Large-Scale Multiple Hypothesis Testing in Information Retrieval: Towards a new approach to Document Ranking

Miles Efron
School of Information, University of Texas, Austin
1 University Station D7000
Austin, TX 78703
512-232-5697
miles@ischool.utexas.edu

Abstract

Information retrieval (IR) may be considered an instance of a common modern statistical problem: a massive simultaneous hypothesis test. Such problems arise often in biostatistics where plentiful data must be winnowed to name a small number of potentially "interesting" cases. For instance, DNA microarray analysis requires researchers to filter thousands of genes, searching for genes implicated in a particular condition. This paper describes a novel approach to IR that is based on the notion of simultaneous hypothesis testing. In this case the test is performed on each document and the null hypothesis is that the document is non-relevant. After a mathematical derivation of the proposed model, we test its performance on three standard data sets against the effectiveness of two baseline IR systems, a vector space model and a language modeling-based system. These preliminary experiments show that the hypothesis testing approach to IR is not only philosophically appealing, but that it also operates at the state of the art in effectiveness.

1. Introduction

Recent results in genomics (especially DNA microarray analysis) have brought the problem of massive, simultaneous hypothesis testing to the forefront of the statistical literature [4, 6, 14]. This paper argues that understanding information retrieval (IR) as a hypothesis-testing problem is not only an apt metaphor but also offers novel approaches to document ranking and IR[1].

In DNA microarray analysis, researchers are presented with a potentially huge body of gene expression data. From these expressions, researchers are interested in finding a relatively small number of genes that are "interesting" with respect to a particular condition. This is similar to the IR problem, where a searcher hopes to find a manageable set of documents that are interesting with respect to his or her information need.

[1] In other work [7] we show that the hypothesis testing framework developed here also provides a novel means of conducting and interpreting IR experimentation.

Due to the natural sciences' increasing concern with problems such as these, modern statistics has developed a body of literature on large-scale multiple hypothesis testing [4, 6]. To capitalize on this research it is fruitful to consider IR as a simultaneous hypothesis test on each document d_i where the null hypothesis is that d_i is not relevant to the query q.

We argue that approaching IR as a massive, large-scale hypothesis test constitutes a novel and effective method of ranking documents against a query. After a brief review of large-scale hypothesis testing in Section 2, Section 3 formally derives method for performing IR in a hypothesis testing framework. In Section 4 we offer preliminary experimental results suggesting that our proposed IR model operates at state of the art levels of accuracy. Finally Section 5 concludes with an interpretation of our findings and a treatment of planned research to help understand the relationship between IR and the statistical results drawn on throughout the paper.

2. Motivation

Information retrieval presents problems that are increasingly common in modern data analysis. Particularly in the area of DNA microarray studies, identifying a small number of "interesting" items in an enormous database has become crucial. This paper argues that IR would be well served by taking account of the sophisticated statistical methods developed for such problems, especially those concerned with multiple simultaneous testing of hypotheses.

For contemporary statistical analysis, massive, sparse data of exceedingly high dimensionality is often the norm. The pervasiveness of massive data sets constitutes a sea change in undertaking statistical analysis [4]. However, in IR outsized, sparse data has been the norm for years. Thus IR researchers have suddenly found themselves on the cutting edge of mathematical statistics.

Nowhere is the change faced by modern statistics more evident than in the development of methodology for analyzing DNA microarrays [1, 2, 5, 6, 8, 9, 11]. In microarray research investigators typically search for a small number of genes that are implicated in the onset of a particular condition. To find these genes, researchers must sift through many thousands of non-interesting (null) gene expressions.

From a statistical standpoint, microarray analysis interrogates each gene g against the null hypothesis H_0*: g is not significantly, differentially expressed in affected patients.* In microarray experiments the number of tests performed is often large, on the order of many thousands. To counter spurious "multiplicity effects," these hypothesis tests must be conducted with care.

This paper argues that the apparatus developed to exercise such care is also appropriate in the context of IR. As in microarray analysis, IR presents a "needle in the haystack" problem. Given a searcher's query *q* we should like to find a relatively small number of documents that are relevant. This typically involves searching over a very large body of data, with correspondingly low probability that any one document is relevant.

We may thus understand IR in terms of the following hypothesis testing framework. Given a query *q* and a document *d* we test the hypothesis H_0: d *is not relevant to* q. Rejecting the null hypothesis corresponds to the decision to retrieve a document. For a large corpus, the number of simultaneous hypothesis tests will thus be on the order of thousands, millions, or even billions.

Treating IR in a hypothesis testing framework yields two benefits. First, it allows us to use false discovery rate (FDR) theory for IR evaluation [3, 14]. As described in [7], FDR theory gives a firm probabilistic basis for traditional Cranfield-style evaluation. But perhaps more importantly, FDR theory allows us to proceed with performance evaluation even in the absence of relevance judgments.

Having explored the use of FDR theory for IR in [7], this paper focuses on another benefit of the proposed theory: hypothesis testing lends itself to a new method of document ranking. In traditional IR models such as the vector space approach [12, 13], query-document similarity is measured strictly as a function of the words shared between a searcher's query and each document. Modern approaches to IR such as the family of language modeling approaches [10, 15, 16] include reference to a "collection language model," in many cases yielding a more nuanced estimate of document relevance. In the proposed model we rely explicitly on the question: given a document *d* are the query term expressions in *d* greater than we expect them to be given that *d* is not relevant? To test this hypothesis, we derive a similarity score based on a modified t-statistic. This is innovative and helpful insofar as it accounts explicitly for the null condition, interrogating the *significance* document-query term co-occurrence information.

3. Ranking Documents in a Hypothesis Testing Framework

We begin with a review of a simple hypothesis test for the equality of means of two populations, *X* and *Y.* For simplicity we assume that our samples *x* and *y* are of the same size, *n*. Let $\bar{x}$ be the sample mean and s_x be the sample standard deviation computed from *x*, with analogous notation for the sample *y*. The test statistic is

(1

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2 + s_y^2}{n}}}$$

where the denominator of *t* is simply the standard error of the means. The statistic *t* follows the *t* distribution with *n*-2 degrees of freedom.

During a standard hypothesis test, the distribution of the statistic given in Equation 1 allows us to quantify the likelihood of finding the observed difference between *x* and *y* given that the means of *X* and *Y* are equal.

The motivation for the hypothesis testing approach to IR is analogous to the test described above. Instead of the equality of means, however, for IR we are interested in the null hypothesis: H_0*: document* d *is not relevant to query* q. Instead of comparing means, in the case of IR we will compare evidence for and against relevance.

Let **A** be an *m* document by *n* term matrix. The matrix **A** may contain word frequencies, weights, etc. For now we assume it contains word frequencies. Let $\mathbf{a}_i$ be the *n*-vector for the *i*th document. Also let **q** be an *n*-dimensional query vector corresponding to a query *q,* with non-zero values corresponding to the query terms. We assume the query vector contains query term frequencies. We first define

(2

$$\mathbf{r}_1 = \mathbf{Aq}$$

which is simply the inner product between the query and each document. The magnitude of the r_i, the *i*th element of **r**, quantifies the evidence against the null hypothesis that document *i* is not relevant to the query. This, of course, is the relevance metric used in the vector space model (VSM), without the VSM's length normalization.

However, we can improve our relevance estimate by considering evidence in favor of the null hypothesis. We operationalize this by defining a "null" vector—an *n*-vector containing the expected frequency of each term in a document without regard for that document's topic. In this paper we define the null vector simply as each term's average document frequency. In matrix notation we have

(3

$$\bar{\mathbf{a}} = \frac{1}{m}(\mathbf{1} \cdot \mathbf{A})$$

where **1** is simply an *n*-vector of ones.

Two points are worth noting regarding Equation 3. First, the null vector is query-independent; we need only compute it once for the document collection. Second, under the hypothesis testing framework introduced here, the null condition could be formulated differently. Such formulations will form the basis of future work. The crucial point for the proposed methodology is the presence of a null condition.

Having defined our null vector, we create a "null query" $\mathbf{q}_0 = \mathbf{q} \times \bar{\mathbf{a}}$. The null query contains zero for all non-query words. Words that appear in the query are weighted proportionally to their expected frequency in a document under the null condition (i.e. non-relevance). Based on this definition we calculate

(4

$$\mathbf{r}_0 = \mathbf{A}\mathbf{q}_0 \; .$$

Equation 4 quantifies the evidence in favor of the null hypothesis (non-relevance).

With these definitions in place, we may state our null hypothesis formally: H_0: $\mathbf{r}_1 - \mathbf{r}_0 = \mathbf{0}$. For each document, we compute a statistic to test the hypothesis that it is not relevant to the query. Those documents that have strong evidence against the null hypothesis, we argue, have high likelihood of relevance based on the query.

To derive a *bona fide* test statistic for the null hypothesis we need to define the standard error of $\mathbf{r}_1$ and $\mathbf{r}_0$**.** The squared standard error of the statistic defined in Equation 2 is derived by the sum of squares

(5

$$s_1^2 = \frac{\sum((\mathbf{A}^2\mathbf{q}_1) + (\mathbf{A}\mathbf{q}_1)^2)^2}{m-2}$$

where the sum is taken over the m documents (rows) in **A.** The sum of squares for the null case is defined analogously to that of Equation 5.

The test statistic for our null hypothesis will thus be

(6

$$\mathbf{t} = \frac{\mathbf{A}\mathbf{q}_1 - \mathbf{A}\mathbf{q}_0}{\sqrt{\frac{(s_1)^2 + (s_0)^2}{m}}} \; .$$

The denominator of Equation 6 is simply the pooled standard error.

Alternative formulations of our *t*-statistic are possible, and in future work we shall pursue them. But for now the simple formulation given in Equation 6 is both principled and sufficient for our purposes.

Since the sums of squares in Equation 5 are scalars computed on the query-corpus level, they are constant across all documents for a particular query. For practical purposes, then, we may ignore the denominator and rank documents by the numerator of Equation 6.

Each element *i* of the *m*-vector **t** quantifies the evidence against the *i*th null hypothesis: document *i* is not relevant to the query. Because the number of tests is high (equal to the number of documents), and because we expect the elements of **A** not to follow a Gaussian distribution, we cannot go so far as to use our t-statistic to derive a traditional *p*-value. Instead, evaluation of the probability of the null hypothesis must be approached with more appropriate inferential tools such as the false discovery rate, as discussed in [7]. But for the purposes of *ad hoc* IR, it suffices to rank documents in decreasing order of their magnitude on *t*.

The ranking statistic *t* defined in Equation 6 is similar to other IR models such as the vector space and language modeling approaches two main senses. In all of the major IR models, evidence of a document's relevance begins by identifying the words it shares with the query. This is also the case in our model, where Equation 2 exerts this influence.

As in the Kullback-Leibler (KL) IR model (a variant of the language modeling framework), the numerator of Equation 6 conditions the query-document similarity against evidence that the document is similar to a null condition. In the KL model we estimate the "distance" between the query model and the "collection model." In the hypothesis testing approach we analyze the magnitude of the difference between query-document similarity and the similarity between the document and the "null query," which is quite similar to the notion of a collection model.

The second function of Equation 6's numerator is as a form of length normalization. If a particular document *i* is long, the *i*th row of **A** is apt to have large values. The simple dot product between the query vector and the *i*th row vector will thus be large (assuming the document and query share terms). Using Equation 2 to rank document thus favors long documents. However, if document *i* has large values on the query terms, the magnitude of those terms in Equation 4 will also be high, and so the subtraction in Equation 6's numerator will dampen the effect of document length.

4. Experimental Evaluation

To test the proposed IR model we compared its performance to two other models on three data sets. Data analysis followed the standard Cranfield model, using precision and recall—especially mean average precision (MAP)—to gauge system effectiveness.

4.1 Data Sets

We compared IR performance over three data sets shown in Table 1. Medline[2] and CACM[3] both consist of academic articles' titles and abstracts. Reuters-21578[4] contains approximately 20,000 newswire articles, each of which is tagged by human editors with topical descriptors. Reuters queries were created by choosing the *topic* tags for each article. Each topic assigned to at least ten documents was used as a query—55 queries in all. Documents in the Reuters corpus were considered relevant to a query if they were tagged with that query term.

Table 1. Test Collections used for Experimentation

	# Docs	# Queries
medline	1033	30
CACM	3204	64
Reuters	21,578	55

4.2 Baseline Models

To assess the performance of the proposed method we compared its effectiveness to two other methods: a simple vector space model using TF-IDF weighting [12, 13] and a language modeling system using Jelinek-Mercer smoothing [10, 15, 16]. Both of these baseline models have been shown to be effective for *ad hoc* retrieval, with the language modeling approach constituting the current state of the art.

No stemming was applied to documents during indexing, nor was a stoplist used. The TF-IDF model was tested using both normalized document vectors and non-normalized vectors. Our TF-IDF results report runs using the best-performing normalization (normalized for medline; non-normalized for CACM and Reuters).

4.3 Results

Figure 1 and Table 2 summarize the results of our initial experiments. Each panel of Figure 1 shows precision-recall curves (over eleven recall points) for the models described above on a given data set. From the figure it is clear that the simple TF-IDF model did not perform as well as the language modeling system or the hypothesis testing model. The difference is especially stark on the larger data sets, CACM and Reuters, where TF-IDF's precision-recall curve is consistently below those of the other models.

Both Figure 1 and Table 2 suggest that there is very little difference between the language modeling system and our hypothesis testing model with respect to

[2] http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/medl/

[3] http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cacm/

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578

precision and recall: the precision-recall curves of these models are nearly identical.

Figure 1. Precision and Recall for Three Models and Three Corpora

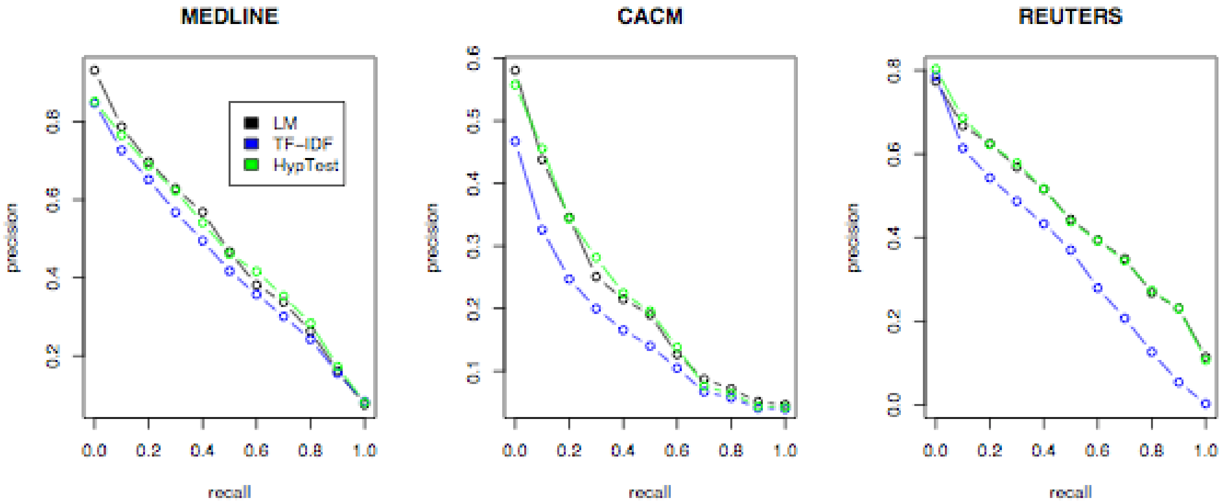

Table 2 shows in boldface the best-performing model for each corpus according to mean average precision (MAP). On CACM and Reuters the hypothesis testing system narrowly outperformed the language modeling system.

Table 2. Mean Average Precision for Three Models and Three Corpora

	Medline	CACM	Reuters
Lang. Model	**0.469**	0.198	0.431
TF-IDF	0.416	0.149	0.34
Hyp. Testing	0.46	**0.2**	**0.438**

The differences between the language modeling system and the hypothesis testing model were narrow, a fact reinforced by statistical testing. Undertaking one-sided, paired (i.e. comparing each query's average precision for each model) t-tests on the precision of the systems yielded *p*-values of 0.263, 0.535, and 0.847 for medline, CACM, and Reuters, respectively. Sign tests on the query by query performance yielded similarly high *p*-values, suggesting that there is no statistically significant difference between the average precision on these corpora using a language modeling framework or the hypothesis testing model proposed here.

5. Discussion and Conclusions

Our experimental results suggest that the hypothesis testing IR model proposed in this paper operates at a level comparable to the state of the art. However, our experiments do not offer the final word on this matter. In future work we propose two experimental changes and several modifications to the model itself.

First, we will compare the hypothesis testing model to other IR models on larger, more heterogeneous data sets (several of the TREC collections, specifically). Though we have no reason to suspect that larger data sets will show defects in the model, more realistic testing will surely point out the model's limitations.

Additionally, we only compared our model to Jelinek-Mercer smoothing with a single mixing parameterization (lambda=0.5). In future work we will test our method against other IR models. With respect to language modeling systems we will add Dirichlet smoothing to the experiment. We will also test the hypothesis testing method against the standard Okapi probabilistic model.

From a theoretical standpoint, we plan to undertake research to define the elements of Equation 6 (our basic model) more rigorously. In this paper we have taken a simple interpretation of the *t*-statistic used to rank documents. In upcoming work, however, we expect to improve the method of defining the null condition. With this will come an alternative formulation of the standard error. We suspect that a more sophisticated null will improve the performance of the proposed model.

With these caveats in mind, we conclude by noting that the proposed model appears to operate at a level of accuracy comparable to the best modern IR methods. The influence of considering the null case during retrieval is the hallmark of the proposed model. Accounting for evidence in favor of the null (non-relevance) conditions relevance predictions on the relative frequency of words in the collection and of the length of documents. But perhaps more provocatively, the model as a whole constitutes a novel way of thinking about IR—as a massive simultaneous hypothesis test. As we continue to work on the model (especially with respect to maturing the null condition and its standard error) we plan to pursue the statistical benefits of this perspective.

6. References

1. Allison, D., et al., *A mixture model approach for the analysis of microarray gene expression data.* Computational Statistics and Data Analysis, 2002. **39**: p. 1-20.
2. Aubert, J., et al., *Determination of the differentially expressed genes in microarray experiments using local FDR.* BMC Bioinformatics, 2004. **5**(25).
3. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society, 1995. **57**(1): p. 289-300.
4. Donoho, D., *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, in *Math Challenges of the 21st Century*, A.M. Society, Editor. 2000.

5. Dudoit, S., J. Shaffer, and J. Boldrick, *Multiple hypothesis testing microarray experiments.* Statistical Science, 2003. **18**: p. 71-103.
6. Efron, B., et al., *Empirical Bayes methods and false discovery rates for microarrays.* Genetic Epidemiology, 2002. **23**: p. 70-86.
7. Efron, M. and B. Efron, *False discovery rates for evaluating information retrieval systems.* Under review, 2008. (available online at http://www.ischool.utexas.edu/~miles/papers/efronAndEfron.010.pdf)
8. Liao, J., et al., *A mixture model for estimating the local false discovery rate in DNA microarray analysis.* Bioinformatics, 2004. **20**: p. 2694-2701.
9. Newton, M., et al., *On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.* Journal of Computational Biology, 2001. **8**: p. 37-52.
10. Ponte, J.M. and W.B. Croft, *A Language Modeling Approach to Information Retrieval.* Research and Development in Information Retrieval, 1998: p. 275--281.
11. Pounds, S. and S. Morris, *Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of the p-values.* Bioinformatics, 2003. **19**: p. 1236-42.
12. Salton, G. and M.J. McGill, *Introduction to Modern Information Retrieval*. 1983, New York: McGraw Hill.
13. Salton, G., A. Wong, and C.S. Yang, *A vector space model for automatic indexing.* Commun. ACM, 1975. **18**(11): p. 613-620.
14. Storey, J., *A direct approach to false discovery rates.* Journal of the Royal Statistical Society, 2002. **Ser. B**(64): p. 479-498.
15. Tao, T., et al. *Language Model Information Retrieval with Document Expansion*. in *Human Language Technology Conference of the North American Chapter of the ACL*. 2006. New York.
16. Zhai, C. and J. Lafferty, *A Study of Smoothing Methods for Language Models Applied to Information Retrieval.* ACM Transactions on Information Systems, 2004. **2**(2): p. 179--214.