Information Routing using a Corpus Distribution

Jeffrey A. Goldman

Computer Science Department
University of California
3436 Boelter Hall
Los Angeles, California 90095-1596
goldman@ucla.edu

Introduction

The research goal of information routing (IR) is to retrieve and rank a collection of text documents that coincide with a user profile (Harman 1995). Ideally, the profile can be derived automatically from a set of documents the user has identified as relevant to a particular topic of interest. The assumption for this work is a user has provided this small set of documents. It is then our goal to rank a previously unseen set.

With other researchers trying to solve this problem, the focus has become a question of which paradigm is able to rank these documents appropriately (Harman 1995). Approaches vary from conventional probabilistic models and vector space models, to Boolean models and complex weighting schemes (Salton 1989). Methods from all fields have been able to effectively retrieve the appropriate documents. The variability in performance, however, comes from their ability to rank them (Harman 1995). This led researchers to focus on weighting methods in order to artificially push more relevant documents toward the top of the list. However, there is no theoretical basis to suggest this approach will generalize.

Our methods, on the other hand, have a theoretical foundation in mathematics. With the exception of the Boolean model, all other approaches rely on term frequency in one form or another. This suggests a statistical distribution comparison of the collection of relevant documents to a general corpus. Words of significance should stand out from the much larger broad topic collection. Preliminary study suggests by using a vast collection of documents covering a range of topics as a baseline for term statistics, we can appropriately weight terms for our routing collection in order to effectively rank and retrieve.

Theoretical Foundation

IR justifies the use of term frequency from the notion that writers normally repeat certain words as they advance their argument. Moreover, words are repeated as the writer elaborates on the subject (Luhn 1958). Salton further appeals to Zipf's law (Salton 1983) to strengthen the position of using term frequencies. Although these notions appeal to our intuition, they are not mathematical principles. Our new approach is to consider the occurrence of a term as a Bernoulli trial and the total number of occurrences within a document becomes a binomial random variable. We can now compare two documents term by term, as a pair of binomial random variables. As a baseline with which to compare, we can use a large collection of documents of varying topics. Such a collection exists and is in fact published (Kucera & Francis 1967).

Future Work

By treating word frequencies as binomial random variables, there are advantages not available to conventional routing methods. Renewal theory is used in communications to estimate failure rates (Gut 1988). Since the theory was developed for binomial random variables, we can apply it to routing. It allows estimates of frequencies to be made by only examining a small part of the document. In a real routing application, this is critical. Even for off-line algorithms that are to evaluate and retrieve documents, text databases are so large, it is not possible to examine them all.

From a linguistic point of view, further improvements can be made by incorporating di-grams and tri-grams to the list of variables. This would allow for better matching. Furthermore, using tri-grams with a thesaurus would start to capture content within a document. Beyond tri-grams, research suggests that word pairs can occur as much as twenty words away. The computation needed to keep track of such variables outweighs the benefits. However, since the occurrence of individual di-grams and tri-grams will be less frequent than single words, we are left with the question of how to fold this information into our paradigm. If treated equally, their contribution will be insignificant. We currently have no basis with which to weight them. We would have to acquire data at the corpus level in order to determine their significance.

References

Gut, A. 1988. *Stopped Random Walks*. New York: Springer-Verlag.

Harman, D. 1995. Overview of the fourth text retrieval conference (TREC-4). *The Fourth Text Retrieval Conference (TREC-4)*.

Kucera, H., and Francis, W. N. 1967. *Computational Analysis of Present-Day American English*. Providence, Rhode Island: Brown University Press.

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2):159–165.

Salton, G. 1983. *Introduction to Modern Information Retrieval* McGraw-Hill.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.