The Use of Inference Network Models in NTCIR-9 GeoTime

Christopher G. Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242

christopher-harris@uiowa.edu

ABSTRACT
We describe our approach in identifying documents in a specific collection that are able to sufficiently answer a series of geo-temporal queries. This year, we submitted four runs for the NTCIR-9 GeoTime task, using the Indri search engine, on a collection of English-language newspaper articles. Our four submitted runs achieved nDCG scores ranging from 0.5275 to 0.6254 and MAP ranging from 0.4164 to 0.4990 across twenty-five separate geo-temporal queries.

Categories and Subject Descriptors
H.3.3 Information Storage and Retrieval – Retrieval Models

General Terms
Algorithms, Performance, Design, Experimentation

Keywords
Geographic Information Retrieval, Temporal Retrieval, Indri, Inference Network Models, Language Models

1. INTRODUCTION
At NTCIR-9, we participated in the GeoTime task for a second year. This task, consisted of providing a ranked document set for a series of geographic and temporal searches in the form of questions. Although the task had both Japanese and English sub-tracks, we participated in the English sub-track only.

For NTCIR-9 GeoTime, both a narrative portion and a descriptive portion of twenty-five distinct XML-formatted queries were provided to participant teams. The narrative provided additional information that was not present in the description. Detailed information about the NTCIR-9 GeoTime task can be found in the task overview [3].

The remainder of this paper is organized as follows. In Section 2, we describe the experimental system we implemented for this task and the basic retrieval models. Section 3 describes the ideas we incorporated in our submitted runs. In Section 4, we present the experimental results for our four runs and discuss how some of the components incorporated into each model affected the results. We conclude our discussion in Section 5.

2. SYSTEM DESCRIPTION
As with our approach in NTCIR-8 GeoTime, at the core of our retrieval system was the Indri search engine [9], an open source component of the Lemur Language Modeling Toolkit. The retrieval model implemented in Indri combines language modeling [7] with an inference network [10].

In the language modeling framework, a document is usually represented as a sequence of tokens or terms. Then a language model estimating the characteristics over the vocabulary can be made. With Indri, documents are represented as binary feature vectors. The features can be nearly as interesting as observations on the underlying text.

Indri's language model assumes a single feature vector for each position within a document, although in general this may not be the case. This allows Indri to model text and, more importantly, *features* of text. Although it is widely viewed that Indri establishes and exclusively makes use of language models, a better description for these models would be language feature models, as first described in [8]. These language models are then combined with inference network models to produce a ranked set of results.

To explain how inference network models work, we provide the following description. Given a query q that consists of several query terms (q_1, q_2, ... q_n) and a document d, the occurrence of each of these individual query terms, q_i, are assumed to be independent from the occurrence of the other query terms. Therefore, the likelihood of the entire query can be calculated as the product of the likelihood of each individual query term appearing in a specific document [2]:

$$P\ (q \mid d) = \prod_{q_i} P\ (\ q_i \mid d)$$

Indri allows us to create a separate index for a defined portion of the document (the portion of a document is called an *extent*). In our work in GeoTime, we specify a separate extent for the article's *title*, the article's *dateline*, and for the article's *body*, allowing us to combine beliefs, or probabilities of term occurrence, on each extent. In addition, these extents can be nested and weighted, permitting substantial flexibility in our retrieval model.

In our model, for all four submitted runs, we evaluated each document in our collection as a single extent, primarily for simplicity - the geo-temporal information could appear in any part of a given document. Earlier experiments demonstrated that using different extents and indexing them separately provided no retrieval benefit for this study.

We eliminate stop words using the standard stop-list in Indri and derive our query terms. The Indri model seeks to determine P(r | θ), or the probability that a particular query term, *r,* occurs in our context language model, θ.

At indexing time, the document is divided into three separate extents: *headline*, *dateline*, and *body*, with the smoothing parameters α,β_{head}, α,β_{date}, and α,β_{body} applied to each, respectively. Feature language models θ_{head}, θ_{date}, and θ_{body} are built specific to each document in our collection. Indri's inference engine assumes r approximates Bernoulli(θ) [6]. We applied different smoothing methods in this year's runs, allowing us to experiment with their ability to retrieve and rank results.

The retrieval examines the representation concept nodes, r_i, constructed over our collection model, C, based on Bernoulli's conjugate prior, with $\alpha_w = \mu P(w \mid C) + 1$ and $\beta_w = \mu P(w \mid C) + 1$ (Note that μ is a Dirichlet smoothing parameter; in Section 3.3 we discuss the application of different smoothing methods and parameters in this year's runs, allowing us to experiment with their ability to retrieve and rank results).

The probability of a representation concept node, r_i, being satisfied by the smoothing parameters α,β_{head}, α,β_{date}, and α,β_{body} in any given document D is therefore:

$$P(r_i \mid \alpha,\beta,D) = \int_{\theta} P(r_i \mid \theta)\, P(\theta \mid \alpha,\beta,D) = \frac{tf_{w,D} + \mu P(w \mid C)}{|D| + \mu}$$

Note that in UIOWAs approach in GeoTime, we use representation concept nodes r_i from the *headline* and *body* extents - but not from the *dateline* extent – to satisfy a given query.

The final inference engine step is the creation of the final 'information need' node, which combines the belief node scores into a single score for ranking the document based on the query terms provided

We present some examples of our queries in the next section. For additional information on the use of specific INQUERY language syntax used by UIOWA for Indri geo-temporal queries, see [5].

3. EXPERIMENTS

UIOWA submitted four runs to NTCIR-9 GeoTime. Each of these runs examined the same collection of English language newspaper articles from four different sources. It represents an expansion over the previous NTCIR GeoTime document collection. Table 1 summarizes the approach of each run. Subsections 3.1 through 3.3 describe the variables used in each of these approaches in more detail.

Table 1. Description of Submitted UIOWA Runs

Run UIOWA-EN-EN-	Use Desc only (D) or both Narr and Desc (DN)?	Stemmer Used	Smoothing Method Used
01-D	D	Krovetz	Single-stage
02-DN	DN	Krovetz	Two-stage
03-DN	DN	Porter	Two-stage
04-DN	DN	Krovetz	Single-stage

We decided to incorporate the use of additional geographic information as synonyms to expand our queries of identified geographic terms. To accomplish this, we first used a lexical parser to separate and identify geographic entities and then incorporated the use of 5.9 million geographic terms obtained from the Alexandria Digital Library[1][1]. We used these terms to expand our queries in Indri for all four of our runs. Although all of our submitted runs used the Alexandria Digital Library, our preliminary examination found that adding these synonyms improved our results significantly.

In each of our runs, we attempted to limit our manual intervention in the creation of our Indri queries as much as possible. We did not add temporal terms (such as specific dates) to our queries unless the information was provided in the narrative or description and important to the topic. For example, in topic GeoTime-0034:

```
When and from what airport was an ANA
plane hijacked and a pilot killed?
```

Our Indri query does not include the date as a term:

```
#combine( <#1(All Nippon Airlines)
ANA> hijack pilot #1(pilot killed))
```

Through preliminary examinations in Indri, we determined that manually adding a date (or set of dates) in the query is unnecessary because (1) ANA planes being hijacked are a relatively rare occurrence and (2) the domain of newspaper articles in this task sufficiently limits our query to the most relevant articles automatically. In contrast, the description of topic GeoTime-0042:

```
Describe the name of the country of
Middle East whose King died in 1999.
```

Indicates that the year '1999' is important to our query, and therefore it is explicitly provided:

```
#combine( <#1(Middle East) Egypt
Lebanon Iran Turkey Israel Kuwait
#1(Saudi Arabia) Algeria Syria
Morocco Iran Iraq Yemen Oman Qatar
Palestine Jordan Tunisia> <king
ruler> <died die #(pass away)> 1999)
```

One topic (GeoTime-0037) produced an additional challenge, since the question only provided geographic coordinates of the event, but did not provide a named location:

```
What fatal accident occurred near
(geographical coordinates 5°52′12″N
5°45′00″E / 5.870°N 5.750°E / 5.870;
5.750), which killed hundreds of
people, and when did it occur?
```

We addressed this topic in the following way. We used an API call to GeoNames[2] - a tool that provides Wikipedia entries that occurred at or near the coordinates given:

http://api.geonames.org/findNearbyWikipedia?lat=5.875&lng=5.75&username=demo

[1] http://alexandria.ucsb.edu/gazetteer/

[2] http://geonames.org

This API call produces an XML-formatted result set that includes the both the Wikipedia article title and a brief summary. The following two titles (along with a one-sentence summary of each) appeared in our result set:

```
1998 Jesse Pipeline Explosion
Sapele, Delta
```

Using the Stanford Parser [3], we parsed the titles and the corresponding summaries returned from this API and extracted the nouns and noun phrases. We used these as inputs for our Indri query.

```
#combine( #1(Jesse pipeline
explosion) #1(October 18 1998)
#1(pipeline explosion) community
coordinates southeast #1(Lagos
Nigeria) cause blast #1(Nigerian
government) place scavengers tools
Sapele city #1(Delta State Nigeria)
#1(Benin River) Confluence #1(Ethiope
River) #1(Jamieson River) Urhobo
people century #1(trading village))
```

Although it may be obvious that the first article is more relevant to the topic than the second article, we included the titles and summaries from both articles in our Indri query. Discerning which articles are relevant and which are not would require manual inspection; as stated earlier in this section, the underlying purpose in following these steps was to limit our manual intervention in the query creation process as much as possible.

3.1 The Use of Description and/or Narrative

Two different XML elements were provided for each query. For each run, participants had the option of using the description only, the narrative only, or both. In last year's NTCIR-8 GeoTime, we achieved our best results using the description only [5]. This would suggest we can achieve our best results by ignoring the narrative field entirely; however, clearly there is valuable additional information in the narrative that may not be have been utilized efficiently in last year's runs. This year, we keep one (Run 1) with description only and three other runs use both the narrative and description.

3.2 Choice of Stemmer

Since Indri allows us to use a different stemmer for each index, we created two separate indexes – one using the default Krovetz stemmer and one using the Porter stemmer, which Indri also supports. This provides us an opportunity to examine the stemmer's effects on retrieval with Indri. Only Run 3 uses the Porter stemmer whereas the other three submitted runs use the default Krovetz stemmer.

3.3 Smoothing Methods and Parameters

Query evaluation proceeds in two stages. In the first stage, statistics about the number of times terms and phrases appear in the collection are gathered. In the second stage, the statistics from the first stage are used to evaluate the query against the collection. Thus, it appears that two-stage smoothing should be better able to filter out many of the noisiest documents when the collection is large enough to permit likelihood estimates, which is indeed the case with the GeoTime document collection.

The default smoothing method in Indri is single-stage Dirichlet, with a parameter of $\mu = 2500$. In Run 1 and Run 4, we kept this default parameter; for Runs 2 and 3, we incorporate two-stage Jelinek-Mercer smoothing with $\mu = 1500$ and $\lambda = 0.4$. Additional information about smoothing methods can be found in [11].

4. RESULTS

Overall, all four of our submitted runs surpassed the English-only subtask averages across all five reported metrics (MAP, Q, and nDCG@10, nDCG@100, nDCG@1000). Table 2 shows three metrics: MAP, Q, and nDCG@1000, for each of our four runs. For comparison we provide the metric from our best run from NTCIR-8 and the NTCIR-9 GeoTime English subtask averages as well. The highest result for each metric is displayed in bold.

Run 1 used only the description only. We added information from the narrative (and the geographic synonyms from the Alexandria Gazetteer) for Runs 2, 3 and 4. For all runs, we used the automatic methods and default weighting provided by Indri. In NTCIR-8 GeoTime, we learned that Indri's native ability to determine term rarity based on our collection is slightly better than our ability to correctly assign weights.

Table 2 shows us that all of our NTCIR-9 runs were better than our best NTCIR-8run. Although this year's dataset was larger than that used in NTCIR-8 and we had to evaluate different queries, we believe our techniques incorporating the geographic information and ability to better tune our queries in Indri were collectively responsible for a large part of this gain. Run 3 - our best run – incorporated the description and narrative, used the Porter stemmer, and used the two-stage smoothing method.

Table 2. Selected Metrics for Submitted UIOWA NTCIR-9 GeoTime Runs

Run **UIOWA- EN-EN-**	**MAP**	**Q**	**nDCG@1000**
01-D	0.4164	0.4372	0.6425
02-DN	0.4955	0.5134	0.6919
03-DN	**0.4990**	**0.5197**	**0.6998**
04-DN	0.4869	0.5069	0.6889
01-D (from NTCIR-8 GeoTime)	0.3971	0.4162	0.6228
NTCIR-9 GeoTime English Subtask Average	0.3517	0.3510	0.5684

By changing one parameter per run, we are able to examine the relative benefits of each technique. In Table 3, we report each of the parameters discussed in Section 3 and the effects of each. Note that our process of determining improvement is not using the same baseline, so is only able to determine the cumulative gain for adding each additional modification. We would need to establish gains from the same baseline run to determine absolute improvements for each change.

Table 3. Changes Implemented and Improvements Observed for each of our Runs

Runs Compared	Change Made	Improvement	
Run 1 - Run 4	Added narrative	MAP:	0.0705
		Q:	0.0697
		nDCG@1000:	0.0464
Run 2 - Run 4	Smoothing method	MAP:	0.0086
		Q:	0.0065
		nDCG@1000:	0.0030
Run 2 - Run 3	Stemmer from Krovetz to Porter	MAP:	0.0035
		Q:	0.0063
		nDCG@1000:	0.0079

In Table 3, we observe that each change made increased our scores across all metrics; this was true for the two nDCG metrics not reported as well. In contrast to our observation in NTCIR-8 GeoTime, the inclusion of the narrative text provides a substantial improvement over using the description text alone. Likewise, using the two-stage Jelinek-Mercer smoothing method also improves our metrics over the default Dirichlet single-stage method. Finally, we observe that using the Porter stemmer increases our score over the default Krovetz stemmer for this newspaper collection dataset.

5. CONCLUSIONS

We applied Indri, an inference network model to the NTCIR-9 GeoTime newspaper collection. Beginning with the approach we used in last year's NTCIR-8 GeoTime, we added additional enhancements to improve our scores across all metrics. Using the description, along with the Porter stemmer and the two-stage smoothing method provided us with our best score. The proper settings for these parameters is collection-dependent, and we note there are a number of parameters in Indri we have yet to explore, which might further improve our results.

6. REFERENCES

[1] Alexandria Digital Library Gazetteer. 1999- . Santa Barbara CA: Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Copyright UC Regents. http://www.alexandria.ucsb.edu/gazetteer

[2] Cao, G, Nie, J, and Shi, L. NTCIR-7 Patent Mining Experiments at RALI. In the *Proceedings of NTCIR-7 Workshop Meeting*, December 16-19, 2008, pp 347-350.

[3] de Marneffe, M., MacCartney B., and Manning, C.D.. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

[4] Gey, F., Larson, R., Machado, J., and Yoshioka, M. NTCIR9-GeoTime Overview - Evaluating Geographic and Temporal Search: Round 2. In the *Proceedings of NTCIR-9 Workshop Meeting*, December 6-9, 2011.

[5] Harris, C.G. Geographic Information Retrieval Involving Temporal Components. In the *Proceedings of NTCIR-8 Workshop Meeting*, June 15-18, 2010, pp. 185-189.

[6] Metzler, D. Indri Retrieval Model Overview. July 2005. http://ciir.cs.umass.edu/indriretmodel.html. Retrieved on September 16, 2011.

[7] Ponte, J. and Croft, B. A language modeling approach to information retrieval. In *Proceedings of SIGIR*, pp.275-281, 1998.

[8] Turtle, H. and Croft, B. Evaluation of an inference network based retrieval model. In *ACM Transactions on Information Systems*, 9(3):187-222, 1991.

[9] Si, L., Jin. R., and Callan, J., A language modeling framework for resource selection and results merging. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, McLean, VA, November 4-9, 2002. pp 391-397.

[10] Strohman, T., Metzler, D., Turtle, H., and Croft, B. Indri: A language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*. McLean, VA, May 2-6, 2005.

[11] Zhai C. and Lafferty, J. Dual role of smoothing in the language modeling approach. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR) 2001*, pp. 31–36.