Chapter 38
Validating a Model Predicting Retrieval Ordering Performance with Statistically Dependent Binary Features

Robert M. Losee
University of North Carolina, USA

ABSTRACT

The Information Retrieval Dependence (IRD) model predicts retrieval performance, with some or all dependencies and where there are binary features. Simulations with the Information Retrieval Validation (IRV) software are described that have been used to validate the IRD predictive model, showing that the IRD model accurately or exactly predicts retrieval performance under a variety of conditions. Instead of using traditional research methods using a sample of realistic documents and realistic queries, the authors exhaustively examine all document, query, and relevance combinations within a certain size range. While each of the numbers of components may be small, going through all the permutations of the relevance judgments, terms, and documents produced in one situation 551370 predictions, all of them matching the empirical ordering, suggesting that the predicting method is valid and accurate.

INTRODUCTION

When improving a science, such as Information Retrieval, models that describe what occurs are developed and compared to data by McCullagh (2002) and Shiflet & Shiflet (2014). Models are used in developing further capabilities of scientists to describe what is happening, predict future occurrences, and to understand why the results are occurring. Below, we describe performance models of information retrieval systems and show how a model can be validated, within certain limits, by generating all possible document orderings or term combinations of a certain size. Empirical results showing a range of binary feature dependencies are compared to what is predicted. This enables one to make a claim that a particular scientific model is correct, given the assumptions of the model and the limitations imposed by the time needed to generate result data. This is different

DOI: 10.4018/978-1-4666-9562-7.ch038

than most experimental studies, which determine performance levels associated with certain assumptions or techniques with existing data sets, such as TREC studies, and try to obtain better performance results than certain other methods have achieved with this same data set. These empirical studies use a relatively small set of queries with a set of real documents and assigned relevance judgments. Our study below will examine exhaustively a large number of generated queries, which can range from the thousands to millions of queries, with similar numbers of generated documents, and all the possible permutations of relevance values. While these generated data sets are artificial, as compared to actual natural language queries and documents, the exhaustive generation of all possible query characteristics, document characteristics, and relevance judgments within certain ranges may provide a more rigorous study of all the possible document sets within certain ranges, while at the same time examining large numbers of generated query and document combinations.

There are several basic models of Information Retrieval Systems. Vector systems may treat both documents and queries as vectors, with retrieval decisions being determined from the angle between the query and document vectors. Probabilistic retrieval often emphasizes the probabilities of various factors used in making decisions, such as the probability a document is relevant. Documents with higher expected probabilities of relevance are ranked ahead of those with lower expected probabilities of relevance. Language models suggest that the ranking of documents may be based upon the probabilities that query features are produced given the ordered set of features in each document. Other models include Boolean retrieval, where statements of the characteristics of documents to be retrieved are combined with Boolean operators of *and*, *or*, and *not*. Support Vector Machines and dimensionality expansion techniques modify the system of features so that an improved separation of classes of documents will occur. These and other models all have a wide range of assumptions and parameters that can be studied in order to implement and to improve the performance of the systems.

In the work below, we examine a probabilistic performance model that predicts a particular measure of retrieval performance developed by Losee (1998). This model can predict performance under feature independence assumptions or under feature dependence conditions. Python 3 code has been written and made available at http://ils.unc.edu/~losee/irv that can predict retrieval performance given certain parameters. This Information Retrieval Validation (IRV) software can iterate through all the possible document and term sets of certain sizes and characteristics, and the performance with these synthesized documents and terms can be evaluated and then compared with the predicted results. The predictions produced by the model match the empirical performance, suggesting that the model is accurate, at least within the size constraints tested. Using this IRV software, the effects of making incorrect statistical dependence assumptions or statistical independence assumptions are studied, when trying to make accurate estimations of the degree of dependence between features.

We present a brief discussion of information retrieval performance measures, with emphasis on search length measures such as Expected Precision of Relevant Documents (EPRD). In the next section, methods for analytically predicting EPRD are discussed. Feature dependence is examined, along with a specific technique for predicting high order binary term dependencies. A section on validation procedures is provided, describing the "methodology" used here to show that the formulas developed are correct. The results of the procedure show that the techniques are correct for hundreds of thousands of cases, suggesting that the formulas and technique are valid. The results are interpreted, and a numeric example is provided to help clarify the methodology. Practical uses for this method of predicting performance are discussed.