CSCI572: Information Retrieval and Search Engines

John J. Tran

June 28, 2011

P. Clough’s “Extracting Metadata for Spatially-Aware Information Retrieval on the Web”
1. Introduction
   - Motivation
   - Constraints
   - Related work
2. Background
   - Data source
   - Parsing
   - Coding
3. Technique and algorithm
   - Experiment
   - Result
4. Analysis
   - Relevancy
   - Contributions
   - Pros and Cons
GOAL: Crawl the web, parse the pages, and contextualize the geospatial information

Develop a prototype system/framework called SPIRIT
Once geospatial information is extracted [presumably indexed and archived], they can search for documents that are relevant to the locales mentioned in the documents....

Perhaps of more significance: users can ask proximity questions: “give me XYZ near [fill-in-the-blank]”
Canonical web search and indexing contrains:

- Speed – 1TB of data within a time constraint $\rightarrow P_t < \tau$
- Reliability – Robustness and error recovery $\rightarrow$ reduce HITL
- Flexibility – customization/extension points – involves HITL
- Multilingualism – English is not the only language on Earth...
The paper is somewhat self-contained – and does not reference competitors – but the research is built on top of the SPIRIT project.

However, if you’re interested in geospatial research, a useful text is *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society* [Springer 2007].
Spatial Data source: a cocktail combination of TGN, OS, and SABE (UK, France, Germany, and Switzerland) – which include both point and polygon footprints.

Web Source: 1TB of $>94$ million webpages: $\sim 10\%$ euro TLDs.
Their approach is very heuristic:

- Employed NER – combine known gazeteers with rules; name and location lookup table implemented with lexical analyzer grammar — side note: not language dependent.
- Restricted geospatial names and locates to the lookup source — constrain problem size.
- Prune HTML garnishing (speed), stop words, e.g. if the stop word has no context relevancy, and proper names; this eliminates search vector from consideration.
Geo-coding

- Use default (most frequently used) context for geospatial terms
- Implement a hierarchical ranking (or precedence of matching)
(1) ambiguous location
(2) get matching locations from sources
(3) compute overlapping score
(4) rank result according to
   → (a) overlap
   → (b) depth of hierarchy
   → (c) resource preference
   → (d) country preference
Only a small subset of the original documents were considered (∼ 169k of 885k).

Rationale: language and available footprint. Then they further down the sample size to only 10%.

Starting to get fishy...Process is now manual using the GUI.

Classified findings against data source (and hybrid or combination of data sources) as missing, correct, false positive, precision, and recall.
Evaluation: geo-coding result

TGN: data identified 68% ambiguous
SABE (UK): 11% ambiguous
Hybrid: Combine everything SABE → OS → TGN

BL: 89% accuracy; in reality only SABE (UK) data is relevant (useful) — others are noises and don’t add up much.
1TB data: Pentium 4 2.0 Ghz with 1GB memory. For 100 [random] documents: 28.5 seconds and 15.4 seconds for markup and 40.3 seconds for coding. Total execution was 100 seconds.

For 10k documents 3 hours.

**Question:** Is that good enough?
A core component of building a search engine and indexing search documents is the ability to extract metadata, e.g. geospatial information.
Leveraging “crawling” and “query models” and “ranking”
The SPIRIT system is not the best out there — but it gives insight to some detection optimization techniques.

For example, use of hybrid source, pruning based on context, and ranking based on overlapping scores.
Pros

- Simple and easy to understand
- Limit their search scope to just the UK
Analysis

Cons

- A little terse on the algorithm...lacks formality or rigor
- Does not provide background or other research
- Does not do sufficient [computation] performance analysis
- Limit their search scope to just the UK
- Sample size for validation is small!
This paper is a report on an exercise P. Clough et al conducted using existing tools.

Their findings are useful and the take aways are:

▶ They can achieve 89% accuracy
▶ High relevance and has practical applications