Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections

Fredric C. Gey, Aitao Chen, Jianzhang He and Jason Meggs
U.C. Data Archive & Technical Assistance (UC DATA)
University of California, Berkeley
e-mail: gey@ucdata.berkeley.edu

ABSTRACT

The Berkeley experiments for TREC4 extend those of TREC3 in three ways: for ad-hoc retrieval we retain the manual reformulations of the topics and experiment with limited query expansion based upon the assumption that top documents are relevant (this experiment was an interesting failure); for routing retrieval we introduce a logistic regression which assumes relevance weights to be only one clue among several in predicting probability of relevance. Finally, for Spanish retrieval we retrain the basic logistic regression equations to apply to the statistical distributions of Spanish words. In addition we apply two approaches to Spanish stemming, one which attempts to resolve verb variants into a standardized form, the other of which eschews stemming in favor of a massive stop word list of variants of common words.

1. Introduction and history of Berkeley TREC participation

For the past several years the UC Berekley Text Retrieval Research Group has been developing an approach to retrieving full-text documents from collections as large as 1 million documents, an approach which returns documents relevant to a query in order of estimated probability of relevance. Using the well-known technique of logistic regression, we predict the dichotomous variable 'relevance' as a function of statistical clues such as query term frequency in document and collection.

The Berkeley group began its participation with TREC1 ad-hoc track using the concept of staged logistic regression (Cooper, Gey, Chen 1993). In this formulation relevance was predicted as a logodds formula for query-document-term triples and statistical clues associated with these triples such as IDF, and term frequencies in document and query. Since the summation of these clues utilized the principle of linked dependence, a second stage regression was introduced between query-document pairs to compensate for known dependencies. The results in TREC1 were compromised by the inadequate training set and incomplete relevance judgments so necessary for adequate statistical fitting. In TREC1 Berkeley submitted no routing entry.

For TREC2 a single stage regression was introduced in the ad-hoc track which was based upon the concepts of "optimized relative frequencies" and the explicit use of number of match terms between document and query as a important statistical clue used in the logistic regression (Cooper Gey Chen 1994). For routing retrieval in TREC2, Berkeley used a logistic regression on relevance history terms weights for term presence or term absence from the document.

In the TREC3 ad-hoc development, Berkeley introduced a concept of averaging statistical clues across terms in common with query and document and performing a single regression against the averaged clues and the number of terms in common between document and query (Cooper, Chen, Gey 1995). The clues were carefully chosen to mirror the statistical weights of the tfidf vector space model. In addition for TREC3 Berkeley introduced the idea of manual reformulation and, where called-for, expansion of the TREC3 topics. Search terms were gathered by search of a similar but independent database, the University of California MELVYL news database. The use of this approach is based upon the idea that there exists some optimal version of a query which will produce the most relevant documents in the best order, and that human intervention based upon search experience can approximate the optimal query. For TREC3 routing, Berkeley used massive query expansion from relevant documents by computing a χ^2 relationship between terms and relevance as the discriminating factor in choice of terms. All terms which passed a five percent significance test for each query were added to the query. Original query terms which passed the test were weighted five times more than other terms of the expansion. This expansion produced from 300 to 4,114 terms, depending on the query.

2. TREC4 ad-hoc methodology

Our review and analysis of the TREC3 results showed some disappointment with the automatic run Brkly6 which used the formula of averaging clues across terms. The Brkly6 automatic run had an overall average precision of 0.2775 whereas the TREC2 fully automatic run Brkly3 had an overall average precision of 0.3270. Independent analysis (Fontaine, 1995) has shown that the Brkly6 algorithm performs significantly better than the vector space model. Although these performance measures were for different query sets and different collections, they were enough to make us reconsider whether the reformulation of TREC3 should be continued. We embarked on a series of experiments comparing the two algorithms which showed that the TREC2 Brkly6 algorithm performed about 10 percent better overall in a number of tests. Thus for TREC4 we made the decision to retreat to the TREC2 formula for our prediction of logodds of relevance between a query Q and document D. In particular the logodds of relevance when there are M match terms in common between query and document, we obtain a formula for M properties $A_1 \cdots A_M$, each associated with a match term, equation (1):

$$\log O(R \mid A_1, \ldots, A_M) \approx c_0 + c_1 f(M) \sum_1^M X_{m,1}$$

$$+ \cdots + c_K f(M) \sum_{m=1}^M X_{m,K}$$

$$+ c_{K+1} M + c_{K+2} M^2$$

where the function $f(M)$ is one that drops gently with increasing M, say $\frac{1}{\sqrt{M}}$ which exerts a damping influence as the number of match terms increases. Statistical tests showed that the M^2 clue did not make a statistically significant contribution and so it was dropped.

The properties chosen for the final equation were what were called in TREC2 "optimized relative frequencies". Thus the final equation after fitting was Equation (2):

$$\log O(R \mid A_1, \ldots, A_M) \approx$$

$$-3.51 + \frac{1}{\sqrt{M} + 1} \Phi + 0.0929 M$$

where Φ is the expression

$$37.4 \sum_{m=1}^M X_{m,1} + 0.330 \sum_{m=1}^M X_{m,2} + 0.1937 \sum_{m=1}^M X_{m,3}$$

Here

$X_{m,1}$ = number of times the m'th stem occurs in the query, divided by (total number of all stem occurrences in query + 35);

$X_{m,2}$ = number of times the m'th stem occurs in the document, divided by (total number of all stem occurrences in document + 80), quotient logged;

$X_{m,3}$ = number of times the m'th stem occurs in the collection, divided by the total number of all stem occurrences in the collection, quotient logged;

M = number of distinct stems common to both query and document.

Equation (2) above remains unchanged from TREC2. In particular, no refitting of logistic coefficients was attempted. However, the Berkeley entries for TREC-4 utilized limited query expansion, described below. For a more complete discussion of optimized relative frequencies, the reader is referred to (Cooper Gey Chen, 1994).

2.1. Manual query reformulation

In TREC3 the Berkeley group introduced a manual reformulation of the topics into new and different queries by having intermediaries search a parallel database, in this case the News database of the University of California's MELVYL electronic catalog. This approach proved so successful that we continued the process for TREC4. This seemed to be very promising when we observed the parsimony of expression in the TREC4 adhoc topic statements. As an example of such a reformulation, here is the original query for topic 241 on professional malfeasance:

> *Find examples of doctor and lawyer groups considering penalties against members of their professions for malfeasance and show the results of such investigations*

The average precision for Brkly9 entry on this query was 0.0120, retrieving only 16 out of 62 relevant documents. The manual reformulation of the same topic leads to the following query:

> *Find examples of doctor and lawyer groups considering penalties against members of their professions for malfeasance and show the results of such investigations doctor physician lawyer ama aba bar bar association license license suspended suspended disciplinary disciplinary action action disbar stealing sanction misconduct misconduct criminal physician license suspended nurse accuse case confidentiality mediations state licensing boards*

This reformulation (augmented by query expansion described below) allowed the Brkly10 run to achieve best average precision of 0.3199 in retrieving 56 of the 62 relevant documents.

Overall, manual reformulation nearly doubled performance from 0.1388 (Brkly9) to 0.2660 (Brkly10). The salient features of manual query development are *specificity* (adding query terms which are examples or instances of events being sought) and *generality* (adding synonymous or closely related terms which might expand the net of documents being ranked).

Finally, we experimented with Boolean filtering on a set of manually chosen negative terms, such that if a negative term occurred within a document, the document would not be retrieved. While this seemed to have great promise for certain queries limited to the United States in geographic scope, experiments showed it to be a two-edged sword such that it increased precision for some queries while reducing it for others.

2.2. A cautionary tale of ad-hoc query expansion

In TREC3 the entries from City University of London (Robertson, Walker, Jones, Hancock-Beualieu, 1995) and Cornell University (Buckley Salton Allan Singhal, 1995) utilized query expansion based upon the assumption that a trial run would produce an initial document ranking from which additional query terms could be harvested out of the top documents. City University's approach was to choose a limited number of expanded terms with a complex weighting formula, also from the top 30 documents of the initial probe. Cornell's approach was to choose the top 30 documents and find the 300 "best" terms and expand the query thusly using Roccio weights. Berkeley felt there was merit in these approaches and began to experiment with past collections to see if the methodology would work within the logistic regression environment.

Our experiments showed that using 30 documents and even 100 additional terms seemed to make performance deteriorate.

Table I: Brkly10 Ad-hoc Entry Average Precision for document/term expansions				
0.2945*	10 terms	20 terms	50 terms	100 terms
10 docs	0.2821	**0.2660****	0.2335	0.1995
20 docs	0.2646	0.2533	0.2263	0.1857
30 docs	0.2589	0.2477	0.2138	0.1885

*** - No expansion, ** - Brkly10 official entry**

Eventually we settled on expansion by twenty 'best' terms from the top 10 retrieved documents. This expansion was applied to both the original query and the manually reformulated queries. Thus the Brkly9 run is obtained by taking the original topic, performing a trial run, and expanding the queries by 20 terms taken from the top 10 ranked documents. A similar process was used for Brkly10, except that the seed query was the manually reformulated one.

Following receipt of the relevance judgments, Berkeley did a more systematic series of runs connected with query expansion. These runs, summarized in Table I, show that Berkeley would have been better served by making its official entry the run without expansion, which at 0.2945 was nearly 10 percent better than the 10 document/ 20 term expansion of the actual entry. Moreover, the same is true of the automatic entry, Brkly9; the official entry using query expansion had a precision of 0.1388. Without query expansion, the automatic methodology of equation (2) achieves average precision of 0.2001, more than 45 percent better than our official entry.

Our experiments show that query expansion must be done judiciously in terms of choice of number of documents and choice of number of terms. Even the top ten documents can produce several hundred additional terms to choose from. Berkeley used the following methodology to choose the top 20 terms from an initial ranking of the top ten documents using the original query -- term weights were calculated according to the following formula:

$$\sum_{i=1}^{10} DAF / NTD * log(NTD / DF)$$

where

DAF is the term frequency in document D_i,
$\sum(DAF)$ is the total number of occurrences of a stem in the NTD top ranked documents.
NTD is the number of top ranked documents (e.g. 10)
DF is the document frequency. (i.e. The number of documents among the NTD top ranked documents that contains the stem)

For our official entries, the twenty terms with the highest weights were added to the original query to form the expanded query.

2.3. Failure analysis

Both the Brkly9 original query and Brkly10 manual query modification stumbled on queries which specified events, actions, or scenarios which were restricted geographically to the United States. For example topic 205 (for which Brkly10 entry retrieved only 3 of 310 relevant documents)

What evidence is there of paramilitary activity in the US

and our manual reformulation

What evidence is there of paramilitary activity in the US paramilitary paramilitary paramilitary militia militia militia group right wing armed guns hate anti-government extremists bomb shooting waco Koernke threat attack idaho michigan keystone montana Koernke Koernke sherwood terrorism

retrieved numerous documents about military and liberation activities of the African

National Congress in South Africa. Similarly topic 206,

> *Prognosis/viability of a political third party in U.S.*

retrieved a number of documents about Venezuelan elections into the top ranks. It is clear that judicious use of a limited scope knowledge-base of geographic locations might considerably improve results of queries of this type.

3. TREC4 routing methodology

To prepare for TREC4 routing, the Berkeley group ran a number of retrospective experiments on the TREC3 collection. We were motivated by the City University approach of limited query expansion which was a marked contrast from the massive query expansion which Berkeley used in TREC3. We ran several experiments which utilized χ^2 tests and relevance history weights (the so-called "Robertson status value" ranking) for choosing terms. In both tests we found the Cornell approach of cutting off term expansion at 300 terms to be effective, however we could not achieve equal performance by utilizing a small number of additional terms of highest precision. In addition our experiments showed that relevance history ranking was better than χ^2 ranking for choice of the 300 expansion terms. This seems to be puzzlingly different from the Xerox-Parc results reported in SIGIR95 (Schutze, Hull, Pedersen 1995).

3.1. Logistic regression in routing

In an endeavor to improve on blind-faith query expansion, the Berkeley group noticed that the City University formulae for TREC3 made use of an additional clues of term frequency in document and query as well as document length. This seemed to indicate that a combination of relevance weighting with other clues similar to the ad-hoc methodology might actually improve performance. For TREC4 routing, Berkeley arrived at the following formula for logodds of relevance, Equation (3):

$$\log O(R \mid Q,D) = -0.2983+0.9309*X_1-0.001622*X_2$$

$$-0.000003059*X_3+0.05316*X_4+0.04613*X_5$$

$$+0.01573*X_6-0.005281*X_7+3.9488*X_8-0.06526*X_9$$

Where the individual clues are defined by

X_1 Query prior probability,
X_2 Number of unique stems in a document,
X_3 Document length (i.e. number of stems in a document),
X_4 Sum of present weight over match stems,
X_5 Sum of absent weight over query stems absent in document,
X_6 Sum of the log of the historical idf over match stems,
X_7 Sum of the log of the document term frequency over match stems,
X_8 Sum of the log of the relative document term frequency over match stems,
X_9 Number of match stems.

This logistic regression on multiple clues where relevance history weights are only one clue among many improved performance considerably in experiments on the TREC3 collections. Our baseline run for routing (Brkly11) consists of the original query terms ranked by the above formula. The Brkly12 run uses the same formula, applied to the top 300 terms for each query chosen from past relevant documents by their relevance weight.

3.2. TREC4 routing performance

The results for TREC4 run Brkly11 are abysmal, with average precision of 0.0345, showing that using only the original query is completely inadequate for the routing retrieval problem. The Brkly12 run which did query expansion by 300 terms chosen in order of relevance history weight improves performance by a factor of seven to 0.2163. However, even this performance is below the median for most queries, puzzlingly low from our point of view, especially since the performance on query 17 ("Measures to control agrochemicals") was

0.8509, the best for that query.

Since receiving the results, we have rerun using the TREC-3 routing formulation with a χ^2 computation to choose the top 1000 expansion terms. This run achieved an average precision of 0.2675 approximately twenty percent better than our official entry.

4. TREC4 Spanish retrieval

In approaching Spanish retrieval the Berkeley group was first faced with the problem of developing a Spanish stemmer. There were two conflicting approaches to stemming of TREC3: Cornell's assertion that only extremely primitive stemming is needed, and that a stop-word list can be developed from the most frequent stems, manually reviewed to choose which stems should be kept, versus Univerity of Massachusetts assertion that sophisticated stemming produces a significant improvement. Mindful that it might not improve performance, Berkeley nevertheless undertook to develop a Spanish stemmer.

4.1. Spanish stemming software

Spanish stemming methodology follows some rules for morphology found in recent computational linguistic literature (Moreno and Goni, 1995). The Berkeley Spanish stemmer begins with removal of longest ending found, if any. If an ending was found, the software converts final c to z on stem then normalizes verb forms found into their normal form. Verbs are the most complex items to stem; they consist of regular, ortho-irregular, and irregular verbs of three classes, -ir, -er and -ar. The classification is made by the ending of the infinitive form of the verb, which is also the name of the verb.

The *regular verbs* have standard endings similar to Spanish nouns, and their stem does not change throughout the conjugation. It is relatively trivial to stem these verbs. *Ortho-irregular verbs* behave like regular verbs in most respects except for the fact that their stem changes throughout the conjugation into one or more new forms. These "stem-changing" or "radical" verbs fortunately follow fairly basic rules which are outlined in standard grammars of Spanish. It was and still is a matter of interest how ambiguous it is to back-track in that respelling. *Irregular verbs* are verbs that may partially follow regular or ortho-irregular patterns, save for one or more forms are totally unpredictable. No rule can be written for a change in a verb from, say, "ser" to "fui". There are unfortunately around 2,750 or more irregular verbs in the Spanish language. The task of normalizing irregular verbs would have been insurmountable had we not found a researcher at the University of Waterloo who was willing to provide us with his laboriously constructed machine-readable file of all verb forms for most irregular verbs (German, 1995). *Non-verbs: nouns, adverbs, adjectives* are generally much easier, since they rarely undergo respelling. They can have occasional irregularities such as the same form for plural as singular, or never assume a singular form (English example: what's the singular of pants?). This is still fairly straightforward unless extreme disambiguation is desired. The basic endings are found in any complete grammar, and nouns, adjectives and adverbs usually resolve to the same stem when they have the same basic meaning.

We also performed two forms of stemming, regular stemming which resolved word variants (mostly verb variants) into their normal form, and no collection stemming, but the construction of a massive stop word list by taking all words of a small stop list (the translation of the SMART list into Spanish) and expanding them into all possible variants. In the latter case we obtained a stop word list of some 10,000 words.

4.2. Logistic regression for Spanish

The Berkeley group methodology for retrieving Spanish was to utilize the proven formula for adhoc logistic regression and retrain on the Spanish training set (Spanish queries 1-25). Doing this resulted in the following Equation (4):

$$\log O(R \mid A_1, \ldots, A_M) \approx$$

$$-3.439 + \frac{1}{\sqrt{M}+1} \Phi - 0.0143 M$$

where Φ is the expression

$$84.86 \sum_{m=1}^{M} X_{m,1} + 0.508 \sum_{m=1}^{M} X_{m,2} - 0.3197 \sum_{m=1}^{M} X_{m,3}$$

where the $X_{m,k}$ are the same clues as in Equation (2) above.

4.3. Spanish entries -- manual reformulation

Mindful of the success of manual reformulation of English ad-hoc queries, we undertook to create manual versions of queries 25-50. Queries 1-25 were considered sufficiently rich in natural language expression as to make manual construction unnecessary. The process was to translate the queries into English, search the MELVYL NEWS database in English, then reformulate manual queries in English and translate them into Spanish.

4.4. Spanish results

The TREC4 official entry had an average precision of 0.1683, about equal to a number of other Spanish entries and considerably below the University of Central Florida entry. Manual reformulation in this case showed no improvement, with average precision of 0.1637. Following submission of the official entry, we discovered an error -- in computing the X_3 component of the regression, we used the collection size from the English collection (140 million stems) rather than the Spanish collection (14 million stems). Rerunning with the correct collection size showed insignificant change (average precision 0.1702) in the result. The most important run was that without stemming and the massively expanded stop word list, whose precision was 0.1009. Our conclusion is that stemming works for Spanish as well as English.

5. Acknowledgments

All Berkeley group research is based upon a modified version of the Cornell SMART system. Berkeley's modification to this software consists of replacement of the vector space similarity computation algorithms with probabilistic algorithms, together with such modifications of the SMART vector and inverted file data structures as necessary to incorporate additional clues. Support for this project came from UC DATA's research and development budget, augmented by research assistant funds from the UC Berkeley School of Information Management and Systems. We also acknowledge the encouragement and moral support of Professor William S. Cooper, whose basic ideas for the past half decade form the core of all our methods.

The generosity of Daniel German of the University of Waterloo in providing his Spanish irregular verb conjugation list made a crucial difference in our ability to deliver probabilistic retrieval of Spanish text.

6. References

Buckley, Chris, Gerard Salton, James Allan, Amit Singhal (1995), "Automatic Query Expansion using SMART: TREC 3," *Proceedings of the Third NIST Text Retrieval Conference (TREC3),* National Institute for Standards and Technology, Washington, DC, November 2-4, 1994, NIST Special Publication 500-225, April 1995, pp. 69-80.

Cooper, W. S.; Gey, F. C.; Chen, A. (1993). "Probabilistic retrieval in the TIPSTER Collections: An Application of Staged Logistic Regression." *The First Text REtrieval Conference (TREC-1).* NIST Special Publication 500-207. Washington, 73-88.

Cooper, W. S.; Gey, F. C.; Chen, A. (1994). "Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression." *The Second Text REtrieval Conference (TREC-2).* Washington, DC, August 31-September 2, 1993, NIST Special Publication 500-215. March 1994, pp.57-66.

Cooper, William S., Aitao Chen, and Fredric Gey (1995), "Experiments in Probabilistic

Retrieval of Full Text Documents," *Proceedings of the Third NIST Text Retrieval Conference (TREC3),* National Institute for Standards and Technology, Washington, DC, November 2-4, 1994, NIST Special Publication 500-225, April 1995, pp. 127-134.

Fontaine, Anne (1995), "Subelement indexing and Probabilistic Retrieval in the POSTGRES Database System", Masters Thesis, Computer Science Department, University of California, Berkeley, May 1995.

German, Daniel M., Private Communication, June 1995.

Moreno, A. and Goni, J (1995), "GRAMPAL: A Morphological Processor for Spanish implemented in Prolog", Technical Report, Dep. de Linguistica, Universidad Autonoma de Madrid, 1995.

Robertson, S.E., S. Walker, S. Jones, M.M. Hancock-Beualieu and M Gatfort (1995), "Okapi at TREC-3," *Proceedings of the Third NIST Text Retrieval Conference (TREC3),* National Institute for Standards and Technology, Washington, DC, November 2-4, 1994, NIST Special Publication 500-225, April 1995, pp. 109-126.

Schutze, Hinrich, David A. Hull, and Jan O. Pedersen (1995), "A Comparison of Classifiers and Document Representations," *Proceedings of SIGIR95, the 18th Annual Conference on Research and Development in Information retrieval*, Seattle Washington, July 9-12, 1995, pp. 229-237.