Blind Relevance Feedback for the ImageCLEF Wikipedia Retrieval Task

Ray R. Larson

School of Information
University of California, Berkeley, CA, USA
ray@ischool.berkeley.edu

Abstract. In this paper we will describe Berkeley's approach to the ImageCLEF Wikipedia Retrieval task for 2010. Our approach to this task was primarily to use text-based searches on the contents of the Wikipedia image metadata records. In addition we submitted one run using a database derived from the provided "bag.xml" set of 5000 descriptor "words" for each image and query example images. We had also intended to combine this one image-based approach to other image-based approaches and to the text-based approaches using fusion methods, but were unable to complete the coding in time.
We submitted 8 runs for ImageCLEF Wikipedia Retrieval this year, of which 6 where monolingual English, German and French with differing search areas in the metadata record, one was multilingual and the remaining one was image-based using the data derived from bag.xml file.
Our best performing run was ranked 24th among the 127 submitted runs by all participants with a MAP of 0.2014, while the image-only approach was ranked dead last (one wonders, in fact, if random results might have done better).

1 Introduction

This paper discusses the retrieval methods and evaluation results for Berkeley's participation in the ImageCLEF Wikipedia Retrieval task. This year we used primarily text-based retrieval methods for ImageCLEF Wikipedia Retrieval, but also attempted to use some of the supplied image-derived information. We did manage to submit a single image-only run, but were not able to use it in combination with our text-based approach (unfortunately we were not able to complete the merger software in time for official submissions, although we hope to have some combined runs to report later).

This year Berkeley submitted 8 runs, of which 2 where English Monolingual, 2 German Monolingual, and 2 were French monolingual. The remaining runs included one multilingual run (using the English German and French topic text and the entire metadata record as a search target), and a single image-based run derived from the "bag.xml" file provided with the database.

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 ImageCLEF Notebook paper, since the basic retrieval algorithms used and the approaches to indexing the content have not been changed since then. This same algorithm and approach have been used in a wide variety of cross-language retrieval experiments in both CLEF and NTCIR (see, for example, [9, 10, 7, 11])

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our text-based submissions was originally developed by Cooper, et al. [6]. 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[4] 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. [5] 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[2]. 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. [5].

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 [3]. 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 [8]. The blind feedback algorithm is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [13].

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.

	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 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 ImageCLEF Wikipedia Retrieval

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

Cheshire II system uses the XML structure of documents and extracts selected portions of those record for indexing and retrieval. In our submitted runs this year we used separate indexes for each of the languages (English, German and French) as well as a global index combining all elements of the metadata records.

Name	Description	Content Tags	Used
names	Image name	names	no
topic	Entire Record	image	yes
texten	English Text	text@lang="en"	yes
textde	German Text	text@lang="de"	yes
textfr	French Text	text@lang="fr"	yes

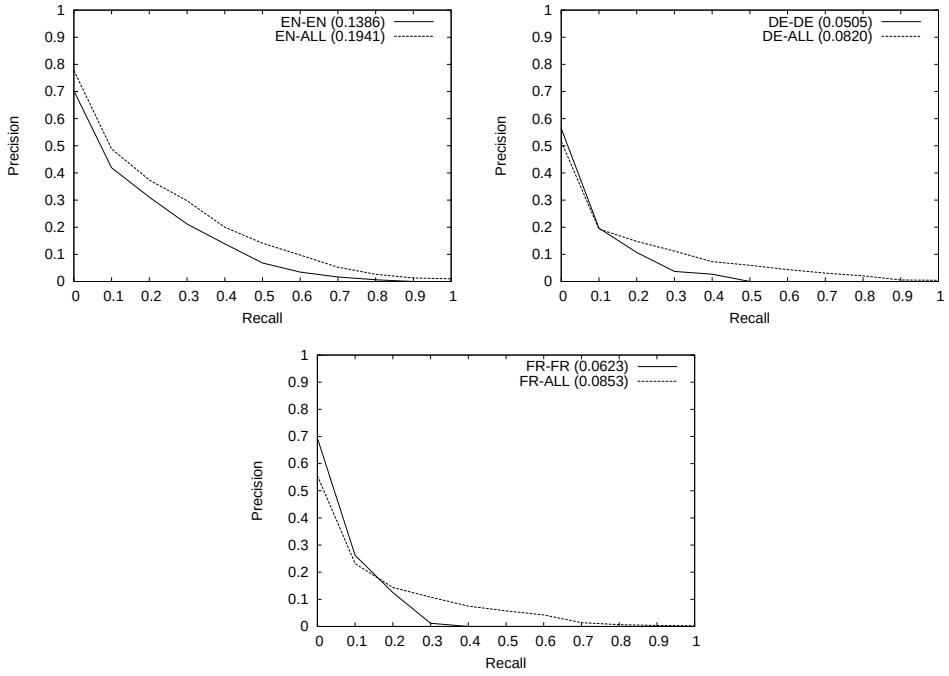
Table 2. Indexes for ImageCLEF Wikipedia Retrieval

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. 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 Image Content Indexing and Processing

In earlier work we used the Berkeley Blobworld algorithms [1] in some digital library retrieval experiments with quite good results (see [12]). In that approach, the blobworld features for each "blob" (a coherent region of color and texture) were quantized and treated as a set of tokens, with token frequency based on weights for the different quanta. Retrieval then treated the tokenized image "blob" information as terms, and used simple text ranking methods to rank blobs and hence images. We had hoped to revive the blobworld segmentation software for this task, but were not able to complete the conversion from MatLab code to C in the time available for the task.

Fig. 1. Berkeley Monolingual Runs – English (top left), German (top right) and French (lower)



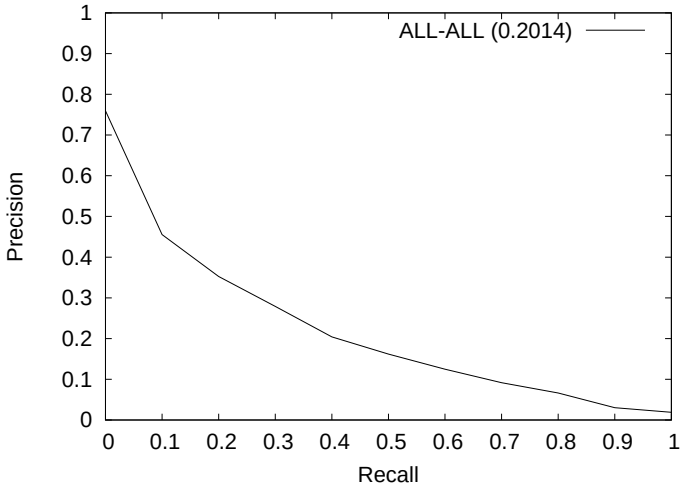
In looking at the image feature data provided with the collection, we realized that a similar approach might be attempted with that data, so as a last-minute attempt, we tokenized the 5000 element "bag" vectors, using a very simple approach (probably too simple, considering the rather terrible results) and used the basic TREC2 algorithm (without blind feedback) to rank the results. The same approach was used on the provided sample images for the topics to provide the queries for processing.

3.3 Search Processing

Searching the ImageCLEF Wikipedia collection used Cheshire II scripts to parse the topics and submit a query to the system using the topic title in a particular language (or for all of the languages in the multilingual case). Depending on settings in the script, the queries were run against the language specific indexes, or the entire document. The TREC2 algorithm with blind feedback used the top 10 terms from the 10 top-ranked documents in the initial retrieval for the blind feedback.

Our single image-based run used the tokenized features derived from the example images to search the tokenized feature vector indices.

Fig. 2. Berkeley Multilingual Run



4 Results for Submitted Runs

The summary results (as Mean Average Precision) for all of our official submitted runs are shown in Table 3, the Recall-Precision curves for the text-based runs are also shown in Figures 1 (for monolingual) and 2 (for multilingual). In Figures 1 and 2 the names are abbrevated as indicated in the "Abbrev." column. 3.

Run Name	Target	Description	Feedback	MAP
BRK-T2BF-DE-DE	DE	Mono. German	Y	0.0505
BRK-T2BF-DE-ALL	ENDEFR	Mono. German	Y	0.0820
BRK-T2BF-EN-EN	EN	Mono. English	Y	0.1386
BRK-T2BF-EN-ALL	ENDEFR	Mono. English	Y	0.1941
BRK-T2BF-FR-FR	FR	Mono. French	Y	0.0623
BRK-T2BF-FR-ALL	ENDEFR	Mono. French	Y	0.0853
BRK-T2BF-ENDEFR-ALL	ENDEFR	Multilingual	Y	0.2014
BRK-FEAT-T3	image	visual features	N	0.0001

Table 3. Submitted ImageCLEF Runs

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

Our officially submitted runs using text retrieval with blind feedback did not perform as well as some of the other participants' text-only runs and definitely lagged behind the best performing mixed image and text runs. What was also

apparent (both from the results and examination of the database itself) was that the multilingual character of the data was very spotty. Most of the metadata entries did not include all three target languages, and many records contained no text descriptions at all other than a caption in a single language, or in some cases terms in the image name itself. Thus, our runs that targetted the entire record did much better than those attempting to access the language-tagged text. Another interesting point is that simply combining the topic titles for each language was our best performing run. This is interesting primarily in comparison with previous CLEF multilingual tasks in other track, where this approach usually made results worse than monolingual approaches. It may be that the sparseness of the metadata records and uneven distribution of language use favors the multilingual approach.

As note above, our single image-only submission was a dismal failure. However, we hope to be able to further test using image element summarization and tokenization, but to do that effectively we first need to know much more about how the supplied feature vectors were created and what each element in those vectors represents. Considering the only real documentation is pointers to journal papers, where the data format is not described at all, it is probably not surprising that our last minute approach made some wrong assumptions and choices in representing the image features. We also still hope to be able to revive the blobworld sofware and use it for future experiments, but at present that is dependent on either getting funding to re-license Matlab, or completing the rewrite of the software in C or some other high-level language.

References

1. Chad Carson, Serge Belongie, Hayit Greenspan, and Jitendra Malik. Color- and texture-based image segmentation using EM and its application to image querying and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page In Press, 1998.
2. 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.
3. Aitao Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.
4. Aitao Chen and Fredric C. Gey. Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7:149–182, 2004.
5. 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.
6. 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.

7. Fredric C. Gey and Ray R. Larson. Patent mining: A baseline approach. In *Proceedings of the NTCIR-7 Workshop Meeting, Tokyo, December 2008*, pages 358–361, 2008.
8. 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.
9. Ray R. Larson. Cheshire at GeoCLEF 2007: Retesting text retrieval baselines. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 811–814, Budapest, Hungary, September 2008.
10. Ray R. Larson. Cheshire at GeoCLEF 2008: Text and fusion approaches for GIR: CLEF working notes, 2008. http://www.clef-campaign.org/2008/working_notes/larson_GeoCLEF.pdf.
11. Ray R. Larson. Logistic regression for ir4qa. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 0–0, 2010.
12. 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.
13. 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.