# Cheshire II at GeoCLEF: Fusion and Query Expansion for GIR

Ray R. Larson
School of Information Management and Systems
University of California, Berkeley, USA
ray@sims.berkeley.edu

**Abstract**

In this paper I will describe the Berkeley (group 1) approach to the GeoCLEF task for CLEF 2005. The main technique we are testing is the fusion of multiple probabilistic searches against different XML components using both Logistic Regression (LR) algorithms and a version of the Okapi BM-25 algorithm. We also combine multiple translations of queries in cross-language searching. Since this is the first time that the Cheshire system has been used for CLEF this approach can, at best, be considered a very preliminary base testing of some retrieval algorithms and approaches. The primary geographically based approaches taken for GeoCLEF were to georeference proper nouns in the text using a gazetteer derived from the World Gazetteer with both English and German names for each place, and to expand place names for regions or countries in the queries by the names of the countries or cities in those regions or countries.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

## General Terms

Algorithms, Performance, Measurement

## Keywords

Cheshire II, Logistic Regression, Data Fusion

# 1 Introduction

For GeoCLEF 2005 the Berkeley IR research group split into two groups (Berkeley 1 and Berkeley 2). Berkeley 2 used the same technques as used in previous CLEF evaluations, while Berkeley 1 tried some alternative algorithms and fusion methods for both the GeoCLEF and Domain Specific task. This paper will focus on the techniques used by the Berkeley 1 group for GeoCLEF and the results of our official submissions, as well as some additional tests using versions of the algorithms employed by the Berkeley 2 group. The main technique being tested is the fusion of multiple probabilistic searches against different XML components using both Logistic Regression (LR) algorithms and a version of the Okapi BM-25 algorithm. We also combine multiple translations of queries in cross-language searching. Since this is the first time that the Cheshire II system has been used for CLEF, this approach can at best be considered a very preliminary base testing of some retrieval algorithms and approaches. This paper is organized as follows: In the next section we discuss the retrieval algorithms and fusion methods used for the submitted runs. We then discuss the specific approaches taken for indexing and retrieval in GeoCLEF and the results of the submitted runs. Then we compare our submitted results to some additional runs with alternate approaches conducted later. Finally we present conclusions and some discussion of the GeoCLEF task.

## 2 The Retrieval Algorithms and Fusion Operators

In [8] we conducted an analysis of the overlap between the result lists retrieved by our Logistic Regression algorithm and the Okapi BM-25 algorithm for the INEX XML Retrieval test collection. We found that, on average, over half of the result lists retrieved by each algorithm in these overlap tests were both non-relevant *and* unique to that algorithm, fulfilling the main criteria for effective algorithm combination suggested by Lee[9]: that the algorithms have similar sets of relevant documents and different sets of non-relevant. This section is largely a repetition of the material presented in [8], with additional discussion of how these algorithms were applied for the CLEF GeoCLEF task.

In the remainder of this section we describe the Logistic Regression and Okapi BM-25 algorithms that were used for GeoCLEF and we also discuss the methods used to combine the results of the different algorithms. The algorithms and combination methods are implemented as part of the Cheshire II XML/SGML search engine [6, 7, 5] which also supports a number of other algorithms for distributed search and operators for merging result lists from ranked or Boolean sub-queries.

### 2.1 Logistic Regression Algorithm

The basic form and variables of the *Logistic Regression* (LR) algorithm used was originally developed by Cooper, et al. [3]. It provided good full-text retrieval performance in the TREC3 ad hoc task and in TREC interactive tasks [4] and for distributed IR [5]. As originally formulated, the LR model of probabilistic IR attempts to estimate the probability of relevance for each document based on a set of statistics about a document collection and a set of queries in combination with a set of weighting coefficients for those statistics. The statistics to be used and the values of the coefficients are obtained from regression analysis of a sample of a collection (or similar test collection) for some set of queries where relevance and non-relevance has been determined. More formally, given a particular query and a particular document in a collection $P(R \mid Q, D)$ is calculated and the documents or components are presented to the user ranked in order of decreasing values of that probability. To avoid invalid probability values, the usual calculation of $P(R \mid Q, D)$ uses the "log odds" of relevance given a set of $S$ statistics, $s_i$, derived from the query and database, such that:

$$\log O(R \mid Q, D) = b_0 + \sum_{i=1}^{S} b_i s_i \tag{1}$$

where $b_0$ is the intercept term and the $b_i$ are the coefficients obtained from the regression analysis of the sample collection and relevance judgements. The final ranking is determined by the conversion of the log odds form to probabilities:

$$P(R \mid Q, D) = \frac{e^{\log O(R|Q,D)}}{1 + e^{\log O(R|Q,D)}} \tag{2}$$

Based on the structure of XML documents as a tree of XML elements, we define a "document component" as an XML subtree that may include zero or more subordinate XML elements or subtrees with text as the leaf nodes of the tree. Thus, a component might be defined using any of the tagged elements in a document. However, *not all possible components are likely to be useful* in content-oriented retrieval (e.g., tags indicating that a word in the title should be in italic type, or the page number range) therefore we defined the retrievable components selectively, including the titles, dates, and document ids.

Naturally, a full XML document may also be considered a "document component". As discussed below, the indexing and retrieval methods used in this research take into account a selected set of document components for generating the statistics used in the search process and for extraction of the parts of a document to be returned in response to a query. Because we are dealing with not only full documents, but also document components (which for some collections include elements such as sections and paragraphs or similar structures) derived from the documents, we will use $C$ to represent document components in place of $D$. Therefore, the full equation describing the LR algorithm used in these experiments is:

$$\begin{aligned}
\log O(R \mid Q, C) &= \\
& \quad b_0 + \left(b_1 \cdot \left(\frac{1}{|Q_c|}\sum_{j=1}^{|Q_c|} \log qtf_j\right)\right) \\
&+ \left(b_2 \cdot \sqrt{|Q|}\right) \\
&+ \left(b_3 \cdot \left(\frac{1}{|Q_c|}\sum_{j=1}^{|Q_c|} \log tf_j\right)\right) \\
&+ \left(b_4 \cdot \sqrt{cl}\right) \\
&+ \left(b_5 \cdot \left(\frac{1}{|Q_c|}\sum_{j=1}^{|Q_c|} \log \frac{N - n_{t_j}}{n_{t_j}}\right)\right) \\
&+ \left(b_6 \cdot \log |Q_d|\right)
\end{aligned} \tag{3}$$

Where:

$Q$ is a query containing terms $T$,

$|Q|$ is the total number of terms in $Q$,

$|Q_c|$ is the number of terms in $Q$ that also occur in the document component,

$tf_j$ is the frequency of the $j$th term in a specific document component,

$qtf_j$ is the frequency of the $j$th term in Q,

$n_{t_j}$ is the number of components (of a given type) containing the $j$th term,

$cl$ is the document component length measured in bytes.

$N$ is the number of components of a given type in the collection.

$b_i$ are the coefficients obtained though the regression analysis.

This equation, used in estimating the probability of relevance in this research, is essentially the same as that used in [2] for TREC3. The $b_i$ coefficients in the "Base" version of this algorithm were estimated using relevance judgements and statistics from the TREC/TIPSTER test collection. For GeoCLEF we used this Base version for our retrieval of all components with the addition of the component fusion methods described later. The coefficients for the Base version were $b_0 = -3.70, b_1 = 1.269, b_2 = -0.310, b_3 = 0.679, b_4 = -0.021, b_5 = 0.223$ and $b_6 = 4.01$.

## 2.2 Okapi BM-25 Algorithm

The version of the Okapi BM-25 algorithm used in these experiments is based on the description of the algorithm in Robertson [11], and in TREC notebook proceedings [10]. As with the LR algorithm, we have adapted the Okapi BM-25 algorithm to deal with document components :

$$\sum_{j=1}^{|Q_c|} w^{(1)} \frac{(k_1+1)tf_j}{K+tf_j} \frac{(k_3+1)qtf_j}{k_3+qtf_j} \tag{4}$$

Where (in addition to the variables already defined):

$K$ is $k_1((1-b)+b \cdot dl/avcl)$

$k_1$**,** $b$ **and** $k_3$ are parameters (1.5, 0.45 and 500, respectively, were used),

$avcl$ is the average component length measured in bytes

$w^{(1)}$ is the Robertson-Sparck Jones weight:

$$w^{(1)} = \log \frac{(\frac{r+0.5}{R-r+0.5})}{(\frac{n_{t_j}-r+0.5}{N-n_{t_j}-R-r+0.5})}$$

$r$ is the number of relevant components of a given type that contain a given term,

$R$ is the total number of relevant components of a given type for the query.

Our current implementation uses only the *a priori* version (i.e., without relevance information) of the Robertson-Sparck Jones weights, and therefore the $w^{(1)}$ value is effectively just an IDF weighting. The results of searches using our implementation of Okapi BM-25 and the LR algorithm seemed sufficiently different to offer the kind of conditions where data fusion has been shown to be be most effective [9], and our overlap analysis of results for each algorithm (described in the evaluation and discussion section) has confirmed this difference and the fit to the conditions for effective fusion of results.

The system used supports searches combining probabilistic and (strict) Boolean elements, as well as operators to support various merging operations for both types of intermediate result sets. However, in GeoCLEF we did not use this capability.

## 2.3 Result Combination Operators

The Cheshire II system used in this evaluation provides a number of operators to combine the intermediate results of a search from different components or indexes. With these operators we have available an entire spectrum of combination methods ranging from strict Boolean operations to fuzzy Boolean and normalized score combinations for probabilistic and Boolean results. These operators are the means available for performing fusion operations between the results for different retrieval algorithms and the search results from different different components of a document. We will only describe two of these operators here, because they were the only type used in the GEOCLEF runs reported in this paper.

The MERGE_CMBZ operator is based on the "CombMNZ" fusion algorithm developed by Shaw and Fox [12] and used by Lee [9]. In our version we take the normalized scores, but then further enhance scores for components appearing in both lists (doubling them) and penalize normalized scores appearing low in a single result list, while using the unmodified normalized score for higher ranking items in a single list.

The MERGE_PIVOT operator is used primarily to adjust the probability of relevance for one search result based on matching elements in another search result. It was developed primarily to adjust the probabilities of a search result consisting of sub-elements of a document (such as titles or paragraphs) based on the probability obtained for the same search over the entire document. It is basically a weighted combination of the probabilities based on a "DocPivot" fraction, such that:

$$P_n = DocPivot * P_d + (1 - DocPivot) * P_s \tag{5}$$

where $P_d$ represents the document-level probability of relevance, $P_s$ represents the subelement probability, and $P_n$ representing the resulting new probability. The "*DocPivot*" value used for all of the runs submitted was 0.64. Since this was the first year for GeoCLEF, this value was derived from experiments on 2004 data for other CLEF collections (which may have been inappropriate for the GeoCLEF data, which further testing will reveal). The basic operator can be applied to either probabilistic results, or non-probabilistic results or both (in the latter case the scores are normalized using MINMAX normalization to range between 0 and 1).

# 3 Approaches for GeoCLEF

In this section we describe the specific approaches taken for our submitted runs for the GeoCLEF task. First we describe the indexing and term extraction methods used, and then the search features we used for the submitted runs.

## 3.1 Indexing and Term Extraction

For both the monolingual and bilingual tasks we indexed the documents using the Cheshire II system. The document index entries and queries were stemmed using the Snowball stemmer, and a new georeferencing indexing subsystem was used. This subsystem extracts proper nouns from the text being indexed and attempts to match them in a digital gazetteer. For GeoCLEF we used a gazetteer derived from the World Gazetteer (http://www.world-gazetteer.com) with 224698 entries in both English and German. The indexing subsystem provides three different index types: verified place names (an index of names which matched the gazetteer), point coordinates (latitude and longitude coordinates of the verified place name) and bounding box coordinates (bounding boxes for the matched places from the gazetteer). All three types were created, but due to time constraints we only used the verified place names in our tests. Text indexes were also created for separate XML elements (such as document titles or dates) as well as for the entire document. It is worth noting that, although the names are compared against the gazetteer, it is quite common for proper name of persons and places to be the same and this leads to potential false associations between articles mentioning persons with such name and particular places.

| Name | Description | Content Tags | Used |
|---|---|---|---|
| docno | Document ID | DOCNO | no |
| pauthor | Author Names | BYLINE, AU | no |
| headline | Article Title | HEADLINE, TITLE, LEAD, LD, TI | yes |
| topic | Content Words | HEADLINE, TITLE, TI, LEAD<br>BYLINE, TEXT, LD, TX | yes<br>yes |
| date | Date of Publication | DATE, WEEK | yes |
| geotext | Validated place names | TEXT, LD, TX | yes |
| geopoint | Validated coordinates for place names | TEXT, LD, TX | no |
| geobox | Validated bounding boxes for place names | TEXT, LD, TX | no |

Table 1: Cheshire II Indexes for GeoCLEF 2005

Table 1 lists the indexes created for the GeoCLEF database and the document elements from which the contents of those indexes were extracted. The "Used" column in Table 1 indicates whether or not a particular index was used in the submitted GeoCLEF runs.

Because there was no explicit tagging of location-related terms in the collections used for GeoCLEF, we applied the above approach to the "TEXT", "LD", and "TX" elements of the records of the various collections. The part of news articles normally called the "dateline" indicating the location of the news story was not separately tagged in any of the GeoCLEF collection, but often appeared as the first part of the text for the story. (In addition, we discovered when writing these notes that the "TX" and "LD" were not included in the indexing for some collections and elements, meaning that the SDA collection was not included in the German indexing for these types).

For all indexing we used English and German stoplists to exclude function words and very common words from the indexing and searching. For the runs reported here we also did not use any decompounding of German terms.

## 3.2 Search Processing

Searching the GeoCLEF collection used Cheshire II scripts to parse the topics and submit the title and description from the topics to one or more indexes. For monolingual search tasks we used the topics in the appropriate language (English or German), for bilingual tasks the topics were translated from the source language to the target language using three different machine translation (MT) systems, the L&H PC-based system, SYSTRAN (via Babelfish at Altavista), and PROMT (also via their web interface). Each of these translations were combined into a single probabilistic query. The hope was to overcome the translation errors of a single system by including alternatives.

We tried two main approaches for searching, the first used only the topic text from the title and desc elements, the second included the spatialrelation and location elements as well. In all cases the different indexes mentioned above were used, and probabilistic searches were carried out on each index, and the results combined using the CombMNZ algorithm, and by a weighted combination of partial element and full document scores. For bilingual searching we used both the Berkeley TREC3 and the Okapi BM-25 algorithm, for monolingual we used only TREC3. For one submitted run in each task we did no query expansion and did not use the location elements in the topics. For the other runs each of the place names identified in the queries were expanded when that place was the name of a region or country. For example when running search against the English databases the name "Europe" was expanded to "Albania Andorra Austria Belarus Belgium Bosnia and Herzegovina Bulgaria Croatia Cyprus Czech Republic Denmark Estonia Faroe Islands Finland France Georgia Germany Gibraltar Greece Guernsey and Alderney Hungary Iceland Ireland Isle of Man Italy Jersey Latvia Liechtenstein Lithuania Luxembourg Macedonia Malta Moldova Monaco Netherlands Norway Poland Portugal Romania Russia San Marino Serbia and Montenegro Slovakia Slovenia Spain Svalbard and Jan Mayen Sweden Switzerland Turkey Ukraine United Kingdom Vatican City", while for searches against the German databases "Europa" was expanded to "Albanien Andorra sterreich Weirussland Belgien Bosnien und Herzegowina Bulgarien Kroatien Zypern Tschechische Republik Dnemark Estland Frer-Inseln Finnland Frankreich Georgien Deutschland Gibraltar Griechenland Guernsey und Alderney Ungarn Island Irland Man Italien Jersey Lettland Liechtenstein Litauen Luxemburg Mazedonien Malta Moldawien Monaco Niederlande Norwegen Polen Portugal Rumnien Russland San Marino Serbien und Montenegro Slowakei Slowenien Spanien Svalbard und Jan Mayen Schweden Schweiz Trkei Ukraine Grobritannien Vatikan". Example queries for monolingual searches are shown in Figure 3

The indexes combined in searching included the headline, topic, and geotext indexes (as described in Table 1) for searches that include the location element, and the headline and topic for the searches without the locations element. For the bilingual tasks, three sub-queries, one for each query translation were run and then the results were merged using the CombMNZ algorithm. For Monolingual tasks the title and topic results were combined with each other using CombMNZ and the final score combined with an expanded search for place names in the topic and geotext indexes.

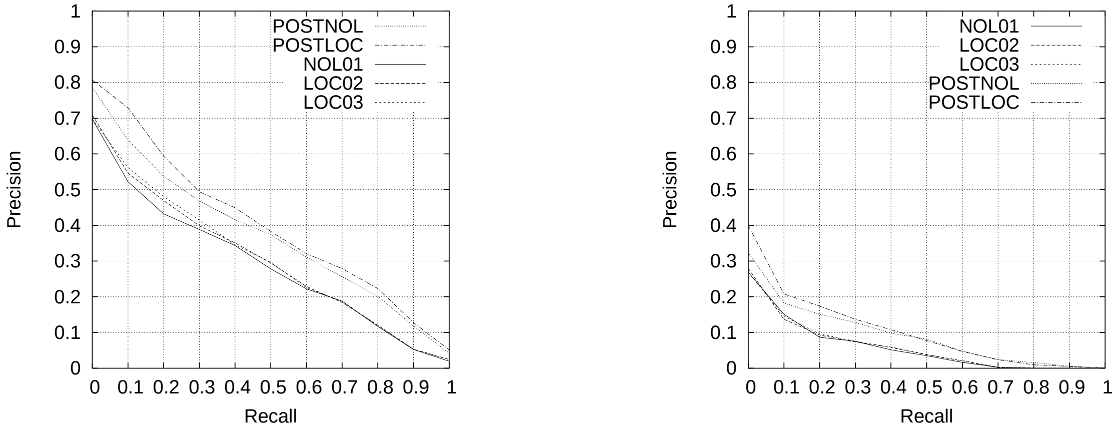

Figure 1: Berkeley1 Monolingual Runs – English (left) and German (right)

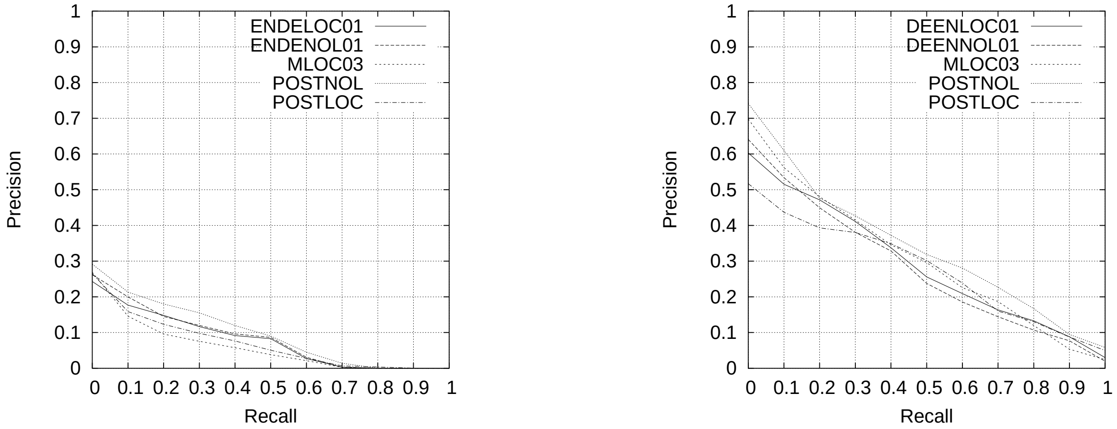

Figure 2: Berkeley1 Bilingual Runs – English to German (left) and German to English (right)

Examples of the queries used are shown in Figures 3 and 4 in Appendix A, as you may observe by close inspection there were some bugs in the scripts used to generate these queries some of which have been removed for this paper. These included things such as including "Kenya" in the expansion for Europe, and including two copies of all expansion names, when a single copy should have been used. We intend (time permitting) to rerun a number of the queries to see if, and how, these errors affected the results.

# 4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the submitted bilingual and monolingual runs for both English and German are shown in Table 2, the Recall-Precision curves for these runs are also shown in Figures 1 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the name are abbrevated to the final letters and numbers of the full name in Table 2, and those beginning with "POST" are unofficial runs described in the next section.

Table 2 indicates some rather curious results that warrant further investigation as to the cause. Notice that the result for all of the English monolingual runs exceed the rsults for bilingual German to English runs - this is typical for cross-langauge retrieval. However, in the case of German this expected pattern is reversed, and the German monolingual runs *perform worse* than either of the bilingual English to German runs. We haven't yet determined exactly why this might be the case,

| Run Name | Description | Location | MAP |
|---|---|---|---|
| BERK1BLDEENLOC01 | Bilingual German⇒English | yes | 0.2753 |
| BERK1BLDEENNOL01 | Bilingual German⇒English | no | 0.2668 |
| BERK1BLENDELOC01 | Bilingual English⇒German | yes | 0.0725 |
| BERK1BLENDENOL01 | Bilingual English⇒German | no | 0.0777 |
| BERK1MLDELOC02 | Monolingual German | yes | 0.0535 |
| BERK1MLDELOC03 | Monolingual German | yes | 0.0533 |
| BERK1MLDENOL01 | Monolingual German | no | 0.0504 |
| BERK1MLENLOC02 | Monolingual English | yes | 0.2910 |
| BERK1MLENLOC03 | Monolingual English | yes | 0.2924 |
| BERK1MLENNOL01 | Monolingual English | no | 0.2794 |

Table 2: Submitted GeoCLEF Runs

but there are number possible reasons (e.g., since a combination of Okapi and Logistic Regression searches are used for the bilingual task this may be an indication that Okapi is more effective for German). Also, in the monolingual runs, both English and German, use of the location tag and expansion of the query (runs numbered LOC02 and LOC03 respectively) did better than no use of the location tag or expansion. For the bilingual runs the results are mixed, with German to English runs showing an improvement with location use and expansion (LOC01) and English to German showing the opposite.

## 5 Additional Runs

After the official submission we used the same version of the Logistic Regression algorithm as the Berkeley2 group (the "TREC2" algorithm), which incorporates blind feedback (which is lacking in the LR algorithm described above). The parameters used for blind feedback were 13 documents and the top-ranked 16 terms from those documents added to the original query. We used essentially an identical algorithm to that defined by Cooper, Gey and Chen in [1]. The results from the bilingual and monolingual runs for both English and German are shown in Table 3, the Recall-Precision curves for these runs are also shown in Figures 1 (for monolingual) and 2 (for bilingual). In Figures 1 and 2 the names abbrevated to the final letters of the full name in Table 3, prefixed by "POST". These are unofficial runs to test the difference in the algorithms in an identical runtime environment.

| Run Name | Description | Location | MAP |
|---|---|---|---|
| POSTBLDEENEXP | Bilingual German⇒English | yes | 0.2636 |
| POSTBLDEENNOL | Bilingual German⇒English | no | 0.3205 |
| POSTBLENDEEXP | Bilingual English⇒German | yes | 0.0626 |
| POSTBLENDENOL | Bilingual English⇒German | no | 0.0913 |
| POSTMLDELOC | Monolingual German | yes | 0.0929 |
| POSTMLDENOL | Monolingual German | no | 0.0861 |
| POSTMLENEXP | Monolingual English | yes | 0.2892 |
| POSTMLENLOC | Monolingual English | yes | 0.3879 |
| POSTMLENNOL | Monolingual English | no | 0.3615 |

Table 3: Additional Post-Submission GeoCLEF Runs

As can be seen by comparing Table 3 with Table 2, all of the comparable runs for show improvement in results with the TREC2 algorithm with blind feedback. We have compared notes with the Berkeley2 group and with minor differences to be expected given the different indexing methods, stoplists, etc. used, these results are comparable to theirs.

The queries submitted in these unofficial runs were much simpler than those used in the official runs. For monolingual retrieval only the “topic” index was used and the geotext index was not used at all, for the bilingual runs the same pattern of using multiple query translations and combining the results was used as in our official runs. This may actually be detrimental to the performance, since the expanded queries perform worse than the unexpanded queries - the opposite behaviour observed in the official runs.

In the monolingual runs there appears to be similar behavior, The best using the topic titles and description along with the location tag provided the best results, but expanding the locations as in the official runs (the English ML run ending in EXP) performed considerably worse than the the unexpanded runs.

# 6 Conclusions

Analysis of these results is still ongoing. There are a number of, as yet, unexplained behaviors in some of our results. We plan to continue working on the use of fusion, and hope to discover effective ways to combine highly effective algorithms, such as the TREC2 algorith, as well as work on adding the same blind feedback capability to the TREC3 Logistic Regression algorithm.

One obvious conclusion that can be drawn is that basic TREC2 is a highly effective algorithm for the GeoCLEF tasks, and the fusion approaches tried in these tests are most definitely NOT very effective (in sprite of their relatively good effectiveness in other retrieval tasks such as INEX).

Another conclusion is that, in some cases, query expansion of region names to a list of names of particular countries in that region is modestly effective (although we haven’t yet been able to test for statistical significance). In other cases, however it can be quite detrimental. However we still need to determine if the problems with the expansion were due the nature of the expansion itself, or errors in how it was done.

# References

[1] Aitao Chen. Cross-language retrieval experiments at clef 2002. pages 28–48, 2003.

[2] William S. Cooper, Aitao Chen, and Fredric C. Gey. Experiments in the probabilistic retrieval of full text documents. In Donna K. Harman, editor, *Overview of the Third Text Retrieval Conference (TREC-3): (NIST Special Publication 500-225)*, Gaithersburg, MD, 1994. National Institute of Standards and Technology.

[3] William S. Cooper, Fredric C. Gey, and Daniel P. Dabney. Probabilistic retrieval based on staged logistic regression. In *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, June 21-24*, pages 198–210, New York, 1992. ACM.

[4] Ray R. Larson. TREC interactive with cheshire II. *Information Processing and Management*, 37:485–505, 2001.

[5] Ray R. Larson. A logistic regression approach to distributed IR. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 399–400. ACM, 2002.

[6] Ray R. Larson. Cheshire II at INEX: Using a hybrid logistic regression and boolean model for XML retrieval. In *Proceedings of the First Annual Workshop of the Initiative for the Evaluation of XML retrieval (INEX)*, pages 18–25. DELOS workshop series, 2003.

[7] Ray R. Larson. Cheshire II at INEX 03: Component and algorithm fusion for XML retrieval. In *INEX 2003 Workshop Proceedings*, pages 38–45. University of Duisburg, 2004.

[8] Ray R. Larson. A fusion approach to XML structured document retrieval. *Information Retrieval*, 8:601–629, 2005.

[9] Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27-31, 1997, Philadelphia*, pages 267–276. ACM, 1997.

[10] Stephen E. Robertson, Stephen Walker, and Micheline M. Hancock-Beauliee. OKAPI at TREC-7: ad hoc, filtering, vlc and interactive track. In *Text Retrieval Conference (TREC-7), Nov. 9-1 1998 (Notebook)*, pages 152–164, 1998.

[11] Stephen E. Robertson and Steven Walker. On relevance weights with little relevance information. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 16–24. ACM Press, 1997.

[12] Joseph A. Shaw and Edward A. Fox. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215*, pages 243–252, 1994.

## A Example Queries submitted

```
search ((headline @ {vegetable exporters of europe what countries are
   exporters of fresh, dried or frozen vegetables? })
!MERGE_CMBZ (topic @ {vegetable exporters of europe what countries
   are exporters of fresh, dried or frozen vegetables? }))
!MERGE_PIVOT/64 (topic @ {vegetable exporters of europe what
   countries are exporters of fresh, dried or frozen vegetables? })
```

```
search ((headline @ {vegetable exporters of europe what countries are
  exporters of fresh, dried or frozen vegetables?  vegetable exporters europe }
!MERGE_CMBZ (topic @ {vegetable exporters of europe what countries are
  exporters of fresh, dried or frozen vegetables?  vegetable exporters europe})
!MERGE_CMBZ ((geotext @ {vegetable exporters of europe what countries are
  exporters of fresh, dried or frozen vegetables? vegetable exporters europe })
!MERGE_CMBZ (topic @ { Albania Andorra Austria Belarus Belgium
  Bosnia and Herzegovina Bulgaria Croatia Cyprus Czech Republic Denmark
  Estonia Faroe Islands Finland France Georgia Germany Gibraltar Greece
  Guernsey and Alderney Hungary Iceland Ireland Isle of Man Italy Jersey
  Latvia Liechtenstein Lithuania Luxembourg Macedonia Malta Moldova Monaco
  Netherlands Norway Poland Portugal Romania Russia San Marino
  Serbia and Montenegro Slovakia Slovenia Spain Svalbard and Jan Mayen
  Sweden Switzerland Turkey Ukraine United Kingdom Vatican City }))
!MERGE_PIVOT/64 (topic @ {vegetable exporters of europe what countries are
  exporters of fresh, dried or frozen vegetables?  vegetable exporters europe })
```

Figure 3: Example Berkeley1 Monolingual Queries with, and without geographic elements

```
PART QUERY1: search (topic @+ { shark attacks against australia and california
 the documents reports over attacks of sharks on people.})
!MERGE_CMBZ (topic @ { shark attacks against australia and california the
 documents reports over attacks of sharks on people.}) RESULTSETID SET1
PART QUERY2: search (topic @+ { shark fish attacks before australia and
 california the documents report?r attacks of shark fish on humans.})
!MERGE_CMBZ (topic @ { shark fish attacks before australia and california
 the documents report?r attacks of shark fish on humans.}) RESULTSETID SET2
PART QUERY3: search (topic @+ { shark fish attacks before australia and
 california the documents report about attacks about shark fishing on person.})
!MERGE_CMBZ (topic @ { shark fish attacks before australia and california
 the documents report about attacks about shark fishing on person.})
 RESULTSETID SET3
FINAL QUERY: search  SET1: !MERGE_CMBZ SET2: !MERGE_CMBZ SET3: RESULTSETID SET4
```

```
PART QUERY1: search (topic @+ { shark attacks against australia and california
 the documents reports over attacks of sharks on people.  shark attacks
 australia : california})
!MERGE_CMBZ (topic @ { shark attacks against australia and california the
 documents reports over attacks of sharks on people.  shark attacks  australia
 : california})
!MERGE_CMBZ (topic @ {  australien californien australien
 californien }) RESULTSETID SET1
PART QUERY2: search (topic @+ { shark fish attacks before australia and
 california the documents report?r attacks of shark fish on humans.
 shark fish attacks  australia : california})
!MERGE_CMBZ (topic @ { shark fish attacks before australia and california the
 documents report?r attacks of shark fish on humans.  shark fish attacks
 australia : california})
!MERGE_CMBZ (topic @ {australien californien australien californien})
 RESULTSETID SET2
PART QUERY3: search (topic @+ {shark fish attacks before australia and
 california the documents report about attacks about shark fishing on person.
 shark fish attacks  australia : california})
!MERGE_CMBZ (topic @ { shark fish attacks before australia and california the
 documents report about attacks about shark fishing on person.  shark fish
 attacks  australia : california})
!MERGE_CMBZ (topic @ {australien californien australien californien})
 RESULTSETID SET3
FINAL QUERY: search SET1: !MERGE_CMBZ SET2: !MERGE_CMBZ SET3: RESULTSETID SET4
```

Figure 4: Example Berkeley1 Bilingual (German to English) Queries with, and without geographic elements