second edition, 1979.