Geographic Information Retrieval (GIR): Algorithms and Approaches

Ray R. Larson
University of California, Berkeley
School of Information

Overview

• What is GIR?
• Spatial Approaches to GIR
• A Logistic Regression Approach to GIR
  – Model
  – Testing and Results
  – Example using Google Earth as an interface
• GIR Evaluation Tests
  – GeoCLEF
  – GikiCLEF
  – NTCIR GeoTime

Geographic Information Retrieval (GIR)

• Geographic information retrieval (GIR) is concerned with spatial approaches to the retrieval of geographically referenced, or georeferenced, information objects (GIOs)
  – about specific regions or features on or near the surface of the Earth.
  – Geospatial data are a special type of GIO that encodes a specific geographic feature or set of features along with associated attributes
    • maps, air photos, satellite imagery, digital geographic data, photos, text documents, etc.

Georeferencing and GIR

• Within a GIR system, e.g., a geographic digital library, information objects can be georeferenced by place names or by geographic coordinates (i.e. longitude & latitude)
GIR is not GIS

- **GIS** is concerned with spatial representations, relationships, and analysis at the level of the individual spatial object or field.
- **GIR** is concerned with the retrieval of geographic information resources (and geographic information objects at the set level) that may be relevant to a geographic query region.

Spatial Approaches to GIR

- A spatial approach to geographic information retrieval is one based on the integrated use of spatial representations, and spatial relationships.
- A spatial approach to GIR can be qualitative or quantitative:
  - **Quantitative**: based on the geometric spatial properties of a geographic information object.
  - **Qualitative**: based on the non-geometric spatial properties.

Spatial Matching and Ranking

- Spatial similarity can be considered as an indicator of relevance: documents whose spatial content is more similar to the spatial content of query will be considered more relevant to the information need represented by the query.
- Need to consider both:
  - Qualitative, non-geometric spatial attributes
  - Quantitative, geometric spatial attributes
    - Topological relationships and metric details
- We focus on the latter…

Spatial Similarity Measures and Spatial Ranking

- Three basic approaches to spatial similarity measures and ranking:
  - Method 1: Simple Overlap
  - Method 2: Topological Overlap
  - Method 3: Degree of Overlap:
Method 1: Simple Overlap

- Candidate geographic information objects (GIOs) that have any overlap with the query region are retrieved.
- Included in the result set are any GIOs that are contained within, overlap, or contain the query region.
- The spatial score for all GIOs is either relevant (1) or not relevant (0).
- The result set cannot be ranked – topological relationship only, no metric refinement.

Method 2: Topological Overlap

- Spatial searches are constrained to only those candidate GIOs that either:
  - are completely contained within the query region,
  - overlap with the query region,
  - or, contain the query region.
- Each category is exclusive and all retrieved items are considered relevant.
- The result set cannot be ranked – categorized topological relationship only, no metric refinement.

Method 3: Degree of Overlap

- Candidate geographic information objects (GIOs) that have any overlap with the query region are retrieved.
- A spatial similarity score is determined based on the degree to which the candidate GIO overlaps with the query region.
- The greater the overlap with respect to the query region, the higher the spatial similarity score.
- This method provides a score by which the result set can be ranked – topological relationship: overlap – metric refinement: area of overlap.

Example: Results display from CheshireGeo:

http://calipr.regis.berkeley.edu/pattyf/mapproxy/cheshire2/cheshire_init.html
Geometric Approximations

- The decomposition of spatial objects into approximate representations is a common approach to simplifying complex and often multi-part coordinate representations.

- Types of Geometric Approximations
  - Conservative: superset
  - Progressive: subset
  - Generalizing: could be either Concave or Convex
    - Geometric operations on convex polygons much faster

Our Research Questions

- Spatial Ranking
  - How effectively can the spatial similarity between a query region and a document region be evaluated and ranked based on the overlap of the geometric approximations for these regions?

- Geometric Approximations & Spatial Ranking:
  - How do different geometric approximations affect the rankings?
    - MBRs: the most popular approximation
    - Convex hulls: the highest quality convex approximation

Spatial Ranking: Methods for computing spatial similarity

<table>
<thead>
<tr>
<th>Reference</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hill, 1990[10]</td>
<td>$Range = \frac{Q_{Q}}{Q_{C}}$</td>
</tr>
<tr>
<td>Walker et al, 1992[19]</td>
<td>$Range = \min \left( \frac{Q_{Q}}{Q_{C}}, 1 \right)$</td>
</tr>
<tr>
<td>Board and Sharmas, 1997[3]</td>
<td>Case 1: $Q$ contains $C$ $Range = \frac{Q_{Q}}{Q_{C}}$</td>
</tr>
</tbody>
</table>

Where:
- $Q$ = area of query region
- $C$ = area of candidate GIS
- $\varnothing$ = no similarity
- $1$ = identical
- $Q_{Q}$ = area of overlap for $Q$, $C$
Proposed Ranking Method

• Probabilistic Spatial Ranking using Logistic Inference
• Probabilistic Models
  – Rigorous formal model attempts to predict the probability that a given document will be relevant to a given query
  – Ranks retrieved documents according to this probability of relevance (Probability Ranking Principle)
  – Rely on accurate estimates of probabilities

Logistic Regression

Probability of relevance is based on Logistic regression from a sample set of documents to determine values of the coefficients.

At retrieval the probability estimate is obtained by:

\[
P(R | Q, D) = c_0 + \sum_{i=1}^{m} c_i X_i
\]

For the \( m \) X attribute measures (on the following page)

Probabilistic Models: Logistic Regression attributes

• \( X_1 \) = area of overlap(query region, candidate GIO) / area of query region
• \( X_2 \) = area of overlap(query region, candidate GIO) / area of candidate GIO
• \( X_3 = 1 – \text{abs(fraction of overlap region that is onshore fraction of candidate GIO that is onshore)} \)
• Where:
  Range for all variables is 0 (not similar) to 1 (same)

Probabilistic Models

Advantages
• Strong theoretical basis
• In principle should supply the best predictions of relevance given available information
• Computationally efficient, straightforward implementation (if based on LR)

Disadvantages
• Relevance information is required -- or is "guestimated"
• Important indicators of relevance may not be captured by the model
• Optimally requires ongoing collection of relevance information
Test Collection

- California Environmental Information Catalog (CEIC)

- Approximately 2500 records selected from collection (Aug 2003) of ~ 4000.

Test Collection Overview

- 2554 metadata records indexed by 322 unique geographic regions (represented as MBRs) and associated place names.
  - 2072 records (81%) indexed by 141 unique CA place names
    - 881 records indexed by 42 unique counties (out of a total of 46 unique counties indexed in CEIC collection)
    - 427 records indexed by 76 cities (of 120)
    - 179 records by 8 bioregions (of 9)
    - 3 records by 2 national parks (of 5)
    - 309 records by 11 national forests (of 11)
    - 3 record by 1 regional water quality control board region (of 1)
    - 270 records by 1 state (CA)
  - 482 records (19%) indexed by 179 unique user defined areas (approx 240) for regions within or overlapping CA
    - 12% represent onshore regions (within the CA mainland)
    - 88% (158 of 179) offshore or coastal regions

CA Named Places in the Test Collection – complex polygons

CA Counties – Geometric Approximations

- MBRs: 94.61%
- Convex Hulls: 26.73%
CA User Defined Areas (UDAs) in the Test Collection

Test Collection Query Regions: CA Counties

42 of 58 counties referenced in the test collection metadata

- 10 counties randomly selected as query regions to train LR model
- 32 counties used as query regions to test model

Test Collection Relevance Judgements

- Determine the reference set of candidate GIO regions relevant to each county query region:
  - Complex polygon data was used to select all CA place named regions (i.e. counties, cities, bioregions, national parks, national forests, and state regional water quality control boards) that overlap each county query region.
  - All overlapping regions were reviewed (semi-automatically) to remove sliver matches, i.e. those regions that only overlap due to differences in the resolution of the 6 data sets.
    - Automated review: overlaps where overlap area/GIO area > .00025 considered relevant, else not relevant.
    - Cases manually reviewed: overlap area/query area < .001 and overlap area/GIO area < .02
- The MBRs and metadata for all information objects referenced by UDAs (user-defined areas) were manually reviewed to determine their relevance to each query region. This process could not be automated because, unlike the CA place named regions, there are no complex polygon representations that delineate the UDAs.
- This process resulted in a master file of CA place named regions and UDAs relevant to each of the 42 CA county query regions.

LR model

- \( X_1 = \frac{\text{area of overlap(query region, candidate GIO)}}{\text{area of query region}} \)
- \( X_2 = \frac{\text{area of overlap(query region, candidate GIO)}}{\text{area of candidate GIO}} \)
- \text{Where:} \quad \text{Range for all variables is 0 (not similar) to 1 (same)}

<table>
<thead>
<tr>
<th>Approximation</th>
<th>Logistic Regression Model Fitted on the Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MBR</td>
<td>( \log OR(Q,C) = -5.0402 + (5.5154 * X_1) + (5.7729 * X_2) )</td>
</tr>
<tr>
<td>Convex Hull</td>
<td>( \log OR(Q,C) = -3.4767 + (7.4536 * X_1) + (5.7569 * X_2) )</td>
</tr>
</tbody>
</table>
Some of our Results

Mean Average Query Precision: the average precision values after each new relevant document is observed in a ranked list.

For metadata indexed by CA named place regions:

<table>
<thead>
<tr>
<th>Ranking Method</th>
<th>MBRs</th>
<th>Convex Hulls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hill, 1990</td>
<td>0.7193</td>
<td>0.8097</td>
</tr>
<tr>
<td>Walker et al., 1992</td>
<td>0.7625</td>
<td>0.8006</td>
</tr>
<tr>
<td>Beard &amp; Sharma, 1997</td>
<td>0.7854</td>
<td>0.8116</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.9339</td>
<td>0.9973</td>
</tr>
</tbody>
</table>

For all metadata in the test collection:

<table>
<thead>
<tr>
<th>Ranking Method</th>
<th>MBRs</th>
<th>Convex Hulls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hill, 1990</td>
<td>0.6722</td>
<td>0.7936</td>
</tr>
<tr>
<td>Walker et al., 1992</td>
<td>0.6630</td>
<td>0.7810</td>
</tr>
<tr>
<td>Beard &amp; Sharma, 1997</td>
<td>0.6523</td>
<td>0.7778</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.9141</td>
<td>0.9409</td>
</tr>
</tbody>
</table>

These results suggest:
- Convex Hulls perform better than MBRs
- Expected result given that the CH is a higher quality approximation

A probabilistic ranking based on MBRs can perform as well if not better than a non-probabilistic ranking method based on Convex Hulls

Interesting
- Since any approximation other than the MBR requires great expense, this suggests that the exploration of new ranking methods based on the MBR are a good way to go.

BUT:
- The inclusion of UDA indexed metadata reduces precision.
- This is because coarse approximations of onshore or coastal geographic regions will necessarily include much irrelevant offshore area, and vice versa.

Results for MBR - Named data

Results for Convex Hulls - Named data
Offshore / Coastal Problem

California EEZ Sonar Imagery Map – GLORIA Quad 13

- PROBLEM: the MBR for GLORIA Quad 13 overlaps with several counties that are completely inland.

Adding Shorefactor Feature Variable

Shorefactor = 1 – abs(fraction of query region approximation that is onshore – fraction of candidate GIO approximation that is onshore)

Onshore Areas

- Candidate GIO MBRs
  - A) GLORIA Quad 13: fraction onshore = .55
  - B) WATER Project Area: fraction onshore = .74

- Query Region MBR
  - Q) Santa Clara County: fraction onshore = .95

Computing Shorefactor:

- Q – A Shorefactor: 1 – abs(.95 - .55) = .60
- Q – B Shorefactor: 1 – abs(.95 - .74) = .79

Even though A & B have the same area of overlap with the query region, B has a higher shorefactor, which would weight this GIO's similarity score higher than A's.

Note: geographic content of A is completely offshore, that of B is completely onshore.

About the Shorefactor Variable

- Characterizes the relationship between the query and candidate GIO regions based on the extent to which their approximations overlap with onshore areas (or offshore areas).

- Assumption: a candidate region is more likely to be relevant to the query region if the extent to which its approximation is onshore (or offshore) is similar to that of the query region's approximation.

About the Shorefactor Variable

- The use of the shorefactor variable is presented as an example of how geographic context can be integrated into the spatial ranking process.

- Performance: Onshore fraction for each GIO approximation can be pre-indexed. Thus, for each query only the onshore fraction of the query region needs to be calculated using a geometric operation. The computational complexity of this type of operation is dependent on the complexity of the coordinate representations of the query region (we used the MBR and Convex hull approximations) and the onshore region (we used a very generalized concave polygon with only 154 pts).
Shorefactor Model

• $X_1 = \frac{\text{area of overlap(query region, candidate GIO)}}{\text{area of query region}}$

• $X_2 = \frac{\text{area of overlap(query region, candidate GIO)}}{\text{area of candidate GIO}}$

• $X_3 = 1 - \text{abs(fraction of query region approximation that is onshore} - \text{fraction of candidate GIO approximation that is onshore)}$

Where: Range for all variables is 0 (not similar) to 1 (same)

Results for All Data - MBRs

Results for All Data - Convex Hull
GIR Examples

- The following screen captures are from a GIR application using the algorithms (2 variable logistic regression model) and data (the CIEC database data)
- Uses a Google Earth network link to provide a GIR search interface
GIR Evaluations

- The GeoCLEF track of CLEF conducted evaluations of GIR systems using text-based queries
  - One finding was that good text retrieval methods may work as well, or better, than more complex geographic modeling and query expansion approaches
- The GikiCLEF track of CLEF
- New NTCIR-GEOTIME track focuses GeoTemporal Information starting -- see http://metadata.berkeley.edu/NTCIR-GeoTime/
GeoCLEF Overview

- Geographical Information Retrieval (GIR) concerns the retrieval of information involving some kind of spatial awareness. Given that many documents (and queries) contain some kind of spatial reference, there are examples where geographical references (geo-references) may be important for IR.
- In addition to this, many documents contain geo-references expressed in multiple languages which may or may not be the same as the query language. This would require an additional translation step to enable successful retrieval.
- Existing evaluation campaigns such as TREC and CLEF do not explicitly evaluate geographical IR relevance.
- The aim of GeoCLEF was to provide the necessary framework in which to evaluate GIR systems for search tasks involving both spatial and multilingual aspects.

Organizers of GeoCLEF

- Fred Gey and Ray Larson, University of California, Berkeley, USA (gey@berkeley.edu, ray@sims.berkeley.edu)
- Mark Sanderson, Department of Information Studies, University of Sheffield, UK (m.sanderson@sheffield.ac.uk)
- Hideo Joho, University of Glasgow, UK (hideo@dcs.gla.ac.uk)
- Thomas Mandl and Christa Womser-Hacker of U. Hildesheim Germany (German language coordinators)
- Diana Santos and Paulo Rocha of Linguateca (Portuguese coordinators)
- Andrés Montoyo of U. Alicante (Spanish coordinator)

GeoCLEF

- Proposed 2004, first evaluation 2005
- The last GeoCLEF was held in 2008, the new GikiCLEF task is taking its place
- This overview will focus on the topics, participants and performance for GeoCLEF 2005 and 2006, with some looks at 2007 and 2008

Topic for GeoCLEF 2005

Topics translated for both English and German

<top>
<num>GC001</num>
<orignum>C084</orignum>
<EN-title>Shark Attacks off Australia and California</EN-title>
<EN-desc>Documents will report any information relating to shark attacks on humans.</EN-desc>
<EN-narr>Identify instances where a human was attacked by a shark, including where the attack took place and the circumstances surrounding the attack. Only documents concerning specific attacks are relevant; unconfirmed shark attacks or suspected bites are not relevant.</EN-narr>
<!-- NOTE: This topic has added tags for GeoCLEF -->
<EN-concept>Shark attacks</EN-concept>
<EN-spatialrelation>near</EN-spatialrelation>
<EN-location>Australia</EN-location>
<EN-location>California</EN-location>
</top>
GeoCLEF 2005 Collections

- The document collections for GeoCLEF 2005 are all newswire stories from the years 1994 and 1995 used in previous CLEF competitions.
- The German document collection consists of 294,809 documents from Der Spiegel (1994/95), the Frankfurter Rundschau (1994) and the Swiss news agency SDA (1994/95).
- The same collections were used for all GeoCLEF evaluations 2005-2008.

GeoCLEF 2005 Documents

- In both collections, the documents have a common structure:
  - newspaper-specific information like:
    - date
    - page
    - issue
    - special filing numbers
    - one or more titles
    - a byline
    - the actual text.
- The document collections were not explicitly geographically tagged or contained any other location-specific information.

GeoCLEF 2005 Runs

<table>
<thead>
<tr>
<th>Group Name</th>
<th>Mono DE</th>
<th>Mono EN</th>
<th>Billin X Giấy</th>
<th>Billin X DE</th>
<th>Total Runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>California State University, San Marcos</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Grupo XLDI (Universidade de Lisboa)</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Linguatex (FernUniversität Hagen)</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>MetaCarta Inc.</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>MIRACLE (Universidad Politécnica de Madrid)</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>NICTA, University of New South Wales</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>TALP (Universitat Politècnica de Catalunya)</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Universidad Politécnica de València</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>University of Alicante</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>University of California, Berkeley (Berkeley 1)</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>University of California, Berkeley (Berkeley 2)</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>University of Hagen (FernUniversität Hagen)</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Total Submitted Runs</td>
<td>53</td>
<td>25</td>
<td>22</td>
<td>17</td>
<td>117</td>
</tr>
<tr>
<td>Number of Groups Participating in Task</td>
<td>11</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

† Linguateca helped with evaluation, but did not submit runs

GeoCLEF 2006 Topics

Topics in English, German, Spanish and Portuguese

- <top><num>GC026</num></top>
  - <EN-title>Wine regions around rivers in Europe</EN-title>
  - <EN-desc>Documents about wine regions along the banks of European rivers</EN-desc>
  - <EN-narr>Relevant documents describe a wine region along a major river in European countries. To be relevant the document must name the region and the river.</EN-narr>
- <top><num>GC027</num></top>
  - <EN-title>Cities within 100km of Frankfurt</EN-title>
  - <EN-desc>Documents about cities within 100 kilometers of the city of Frankfurt am Main Germany, latitude 50.112222, longitude 8.68194. To be relevant the document must describe the city or an event in that city. Stories about Frankfurt itself are not relevant.</EN-narr>
GeoCLEF 2006 Topics

<top>
<num> GC034 </num>
<EN-title> Malaria in the tropics </EN-title>
<EN-desc> Malaria outbreaks in tropical regions and preventive vaccination </EN-desc>
<EN-narr> Relevant documents state cases of malaria in tropical regions and possible preventive measures like chances to vaccinate against the disease. Outbreaks must be of epidemic scope. Tropics are defined as the region between the Tropic of Capricorn, latitude 23.5 degrees South and the Tropic of Cancer, latitude 23.5 degrees North. Not relevant are documents about a single person's infection. </EN-narr>
</top>

<top>
<num> GC042 </num>
<EN-title> Regional elections in Northern Germany </EN-title>
<EN-desc> Documents about regional elections in Northern Germany </EN-desc>
<EN-narr> Relevant documents are those reporting the campaign or results for the state parliaments of any of the regions of Northern Germany. The states of northern Germany are commonly Bremen, Hamburg, Lower Saxony, Mecklenburg-Western Pomerania and Schleswig-Holstein. Only regional elections are relevant; municipal, national and European elections are not. </EN-narr>
</top>

GeoCLEF 2006 Collections

- Same English and German documents as 2005
- Added Spanish and Portuguese collections
  - Spanish: EFE 1994-1995
- For 2007 and 2008 the Spanish collection was dropped

GeoCLEF 2006 Runs

<table>
<thead>
<tr>
<th>NAME</th>
<th>DE</th>
<th>EN</th>
<th>ES</th>
<th>PT</th>
<th>X2DE</th>
<th>X2EN</th>
<th>X2ES</th>
<th>X2PT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>alicante</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>berkeley</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>daedalus</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>hagen</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>hildesheim</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>imp-coil</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>jaen</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>js-n-china</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>nicta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>ria-upv</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>sanmarcos</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>talpi</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>u.buffalo</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>u.groningen</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>u.twente</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>unsr</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>xldb</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>TOTALS (17)</td>
<td>16</td>
<td>73</td>
<td>15</td>
<td>11</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td></td>
<td>137</td>
</tr>
</tbody>
</table>

Techniques used by various groups in 2005 and 2006

- Ad-hoc text retrieval techniques (blind feedback, German word de-compounding, etc.)
- Question-answering modules
- Gazetteer construction (GNIS, World Gazetteer)
- Toponym Named Entity Extraction
- Term expansion using Wordnet, geographic thesauri
- Toponym resolution
- NLP – Geofiltering predicates
- Latitude-longitude assignment
- Gazetteer-based query expansion
Best-Performing Monolingual Runs:
GeoCLEF 2005

<table>
<thead>
<tr>
<th>Best monolingual-English-run</th>
<th>MAP</th>
<th>Best monolingual-German-run</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>berkeley-2_BKGeoE1</td>
<td>0.3936</td>
<td>berkeley-2_BKGeoD3</td>
<td>0.2042</td>
</tr>
<tr>
<td>cuu-sammarcos_csusm1</td>
<td>0.3613</td>
<td>alicante_irua-de-tituldesgeotags</td>
<td>0.1227</td>
</tr>
<tr>
<td>alicante_irua-en-net</td>
<td>0.3495</td>
<td>miracle_GCdeNOR</td>
<td>0.1163</td>
</tr>
<tr>
<td>berkeley_BERK1MLENLOC03</td>
<td>0.2924</td>
<td>xldb_XLDBDEManTDGKBm3</td>
<td>0.1123</td>
</tr>
<tr>
<td>miracle_GCdeNOR</td>
<td>0.2653</td>
<td>hagen_FUHo14d</td>
<td>0.1053</td>
</tr>
<tr>
<td>nicta_i2d2Run1</td>
<td>0.2514</td>
<td>berkeley_BERK1MLDELOC02</td>
<td>0.0535</td>
</tr>
<tr>
<td>linguat_LITLE</td>
<td>0.2362</td>
<td></td>
<td></td>
</tr>
<tr>
<td>xldb_XLDBENManTDL</td>
<td>0.2253</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tulp_geotapplR4</td>
<td>0.2231</td>
<td></td>
<td></td>
</tr>
<tr>
<td>metacarta_run0</td>
<td>0.1496</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u.valencia_dsic_gc052</td>
<td>0.1464</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bilingual English Performance

GeoCLEF 2006 Top Mono. Runs

<table>
<thead>
<tr>
<th>Track</th>
<th>Participant Rank</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Berkeley</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td></td>
<td>30.34%</td>
<td>27.23%</td>
<td>26.37%</td>
<td>26.22%</td>
<td>26.11%</td>
<td>16.20%</td>
</tr>
<tr>
<td>Monolingual German</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td></td>
<td>22.29%</td>
<td>21.51%</td>
<td>15.58%</td>
<td>10.01%</td>
<td>122.68%</td>
<td></td>
</tr>
<tr>
<td>Monolingual Portuguese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td></td>
<td>30.12%</td>
<td>16.92%</td>
<td>13.44%</td>
<td>124.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monolingual Spanish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prec.</td>
<td></td>
<td>35.08%</td>
<td>31.82%</td>
<td>16.12%</td>
<td>14.71%</td>
<td>138.48%</td>
<td></td>
</tr>
</tbody>
</table>
GeoCLEF 2006 Top Biling. Runs

<table>
<thead>
<tr>
<th>Track</th>
<th>Participant Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual English</td>
<td>Part.</td>
<td>jaen* sanmarcos</td>
<td>Run</td>
<td>sinai</td>
<td>ESENE</td>
<td>XP2</td>
<td>pooled</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>Prec.</td>
<td>22.56%</td>
<td>22.46%</td>
<td>16.03%</td>
<td>40.74%</td>
<td></td>
</tr>
<tr>
<td>Bilingual German</td>
<td>Part.</td>
<td>jaen* sanmarcos</td>
<td>Run</td>
<td>HIGeodeen</td>
<td>run12</td>
<td>pooled</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>Prec.</td>
<td>15.61%</td>
<td>12.80%</td>
<td>11.86%</td>
<td>31.62%</td>
<td></td>
</tr>
<tr>
<td>Bilingual Portuguese</td>
<td>Part.</td>
<td>sanmarcos</td>
<td>Run</td>
<td>BKGeoES1</td>
<td>pooled</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>Prec.</td>
<td>14.10%</td>
<td>12.80%</td>
<td>12.30%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilingual Spanish</td>
<td>Part.</td>
<td>sanmarcos</td>
<td>Run</td>
<td>BKGeoEP1</td>
<td>pooled</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AVG</td>
<td>Prec.</td>
<td>25.71%</td>
<td>12.42%</td>
<td>108.55%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GeoCLEF Collections 2007

Table 1. GeoCLEF test collection – collection and topic languages

<table>
<thead>
<tr>
<th>GeoCLEF Year</th>
<th>Collection Languages</th>
<th>Topic Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005 (pilot)</td>
<td>English, German</td>
<td>English, German</td>
</tr>
<tr>
<td>2006</td>
<td>English, German, Portuguese, Spanish</td>
<td>English, German, Portuguese, Spanish, Japanese</td>
</tr>
<tr>
<td>2007</td>
<td>English, German, Portuguese</td>
<td>English, German, Portuguese, Spanish, Indonesian</td>
</tr>
</tbody>
</table>

Example Topics 2007

<num>10.2452/58-GC</num>
<title>Travel problems at major airports near to London</title>
<desc>To be relevant, documents must describe travel problems at one of the major airports close to London.</desc>
<narr>This includes all reported violations of human rights in Burma, no matter when (not only by the present government). Declarations (accusations or denials) about the matter only, are not relevant.</narr>
</top>

<top>
<title>Violation of human rights in Burma</title>
<desc>Documents are relevant if they mention actual violation of human rights in Myanmar, previously named Burma.</desc>
<narr>Major airports to be listed include Heathrow, Gatwick, Luton, Stansted and London City airport.</narr>
</top>

Participan Approaches 2007

- Ad-hoc techniques (weighting, probabilistic retrieval, language model, blind relevance feedback)
- Semantic analysis (annotation and inference)
- Geographic knowledge bases (Gazetteers, thesauri, ontologies)
- Text mining
- Query expansion techniques (e.g. geographic feedback)
- Geographic Named Entity Extraction (LingPipe, GATE, etc.)
- Geographic disambiguation
- Geographic scope and relevance models
- Geographic relation analysis
- Geographic entity type analysis
- Term expansion using WordNet
- Part-of-speech tagging

Fig. 1: Topics GC058 and GC075
## Monolingual Results 2007

<table>
<thead>
<tr>
<th>Track</th>
<th>Rank</th>
<th>Partner</th>
<th>Experiment DOI</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono-lingual English</td>
<td>1st</td>
<td>catalunya</td>
<td>10.2415/GC-MONO-EN-CLEF2007,CATALUNYA,TALPEEGHRT07</td>
<td>28.5%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>cheshire</td>
<td>10.2415/GC-MONO-EN-CLEF2007,CHESHIRE,BERKMOENBASE</td>
<td>26.4%</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>valencia</td>
<td>10.2415/GC-MONO-EN-CLEF2007,VALENCIA,RTIAAPYV07</td>
<td>26.4%</td>
</tr>
<tr>
<td></td>
<td>4th</td>
<td>roningen</td>
<td>10.2415/GC-MONO-EN-CLEF2007,ROHINESS,GEOM0TREED07</td>
<td>25.3%</td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>csusm</td>
<td>10.2415/GC-MONO-EN-CLEF2007,CSUSM,GEOM0EN5</td>
<td>21.3%</td>
</tr>
<tr>
<td>Mono-lingual German</td>
<td>1st</td>
<td>hagen</td>
<td>10.2415/GC-MONO-DE-CLEF2007,HAGEN,FUHTDN5DE</td>
<td>33.7%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>csusm</td>
<td>10.2415/GC-MONO-DE-CLEF2007,CSUSM,GEOM0DE4</td>
<td>25.8%</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>hildesheim</td>
<td>10.2415/GC-MONO-DE-CLEF2007,HILDIESHEIM,HILDERDE3MA</td>
<td>20.7%</td>
</tr>
<tr>
<td></td>
<td>4th</td>
<td>cheshire</td>
<td>10.2415/GC-MONO-DE-CLEF2007,CHESHIRE,BERKMOENBASE</td>
<td>13.9%</td>
</tr>
<tr>
<td>Mono-lingual Portuguese</td>
<td>1st</td>
<td>csusm</td>
<td>10.2415/GC-MONO-PT-CLEF2007,CSUSM,GEOM0PT3</td>
<td>85.3%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>cheshire</td>
<td>10.2415/GC-MONO-PT-CLEF2007,CHESHIRE,BERKMOENBASE</td>
<td>17.4%</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>xldb</td>
<td>10.2415/GC-MONO-PT-CLEF2007,XLDB,XXBPT_1</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

---

### Monolingual English 2007

![Graph showing performance of various teams in Monolingual English 2007](image1.png)

### Monolingual German 2007

![Graph showing performance of various teams in Monolingual German 2007](image2.png)

### Monolingual Portuguese 2007

![Graph showing performance of various teams in Monolingual Portuguese 2007](image3.png)
### Bilingual results 2007

<table>
<thead>
<tr>
<th>Track</th>
<th>Rnk.</th>
<th>Partner</th>
<th>Experiment DOI</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilingual English</td>
<td>1st</td>
<td>cheshire</td>
<td>10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDENBASE</td>
<td>22.1%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>depok*</td>
<td>10.2415/GC-BILI-X2EN-CLEF2007.DEPOK.UUBITDGP</td>
<td>21.0%</td>
</tr>
<tr>
<td></td>
<td>3rd</td>
<td>csusm</td>
<td>10.2415/GC-BILI-X2EN-CLEF2007.CHESHIRE.BERKBIDENBASE</td>
<td>19.6%</td>
</tr>
<tr>
<td></td>
<td>Diff.</td>
<td></td>
<td></td>
<td>12.5%</td>
</tr>
<tr>
<td>Bilingual German</td>
<td>1st</td>
<td>hagen</td>
<td>10.2415/GC-BILI-X2DE-CLEF2007.HAGEN.FUHTDNPEN</td>
<td>20.9%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>cheshire</td>
<td>10.2415/GC-BILI-X2DE-CLEF2007.CHESHIRE.BERKBIDENBASE</td>
<td>11.1%</td>
</tr>
<tr>
<td></td>
<td>Diff.</td>
<td></td>
<td></td>
<td>88.6%</td>
</tr>
<tr>
<td>Bilingual Portuguese</td>
<td>1st</td>
<td>cheshire</td>
<td>10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIDENBASE</td>
<td>20.1%</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>csusm</td>
<td>10.2415/GC-BILI-X2PT-CLEF2007.CHESHIRE.BERKBIDENBASE</td>
<td>5.3%</td>
</tr>
<tr>
<td></td>
<td>Diff.</td>
<td></td>
<td></td>
<td>277.5%</td>
</tr>
</tbody>
</table>
GeoCLEF 2008

- The 2008 evaluation continued the same basic approach to topics and results with the same test collections.
- In 2008 more of the topics were originally formulated in Portuguese, and then translated to English and German.

Example Topics 2008

Tab. 3: Topics GC08958 and GC08475

<table>
<thead>
<tr>
<th>Document type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trade fairs in Lower Saxony</td>
<td>Documents reporting about industrial or cultural fairs in Lower Saxony.</td>
</tr>
<tr>
<td>Atentados bomba na Irlanda do Norte</td>
<td>Os documentos relevantes mencionem atentados bombistas em localidades da Irlanda do Norte.</td>
</tr>
</tbody>
</table>

Monolingual English 2008

Monolingual German 2008
Cheshire Results 2007-2008

- The good results obtained in 2007 and 2008 by our system were not due to explicit geographic processing (such as explicit geographic query expansion or geometric approaches)
- We used only text retrieval methods as used in other text retrieval tasks
  - Logistic regression text retrieval with pseudo relevance feedback
- For GeoCLEF type queries, place names searched as text appears to perform as well or better than more complex geographic processing (but good machine translation software is essential)

Comparison of Cheshire Runs

<table>
<thead>
<tr>
<th>TASK</th>
<th>MAP '06</th>
<th>MAP '07</th>
<th>MAP '08</th>
<th>Diff '06/'07</th>
<th>Diff '07/'08</th>
<th>Diff '08/'09</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual English</td>
<td>0.250</td>
<td>0.264</td>
<td>0.268</td>
<td>5.303</td>
<td>1.493</td>
<td>6.716</td>
</tr>
<tr>
<td>Monolingual German</td>
<td>0.215</td>
<td>0.139</td>
<td>0.230</td>
<td>-54.676</td>
<td>38.565</td>
<td>6.522</td>
</tr>
<tr>
<td>Monolingual Portuguese</td>
<td>0.162</td>
<td>0.174</td>
<td>0.231</td>
<td>6.897</td>
<td>24.675</td>
<td>29.870</td>
</tr>
<tr>
<td>Bilingual English→German</td>
<td>0.156</td>
<td>0.090</td>
<td>0.225</td>
<td>-73.333</td>
<td>60.00</td>
<td>30.667</td>
</tr>
<tr>
<td>Bilingual English→Portuguese</td>
<td>0.120</td>
<td>0.201</td>
<td>0.207</td>
<td>37.313</td>
<td>2.899</td>
<td>39.130</td>
</tr>
</tbody>
</table>

GikiCLEF 2009

- GikiCLEF has replaced GeoCLEF for GIR-related retrieval in the 2009 CLEF Evaluation
- GikiCLEF uses the Wikipedia database in 10 different languages
  - Bulgarian, Dutch, English, German, Italian, Norwegian (Bokmål and Nynorsk), Portuguese, Romanian and Spanish

GikiCLEF 2009

- For GikiCLEF, systems need to answer or address geographically challenging topics, on the Wikipedia collections, returning Wikipedia document titles as list of answers
- The user model for which GikiCLEF systems intend to cater for is anyone who is interested in knowing something that might be already included in Wikipedia, but has not enough time or imagination to browse it manually
GikiCLEF 2009 Example Topics

- <topic id="GC-2009-01">List the Italian places where Ernest Hemingway visited during his life.</topic>
- <topic id="GC-2009-07">What capitals of Dutch provinces received their town privileges before the fourteenth century?</topic>
- <topic id="GC-2009-21">List the left side tributaries of the Po river.</topic>

NTCIR GeoTime 2010

- The introductory NTCIR GeoTime track will explore GIR with the added complexity of temporal (time-based) elements
- Will use both English and Japanese collections
- Still open for participation

NTCIR GeoTime Example Topics

```xml
<TOPIC ID="ACL11-JA-7119"/>
<QUESTION LANG="EN">
  <HEADER>
    What is the controversy surrounding the use of the Stealth Fighter in Yugoslavia?</HEADER>
  <QUESTION/>
</QUESTION>
<QUESTION LANG="JA">
  ユーロスラビアにわるステルス戦闘機の話題にはどんなものがありますか？
</QUESTION>
<NARRATIVE LANG="EN">
  <DOCUMENT>
    I would like to know about the dates and times of events and places in which there was a controversy surrounding the use of the Stealth Fighter in Yugoslavia.
  </DOCUMENT>
</NARRATIVE>
<NARRATIVE LANG="JA">
  <DOCUMENT>
    ユーロスラビアにわるステルス戦闘機の話題について日時、場所なども含めたい。
  </DOCUMENT>
</NARRATIVE>
```

GeoTime Web Site: http://metadata.berkeley.edu/NTCIR-GeoTime
Thank you.
ありがとうございます。

Questions?