Bayesian Model for Information Retrieval

Jaime B. Teevan

Artificial Intelligence Laboratory
Massachusetts Institue 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. 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 is 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 it. We are in the process of developing a model for information retrieval.

Motivation: By developing a good model of text documents and their relationships to each other and to a query, it is 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.

Previous Work: There have been a number of previously developed models for the purpose of indexing and retrieving text documents. The most widely used model for information retrieval uses boolean logic, discussed by Cooper [1] and van Rijsbergen [3]. Another standard model is the vector space model, described by Salton et al. [5].

We are investigating a probabilistic model, where documents are ranked according to the probability that they contain information relevant to the user's information need. Previous work on probabilistic models include the standard 2-Poisson model developed by Harter [2]. Other interesting probabilistic models include work by Robertson [4] and Turtle and Croft [6].

Approach: The ideal model to use in information retrieval is, by definition, the one that works the best. Of course, if we could find the model that produced what we knew to be the "right" answers, then we would know the right answers, and should just return them. So instead we try to approximate the ideal model and look for the one that best fits the data we have.

Something fundamental that seems to be left out of many of the previous text retrieval models models are the terms that are *not* in the user's query that also describe the user's information need. The standard user of a text retrieval system uses just a few words to specify what he or she is looking for. However, this need is likely to be more accurately described by the inclusion of many words and possibly even the exclusion of others. Perhaps a better model also includes some form of probabilistic relevance feedback, where relevant documents reinforce each other.

We make the assumption that the documents comprising the corpus fall into one of two categories: those that are relevant to the user's information need, and those that are irrelevant. The documents that are relevant to the query share certain distinguishing features, namely the presence of those the features that make them relevant to the users needs and the absence of those which don't. The same holds for irrelevant documents. This results in what is essentially a clustering problem that involves finding the set of documents that share similar identifying features with each other and with the query, and don't share features with the irrelevant documents. A benefit of a model based on these assumption is that it can incorporate relevance feedback in a consistent manner.

We are in the process of investing several Bayesian frameworks to find the model that best fits the data. We assume that there exists some underlying feature distribution from which words are selected to generate a document. Furthermore, we assume that the corpus is entirely generated by selecting words from one of the two distributions, the relevant distribution, and the irrelevant distribution. These distributions are modeled as Dirichlet distributions. We are empirically testing the models we develop on TREK data and comparing the results to standard tf.idf results as well as the results from similar models.

Difficulty: There are several difficulties involved in developing a model for text retrieval. One is in finding a statistical model that accurately describes natural language. If the assumptions that are made in the similarities between relevant documents as compared to irrelevant documents are bad, then no model using those assumptions can be good. Associated with this are difficulties such as selecting the features to represent the corpus, and determining the representation of the query.

Another difficulty is making the model computationally feasible. Because we are saying that the interaction between documents is important, and because as a corpus grows, so do the number of relationships, it is likely that the best model is not one which can be solved quickly.

Impact: However, if these difficulties can be surmounted, improved information retrieval has ramifications in many areas, most obviously in Web search but also in any application that deals with large quantities of text. Having a good model for information retrieval in which relevance feedback can be incorporated will allow people to more easily talk about and develop new techniques.

Future Work: We are still in the process of developing the system. There are a number of models we are still investigating, and more implementation and testing to be done. Additionally, as mentioned above, the successful models must be simplified to be feasible for retrieval on large corpora.

Research Support: This material is based upon work supported by Merrill Lynch under the Financial Technology Education Initiative, by NTT, by the Oxygen project, and by the National Science Foundation. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References:

[1] W.S. Cooper. Getting beyond Boole. *Information Processing and Management*, 24:243-248, 1988.

[2] S.P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 35:285-295, 1975.

[3] C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481-485, 1986.

[4] S.E. Robertson. The Probability Ranking Principle in IR. Journal of Documentation, 33(4):294-304, December 1977.

[5] G. Salton, A. Wong, and C.S. Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[6] H. Turtle and W.B. Croft. Inference Networks for Document Retrieval. *Proceedings of the 13th International Conference on Research and Development of Information Retrieval*, pp. 1-24, 1990.