Geotemporal Access to Multilingual Documents

Fredric C. Gey
UC Data Archive
University of California
Berkeley, CA 94720
510-643-1298
gey@ucdata.berkeley.edu

Kim Carl
Electronic Cultural Atlas Initiative
University of California
Berkeley, CA 94720
kcarl@uclink.berkeley.edu

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval H.5 Information Interfaces and Presentation

General Terms

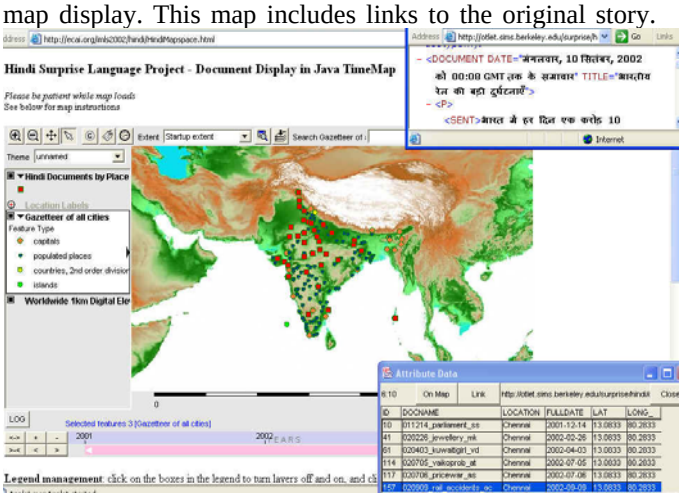
Design, Experimentation

Keywords

Cross Language Information Retrieval, Geographic Information Retrieval, Gazetteers

1. GEOTEMPORAL ACCESS TO TEXT

Most news documents are about something which happened in some place at some particular time. Thus news documents can, in principle, be searched on these dimensions. The reality, as with all information retrieval, is more complex. Place names are ambiguous (e.g. is Aberdeen in Scotland, Maryland or Sri Lanka?) and time periods are described with indistinctness (e.g. "Gulf War Period"). Even more complex is text in an unfamiliar natural language. However if these problems can be overcome, *geographic and temporal querying* can be done using a map paradigm as the interface. A capability to be demonstrated at SIGIR 2004 shows querying for Hindi documents for the TIDES Surprise Language Exercise of June 2003 using a geographic and temporal editing and browsing interface, TimeMap developed and moved to open source by University of Sydney and the Electronic Cultural Atlas Initiative group at UC Berkeley. Place name entity extraction and gazetteer entries locate the latitude/longitude of the place, then load the time and place of the story into a map display. This map includes links to the original story.



The mapping capability also incorporates a time window which shows where stories are located within a moving time frame (say, weekly or daily). The temporal mapping software was adapted to interface to Hindi data in the DARPA TIDES Surprise Language exercise (Gey & Chen 2003, Oard 2003) and enabled a display which coordinates *story, time-of-story and location of story.* See: http://ecai.org/imls2002/hindi/HindiMapspace.html This demonstration maps 863 documents from the BBC test corpora for 178 cities in India by extracting Hindi place names found in each document and matching to a mini-gazetteer created for this list of cities.

2. DEPLOYMENT TO A NEW LANGUAGE

In order to make this work for other languages and areas, one needs to develop two items: 1) A gazetteer for the region being discussed in the documents and 2) place name extraction software which can identify places within the full text of the documents. We assume (since we are working with news stories) that each story is given a time stamp. This finesses the problem of the document discussing events at different dates, or even the vagueness of temporal expressions in text (e.g. "Gulf War Period") but we will approach the problem one step at a time.

2.1 Mapping Russian Documents

Among the collections of the CLEF European language evaluation campaigns are is the CLEF Russian collection consisted of 16,716 articles from *Izvestia* newspaper during 1995, first introduced at CLEF 2003 workshop in Trondheim Norway (Peters 2003). Each document is tagged in XML format using the Unicode UTF-8 encoding. Among the tags within most documents is one for *geography*, e.g.

<GEOGRAPHY>**Санкт-Петербург**</GEOGRAPHY>

(Saint Petersburg).

We extracted and summarized 619 unique tags, including country names around the world, such as **Гаити** (Haiti), as well as Russian cities and districts (Oblasts). Separately, the World Gazetteer (http://www.world-gazetteer.com) has a downloadable file of Russian cities with latitude and longitudes attached. However the city utilizes only the Romanized transliterated version of each city name(e.g. Sankt Peterburg) rather than the Cyrillic. Thus we have to match this name with another city population table found at the Center for Information Research (CIR) web site (http://www.cir.ru) which has city names in Cyrillic. In order to do this, we implemented the Library of Congress (LOC) transliteration algorithm for Russian (Berry 1997) to transliterate the Cyrillic names and then to do an alphabetic sort and match the two resultant lists within Excel, as in the short fragment below (data from the World Gazetteer highlighted in blue and from CIR in yellow).

Note the difference in transliteration schemes – this is a significant problem to be overcome in merging multilingual information from multiple knowledge sources. Note also that one city is available in one source and not the other. Successful merging of these datasets will allow us to utilize the Cyrillic name for searching the text and to map the resulting documents to the proper latitude/longitude location. A preliminary map of this population data may be found at:

http://ecai.org/samples/Russian/Standalone/tmj_RussiaMapspace.html

By the time of the workshop at SIGIR-2004 we expect to have geotemporal query interface to the Izvestia news collection to query stories in Russian in time and space.

REFERENCES:

Barry, R. K., editor (1997). **ALA-LC romanization tables : transliteration schemes for non-Roman scripts approved by the Library of Congress and the American Library Association.** Library of Congress, Washington.

Gey, F & Chen, A. (2003). Searching Hindi using Statistical Stemming. *Team TIDES newsletter*, (October 2003), available at http://language.cnri.reston.va.us/TeamTIDES/tt02e3-final.pdf.

Oard, D, editor (2003), Special Issue on the Surprise Language Exercises, ACM Transactions on Asian Language Processing, Vol. 2, No. 2, June 2003.

Peters, C, (2003). Introduction to the CLEF 2003 Working Notes: http://clef.iei.pi.cnr.it:2002/2003/WN_web/00.2%20-%20intro.pdf, August 2003.

Gazetteer place name	current population [1000]	latitude	longitude	LOC Transliterated name	Cyrillic name	1985 Population
Almetjevsk	140.4	54.90°N	52.31°E	Al'met'evsk	Альметьевск	124.9
Angarsk	233.5	52.57°N	103.91°E	Angarsk	Ангарск	259.3
				Anzhero-Sudzhensk	Анжеро-Судженск	111.3
Arhangelsk	343.5	64.54°N	40.54°E	Arkhangel'sk	Архангельск	411.7