STATISTICAL MODELS FOR UNFORMATTED TEXT

Christopher Landauer
System Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406

0. Abstract

In this note, we will describe some of the outstanding problems concerning statistical information retrieval models, and the underlying stochastic language production models they assume. The problems can be separated into classes according to the underlying language model, which can be either a sequence model or a grammar model. Both kinds of model are based on a stochastic process, but there is a different filter for the realization. The grammar models use a stochastic context sensitive grammar, and the sequence models use a high order Markov chain.

Most of these problems cannot be solved without experimentation with information retrieval concepts and systems. Most information retrieval systems that currently exist have had to make operational assumptions about the answers to these questions. It is expected that more precise knowledge of solutions for these problems will simplify the design and improve the effectiveness of statistical information retrieval systems.

ACM Categories: 3.79, 5.39
key phrases: statistical information retrieval, discrete probability structures, stochastic language description models

1. Introduction

This note describes some of the outstanding problems in the area of statistical information retrieval from unformatted natural language text, including the language production models that form the foundation of many information retrieval systems. A segment of unformatted text is a sequence of sentences, or paragraphs, or even chapters, that does not contain any content labelling information beyond the text words themselves. We will allow special words to be marked, such as proper names, geographical names, or foreign words.

We are mainly concerned with modelling the information contentof unformatted text, according to certain models derived from some simple language production models. These models are known not to be sufficient for detailed language analysis, but it has been shown that they can be successfully applied to information retrieval problems. We will concentrate on statistical approaches to text retrieval, in order to avoid dealing with hard problems of semantics. This restriction of the approach provides a flexibility that other approaches do not, especially in cases where the language domain cannot be limited in advance. The relation of these problems to current work on structured databases is beyond the scope of this note.

We will begin with a note on our terminology for models. A framework is a mathematical object, together with various functions that operate on such objects. On the other hand, a model includes an object as well as one or more hypotheses on its behavior for the application. In other words, a mathematical object can be a framework, and a process can be a model.

Each of our models for information retrieval automatically implies certain performance predictions, either in terms of the statistical distributions of various text features, or in terms of specific retrieval or implementation strategies. Where it is convenient, we also try to state these predictions, in order to suggest that they be formally tested as experimental hypotheses.

Reference [LM] describes a particular information retrieval system, in which these considerations arose. It is expected that the questions have a rather wider applicability.

2. Language Production Models

We discuss two types of language production models, stochastic and grammatical. Each model can be used at any one of several levels of detail, including letters, syllables, or words. For convenience, we will call the basic object an item. In both cases, the underlying process driving the model is stochastic. Note that we do not assume that language is actually generated by stochastic processes, only that certain aspects of the syntax of language can be described by stochastic processes.

2.1. Window or Sequence Models

The sequence model assumes that the production of language can be described by a stochastic process, in particular a Markov chain with a finite number of states. Given the relative frequencies of occurrence of all sequences of n items, we use an (n-1)-order Markov chain, where the states are the items generated. Thus each item is chosen according to its actual distribution among all sequences with the same previous (n-1) items. The sequence length n is called the window size of the model. The computation of the appropriate frequencies is a simple problem, though time consuming, given a database of item sequences.

Various extensions of the sequence model can be formulated, including one with variable length sequences. In this case, some care must be taken to insure that the set of sequences for which frequencies are computed is sufficiently large to define the stochastic process.

2.2. Grammar Models

The grammatical models assume that the production of language can be described by a context sensitive grammar (or a context free grammar in some cases). The production of item sequences is governed by a stochastic process that chooses one element of a set of applicable production rules, expanding each remaining nonterminal symbol. The computation of frequency statistics for these models can either assume that the grammar is given, and collect statistics from parses of elements of a database of item sequences, or it can attempt to infer the grammar also.

A stochastic context free grammar produces the same language as its underlying context free grammar, with the same derivation sequences. The rules are applied to the remaining nonterminal symbols independently of previous rule applications, and are chosen according to the given probabilities. The rule applications are therefore not statistically independent, since a rule may produce a new nonterminal symbol that did not occur in the partially generated sequence.

2.3. Language Model Problems

Problem: using sequence model

A sequence model seems to produce meaningful sequences of items with lengths greatly exceeding the window size, when they are used with a sufficiently large database. However, as the size of the database continues to grow, the lengths of these meaningful sequences of items decrease to some (limiting?) average value.

How close is this limit to the original window size? We claim that it is not much more than the window size.

What is the best mathematical estimate for this limiting value?

Problem: using grammar model

Independent rule application in a stochastic context free grammar does not adequately model either natural or most artificial languages. One way to introduce dependence in the rule applications is to consider partial derivation sequences (independent rule application is the case when sequences are limited to length 1). Then a sequence model in the derivation

sequence space produces a set of rule applications that depend on previous rule applications.
The collection of statistics for this model is also easy, given a sufficiently large set of derivations from the language.
How much better does this model perform when the window size for derivation strings increases?

3. Information Retrieval Models

For the purposes of this note, an information retrieval model relates the information content of a segment of text to a language production model. We will call the basic unit of text a message, and assume as an approximation that each message is independently produced. For example, we might consider a message to be a single sentence from a long segment of text, or it might be as much as a complete newspaper article (say). The choice of message size will certainly have a nontrivial effect on reasonable retrieval strategies, and this fact must be considered for each implementation.

The purpose of the information retrieval system is to accept a query from a user, and return those messages that are estimated by the system to be relevant to the user query, ranked by an estimated value for the relevance. The algorithm used to compute estimated relevance will be based on its value as a predictor of information content, according to the underlying model.

There are basically two kinds of retrieval models, just as there are two kinds of language production models (of concern in this note). The statistical models assume that the information content of a message can be derived from the distributions of its words (or item sequences). The grammatical models augment the same assumption with a more detailed description of the important sequences of items.

We will use the phrase 'content term' to denote any feature of a message that is used to represent its meaning to the retrieval system. Such a feature is usually a word, word stem, or phrase derived from the words in the message, but it may occasionally be some other feature, identified (perhaps) by a special process. It is beyond the scope of this note to describe all the possibilities here. It will suffice to consider the content terms to be particular words from the message.

3.1. Statistical Models

The simple statistical models for information retrieval are based on a stochastic analysis of the query. It is assumed that the query can be adequately described by its set (or sequence) of content terms. It is further (usually) assumed that the word occurrences are independent, so that the comparison of messages to queries is simpler.

The comparisons are statistical, according to one (or more) of a bewildering array of similarity measures. Many of these measures have statistical or computational properties that recommend them, but the most important of these properties is a relation to a classical correlation computation (the vector space terminology of many of the definitions simply obscures this connection, as discussed in reference [LM]). In reference [NM], the reader may find a detailed comparison of many of these measures, and in reference [RB], another interpretation of the connection to statistical correlations.

3.2. Grammatical Models

The grammatical models are distinguished from the sequence models by the application of some kind of grammatical analysis of the query. However, since we are still trying to avoid the semantics, we will restrict our grammatical analysis to a feature recognition role. In other words, the grammatical analysis is used only to identify certain syntactic constructions within the query, so that they can be used as terms in the comparison process.

We are not ruling out a linguistic analysis of a query. We are simply describing some things that can be done without it. We expect that a radical improvement in statistical information retrieval systems will only come by incorporating both general and specific linguistic knowledge, as well as specific world knowledge of the application domain. However, we think that the statistical methods have something to offer in their generality (which is precisely their weak point as well).

3.3. Information Retrieval Problems

Problem: generating simulated databases
Most simulated databases are generated by collecting a set of independent word distribution models together, or assuming a

particular form for the dependencies. Does a stochastic context free grammar provide a better model of the distribution of word occurrences?

Problem: selecting terms
Words make good terms for many kinds of retrieval system, since they are (relatively) easily recognized, and are known to have some relation to actual information content. Are there any other easily described and recognized structures that might make good terms (such as certain kinds of phrases)?

Problem: selecting content terms
Are statistical features sufficient to allow an effective choice of 'important' terms from a preselected set of possible terms?

Problem: computing term associations
Computing relations between terms can often improve retrieval performance. The most common measures of relation compare the distributions of two terms according to various similarity measures.
Is it enough to use occurrence data? What comparison measures for distributions are most effective?
Word order within a message is not required for small message databases. Is it required, or even helpful, for large message sizes?
Can we distinguish different types of association, such as dependencies and adjacencies, from other common occurrence patterns?
For large databases, the comparison of all pairs of content terms can take too much time. How much can the computation be reduced by approximations or other assumptions?

4. Summary

In this note, we have described some current problems with statistical methods for information retrieval from unformatted English text. Some of the problems represent serious obstacles to the improvement of retrieval effectiveness, and some are only curiosities (at present). Most of these problems are derived from a particular retrieval strategy, but they relate to many other strategies based on statistical analysis.

Space-time limitations preclude discussion of several other problems that face designers and implementors of these systems. We do not know how best to access large message collections, though it seems to be agreed that partitioning (but how?) is a useful concept. We do not know how best to relate words to each other at a lexical level, with stemming, or some other inexact matching techniques. We do not know how best to incorporate knowledge as part of the processing, either as data, as models, or as procedures.

Finally, it should be noted that even as the questions were posed from many points of view, so also the answers may be valuable from different points of view, ranging from implementations on one side to theoretical foundations on the other.

5. References

[LM] C. Landauer, C. Mah
Message Extraction Through Estimation of Relevance
Chapter 8 of [OR]

[NM] T. Noreault, M. McGill, M. Koll
A Performance Evaluation of Similarity Measures, Document Weighting Schemes, and Representations in a Boolean Environment
Chapter 5 of [OR]

[OR] R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen, P. Williams (eds.)
Information Retrieval Research
Proceedings of a Conference on Research and Development in Information Retrieval, Cambridge University, 1980
Butterworths, London (1981)

[RB] C. J. van Rijsbergen
A theoretical basis for the use of co-occurrence data in information retrieval
J. Documentation 33, p. 106-119 (1977)