Text Retrieval and Blind Feedback for the ImageCLEF Photo Task

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

Abstract

In this paper we will describe Berkeley's approach to the ImageCLEF "Photo" task for CLEF 2006. This year is the first time that we have participated in ImageCLEF, and we chose to primarily establish a baseline for the Cheshire II system for this task, while we had originally hoped to use GeoCLEF methods for this task, in the end time constraints led us to restrict our submissions to the basic required runs for the task.

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

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the ImageCLEF Photo task. Our submitted runs this year are intended to establish a simple baseline for comparision in future ImageCLEF tasks. This year we used only text-based retrieval methods for Imageclef, totally ignoring the images themselves (and the reference images specified in the queries). We hope, in future years, to be able to use combined text and image processing approaches similar to those used in some earlier work combining Berkeley's BlobWorld image retrieval system with the Cheshire II system (see [7]).

This year Berkeley submitted 7 runs, of which 4 where English Monolingual, 2 German Monolingual, and 1 Bilingual English to German.

This paper first describes the retrieval methods used, including our blind feedback method for text, followed by a discussion of our official submissions and the results obtained from other methods after official submissions had been closed. Finally we present some discussion of the results and our conclusions.

2 The Retrieval Algorithms

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions was originally developed by Cooper, et al. [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:

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

2.1 TREC2 Logistic Regression Algorithm

For all of our ImageCLEF submissions this year we used a version of the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[3] and which is also used in our GeoCLEF and Domain Specific submissions. For the ImageCLEF task we used the Cheshire II information retrieval system implementation of this algorithm. One of the current limitations of this implementation is the lack of decompounding for German documents and query terms in the current system. As noted in our other CLEF notebook papers, the Logistic Regression algorithm used was originally developed by Cooper et al. [4] for text retrieval from the TREC collections for TREC2. The basic formula is:

$$\begin{aligned}
\log O(R|C,Q) &= log\frac{p(R|C,Q)}{1-p(R|C,Q)} = log\frac{p(R|C,Q)}{p(\overline{R}|C,Q)} \\
&= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \frac{qtf_i}{ql+35} \\
&+ c_2 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \log \frac{tf_i}{cl+80} \\
&- c_3 * \frac{1}{\sqrt{|Q_c|}+1} \sum_{i=1}^{|Q_c|} \log \frac{ctf_i}{N_t} \\
&+ c_4 * |Q_c|
\end{aligned}$$

where C denotes a document component (i.e., an indexed part of a document which may be the entire document) and Q a query, R is a relevance variable,

$p(R|C,Q)$ is the probability that document component C is relevant to query Q,

$p(\overline{R}|C,Q)$ the probability that document component C is *not relevant* to query Q, which is 1.0 - $p(R|C,Q)$

$|Q_c|$ is the number of matching terms between a document component and a query,

qtf_i is the within-query frequency of the ith matching term,

tf_i is the within-document frequency of the ith matching term,

ctf_i is the occurrence frequency in a collection of the ith matching term,

ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),

cl is component length (i.e., number of terms in a component), and

N_t is collection length (i.e., number of terms in a test collection).

c_k are the k coefficients obtained though the regression analysis.

If stopwords are removed from indexing, then ql, cl, and N_t are the query length, document length, and collection length, respectively. If the query terms are re-weighted (in feedback, for example), then qtf_i is no longer the original term frequency, but the new weight, and ql is the sum of the new weight values for the query terms. Note that, unlike the document and collection lengths, query length is the "optimized" relative frequency without first taking the log over the matching terms.

The coefficients were determined by fitting the logistic regression model specified in $\log O(R|C,Q)$ to TREC training data using a statistical software package. The coefficients, c_k, used for our official runs are the same as those described by Chen[1]. These were: $c_0 = -3.51$, $c_1 = 37.4$, $c_2 = 0.330$, $c_3 = 0.1937$ and $c_4 = 0.0929$. Further details on the TREC2 version of the Logistic Regression algorithm may be found in Cooper et al. [4].

2.2 Blind Relevance Feedback

In addition to the direct retrieval of documents using the TREC2 logistic regression algorithm described above, we have implemented a form of "blind relevance feedback" as a supplement to the basic algorithm. The algorithm used for blind feedback was originally developed and described by Chen [2]. Blind relevance feedback has become established in the information retrieval community due to its consistent improvement of initial search results as seen in TREC, CLEF and other retrieval evaluations [6]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [8].

Blind relevance feedback is typically performed in two stages. First, an initial search using the original topic statement is performed, after which a number of terms are selected from some number of the top-ranked documents (which are presumed to be relevant). The selected terms are then weighted and then merged with the initial query to formulate a new query. Finally the reweighted and expanded query is submitted against the same collection to produce a final ranked list of documents. Obviously there are important choices to be made regarding the number of top-ranked documents to consider, and the number of terms to extract from those documents. For ImageCLEF this year, having no prior data to guide us, we chose to use the top 10 terms from 10 top-ranked documents. The terms were chosen by extracting the document vectors for each of the 10 and computing the Robertson and Sparck Jones term relevance weight for each document. This weight is based on a contingency table where the counts of 4 different conditions for combinations of (assumed) relevance and whether or not the term is, or is not in a document. Table 1 shows this contingency table.

The relevance weight is calculated using the assumption that the first 10 documents are relevant and all others are not. For each term in these documents the following weight is calculated:

$$w_t = log \frac{\frac{R_t}{R - R_t}}{\frac{N_t - R_t}{N - N_t - R + R_t}} \tag{3}$$

	Relevant	Not Relevant	
In doc	R_t	$N_t - R_t$	N_t
Not in doc	$R - R_t$	$N - N_t - R + R_t$	$N - N_t$
	R	$N - R$	N

Table 1: Contingency table for term relevance weighting

The 10 terms (including those that appeared in the original query) with the highest w_t are selected and added to the original query terms. For the terms not in the original query, the new "term frequency" (qtf_i in Equation 3 above) is set to 0.5. Terms that were in the original query, but are not in the top 10 terms are left with their original qtf_i. For terms in the top 10 and in the original query the new qtf_i is set to 1.5 times the original qtf_i for the query. The new query is then processed using the same LR algorithm as shown in Equation 3 and the ranked results returned as the response for that topic.

3 Approaches for ImageCLEF

In this section we describe the specific approaches taken for our official submitted runs for the ImageCLEF photo 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

Although the Cheshire II system uses the XML structure of documents and extracts selected portions of the record for indexing and retrieval, for the submitted runs this year we used only a single one of these indexes that contains the entire content of the document.

Name	Description	Content Tags	Used
docno	Document ID	DOCNO	no
title	Article Title	TITLE	no
topic	All Content Words	DOC	yes
date	Date of Image	DATE	no
geoname	Image Place names	LOCATION	no

Table 2: Cheshire II Indexes for ImageCLEF 2006

Table 2 lists the indexes created for the ImageCLEF database and the document elements from which the contents of those indexes were extracted. The "Used" column in Table 2 indicates whether or not a particular index was used in the submitted ImageCLEF runs. Although we had hoped to use more elements (especially geographic names) we ran out of time due to work on other CLEF tracks. We hope to be able to do some additional runs for discussion at the Meeting.

For all indexing we used language-specific stoplists to exclude function words and very common words from the indexing and searching. The German language runs, however, did *not* use decompounding in the indexing and querying processes to generate simple word forms from compounds (actually we tried, but there was a bug that failed to match any compounds in our runs).

3.2 Search Processing

Searching the ImageCLEF collection used Cheshire II scripts to parse the topics and submit the title or title and narrative from the topics to the "topic" index containing all terms from the documents. For the monolingual search tasks we used the topics in the appropriate language

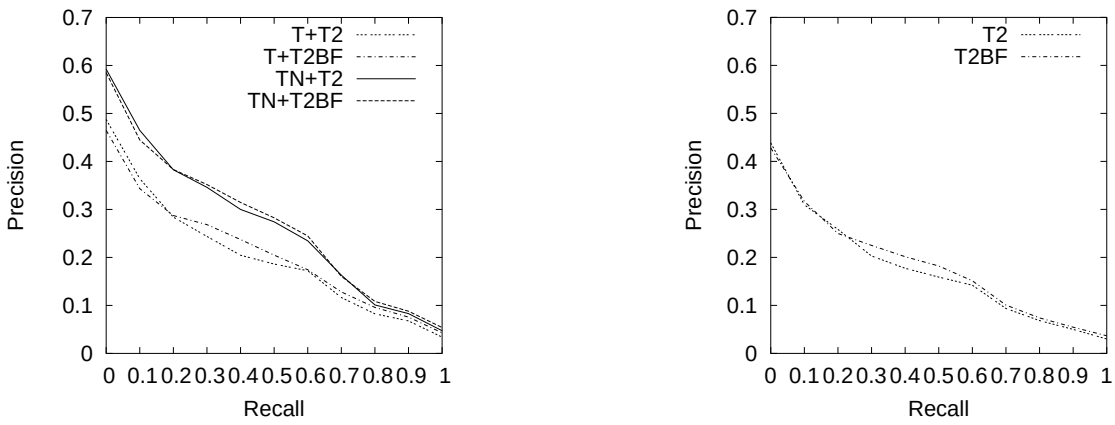

Figure 1: Berkeley Monolingual Runs – English (left) and German (right)

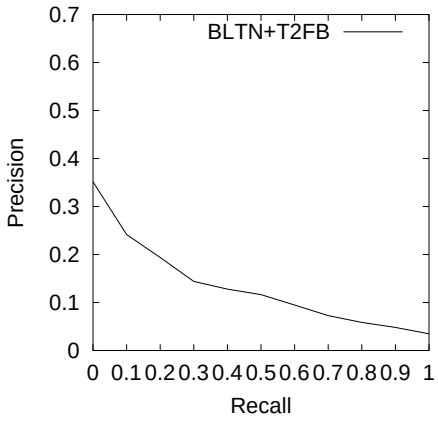

Figure 2: Berkeley Bilingual Run – English to German

(English or German), and for bilingual tasks the topics were translated from the source language to the target language using SYSTRAN (via Babelfish at Altavista). We believe that other translation tools provide a more accurate representation of the topics (like the L&H P.C. translator used in our GeoCLEF entries. However since the runs and submission were done from a hotel room over a highly variable wireless connection we did not have access to all of the tools we would have normally used. We also did a German to English translation, but the quality of translation was so poor (partially because only short titles were available) that we decided not to submit that run.

We tried two main approaches for searching, the first used only the topic text from the title element, the second included both titles and the narrative elements. In all cases the "topic" index mentioned above was used, and probabilistic searches were carried out. Two forms of the TREC2 logistic regression algorithm were used. One used the basic algorithm as described above, and the other used the TREC2 algorithm with blind feedback using the top 10 terms from the 10 top-ranked documents in the initial retrieval.

Our official runs did not make use of the example images in queries, but after our official submissions we realized that this information could be used in a text-only system as well. This involved retrieving the associated metadata records from the example images included in the queries and using their titles and location information to expand the basic query. We tested this in unofficial "POST" runs that are also described below following the discussion of the official

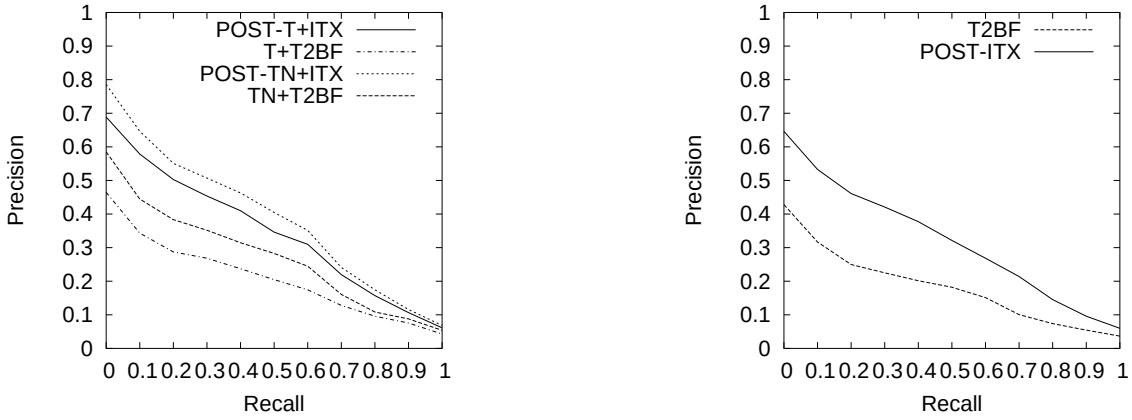

Figure 3: Berkeley Monolingual POST Runs – English (left) and German (right)

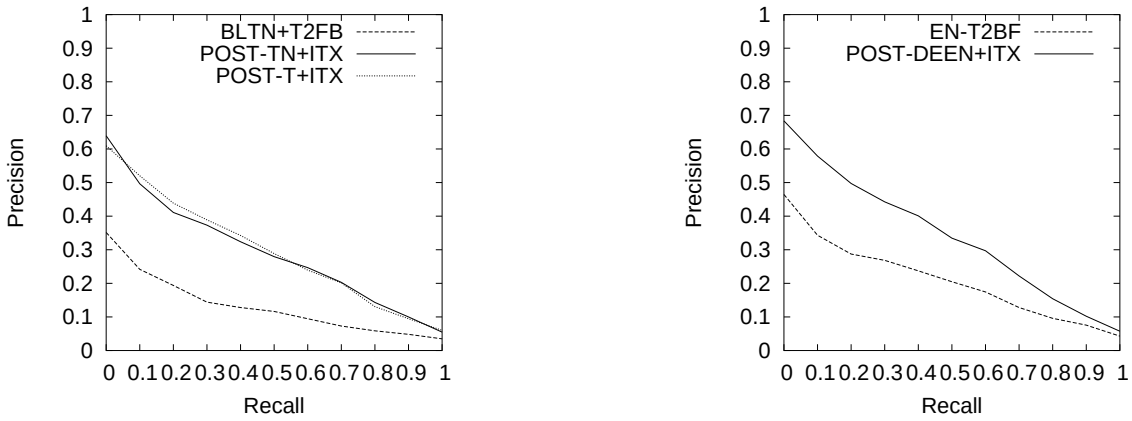

Figure 4: Berkeley Bilingual POST Runs – English to German (left) and German to English (right)

results.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the official submitted 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 name are abbrevated to the final letters and numbers of the full name in Table 3. Table 4 has a number of unofficial runs described in the next section, and Figures 3 and 4 show the Precision and Recall curves for those runs (with some of the runs from Figures 1 and 2 for comparison. (Note also the differences of the Y scale in these figures).

Table 3 shows all of our submitted runs for the ImageCLEF Photo task. Precision and recall curves for the runs are shown in Figures 1 and 2.

5 Discussion and Conclusions

The officially submitted runs described above, when compared to participants using combined text and image methods, are not particularly distinguished. The German monolingual and bilingual

Run Name	Description	Query	Feedback	MAP
BERK_BI_ENDE_TN_T2FB	Bilingual English⇒German	Title+Narr	Y	0.1054
BERK_MO_DE_T_T2	Monolingual German	Title	N	0.1362
BERK_MO_DE_T_T2FB	Monolingual German	Title	Y	0.1422
BERK_MO_EN_T_T2	Monolingual English	Title	N	0.1720
BERK_MO_EN_T_T2FB	Monolingual English	Title	Y	0.1824
BERK_MO_EN_TN_T2	Monolingual English	Title,Narr	N	0.2356
BERK_MO_EN_TN_T2FB	Monolingual English	Title,Narr	Y	0.2392

Table 3: Submitted ImageCLEF Runs

Run Name	Description	Query	Feedback	MAP
POST_BI_DEEN_T_EXP-TL	Bilingual German⇒English	Title	Y	0.3114
POST_BI_ENDE_T_EXP-TL	Bilingual English⇒German	Title	Y	0.2738
POST_BI_ENDE_TN_EXP-TL	Bilingual English⇒German	Title+Narr	Y	0.2645
POST_BI_FREN_T_EXP-TL	Bilingual French⇒English	Title	Y	0.2971
POST_MO_DE_T_EXP-TL	Monolingual German	Title	Y	0.2906
POST_MO_EN_T_EXP-TL	Monolingual English	Title	Y	0.3147
POST_MO_EN_TN_EXP-TL	Monolingual English	Title+Narr	Y	0.3562

Table 4: Unofficial POST ImageCLEF Runs (using image metadata query expansion)

runs suffer from the lack of decompounding in this version of the system. For the purpose of establishing a baseline performance for our system the runs are adequate. One interesting observation for is that, although the MAP for runs with blind feedback was higher than matching runs without it, the precision at low recall levels was lower than runs without feedback. We suspect that this is an effect of the small size of the metadata records available for this task when compared to the full-text in other collections. Use of the narrative for English topics allowed us to further improve performance in the official runs, although the same pattern for feedback versus no feedback performance as mentioned above was observed, indicative that the pattern is likely a function of the record size and not query size.

Although our official runs for ImageCLEF Photo 2006 are in no way outstanding, they do provide a "foot in the door" and a baseline for tuning our future performance for this task.

Quite a different picture appears in our "POST" runs, shown in Table 4 and in Figures 3 and 4. All of these runs use expansion of the queries based on the "TITLE" and "LOCATION" elements of the metadata annotations associated with images included in the "image" element of the queries (e.g. rather like image processing approaches, but using only text). All of these runs, had they been official runs, would have ranked at or near the top of the evaluation pools. (The similarity in between our MAP results and some others would seem to indicate that others were using a similar method of query expansion, perhaps in addition to image processing).

In Table 5 we compare the best performing runs, all using blind relevance feedback, for matching tasks in our official and unofficial runs. For Bilingual English to German we see a dramatic 146% improvement in MAP using this query expansion method. For the comparable Monolingual tasks we see improvements ranging from almost 50% to over 80%. This is strong indication that expanding queries using the metadata of relevant images is a very good strategy for the ImageCLEF Photo task.

As an further experiment, we decided to try this expansion method for Bilingual retrieval *without any translation* of the original topic, that is, we used the title in the original language (in this case French) and did our expansion using appropriate metadata for the target language (English). The result from this experiment is included in Table 4 as run "POST_BI_FREN_T_EXP-

Description	Query	Official MAP	POST MAP	Percent Improv.
Bilingual English⇒German	Title+Narr	0.1072	0.2645	146.74
Monolingual German	Title	0.158	0.2906	83.92
Monolingual English	Title	0.1824	0.3147	72.53
Monolingual English	Title+Narr	0.2392	0.3562	48.91

Table 5: Comparison of Official and Unofficial POST ImageCLEF Runs

TL". If that run had been submitted as an official run, it would have been a 186.5% improvement in MAP over the best official French to English submitted for the ImageCLEF Photo task.

References

[1] Aitao Chen. Multilingual information retrieval using english and chinese queries. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum, CLEF-2001, Darmstadt, Germany, September 2001*, pages 44–58. Springer Computer Scinece Series LNCS 2406, 2002.

[2] Aitao Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.

[3] Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.

[4] W. S. Cooper, A. Chen, and F. C. Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients fitted by Logistic Regression. In *Text REtrieval Conference (TREC-2)*, pages 57–66, 1994.

[5] 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.

[6] Ray R. Larson. Probabilistic retrieval, component fusion and blind feedback for xml retrieval. In *INEX 2005*, pages 225–239. Springer (LNCS #3977), 2006.

[7] Ray R. Larson and Chad Carson. Information access for a digital library: Cheshire II and the Berkeley environmental digital library. In Larry Woods, editor, *Knowledge: Creation, Organization and Use: Proceedings of the 62nd ASIS Annual Meeting, Medford, NJ*, pages 515–535. Information Today, 1999.

[8] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, pages 129–146, May–June 1976.