Geographic Information Retrieval Involving Temporal Components

Christopher G. Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242-1420

christopher-harris@uiowa.edu

ABSTRACT
Searches for information on the web involving both geographic ("where") and temporal ("when") components comprise a non-trivial percentage of overall searches. In this paper, we describe an approach to identifying specific documents within a collection that satisfy a set of geo-temporal queries. To test our approach, we submitted five runs to NTCIR-8 GeoTime, using the Indri search engine, on a three-year collection of English-language newspaper articles. Our five submitted runs achieved nDCG scores ranging from 0.5758 to 0.6233 and MAP ranging from 0.3517 to 0.3951 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

1. INTRODUCTION
At NTCIR-8, we participated in the GeoTime task, which involved geographic and temporal searches on a named entity or set of named entities. Although the task had both Japanese and English sub-tracks, we participated in the English sub-track only.

Queries that request geographic-based information, such as a specific location, are relatively common on the web. These queries can be formulated to inquire where an event or entity was located, is located, or will be located in the future. A temporal aspect to the query - involving the specific time (when) an event occurred is occurring, or will occur - was added as a component to the search criteria; thus many queries were provided in the format of "*where and when did <entity><action clause>*?"

There are several challenges with identifying Geo-temporal information in documents. Many of these have to do with the inconsistency of how geographic information and temporal information is worded in documents. Articles that mention the city of Chicago, for example, could describe it as "Chicago", "Chicago, Illinois", "The Windy City", "Second City", "the most populous city in Cook County", "Chi-town", "Chicagoland", or by geographic features, such as the coordinates of its centroid ("41.840675 N, 87.679365 W"), or by its landmarks ("The Loop", "McCormick Place", "Soldier Field") – however, some of these landmark names could conceivably be shared with other locations and require further disambiguation. Temporal information can be very concrete ("Tuesday, April 6, 2010, 12:00:00 noon CDT"), vague ("late last month"), or ambiguous ("the best time of the year"), creating its own challenges as well. For this year's task, we were asked to identify a set of documents containing the correct geographic and temporal identifiers to satisfy a set of queries. Some of these queries required some disambiguation of named entities, geographic components, temporal components, or a combination involving all three.

For NTCIR-8 GeoTime, both a narrative portion and a descriptive portion of twenty-five distinct XML-formatted queries were provided to participants. The narrative often provided additional useful information not present in the description. As mentioned, the objective was to return a ranked list of documents that satisfied both components of the given query. Detailed information about the NTCIR-8 GeoTime task can be found in the task overview [2].

The remainder of this paper is organized as follows: in Section 2, we will 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 five runs and discuss how some of the components incorporated into each model affected the results. Section 5 summarizes this paper and briefly discusses future directions we envision for geo-temporal searching.

2. SYSTEM DESCRIPTION
At the core of our system was the Indri search engine [8], which is an open source component of the Lemur Language Modeling Toolkit. The retrieval model implemented in Indri combines language modeling [5] with an inference network [4, 6, 9]. See Figure 1 for a pictorial description of how Indri satisfies a given query. If we have 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 [10]. 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 [1]:

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

Indri allows us to create a separate index for a defined portion of the document (the portion of a document is called an *extent*). For example, we could 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. This allows for substantial flexibility in our retrieval model.

In our model, for all five 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. Figure 1 illustrates how the Indri inference model could retrieve and score results for Query GeoTime-0001:

```
When and where did Astrid Lindgren
die?
```

For this query illustration, it is conceivable to search using only two query terms, q_1 = 'Astrid' and q_2 = 'Lindgren' - other terms are either stop-listed or ignored. 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.

In the example shown in Figure 1, 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) [3].

The retrieval examines the representation concept nodes, r_i, constructed over our collection model, *C*, based on Bernoulli's conjugate prior, with α_w = μP(w | C) +1 and β_w = μP(w | C) +1 (Note that μ is a Dirichlet smoothing parameter). 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:

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

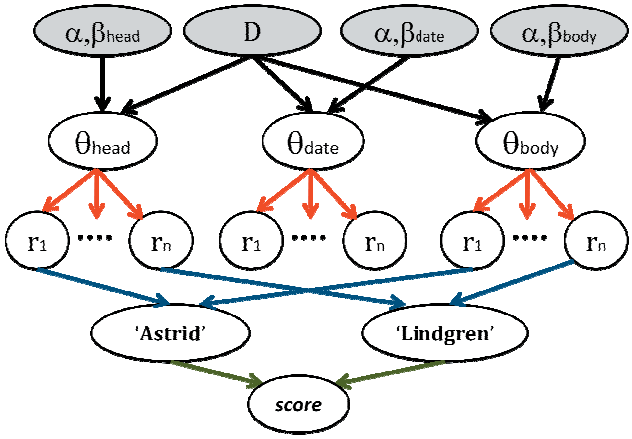

Figure 1. The Indri retrieval model arriving at a final score for a given document, *D*, based on the query GeoTime-0001 ("When and where did Astrid Lindgren die?")

In Figure 1, which illustrates an Indri retrieval that could be applied to Query GeoTime-0001, it can be observed that representation concept nodes r_i from the *headline* and *body* extents - but not from the *dateline* extent – are used to satisfy this given query. The final 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.

3. EXPERIMENTS

Five runs were submitted to NTCIR-8 GeoTime. These runs examined a collection of 315,417 New York Times newspaper articles from January 1, 2002 to December 31, 2005 provided in XML format by the Linguistic Data Consortium [7]. 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 the Five Submitted UIOWA Runs to the NTCIR-8 GeoTime Task

Run UIOWA-	Uses Description (D), Narrative (N) or Both (DN)?	Automatic (A) or Manual (M)?	Uses Probabilistic Weighting?
EN-01-D	D	A	No
EN-02-D	D	M	No
EN-03-DN	DN	A	No
EN-04-DN	DN	M	No
EN-05-DN	DN	M	Yes

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. We discovered that in some queries, the description alone was sufficient, particularly when the entity was clearly identified. For example, the description of Query GeoTime-0005 asks:

```
When and where did Katharine Hepburn
die?
```

Whereas the narrative for Query GeoTime-0005 provides little additional information to aid in constructing our query:

```
The user is investigating the actress
Katharine Hepburn and wants to know
when and where she died.
```

There are several queries where information provided in the narrative is important. To illustrate, the description of Query GeoTime-0010 asks:

```
When was the decision made on siting
the ITER and where is it to be built?
```

The narrative for Query GeoTime-0010 provides additional relevant details on the entity ITER not conveyed in the description:

> The ITER (International Thermonuclear Experimental Reactor) is an experimental facility for conducting international joint research on the feasibility of fusion energy. When was the decision made on where to build the facility and where is it to be sited?

In our first two runs, we included only the description in our query; in the last three runs, we included information from both the description and the narrative.

3.2 Automatic or Manual

For this task, we had the option of constructing our query automatically by simply putting the text of the description and/or narrative into the our model and allowing Indri to construct the query based on default parameters, or by manually specify the query construction such as manual specification of synonyms for specific terms or manually assigning relative weights to each term. In our manual runs (Runs 2, 4, and 5), we manually specified synonyms for several key terms. For example, the description for Query GeoTime-0006 asks:

> When and where did anti-government demonstrations occur in Uzbekistan?

The narrative for Query GeoTime-0006 is:

> The user wants to know what month and year an anti-government riot took place in Uzbekistan that was put down by military force. The user also wants to know where in Uzbekistan this took place.

The Indri query submitted in Run 4 for Query GeoTime-0006 is:

```
anti-government {riot uprising} {Uzbekistan Uzbek Uzbeki} {suppress put down} military force
```

The curly braces {} group synonyms together as a single query term much like an OR comparison; for example, {riot uprising} indicate synonyms for the same type of event. There are many other terms that could possibly refer to a riot, but we used some domain knowledge to understand which terms would be likely and which would be exotic to a newspaper article describing a riot. Terms such as "the user wants to know" and "where and when" are very unlikely to appear in the text of a news article containing the required geo-temporal information, so they are removed from our query as well.

Indri allows for terms to appear in a specific order, or for terms to appear within *n* terms of another term (unordered weighting). For example, the ordered query `#1(white house)` only returns values where the word "house" immediately follows "white", but `#uw2( white house)` allows results where the terms "house" and "white" appear in any order, but within two terms of each other, such as "white house", "house painted white", "white boat house", etc. We applied both types of constructs in our manual queries - depending on how likely a given term is to appear in proximity of another term in the document.

3.3 Probabilistic Weighting

For Run 5, we used a probabilistic weighting scheme to provide a relative weighting of query terms. The query submitted in Run 5 for Query GeoTime-0005 was:

```
#weight(2.0 actress 2.0 Katharine 5.0 Hepburn 1.0 {died dead}
```

Here we apply relative weighting to each of the four query terms. "Hepburn" has a weight that is 2.5 times the weight of "Katharine", and was determined by the anticipated rarity of the term (in other words, "Hepburn" is a rarer term than "Katharine" by an estimated factor of 2.5). Estimating these weights required us to apply some prior domain knowledge of term rarity.

4. RESULTS

Overall, all five of our submitted runs surpassed the English-only subtask averages across all three metrics (MAP, Q, and nDCG). Table 2 shows the metrics for each of our runs and the NTCIR-8 GeoTime English subtask average (the highest result for each metric is displayed in bold).

Table 2. Overall Metrics for Submitted NTCIR GeoTime Runs

Run UIOWA-	MAP	Q	nDCG
EN-01-D	**0.3971**	**0.4162**	0.6228
EN-02-D	0.3605	0.3765	0.5758
EN-03-DN	0.3800	0.3933	**0.6233**
EN-04-DN	0.3517	0.3689	0.5931
EN-05-DN	0.3659	0.3834	0.5849
NTCIR GeoTime English Subtask Average	0.3173	0.3329	0.5317

Table 2 shows that Runs 1 and 3 - our automatic runs - slightly outperformed Runs 2, 4 and 5, where we manually modified the query terms and (in the case of Run 5) applied probabilistic weighting. This either indicates the strength of Indri's native retrieval engine or the sub-optimal performance of the relative weighting, proximity measures, and selected synonyms as applied to Runs 2, 4 and 5.

Although Run 1 outperformed Run 3 overall, indicating that including the narrative may have actually hampered performance, the difference was subtle compared with the difference between the automatic and manual runs. Although there is no observed difference between the MAP, Q, or nDCG scores of our runs at the $p = 0.05$ level of significance, the lack of improvement using probabilistic weighting (Run 5) over our other runs may indicate that Indri's native ability to determine term rarity based on our collection is slightly better than our ability to correctly assign weights.

To see where our models performed well and where they performed poorly, it is helpful to identify those queries for each run that show these extremes. Table 3 shows the highest and lowest nDCG scores for each run, as well as for the NTCIR-8 GeoTime English-only subtask average. With the exception of Runs 2 and 5, all runs had some difficulty with Query GeoTime-0021 - the narrative of which appears below:

> The International Olympic Committee decides when and where the next Winter Olympics is held. When was this announcement made for the next Winter Olympics, and from what city was it made?

The challenge is that the date and the site of the 2010 Winter Olympics was *not* being requested, but the location and the date of the decision. Understandably, this query provides a challenge for information retrieval methods.

For most of the runs, as well as for the task as a whole, Query GeoTime-0001 was the most straightforward. For example, consider the narrative for Query GeoTime-0001 (the description for Query GeoTime-0001 was provided in Section 2):

> The user wants to know when and in what city the children's author Astrid Lindgren died.

The relative rarity of the query terms (it is unlikely much news appears in our collection containing both of the terms "Astrid" and "Lindgren", except for the notice of her death), as well as the fact that her death appeared at the very beginning of our time-ordered window of documents (she passed away on January 28, 2002), limit the number of other news articles where potentially confounding information is likely to appear. Therefore, only truly relevant documents are likely to be retrieved.

Table 3. Overall Best and Worst nDCG Scores for Each Run

Run UIOWA-	Best Query (nDCG)	Worst Query (nDCG)
EN-01-D	GeoTime-0001 (1.0)	GeoTime-0021 (0.1861)
EN-02-D	GeoTime-0001 (0.9691)	GeoTime-0022 (0.0)
EN-03-DN	GeoTime-0008 (0.9256)	GeoTime-0021 (0.0841)
EN-04-DN	GeoTime-0001 (0.9612)	GeoTime-0021 (0.0696)
EN-05-DN	GeoTime-0001 (0.9970)	GeoTime-0010 (0.0251)
NTCIR GeoTime English Subtask Average	GeoTime-0001 (0.9493)	GeoTime-0021 (0.1192)

5. CONCLUSIONS

Although our methods performed relatively well in the English-language only subtask of NTCIR-8 GeoTime, we clearly see that there is room for improvement in geo-temporal search methods. This is particularly true for queries that involve more complex queries (such as Query GeoTime-0021), where the named entity cannot be clearly identified, or where geographic or temporal aspects are defined relative to another place or time. We discuss some of these challenges in Section 5.1 and then consider some potential future directions for geo-temporal search in Section 5.2.

5.1 Some Challenges in Geo-temporal Search Techniques

One of the challenges of geo-temporal retrieval systems includes the ability to identify locations relative to an understood geographic entity; for example, "in this town" would serve as a inferred reference to New York City if it were to appear in a New York Times newspaper article. However, it is unlikely the article will appear in our retrieved list of documents unless it is possible for our language model to map "in this town" to "New York City" for New York-specific articles.

One benefit of searching news articles is that news articles are all dated and thus contain a specific time in a defined format. However, this may not be the case if the collection includes a collection of notes, blog entries, or other documents that do not contain a timestamp. Likewise, articles that refer to relative dates require additional resolution by the search engine.

Additionally, we must know some details about the constraints of our collection. To illustrate, consider the narrative of Query GeoTime-0015:

> The Winter Olympics are held every four years. In which year and in what city were the last three olympics held?

Satisfying this query requires us to know the starting, or "anchor", point of our query window so that we can search forward or backwards relative to this anchor point. For example, it could be when we issued the query (approximately February 1, 2010), the end of our document collection (December 31, 2005) or some other specified date? Since the query is linked to an unspecified point in time (it only specifies the "last three Olympics"), the determination of this anchor point has the potential of affecting which documents are retrieved.

One last issue is determining precedence in document rank. When a given document only satisfies the temporal, or "when" component whereas a second document only satisfies the geographical, or "where" component, which should be ranked higher in our retrieval list? It can be argued that the rarer of the two should be ranked higher, but determining this trade-off for each query may be cumbersome to perform quickly.

5.2 Possible Future Directions of Geo-temporal Search

One future direction to geo-temporal search is the ability to apply diversity measures to the results (so that searches on "Portland" will include results for both Portland, Oregon, and Portland, Maine, and other geographic entities named "Portland", such as Portland, Dorset, UK, Portland, Texas, the Isle of Portland, and

Portland, Connecticut, or several dozen other geographic entities containing 'Portland'). This diversity metric will keep one larger geographic entity from completely "crowding out" another one of a similar name, permitting the user to observe results from both or apply a diversity weighting measure to permit more of a balance between entities. Another direction is to provide a three-dimensional visual representation of the retrieval set (document rank or relevance, granularity of geographic location, and time could be the three axes). Additionally, the integration of community-based tools, such as Wikipedia, into retrieval methods could aid in disambiguation of geographic entities.

6. ACKNOWLEDGMENTS

We thank Padmini Srinivasan for her guidance, Yelena Mejova for her useful comments and feedback, the NTCIR-8 Geotime organizers for providing an interesting yet challenging task, and the Linguistic Data Consortium for making the document collection available for research.

7. REFERENCES

[1] 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.

[2] Gey., F., Kando, N., Machado, J., Sakai, T. NTCIR-GeoTime Overview: Evaluating Geographic and Temporal Search. In the *Proceedings of NTCIR-8 Workshop Meeting*, June 15-18, 2010.

[3] Metzler, D. Indri Retrieval Model Overview. July 2005. http://ciir.cs.umass.edu/indriretmodel.html. Retrieved on April 10, 2010.

[4] Petkova, D. and Croft, W. B. 2007. Proximity-based document representation for named entity retrieval. In *Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management*. Lisbon, Portugal, November 06 - 10, 2007. pp 731-740.

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

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

[7] Sandhaus, E. *The New York Times Annotated Corpus 2002-2005*. LDC2008T19. DVD-Rom. Philadelphia: Linguistic Data Consortium, 2008.

[8] 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, 2002.

[9] 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.

[10] Xu, J. and Croft, B. Query expansion using local and global document analysis. In the *Proceedings of SIGIR*, pp.4-11, 1996.