Linked Relevance 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 ImageCLEFphoto task for CLEF 2007. Once again (as in ImageCLEFphoto for CLEF 2006) we used entirely text-based methods for retrieval. For some runs this year, however, we exploited the basic similarity of the topics and database from 2006 to acquire the metadata descriptions of the "example images" in the 2007 queries, and used that metadata to expand the query content for each topic. The results speak for themselves: use of what amounts to relevance feedback based on image metadata is much more effective than use of unexpanded queries, and even provides a method of cross-language retrieval for unknown languages when parallel topics and example images can be established.
We submitted 19 runs for ImageCLEFphoto this year, of which 8 where monolingual English, German and Spanish, and the remaining 11 where bilingual from various languages to English, German and Spanish.

Categories and Subject Descriptors

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

General Terms

Algorithms, Performance, Measurement

Keywords

Cheshire II, Logistic Regression, Relevance Feedback

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the ImageCLEFphoto task. This year we used only text-based retrieval methods for ImageCLEFphoto, ignoring the images themselves, but using, for some runs, metadata associated with the reference images specified in the queries from 2006. We have not yet been able to convert the BlobWorld software that we wanted to use for combined text and image processing approaches, as used in some previous work (see [7]), but hope to be able to do so for future work.

This year Berkeley submitted 19 runs, of which 3 where English Monolingual, 3 German Monolingual, and 2 were Spanish monolingual. Of the remaining 11 runs, all were bilingual, including German⇒English, German⇒Spanish, English⇒German, English⇒Spanish, Spanish⇒German,

Spanish⇒English, French⇒English, Russian⇒German, Russian⇒English, Chinese⇒German, and Chinese⇒English.

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 methods used for query expansion. Finally we present some discussion of the results and our conclusions.

2 The Retrieval Algorithms

(Note, this section repeats information provided in our 2006 Notebook paper, since the basic retrieval algorithms used and the approaches to indexing the content have not been changed since then.)

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 document and query terms. 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:

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

$$+ \quad c_4 * |Q_c|$$

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

	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

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:

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

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 ImageCLEFphoto

In this section we describe the specific approaches taken for our official submitted runs for the ImageCLEFphoto 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 Terms	DOC	yes
date	Date of Image	DATE	no
geoname	Image Place names	LOCATION	no

Table 2: Cheshire II Indexes for ImageCLEF 2007

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. Note that the database from 2006 was also maintained with the same indexes, although it was not used directly for retrieval since the full extent of changes to that database was not known. The metadata from 2006 database was, however used indirectly for query expansion as described in Section 3.2 below.

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.

3.2 Query Expansion

Last year (after the official runs were submitted) we found that using the metadata associated the "relevant images" included in the topics could be used to expand the query with very good results.

This year, metadata associated with the topic "relevant images" was removed, but since the image ids had not changed from the 2006 collection for the same images, we were able to extract the 2006 metadata for the same relevant images and use it for query expansion. It should be noted that we *did not* use any actual relevance data from 2006, and only used the 2006 metadata associated with the images provided with the ImageCLEFphoto 2007 images in the topic "image" tags.

For the runs that use this form of query expansion we add to the queries data from the "TITLE" and "LOCATION" elements of the 2006 metadata annotations, in the appropriate target language, associated with images included in the "image" element of the 2007 topics (e.g. rather like image processing approaches, but using only the associate metadata text). We did not use the Description or Notes fields of the metadata, since testing with 2006 data showed that including them in the query expansion tended reduce performance when compared to title and location alone.

3.3 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 (English or German), and for bilingual tasks the topics were translated from the source language to the target language using LEC Power Translator (a PC and web-based program and service). This was the first time that we have used this translation tool, which was primarily selected for the broad coverage of languages (including Russian and Chinese), and the apparent accuracy of translation using 2006 data. Because of the method used for query expansion, the expanded queries in target languages were largely the same, with variation only in the translated titles.

Because no narrative was provided for the topics we used only the titles for searching in unexpanded queries. 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.

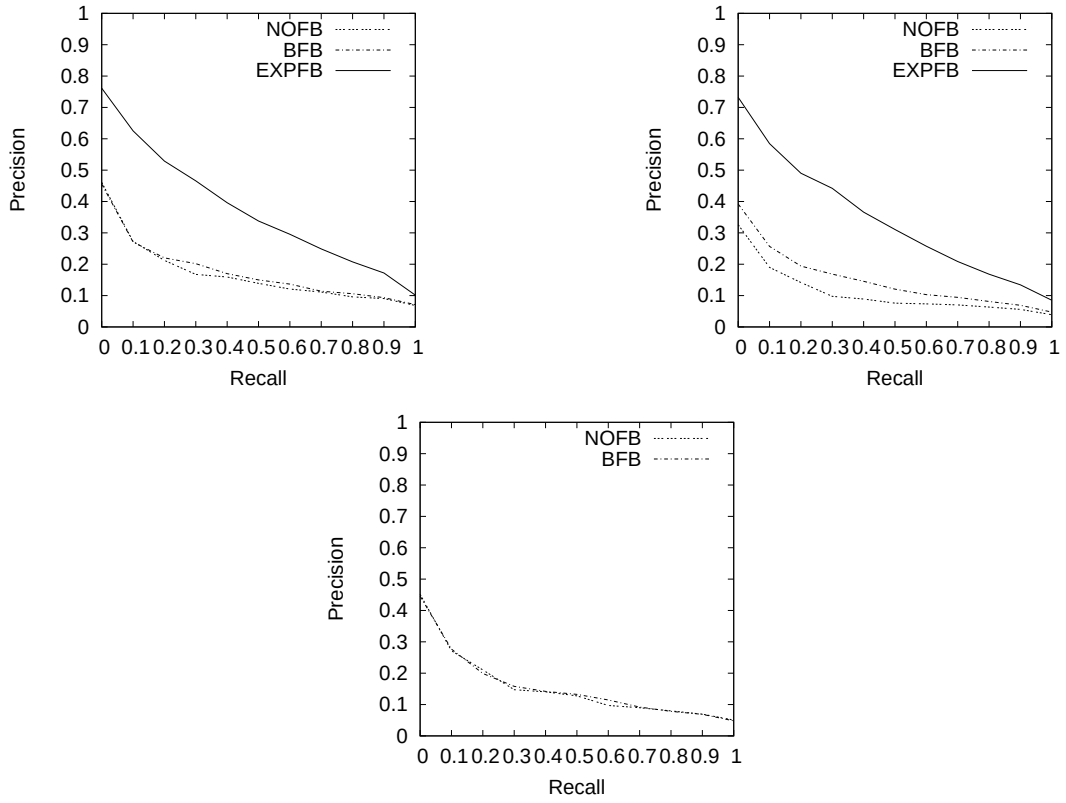
Our expanded query runs did use the example images in topics, this involved retrieving the associated metadata records from the 2006 database for the example image ids included in the queries and using their titles and location information to expand the basic query.

4 Results for Submitted Runs

The summary results (as Mean Average Precision) for the official submitted monolingual and bilingual runs for both English, German and Spanish target languages 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 are abbrevated as indicated in the "Abbrev." column. 3.

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.

Figure 1: Berkeley Monolingual Runs – English (top left), German (top right) and Spanish (lower)



5 Discussion and Conclusions

Our officially submitted runs using query expansion as described above were presented separately in the provided evaluations along with others who apparently used 2006 topic descriptions, notes and qrels. As noted above we used ONLY the provided 2007 topics, but looked up the example images provided in the 2007 topics in the 2006 metadata. It is worth noting that this approach is related to our Digital Library work where we link to multiple resources based on metadata provided in other databases (e.g., linking library catalog records to Wikipedia articles based solely on the catalog data).

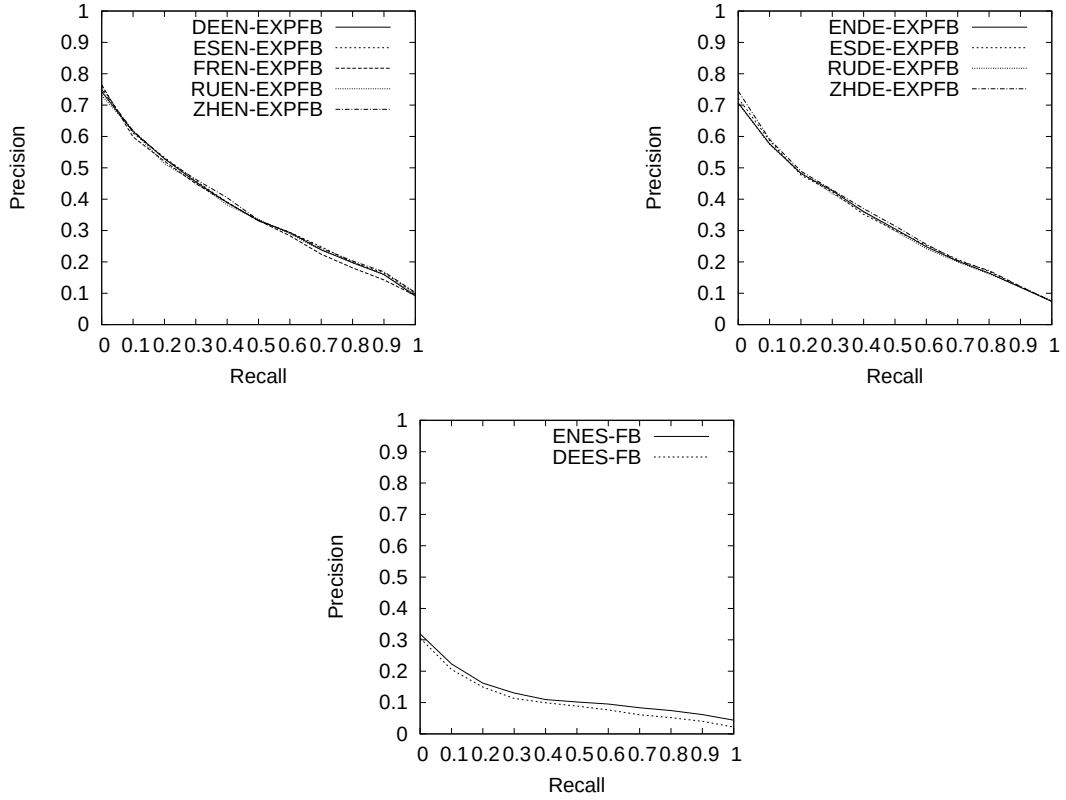
If all runs were to be considered on the basis of effectiveness alone then the query expansion runs listed above would be top-ranked runs of all types.

In Table 4 we compare our best performing runs, all using blind relevance feedback, for monolingual English and German search tasks with and without query expansion. As the "Percent Improv." column shows query expansion provided a 140% and 132% boost for German and English, respectively. This is strong indication that expanding queries using the metadata of relevant images is a very good strategy for the ImageCLEFphoto task.

We would argue that this query expansion approach is exactly comparable to image based approaches that use the example images as the basis for their queries.

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).

Figure 2: Berkeley Bilingual Runs – X to English (top left), X to German (top right) and X to Spanish (lower)



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.

Run Name	Abbrev.	Description	Feedback	MAP
BERK-DE-DE-AUTO-FB-TXT	FB	Mono. German	Y	0.1291
BERK-DE-DE-AUTO-NOFB-TXT	NOFB	Mono. German	N	0.0897
BERK-DE-DE-AUTO-QEFB-TXT	EXPFB	Mono. German	Y	0.3111
BERK-EN-EN-AUTO-FB-TXT	FB	Mono. English	Y	0.1493
BERK-EN-EN-AUTO-NOFB-TXT	NOFB	Mono. English	N	0.1436
BERK-EN-EN-AUTO-QEFB-TXT	EXPFB	Mono. English	Y	0.3467
BERK-ES-ES-AUTO-FB-TXT	FB	Mono. Spanish	Y	0.1280
BERK-ES-ES-AUTO-NOFB-TXT	NOFB	Mono. Spanish	N	0.1244
BERK-DE-EN-AUTO-QEFB-TXT	DEEN-EXPFB	German⇒English	Y	0.3369
BERK-DE-ES-AUTO-FB-TXT	DEES-FB	German⇒Spanish	Y	0.0910
BERK-EN-DE-AUTO-QEFB-TXT	ENDE-EXPFB	English⇒German	Y	0.3024
BERK-EN-ES-AUTO-FB-TXT	ENES-FB	English⇒Spanish	Y	0.1046
BERK-ES-EN-AUTO-QEFB-TXT	ESEN-EXPFB	Spanish⇒English	Y	0.3403
BERK-ES-DE-AUTO-QEFB-TXT	ESDE-EXPFB	Spanish⇒German	Y	0.3022
BERK-FR-EN-AUTO-QEFB-TXT	FREN-EXPFB	French⇒English	Y	0.3314
BERK-RU-EN-AUTO-QEFB-TXT	RUEN-EXPFB	Russian⇒English	Y	0.3342
BERK-RU-DE-AUTO-QEFB-TXT	RUDE-EXPFB	Russian⇒German	Y	0.2987
BERK-ZH-EN-AUTO-QEFB-TXT	ZHEN-EXPFB	Chinese⇒English	Y	0.3410
BERK-ZH-DE-AUTO-QEFB-TXT	ZHDE-EXPFB	Chinese⇒German	Y	0.3076

Table 3: Submitted ImageCLEF Runs

Table 4: Comparison of ImageCLEFphoto Monolingual Runs with and without query expansion.

Description	No Exp. MAP	Query Exp. MAP	Percent Improv.
Mono. German	0.1291	0.3111	140.97
Mono. English	0.1493	0.3467	132.21

[6] Ray R. Larson. Probabilistic retrieval, component fusion and blind feedback for XML retrieval. In *INEX 2005*, pages 225–239. Springer (Lecture Notes in Computer Science, 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.