A Logistic Regression Approach for NTCIR-11 Temporalia

Ray R. Larson
School of Information
University of California, Berkeley
Berkeley, California, USA, 94720-4600
ray@ischool.berkeley.edu

ABSTRACT

Berkeley's approach to the Temporalia TIR retrieval task for NTCIR-11 has been, as is our custom with new tasks, to use our probabilistic text retrieval methods to establish an in-house baseline for future experiments. For our initial experiments we used only the Logistic Regression ranking both with and without pseudo relevance feedback. We have previously used these algorithms in the NTCIR-8 and NTCIR-9 GeoTime tasks, as well as in many other evaluations at CLEF and INEX. This brief paper describes the submitted runs and the methods used for them.

Keywords: Logistic Regression, Probabilistic Retrieval.

1. INTRODUCTION

The NTCIR-11 Temporal Information Access task, or Temporalia, further explores the use of time elements in many of the searches performed performed in IR evaluations, as well as on the internet, as begun in the NTCIR GeoTime tracks for NTCIR-8 and NTCIR-9 [5, 6]. The details of the Temporalia track, its focus, structure, and the results are discussed in the track overview paper [7]. In this paper we describe our submissions for the Temporalia track and consider possible improvements to the retrieval approach for the various temporal aspects of the queries. We used, essentially, the same search tools and methods described in our IR4QA paper for NTCIR-8[12] and GeoTime paper for NTCIR-9[13].

2. DATABASE AND INDEXING

Our document ranking algorithm is a probability model based using the technique of logistic regression [4]. For all of our runs we used the TREC2 logistic regression model, described below, with and without blind or pseudo relevance feedback. The database for the Temporalia TIR sub-track consisted of the "LivingKnowledge news and blogs annotated sub-collection" [14] representing about 20GB of web and blog data. Each file of collections consisted of the items for a particular day, and also included tags indicating the date of the story or blog entry, as well as the full text and other metadata associated with particular terms or names within each document. The original version of collection contained invalid XML structures, but these were corrected in the version that we used by a conversion program provided by the track organizers (the Cheshire system is fairly strict as to XML/SGML syntax and will halt indexing when a invalid construct or undefined tag is encountered). The indexes created for this task are shown in Table 1.

This collection covers materials ranging in date from May 2011 to March 2013 from a large variety of sources. For the indexing process we used the Cheshire version of the Porter stemmer and a stoplist that we had used previously for English language databases. A number of separate indexes were created, although the only index used in our submitted runs for the Temporalia TIR sub-track was an index that contained all of the words from the entire record. This approach was the same that we used in the NTCIR-8 and NTCIR-9 GeoTime tracks. In addition, we created a number of indexes that extracted elements from different data sources that shared the same basic XML structure across the collection, with minor variations. Because of the overall structural similarities of the data we could treat all of the records as if they were a single collection, even when drawn from difference sources with differing internal structure. However, as noted in the abstract the current set of submissions is intended to form a baseline for future evaluation, and as such doesn't make explicit use of temporal elements in the data or in the queries.

3. RETRIEVAL APPROACH

Note that much of this section is based on one that appears in our papers from CLEF participation[10, 9].

The basic form and variables of the *Logistic Regression* (LR) algorithm used for all of our submissions were originally developed by Cooper, et al. [4]. 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

Copyright is held by the author/owner.
NTCIR-11 December, 2014, Tokyo
ACM 0-89791-88-6/97/05.

Table 1: Indexes Created for Temporalia

Type	Name	Content	Stoplist
ExactKey	Doc	Doc ID number	n
Keyword	Title	Doc Title	y
Keyword	Topic	Full doc content	y
Date	ParsedDate	Doc Dates - Time expressions	n
ExactKey	TypeName	Named Entity Types	n

that:

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

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:

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

Of course, this last transformation is not actually necessary since the log odds could also be used directly to rank the results, but we do it in the cheshire system so that the result of any operation is a probability value for each item retrieved.

3.1 TREC2 Logistic Regression Algorithm

For NTCIR9 GeoTime we used a version the Logistic Regression (LR) algorithm that has been used very successfully in Cross-Language IR by Berkeley researchers for a number of years[3]. The formal definition of the TREC2 Logistic Regression algorithm used is:

$$\begin{aligned} \log O(R|C,Q) &= log\frac{p(R|C,Q)}{1-p(R|C,Q)} \\ &= c_0 + c_1 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\frac{qtf_i}{ql+35} \\ &+ c_2 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{tf_i}{cl+80} \\ &- c_3 * \frac{1}{\sqrt{|Q_c|}+1}\sum_{i=1}^{|Q_c|}\log\frac{ctf_i}{N_t} \\ &+ c_4 * |Q_c| \end{aligned} \quad (3)$$

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.

When 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 $c_4 = 0.0929$. We have found over time that these coefficients, trained on TREC data and relevance judgements have proved remarkably stable and effective in retrieval from a variety of collections. Although we have occasionally retrained on different corpora we have found only very minor differences in the coefficients produced for these variables. Naturally the same method could be applied with different variables and could potentially, for tasks like Temporalia, include some indication of the temporal relationship between the query and the document. However, this would need to be calculated for each document at retrieval time, since it would depend on the query, and might therefore be better implemented as post-processing of potential results.

3.2 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 "pseudo (or 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 or pseudo relevance feedback has become established in the information

Table 2: Submitted TIR Runs

RunID	Type	Subquery type	P@20	AP@20	Ms nDCG@20	Mean nDCG@20
TIR_BRKLY_TDS_T2	No PRF	all	0.4584	0.3220	0.3383	0.3409
TIR_BRKLY_TDS_T2FB	PRF	all	0.4805	0.3481	0.3566	0.3585
TIR_BRKLY_TS_T2FB	PRF	all	0.5116	0.3811	0.3858	0.3870
TIR_BRKLY_TDS_T2	No PRF	atemporal	0.4606	0.3326	0.3376	0.3395
TIR_BRKLY_TDS_T2FB	PRF	atemporal	0.4606	0.3445	0.3454	0.3472
TIR_BRKLY_TS_T2FB	PRF	atemporal	0.5351	0.4231	0.4139	0.4150
TIR_BRKLY_TDS_T2	No PRF	future	0.5061	0.3573	0.3681	0.3728
TIR_BRKLY_TDS_T2FB	PRF	future	0.5184	0.3792	0.3793	0.3836
TIR_BRKLY_TS_T2FB	PRF	future	0.5357	0.3870	0.3907	0.3927
TIR_BRKLY_TDS_T2	No PRF	past	0.3881	0.2662	0.2872	0.2880
TIR_BRKLY_TDS_T2FB	PRF	past	0.4357	0.3042	0.3191	0.3191
TIR_BRKLY_TS_T2FB	PRF	past	0.4619	0.3441	0.3415	0.3413
TIR_BRKLY_TDS_T2	No PRF	recency	0.4691	0.3245	0.3537	0.3565
TIR_BRKLY_TDS_T2FB	PRF	recency	0.5011	0.3587	0.3774	0.3788
TIR_BRKLY_TS_T2FB	PRF	recency	0.5074	0.3661	0.3920	0.3940

retrieval community due to its consistent improvement of initial search results (in terms of mean average precision) as seen in TREC, CLEF and other retrieval evaluations [8]. The blind relevance feedback algorithm that we use is based on the probabilistic term relevance weighting formula developed by Robertson and Sparck Jones [15].

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 the Temporalia TIR task, 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 3 shows this contingency table.

Table 3: Contingency table for term relevance weighting

	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

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:

$$w_t = log \frac{\frac{R_t}{R-R_t}}{\frac{N_t-R_t}{N-N_t-R+R_t}} \quad (4)$$

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 main LR equation 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.

Note that for this preliminary evaluation we used only the Topic full-text index for Temporalia, and did not attempt to filter results based on date information in relation to the temporal orientation of the various topic statements. The justification for taking no specific actions based on the temporal orientation of the query was that this orientation is already inherently part of the language in the topic content. This assumption has worked in geographic text search, where a similar blind feedback approach achieved better results than systems incorporating explicit geographic methods[11]. However, given the results discussed below, it appears that purely temporal phrasing may require different interventions than those that seemed to be effective for geographic search. In principle, it might be possible to generate temporal constraints that would limit results in the feedback stage to those that are most appropriate for the relative temporal orientation of the query with respect to the documents.

4. SUBMISSIONS AND RESULTS FOR OFFICIAL RUNS

Table 2 shows the results for our three official submitted runs for the Temporalia TIR task. In Table 2 we report the precision at 20 and average precision at 20, the MS normalized discounted cumulative gain at 20, and the mean nor-

malized discounted cumulative gain at 20 each of the query subtypes and combined subtypes for each of the submitted runs.

All of our submitted runs for the GeoTime track used probabilistic retrieval using TREC2 logistic regression algorithm described in detail above. Two of our submitted runs also used pseudo or blind relevance feedback along with the TREC2 algorithm, indicated by "PRF" in the Feedback Type column. For each runid in table 2 those with "TDS" in the runid name used the Title, Description and Subquery elements of the topics, and those with "TS" did not use the Description. As the scores in table 2 show, using both the title and subquery elements along with blind feedback gives the best results for this collection, and including the description leads to slightly worse results. In all cases

In comparing the results for these runs compared to those reported for other participating groups we see that our baseline results are marginally better or worse than the organizer's default baseline systems, depending on the temporal orientation of the topic, but definitely worse than other participants' systems. We assume that all of the other systems (with the exception of the organizers' default systems) make use of explicit temporal information in processing their queries.

4.1 Analysis of Search and Feedback Failures

We examined the results for individual queries to see if our approach was successful for particular types of topic, and found a (very) few situations where one of the other participant's approaches were not more successful in retrieving and ranking the results. We primarily used nDCG@20 for this analysis, since it is based on the "gold standard" of an optimal ranking of results, and takes both relevance and ranking into account. In addition we looked at Precision@20 to attempt to discover differences in the results due to minor variations in rank ordering when compared to nDCG@20.

One discovery in this analysis was that there were a number of topics where no results were found, and where there was no entry at all in the submitted runs - possibly due to a query parsing error. However for all topics where we had a relatively high score for nDCG@20, usually many other systems performed better.

None of our runs achieved the highest nDCG@20 score for any individual topic regardless of temporal orientation. The best scores we achieved were the second highest scores in both topic 010f and 017f. Among our runs, as indicated by the average results shown in Table 2, the TIR_BRKLY_TS_T2FB run was the best-performing run of those submitted with 74 out of the 200 topics exceeding the mean.

In the analysis we found that only 175 out of the 600 topics submitted for all of our runs exceeded the mean nDCG@20 score, and a similar result of 182 out of 600 of our scores exceeded the mean Precision@20 over all participants.

5. CONCLUSION

This paper has described Berkeley's submissions to NTCIR-11 Temporalia TIR task. We intend to conduct a number of further experiments with the data and relevance judgements, and to see how temporal filters and restrictions affect our results. As noted in the begining this submission was intended to provide a baseline for further evaluation. We fully intend to exploit some of special indexing tools developed for the Cheshire system in the future that can take advantage of time differentials and other approaches for temporal retrieval in the future.

Additionally, as we noted in the discussion of the ranking algorithm and relevance feedback steps above, we used only our logistic regression-based ranking method, with and without blind feedback for this preliminary evaluation. The only index used was the Topic full-text index for Temporalia, and we did not make any attempt to filter results based on date information in relation to the temporal orientation of the various topic statements. As also noted above, it might be possible to generate temporal constraints that would limit results in the feedback stage to those that are most appropriate for the relative temporal orientation of the query with respect to the documents, but this has not been implemented or tested yet.

Overall, given the goals of the Temporalia task, our approach of attempting to use only the text of the topics, for fairly standard text retrieval with blind relevance feedback was not able to compete with methods that did attempt special processing for temporal constraints in the topics. We look forward to hearing how these other methods worked and how they were able to achieve their much better results.

6. REFERENCES

[1] A. Chen. Multilingual information retrieval using english and chinese queries. In C. Peters, M. Braschler, J. Gonzalo, and M. 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] A. Chen. *Cross-Language Retrieval Experiments at CLEF 2002*, pages 28–48. Springer (LNCS #2785), 2003.

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

[4] W. S. Cooper, F. C. Gey, and D. 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.

[5] F. Gey, R. Larson, N. Kando, J. Machado, and T. Sakai. NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 0–0, 2010.

[6] F. Gey, R. Larson, J. Machado, and M. Yoshioka. NTCIR9-GeoTime overview - evaluating geographic and temporal search: Round 2. In *Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011, Tokyo, Japan*, pages 9–17, 2011.

[7] H. Joho, A. Jatowt, R. Blanco, H. Naka, and S. Yamamoto. Overview of ntcir-11 temporal information access (temporalia) task. In *Proceedings of the NTCIR-11 Conference, Tokyo, Japan*, 2014.

[8] R. 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] R. 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, Sept. 2008.

[10] R. R. Larson. Experiments in classification clustering and thesaurus expansion for domain specific cross-language retrieval. In *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, LNCS 5152, pages 188–195, Budapest, Hungary, Sept. 2008.

[11] R. R. Larson. Cheshire at GeoCLEF 2008: Text and fusion approaches for GIR. In *Evaluating Systems for Multilingual and Multi-modal Information Access: 9th Workshop of the Cross-Language Evaluation Forum: CLEF 2008*, pages 830–837. Springer (Lecture Notes in Computer Science LNCS 5706), 2009.

[12] R. R. Larson. Logistic regression for IR4QA. In *Proceedings of the NTCIR-8 Workshop, Tokyo, June 2010*, pages 0–0, 2010.

[13] R. R. Larson. Probabilistic text retrieval for NTCIR9 GeoTime. In *Proceedings of NTCIR-9 Workshop Meeting, December 6-9, 2011, Tokyo, Japan*, pages 33–37, 2011.

[14] M. Matthews, P. Tolchinsky, R. Blanco, J. Atserias, P. Mika, and H. Zaragoza. Searching through time in the new york times. In *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval, NJ, USA*, pages 41–44, 2010.

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