Inferring Probability of Relevance Using the Method of Logistic Regression

Fredric C. Gey
UC Data Archive and Technical Assistance
University of California, Berkeley USA
e-mail: gey@sdp2.berkeley.edu

Abstract

This research evaluates a model for probabilistic text and document retrieval; the model utilizes the technique of *logistic regression* to obtain equations which rank documents by probability of relevance as a function of document and query properties. Since the model infers probability of relevance from statistical clues present in the texts of documents and queries, we call it *logistic inference*. By transforming the distribution of each statistical clue into its standardized distribution (one with mean $\mu=0$ and standard deviation $\sigma=1$), the method allows one to apply logistic coefficients derived from a training collection to other document collections, with little loss of predictive power. The model is applied to three well-known information retrieval test collections, and the results are compared directly to the particular vector space model of retrieval which uses term-frequency/inverse-document-frequency (tfidf) weighting and the cosine similarity measure. In the comparison, the logistic inference method performs significantly better than (in two collections) or equally well as (in the third collection) the tfidf/cosine vector space model. The differences in performances of the two models were subjected to statistical tests to see if the differences are statistically significant or could have occurred by chance.

1. Introduction

We can consider the central task of Information Retrieval to be that of providing intellectual access to the contents of a collection of texts or documents. The heart of this task is the concise representation of the meanings of the texts contained in the collection. In the vector space model developed by Gerard Salton and associates at Cornell University [1] [2] [3] both documents and queries are represented as vectors in the "m"-dimensional term space assuming that the indexing vocabulary of nontrivial terms is of size "m". For this model, we have a clear geometric interpretation for both the query and document and can choose distance measures which measure the degree to which the document vector deviates from the query vector. One such well-know measure computes the cosine of the angle between document and query vector. In the sparsest form of the vector space model, query vectors and document vectors only indicate the presence or absence of a vocabulary term in either the query or the document. This can be easily generalized by adding weights to the query or document. These weights, interpreted geometrically, merely show the magnitude and direction of the appropriate vectors.

Weights may be mechanically derived from attributes of terms. For example, for terms in the query or the document, the attribute 'absolute frequency' is the count of occurrences of that term in either the query (QAF) or the document (DAF). For terms t_j used in documents, the attribute 'inverse document frequency' (IDF) is the ratio, $\frac{N}{n_{t_j}}$ where N is the number of documents in the collection and n_{t_j} is the number of documents for which term t_j occurs in the document's text or has been used to index the document. The attribute IDF, which is usually logged, was first suggested by Karen Sparck-Jones [4]. Two well-known weights, used in the SMART retrieval system, are the term frequency for the query terms (QAF), and, for the document terms, the term frequency- inverse document frequency product $(DAF \cdot IDF)$. Salton and Buckley in 1988 identified a number of these factors and computed performance measurements for multiple weighting schemes against several test collections. Their experiments show that a simple retrieval weight, which only accounted for the occurrence of the term in the document and the occurrence of the term in the

By a probabilistic search system we mean one whose query specification, indexing processes, and retrieval rules are derived from the formal application of the theory of probability to the logic of the information retrieval problem. Such a system will return documents to the user in an order ranked by probability of satisfying the user's information need. The primary focus of this paper is to test a probabilistic search system which is both theoretically sound and practical. Probabilistic models of text and document retrieval attempt to place the computation of retrieval status value on a sound theoretical footing. If retrieval status value can be computed as the probability of satisfying the user's information need, then the probability

query, performed worst in their retrieval comparisons [5].

ranking principle [6] states that the optimal retrieval will be achieved if documents are returned to the user in order of their probability of relevance. This leads to the so-called "Binary Independence" model of probabilistic retrieval [7], which was generalized to the "Linked Dependence model by Cooper [8] Unfortunately experiments with this relevance feedback approach have encountered problems of insufficient statistical evidence to predict relevance, even when half the collection is used as a training set. Improved methods for statistical estimation in the face of this insufficient sample size have only been partially successful [9].

2. The logistic inference model

If no assumptions are made about the character of the probabilities of relevance which associate between a query-document pair $r=f(q_i,d_j)$, but we merely allow that a text retrieval system exists and that associated with each query term t_{q_i} and document term t_{d_j} there are a finite set of attributes about the term (v_1, \cdots, v_n) from the query (such as count of occurrences of the term in the query), from the document (such as count of occurrences of the term in the document), and from the collection (such as the total number of documents in the collection divided by the number of documents indexed by the term), and that attribute values vary statistically with the collection, then we have a "Model 0" system, i.e. one which exists before introducing particular models.

The logistic inference model says that we can use a random sample of query-document-term triples for which binary relevance judgments have been made, and compute the logarithm of the odds of relevance for term t_k which is present in both document d_i and query q_i by the formula:

$$\log O(R \mid q_i d_i t_k) = c_0 + c_1 v_1 + \cdots + c_l v_n$$

and further that the logarithm of the odds of relevance for the i^{th} query $q_i = \langle t_1, \cdots, t_q \rangle$ is the sum of the logodds for all terms:

$$\log O(R \mid q_i, d_j) = \sum_{k=1}^{q} [\log O(R \mid q_i, d_j, t_k) - \log O(R)]$$

where O(R), known as the *prior odds of relevance* is the odds that a document chosen at random from the collection will be relevant to query q_i . The coefficients of individual query and document term properties are derived by using the method of *logistic regression* [10] which fits an equation to predict a dichotomous independent variable as a function of (possibly continuous) independent variables which show statistical variation.

Once log odds of relevance is computed from a simple linear formula, the inverse logistic transformation may be applied to directly obtain the probability of relevance of a document to a query:

$$P((R \mid q_i, d_j) = \frac{1}{1 + e^{-\log(O(R \mid q_i, d_j))}}$$

Once the coefficients of the equation for logodds of relevance have been derived within a document retrieval system for a particular collection from a random sample of query-document-term-relevance quadruples, these coefficients may be used to predict odds of relevance for other query-document pairs (past or future).

Regression models which approximate the dependent variable, document *relevance* by linear combinations of multiple clues such as term frequency, authorship, and co-citation, were first introduced, with some success by Ed Fox [11]. More recently Fuhr and Buckley [12] [13] used polynomial regression to approximate relevance. There are two well-known major problems with the application of ordinary regression approaches to probability of relevance approximation:

- The outcome variable is dichotomous rather than continuous, hence
- the error distribution is binomial rather than normal, as is usually assumed by ordinary regression.

An alternative approach is to use Bayesian inference networks as developed by Turtle and others [14] [15] [16].

3. Sampling for logistic regression

One of the computational problems in using logistic regression is the computational size necessary to compute logistic regression coefficients. Even supposing that only half the document collection is used as a

training set the number of query-document-term triples which are to be used for computation becomes astronomical. Naturally, a sampling approach is necessary. Since the number of relevant documents is, for each query, naturally a very small fraction of the total documents in the collection, it is incumbent upon us to attempt to construct a sample which encompasses a significantly higher proportion of relevant documents than non-relevant ones. To assure adequate representation of relevant documents, we would sample most, if not all, of the relevant documents while sampling a significantly smaller proportion of non-relevant documents, and adjusting for non-relevant documents in the regression process by a weight factor which compensates for the differential. Since relevant documents are typically 1/100th fewer than non-relevant documents, it is reasonable to take into the sample, say, every thirtieth non-relevant query-document-term triple and factor it into the logistic regression computation by applying a weighting factor of 30.

We have chosen the Cranfield test collection upon which to fit our model. The Cranfield collection is a long-standing collection used in information retrieval for numerous experiments beginning with the Cranfield's experiments. It does have some deficiencies as outlined by Swanson [17], but it offers the following desirable properties. First, it has 225 queries, which seems to be a sufficient number to afford a reasonable fit to our logistic model. Second, while it only has 1,400 documents, both the documents and the queries seem to possess a better fit to the characteristics of an ideal test collection.

The characteristics of an ideal test collection should be:

- queries are in natural language form, so that query statistics can be collected as well as document statistics,
- the documents have both titles and abstracts present (abstracts are missing from some of the CACM collection),
- The queries are a random sample from some larger query population. There must be a sufficiently large number of queries in order to achieve some statistical significance.
- Documents are a random sample from a larger document collection and,
- That relevance judgments be made for all query-document pairs in the collection. If this is not possible (as it cannot be for million document collections), then the relevance judgments are a carefully crafted sample of all possible relevance judgments, and that the sample size be known.

We will later apply our fit to the CACM collection (52 queries, 3204 documents) and the CISI collection (76 queries, 1460 documents)

4. Logistic regression for six term attributes

Our particular logistic inference model has the following additional elementary clues: relative frequency in the query (QRF), relative frequency in the document (DRF) and relative frequency of the term in all documents (RFAD). The complete formula for logodds of relevance, given the presence of term t_i is then:

$$\begin{split} Z_{t_j} &= \log O\left(R \mid t_j\right) = c_0 + c_1 log\left(QAF\right) + c_2 log\left(QRF\right) \\ &+ c_3 log\left(DAF\right) + c_4 log\left(DRF\right) + c_5 log\left(IDF\right) + c_6 log\left(RFAD\right) \end{split}$$

Query absolute frequency (QAF) has been previously defined. Query relative frequency (QRF) is equal to query absolute frequency (QAF) query divided by the total number of term occurrences of all terms in the query. Similarly, document relative frequency (DRF) is document absolute frequency divided by the total number of occurrences of all terms in the document. Relative frequency in all documents (RFAD) is the total number of term occurrences of the term in the collection divided by the total number of occurrences of all terms in the entire collection, i.e., collection length.

You will note that all terms in the formula are logged. One reason for using logarithms is to dampen the influence of frequency information. It is unrealistic to assume that 50 occurrences of a term in a document is five times more important than 10 occurrences. Another reason for taking the logs is that the logarithm, in general, smooths out a skewed distribution to one which has nice behavioral properties. Finally, sums of logged variables behave as products of the same variables after an anti-logarithm is applied. Indeed, in comparing the fit for logged and non-logged items, in general (at least for these clues) a higher maximum likelihood is attained for logged clues than for raw clues.

The fit for this extended clues logistic model is done against the same 20,246 query-document-term sample taken from the Cranfield collection. The following coefficients are then derived from this fit:

$$Z_{t_i} = \log O(R \mid t_i) = -0.2085 + -0.2036 \log(QAF) + 0.19143 \log(QRF)$$

$$+0.16789log(DAF) + 0.57544log(DRF) - 1.5967log(IDF) + 0.75033log(RFAD)$$

What remains is to compute log(O(R)) the prior odds of relevance for this collection. This is estimated by counting the number of query-document pairs for which a judgment of relevance has been assigned (for the Cranfield collection this is 1838 pairs) and then dividing by the total number of query-document pairs (for Cranfield 225 queries times 1400 documents). Hence:

$$prior = \frac{1838}{225 \cdot 1400} = .005835$$

becomes the prior probability of relevance, and

$$logprior = log \frac{prior}{1-prior} = -5.138$$

and the final formula for ranking is:

$$log(O(R \mid \overrightarrow{Q})) = -5.138 + \sum_{j=1}^{q} (Z_{t_j} + 5.138)$$

5. Performance evaluation: statistical significance tests

A key part of this paper is to compare performances of two major methods: our logistic regression probabilistic method, which ranks documents according to probability of relevance against the vector space model, which ranks documents according to the cosine of the angle between the query vector and the document vector. Performance evaluation is usually described in terms of the recall and precision measures. **Recall** is the fraction of relevant documents retrieved at a certain retrieval point in the retrieval process while **precision** is the fraction of retrieved documents which are relevant. These measures can be computed for a single query but are usually averaged over all queries in the collection. It is our intent to use recall and precision as one measure of performance; however, it is our intention to also subject performance differences to a statistical test to determine whether the difference is significant or could have occurred by chance.

The reason one wishes to test statistical significance is that on a particular query one method may be better than another and yet the performance could be reversed for the next query posed. In our model of retrieval, we assume that a query is chosen at random, and our method ranks documents according to probability of relevance to the randomly chosen query. It might be that next set of randomly chosen queries would have an entirely different performance profile. Thus, if we are attempting to compare the performance of two methods (such as logistic inference versus tfidf/cosine vector space) we need to establish a statistical test which will show that the results obtained for the randomly chosen set of queries are statistically significant.

Such a test, well-known in the literature, would be a T-test which would compare the differences between performance for each query, and we determine whether these differences in ranking or in precision are statistically significant. Our approach, therefore, for testing for statistical significance will be:

- For each query take the ranking of relevant documents by each method and compute an average precision value over all ranks (all levels of recall). This will give a unique number for each query for each method.
- Compute the difference in average precision between the two numbers found with each method for the two methods being compared. Thus, in applying the T-test, the sample size will equal the number of queries in the test collection.
- Apply a T-test to these differences under the null hypothesis that the methods perform identically and hence their mean difference should be zero and their standard deviation not significantly different.

[18] provides a summary of possible statistical tests which might be used to evaluate retrieval experiments.

6. Logistic model performance for the Cranfield collection

For the logistic inference model on Cranfield the following table

Logistic Inference versus tfidf/cosine Vector Space Performance				
Cranfield Collection: Averages over 225 queries				
	Logistic	Vector Space		
Recall	Precision	Precision		
0.00	0.8330	0.7787		
0.10	0.8116	0.7440		
0.20	0.7129	0.6434		
0.30	0.6021	0.5301		
0.40	0.5161	0.4380		
0.50	0.4503	0.3814		
0.60	0.3698	0.2994		
0.70	0.2859	0.2267		
0.80	0.2280	0.1882		
0.90	0.1640	0.1379		
1.00	0.1464	0.1251		
11-pt Avg:	0.4655	0.4084		
% Change:		-12.3		

displays the recall-precision performance results. If we further compute the difference in average precision for each of the 225 queries for the two methods (cosine and logistic) we find that the results are statistically significant at the one tenth of one percent significance level:

Hypothesis Test: two-tail t on Precision Differences between Logistic inference and Tfifdf/cosine methods						
N	null mean	sample mean	sample SE	df stat	test	P
225	0.0000	-0.059885	0.0072063	224	-8.31	.0000

Thus we can see that making use of the full probabilistic model of logistic regression on the set of six term property clues, we obtain a statistically significant improvement over the tfidf/cosine vector space model.

7. A criticism of the Cranfield results

The above results obtained for the logistic inference model have a fundamental difficulty. The entire set of queries, documents and terms from the relevant set is included within the sample used as the training set for the logistic regression. Naturally, when the resulting logistic coefficients are used to rank documents based on relevance, we have a chicken and egg process, whereby the test is applied to the sample which was used for fitting. In past tests of probabilistic models where this approach has also been used [19], the next step in testing such models has been to train on half the number of queries and then use that training to predict the results of the other half of the queries. We feel that a large number of queries is necessary in order to train for a decent logistic regression approach to document retrieval. Our approach, instead, will be to apply the coefficients obtained from using the Cranfield collection as a training set to other collections in our test repertoire, in particular the CACM and CISI collections.

8. Logistic inference for CACM

If we apply the logistic regression equation fitted for the Cranfield sample to the CACM collection, we obtain the following result:

Cranfield Logistic versus Tfidf/cosine Vector Space Performance CACM Collection: Averages over 52 Queries				
Recall Logistic Vector Precision Prec				
11-pt Avg: % Change:	0.3322	0.3148 -5.2		

While it may seem that the extended logistic model performs 5.2 percent better than the tfidf/cosine vector space model for the CACM collection, if we apply the same hypothesis test:

CACM Hypothesis test: 2-tail T on Precision Differences between Logistic inference and Tfidf/cosine methods						
N	N null sample sample df test P					
	mean	mean	SĒ		stat	value
52	0.0000	-0.020528	0.015866	51.0	-1.29	.2016

we find again that this difference of performance is not statistically significant and hence we cannot reject a null hypothesis that, in fact, given another set of queries the tfidf/cosine vector space model might perform as well or better than the extended logistic model.

9. Cranfield logistic model for CISI

Applying the Cranfield fitted regression equation to the CISI collection, we find:

Cranfield Logistic versus Tfidf/cosine Vector Space Performance CISI Collection: Averages over 76 Queries				
Logistic Vector Space Recall Precision Precision				
11-pt Avg: % Change:	0.2088	0.2137 4.4		

While this seems to indicate better vector space performance, when we apply the statistical test on the average precision and recall measures for 77 CISI queries:

Hypothesis Test: 2-tail T on Precision Differences between Logistic inference and Tfidf/cosine for CISI collection						
N	null	sample	sample	df	test	P
	mean	mean	SE		stat	value
76	0.0000	0.0056973	0.0087791	75.0	0.65	.5183

we find that there can be no statistically significant difference between performance on the vector space model as against the extended logistic model. While the application of the Cranfield-derived coefficient directly to both the CACM collection and the CISI collection produce comparable results on these collections to the vector space model, for neither collection can we find the difference between vector space and extended logistic to be statistically significant.

The blind application of coefficients derived from the statistical distribution of one collection to another collection clearly leaves something to be desired. In the next section, we apply some clever thought to the transformation of coefficients from one collection, in order to adapt them to the statistical clues of the other collections.

10. Standardized variables

The direct application of coefficients derived from a query-document-term sample from one document collection and one set of queries to another document collection and yet another set of queries makes the following assumption. It assumes that all the clues which are being used as sensitive indicators of probability of relevance have the identical statistical distribution from collection to collection. By identical statistical distribution we mean not only do the clues come from the same underlying probability distribution, but they also have the identical mean and standard deviation for each clue as the original collection on which the logistic regression fitting has been done.

Even though each clue might derive from the same underlying probability distribution, it seems highly unlikely that, from collection to collection, there would be no variation in either the mean and standard deviation of the clue being used for indication of probability of relevance. This gives us insight into the possible adaptation of coefficients which fit one collection to another collection. It is well known that one can take the mean and standard deviation of any statistical distribution and transform that distribution (even though the actual underlying probability density function is not known) into a standardized distribution, i.e., one which has mean zero ($\mu = 0$) and standard deviation one ($\sigma = 1$).

If we obtain coefficients for such a standardized distribution, then we can also apply the coefficients of this standardized distribution to compute coefficients for the standardized distribution of yet another collection which uses the same clues to indicate probability of relevance. The underlying assumption of such an adaptation is the following:

• For each clue the underlying probability distribution of that clue remains identical from

collection to collection changing only in its mean and standard deviation, i.e., only the mean and standard deviation of the particular collection-dependent distribution will change while the underlying standardized probability distribution remains constant over all collections.

One might well question this assumption on the following grounds:

• Suppose the mean μ_i and standard deviation σ_i for clue i varies meaningfully from collection to collection, and that these differences have predictive power with respect to relevance. Then the application of standardization of distributions will remove this predictive power.

In the face of such thinking, the best approach is to experimentally test the assumption. If we accept, for purposes of testing, the assumption and we derive the standardized coefficient for a first collection (in our case the Cranfield collection) we can directly compute the adjusted coefficient for the standardized distribution of any other collection to which we wish to apply our logistic regression technique. In terms of formulae, we obtain the coefficients c_0 , c_1 , and c_2 for the equation (here we assume two clues, x and y):

$$log(O(R \mid x_1, y_1)) = c_0 + c_1(\frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}) + c_2(\frac{y_1 - \mu_{y_1}}{\sigma_{y_1}})$$

Given the equation above, for the source collection, if we then look at the equation for the target collection, assuming that we would apply the same coefficients to a standardized distribution, we can then derive the equation for the second collection. If the standardized distribution remains constant from collection to collection, we have for collection two, the same equation,

$$log(O(R \mid x_2, y_2)) = c_0 + c_1(\frac{x_2 - \mu_{x_2}}{\sigma_{x_2}}) + c_2(\frac{y_2 - \mu_{y_2}}{\sigma_{y_2}})$$

and we wish to derive the coefficients c_0' , c_1' , and c_2' for the equation:

$$log(O(R \mid x_2, y_2)) = c_0' + c_1'x_2 + c_2'y_2$$

but if we multiply out the first equation above we obtain

$$log(O(R \mid x_2, y_2)) = c_0 - c_1 \frac{\mu_{x_2}}{\sigma_{x_2}} - c_2 \frac{\mu_{y_2}}{\sigma_{y_2}} + \frac{c_1}{\sigma_{x_2}} x_2 + \frac{c_2}{\sigma_{y_2}} y_2$$

and hence

$$c_0' = c_0 - c_1 \frac{\mu_{x_2}}{\sigma_{x_2}} - c_2 \frac{\mu_{y_2}}{\sigma_{y_2}}$$
$$c_1' = \frac{c_1}{\sigma_{x_2}}$$

and

$$c_2' = \frac{c_2}{\sigma_{y_2}}$$

Thus to find the logistic coefficients for the target collection, we need only determine the mean, μ_{ν} , and standard deviation, σ_{ν} for each clue ν , compute the new coefficients, and apply them directly to clue values for the new collection.

This methodology was first applied by Cooper, Gey and Chen [20] to the queries of the NIST text retrieval conference (TREC1) [21]. This paper provides a more complete analysis of the method.

11. Applying to CACM and CISI collections

In order to obtain means and standard deviation for all clues for the new collection, we must obtain a new sample from each of those collections. If we obtain a small sample (of 1 in 125) query document term

triples and all associated clues values, we can compute the following standard deviations for CACM and CISI.

Coll.	st.	log q.af	log d.af	log q.rf	log d.rf	log idf	log rfad
CISI	mean	0.3031	0.4093	-3.5450	-3.7922	1.9045	-5.5303
CISI	std dev	0.5153	0.5795	0.9110	0.6944	0.8333	1.0675
CACM	mean	0.1763	0.3900	-2.7644	-3.4976	2.0968	-5.2434
CACM	std dev	0.3347	0.5855	0.5612	0.7999	1.1378	1.2146

We then use these means and standard deviations together with the Cranfield coefficients to obtain the following new logistic or logodds coefficients for the CACM and CISI collections.

Coll.	const.	log q.af	log d.af	log q.rf	log d.rf	log idf	log rfad
CRAN	-4.125	-0.03229	0.1059	0.06910	0.4193	1.4515	0.7734
CACM	-1.341	-0.08412	0.1520	0.11993	0.5391	1.0933	0.5507
CISI	-0.934	-0.061660	0.1622	0.0663	0.5572	1.5325	0.6680

If we then apply the fit in order to obtain rankings for all queries for both the CACM and CISI collections, we obtain first for CACM the following recall-precision results:

Standardized Logistic vs. Tfidf/cosine Vector Space Performance CACM Collection: Averages over 52 Queries					
	Standardized Vector Space				
Recall	Logistic Precision	Precision			
0.00	0.7452	0.6985			
0.10	0.6040	0.5683			
0.20	0.5338	0.4830			
0.30	0.4392	0.3887			
0.40	0.3683	0.3210			
0.50	0.3128	0.2761			
0.60	0.2682	0.2306			
0.70	0.1846	0.1770			
0.80	0.1388	0.1441			
0.90	0.0961	0.1018			
1.00	0.0694	0.0735			
11-pt Avg:	0.3419	0.3148			
% Change:		-7.9			

Note that the average precision has increased approximately another 3 percent over the tfidf cosine vector space method of ranking documents. The recall-precision table makes it quite clear that except for the tail of recall-precision, the standardized extended logistic model performs substantially better than the tfidf/cosine vector space model for the CACM collection.

Indeed, if we once again go through our process of computing average precision from all points of recall for each of the 52 significant queries for CACM, and then do a statistical test of the difference from standardized logistic method and tfidf/cosine vector space average precision, we find this improvement is now statistically significant at the 2 percent level.

	CACM Hypothesis Test: 2-tail T on Precision Differences					
	between Standardized Logistic and Tfidf/cosine					
N	null	sample	sample	df	test	P
	mean	mean	SE		stat	value
52	0.0000	-0.030180	0.012335	51.0	-2.45	.0179

That is to say, we would only expect in less than two percent of all samples of queries applied to the CACM collection, for the vector space model to perform better than the standardized logistic model. In this way, we can reject the null hypothesis that there is no difference in the performance in the CACM collection and accept the clear indication that standardized logistic regression performs better than the vector space model for the CACM collection.

On the other hand, if we apply this correction factor and standardization to the CISI collection, we find that there is a slight worsening in the performance results.

Standardized Logistic vs. Tfidf/cosine Vector Space Performance CISI Collection: Averages over 76 Queries			
Standardized Vector Space Recall Logistic Precision Precision			
11-pt Avg: % Change:	0.2042	0.2137 4.7	

When comparing this result with the non-standardized extended logistic model for CISI we see that the precision changes occur in the third decimal place. We can conclude that the difference between the two is certainly not statistically significant, nor is that between CISI standardized and tfidf/cosine. Among reasons for this failure to achieve performance improvement in the CISI collection are the different prior probability of relevance for the collections. [22] gives more detail on the sensitivity of the model to the proper estimation of prior probability of relevance.

12. Conclusions

In this research we have investigated a new probabilistic text and document search method based upon logistic regression. This *logistic inference* method estimates probability of relevance for documents with respect to a query which represents the user's information need. Documents are then ranked in descending order of their estimated probability of relevance to the query.

- 1. The *logistic inference* method has been subjected to detailed performance tests comparing it (in terms of recall and precision averages) to the traditional tfidf/cosine vector space method, using the same retrieval and evaluation software (the Cornell SMART system) for both methods. In this way the test results remain free of bias which might be introduced from different software implementations of evaluation methods. In terms of recall and precision averages, the logistic inference method outperforms the tfidf/cosine vector space method on the Cranfield and CACM test collections. The methods seem to perform equally well on the CISI test collection (for an appropriately estimated prior probability).
- 2. Statistical tests have been applied to ascertain whether these performance differences between the two methods are statistically significant. The performance improvement of the logistic inference method over the tfidf/cosine vector space method for the Cranfield and CACM collection is statistically significant at the five percent level. Performance differences for the CISI collections are not statistically significant, for the most plausible estimate of prior probability.
- 3. The use of standardized variables (statistical clues standardized to mean $\mu=0$ and standard deviation $\sigma=1$) seems to enable the training of and fitting for logistic regression coefficients to take place on the queries and documents of one collection and to be applied directly to the queries and documents of other collections.

13. Acknowledgments

The work described was part of the author's dissertation research at the School of Library and Information Studies at the University of California, Berkeley. Many of the ideas contained herein were jointly formulated with my dissertation advisor, Professor William S. Cooper, whose clarity of thinking and relentless persistence made this research both possible and doable. My outside committee member, Professor of Biostatistics Steve Selvin, provided much helpful advice on statistical techniques.

References

- 1. Salton G et al. The SMART retrieval system: Experiments in automatic document processing. Prentice-Hall, Englewood Cliffs, NJ, 1971
- 2. Salton G. Text processing: the transformation, analysis and retrieval of information by computer. Addison Wesley, Reading, MA-Menlo Park, CA, 1989
- 3. Salton G, McGill M. Introduction to modern information retrieval. McGraw-Hill, New York, 1983
- 4. Sparck-Jones K. A statistical interpretation of term specificity and its application in retrieval. Journal

- of Documentation 1972; 28:11-21
- 5. Salton G Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management 1988; 24:513-523
- 6. Robertson, S. The probability ranking principle in IR. Journal of Documentation 1977; 33:294-304
- 7. Robertson S Sparck-Jones K. Relevance weighting of search terms. Journal of the ASIS 1976; 27:129-145
- 8. Cooper W. Inconsistencies and misnomers in probabilistic IR. In: Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, Ill, Oct 13-16, 1991, pp 57-61
- 9. Fuhr N Huther H. Optimum probability estimation from empirical distributions. Information Processing and Management 1989; 25:493-507
- 10. Hosmer D Lemeshow S. Applied logistic regression. John Wiley & Sons, New York, 1989
- 11. Fox E. Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queries and Multiple Concept Types. PhD dissertation, Computer Science, Cornell University, 1983
- 12. Fuhr N. Optimal polynomial retrieval functions based on the probability ranking principle. ACM Transactions on Informations Systems 1989; 7:183-204
- 13. Fuhr N Buckley C. A probabilistic learning approach for document indexing. ACM Transactions on Informations Systems 1991 9:223-248
- 14. Haines D Croft B. Relevance feedback and inference networks. Proceedings of the 1993 SIGIR International Conference on Information Retrieva 1, Pittsburgh, Pa, June 27-July 1, 1993, pp 2-12
- 15. Turtle H. Inference networks for document retrieval. PhD Dissertation, University of Massachusetts, COINS Technical Report 90-92, February, 1991
- 16. Fung R Crawford S Appelbaum L Tong R. An architecture for probabilistic concept-bases information retrieval. In: Proceedings of the 13th international conference on research and development in information retrieval. Brussels, Belgium, September 5-7, 1990, pp. 455-467
- 17. Swanson D. Information retrieval as a trial-and-error process. Library Quarterly 1977; 47:128-148
- 18. Hull D. Using statistical testing in the evaluation of retrieval experiments. Proceedings of the 1993 SIGIR international conference on information retrieval. Pittsburgh, Pa, June 27-July 1, 1993, pp.329-338
- 19. Yu C Buckley C Lam H Salton G. A generalized term dependence model in information retrieval. Information Technology: Research and Development 1983; 2:129-154
- 20. Cooper W Gey F Chen A. Information retrieval from the TIPSTER collection: an application of staged logistic regression. In: Proceedings of the First NIST Text Retrieval Conference, National Institute for Standards and Technology, Washington, DC, November 4-6, 1992, NIST Special Publication 500-207, March 1993, pp 73-88
- 21. Harman, D. Overview of the first TREC conference. In: Proceedings of the 1993 SIGIR international conference on information retrieva l, Pittsburgh, Pa, June 27-July 1, 1993, pp 36-47
- 22. Gey F. Probabilistic dependence and logistic inference in information retrieval. PhD dissertation, University of California, Berkeley, 1993