Proposal for a geographical IR track

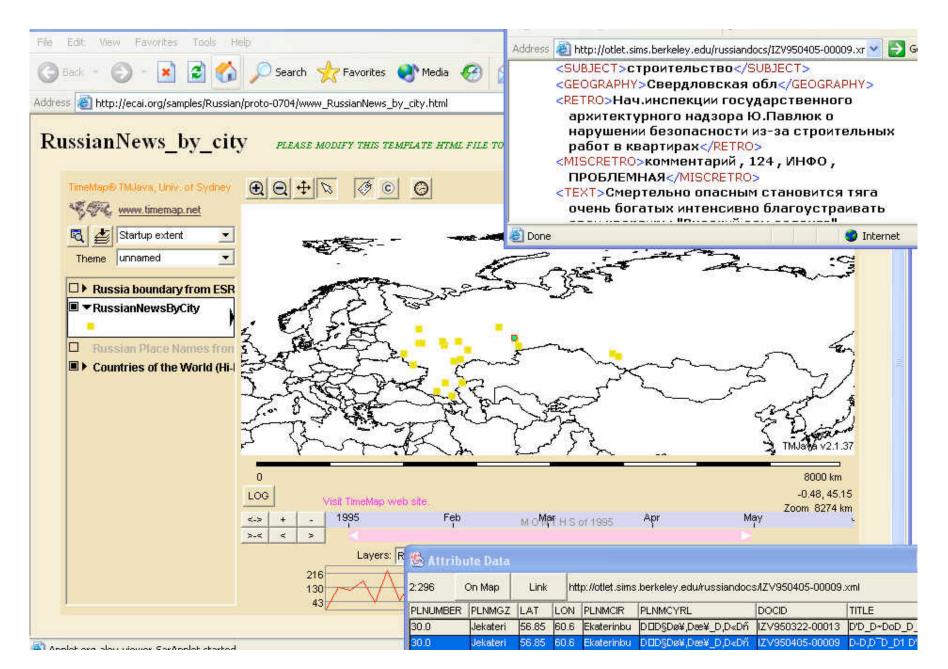Fred Gey, Ray Larson, Hideo Joho, Paul Clough and Mark Sanderson

A definition of GIR

- The retrieval of information which involves some kind of spatial awareness

- There are examples where geographical locations may be important for IR

- For example, to retrieve, re-rank and visualise search results based on a spatial dimension

- Standard IR systems pay little attention to the spatial aspects of queries and documents

Motivation

- Many collections contain lots of geographical references (e.g. Web, news)
- Existing IR evaluation campaigns do not explicitly evaluate geographical relevance
    - **except maybe the NTCIR Web track: http://research.nii.ac.jp/ntcweb/cfp-ntcir4web-en.html#geo**
- The requirements and evaluation methods for GIR may be different from standard IR
- Encourage the collaboration of GIS, NLP and IR communities for GIR research
- Combination of spatial and cross-language problems/ user needs
- Queries and results can be displayed on maps3

Example from the CLEF Russian News Collection

Cross-lingual aspects

- Documents
  - many multilingual sources exist for GIR (document translation?)
  - geographical references may be expressed in various languages
- User needs
  - can be expressed in different languages (query translation?)
- Multilingual geographic resources (e.g. multilingual gazetteer lists) are required for GIR
- Cross-language GIR will need to consider translation of geographical references

Geographical aspects

- Identifying geographic references
  - distinguishing from other named-entities
  - use of gazetteer lists and/or context rules
- Assigning spatial coordinates
  - Ambiguity
    - Reference - same name maps to multiple places (e.g. Lancaster, UK and Lancaster, USA)
    - Referent – different names refer to the same place (e.g. North America and N. America)
  - Spatial reference (point, bounding box or polygon)

Task description

- Possible tasks
  - *Ad hoc retrieval task* – system-centred evaluation to compare GIR methods
  - *User interactive task* – determine the facilities of a GIR system which meet user needs
- Constraints
  - Search results are related both to specified location and topic (spatial and topical relevance)
  - Multilingual topics (yes) and search results (?)

Test collection

- Documents
  - Existing CLEF collection (multilingual newspaper articles)
  - Web collection (?)
- Topics
  - Contain topical and geographic contexts (e.g."Chechen rebels", "stories about London", "hotels in Bath")
  - Involve spatial operators (e.g. "in", "near to", "south of" etc.) or precise geo-location ("stories about places within 250km of Zurich")
  - Topics defined by native inhabitants of geographic region (stories about cities in southern Germany?)
  - Multilingual: English, French, German, Italian and Russian (?) & Spanish (?)

Information need (topics)

<title> museums near Cardiff city

<description> find museums which exhibit Roman artefacts near to Cardiff, a city in South Wales, UK

<narrative> a relevant document will contain information regarding the museum such as opening times, location and the collections exhibited. Locations near to Cardiff include Llandaf, nant garw and Penarth. Locations must be within Wales.

<concept> museums </concept>

<spatial relation> near to

<location> Cardiff city, Wales, UK

(??) <spatial coordinate> latitude, longitude

Evaluation

- Relevance assessment
  - Assessors familiar with geographic region (language?)
  - Special topical (e.g. a story about Jack London is irrelevant) and spatial relevance
  - Binary (R vs. NR) or ternary (R vs. PR vs. NR) judgements'

- Evaluation measures
  - Precision – the proportion of documents retrieved which are relevant
  - Recall – the proportion of all relevant documents retrieved