Web-a-Where: Geotagging Web Content

Einat Amitay
Nadav Har’El
Ron Sivan
Aya Soffer

Feb 11, 2004
What is the geography of a page?
What is the geography of a page?
What is the geography of a page?
What is the geography of a page?

Environment:
- URL, domain
- Owner
- Server locale, IP address
What is the geography of a page?
What is the geography of a page?

<html>
<head>
<title>iGuide - Israeli Internet Guide</title>
<meta name=keywords content="iGuide, israel, guide, classified, hebrew, news, il, tel-aviv, jerusalem, haifa, internet, www, Neystadt, map, physical, logical, line, router, bandwidth, site, category">
<meta name=description content="Israeli Internet Guide and Search engine in English and Hebrew. Provides a updated complete and categorical lists of Israeli sites, maps of ISPs, news.">
</head>
<body>
What is the geography of a page?

- **Environment**: external to page
  - URL, domain
  - Owner
  - Server locale, IP address
- **Content**: text and metatext
  - Place names
  - Adjectives (*Israeli*), language (*Hebrew*), currency (*NIS*)
  - Link analysis
What is the geography of a page?

- **Environment:** external to page
  - URL, domain
  - Owner
  - Server locale, IP address

- **Content:** text and metatext
  - Place names
  - Adjectives *(Israeli)*, language *(Hebrew)*, currency *(NIS)*
  - Link analysis
The process

- **Spotting**
  - Takes names to spot from the gazetteer

- **Disambiguate individual instances**
  - Takes taxonomies from the gazetteer

- **Find page foci**
The problem

- **Find names of places**
  - Must distinguish from non-geo
    - Is “London” a place or a person? (e.g., Jack London)

- **Determine the meaning**
  - Must arbitrate between competing senses
    - Is “London” in England or Ontario, Canada?

- **Finding the focus**
  - Must combine geotags meaningfully
The gazetteer

Contents

Name, synonyms, abbreviations
Population, coordinates, is-capital

Sources we used

GNIS – 2M features in the US
World-gazetteer – 50K cities worldwide
UNSB, ISO-3166, USPS

XML format

Imposes a hierarchical structure

```
<geographic-location level="city" label="San Jose">
  <population>816884</population>
  <coordinates>
    <latitude>37.33944</latitude>
    <longitude>-121.89389</longitude>
  </coordinates>
  <names-to-spot>
    <name>San Jose</name>
  </names-to-spot>
</geographic-location>
```
Disambiguation

- A bag of tricks
  - Qualified names
  - Default name sense
  - Single sense per discourse
  - Names with strong non-geo sense
  - Unifying locality
Qualified names

Names qualified by their next higher taxonomy level

- Haifa, Israel
- San Jose, California, USA
  - Definitely not the San Jose of Costa Rica, Guatemala, Bolivia, Peru or the Philippines
- Sacramento Calif.
The place having the highest population

- Portland, IN: 6,384
- Portland, ME: 61,982
- Portland, OR: 450,777
- Portland, TN: 5,708
- Portland, TX: 13,642
- Portland, Victoria, Australia: 10,455

---

- Portland, OR
A qualified instance induces its meaning on other, unqualified instances in the same context.
Single sense per discourse

- Single sense per discourse
  - A qualified instance induces its meaning on other, unqualified instances in the same context (page)

- Can lead to difficulties
  - As we shall soon see
Some place names are also common words

- Mobile, Al or Reading, GB
- Are ignored unless qualified

Some names are also very very common words

- To, Myanmar, Of, Turkey or As, Belgium
Historically, effective management of the U.S. Government has been hampered by a lack of reliable financial information. To help improve the integrity of financial information, in 1990 the Office of Management and Budget (OMB), the Department of the Treasury (Treasury) and the General Accounting Office (GAO) established the Federal Accounting Standards Advisory Board (FASAB) to develop accounting standards for the U.S. Government. The work of FASAB will enhance the acceptability of our financial reports and add to the level of financial professionalism throughout the U.S. Government.

Working to issue agency financial reports consistent with GAAP and to obtain clean audit opinions, the Administration is committed to improving financial reporting requirements. Financial reports required by 31 U.S.C. 3524 consist of the Management Discussion and Analysis (MD&A), Statement of Operations and Changes in Net Position, Statement of Net Cost, Balance Sheet, Stewardship Information, Notes to the Financial Statements and Supplemental Information. Each section is preceded by a description of its contents.

### Basis of Accounting and Reporting Entity

#### Accounting Standards

The accompanying financial statements generally were prepared based on GAAP standards developed by FASAB. The recent recognition of Federal accounting standards by the AICPA as GAAP enhances their acceptability. These standards form the foundation for preparing consistent and meaningful financial statements both for individual Federal agencies and the Government as a whole.

---

**“The recent recognition of Federal accounting standards by the AICPA as GAAP enhances their acceptability...”**

GAAP for the Federal Government is tailored to the U.S. Government's unique characteristics and special needs. For example, land not used in U.S. Government operations (stewardship land), weapons systems and support property used in the performance of military missions, and vessels held as part of the National Defense Reserve Fleet (national defense assets) are reported in the Stewardship Information section rather than valued on the Balance Sheet. The Government's responsibilities and policy commitments are much broader than the reported Balance Sheet liabilities. They include the social insurance programs disclosed in the Stewardship Information section, as well as a wide range
Some place names are also common words
- Mobile, AL or Reading, GB
- Are ignored unless qualified

Some names are also very very common words
- To, Myanmar, Of, Turkey or As, Belgium
- Excluded altogether
Names with strong non-geo sense

- Can be found from analyzing well-edited text
  - Count capitalized and non-capitalized occurrences
  - For most unambiguous place names –
    - Frequency is related to importance (population)
    - Cap to non-cap ratio is high
  - Names with strong non-geo sense stick out
Currently, a manual pass is required, however

- False positive: Aspen, CO
  - Appears frequently
  - Has only 5,000
  - Therefore included in the list
  - But most occurrences actually refer to it…

- False negative: Metro, Indonesia
  - Appears frequently
  - Has 100,000 inhabitants
  - Therefore excluded from the list
  - But most occurrences of “Metro” refer to Metro Area…
In addition to my house in Vienna, I also own a townhouse in Alexandria. Now, why do I own it? Is it a vacation home? Hardly. I bought it when I worked for R&K Engineering, and had an assignment to build a new factory in Alexandria, Virginia, and Alexandria was too far.

Well, as it turned out, while I had decided that about a mile down the road would do, the mortgage interest and most other expenses, plus depreciation, it looked like pretty much a wash, if I could get it rented out (plus some of that money went into equity, so it was a net win).
Individual spot disambiguation

- Three collections of Web pages
  - Gov: pages from .gov (1.2M)
  - ODP: the Regional tab of the ODP directory (600K)
  - Arb: What Google returned for “+the +and” (1.2K)
- Randomly selected 200 pages from each
- Ran the GeoMiner
  - Medium-sized gazetteer
- Tags given to spots were checked manually
Results of individual spots

Gov

ODP

Arb

Correct
73%

Correct
63%

Correct
82%

Geo/non-geo
20%

Geo/geo
7%

Geo/non-geo
21%

Geo/geo
16%

Geo/non-geo
11%

Geo/geo
7%
Page focus

- Contributions from each disambiguated mention
  - Depends on confidence
  - Contribution to each hierarchy level
- Highest scoring levels selected as foci
  - If they exceed a threshold
  - Are not part of or include another focus
Results of focus accuracy

- Focus given to 55% of the pages
  - 75% of the pages having at least 1 geotag
- Of the focused pages, 80% had but one focus
Results of focus accuracy

- Precise: 39%
- Too broad: 45%
- Too narrow: 4%
- Wrong: 12%
Precision vs. gazetteer size

- pages w/spots
- pages w/focus
- correct
- missing 1

Number of pages vs. Size of smallest cities in gazetteer.
Search results (Reuters corpus 1996)
Search results (pharmaceutical corpus)

AIDS

SARS
Future work

- Reduce geo/non-geo error rate
  - Ignore person names
- Use coordinate data
  - Newark in the context of Manhattan is in NJ, not NY
- Expand the gazetteer
  - Non-incorporated places
  - Counties, shires, cantons, etc.
  - Even smaller places
- Grammatical text analysis
Conclusions

- Web-a-where does a reasonable job of identifying place names in noisy Web pages
  - Some novel heuristics
  - Could do better…
  - …and run faster
- Performance tests
  - Not found in most publications (including MetaCarta)
  - Can help in designing other disambig algorithms
My gourmet friend went to India because he heard there was a new deli there.
What is the geography of a page?

ronny@asgard=> whois -h whois.isoc.org.il

person: John Neystadt
address: HaRav-Mimon 18/5
address: Bat-Yam
address: 59626
address: Israel
phone: +972 52 568-542
fax-no: +972 3 752-6012
e-mail: john@neystadt.org

% Rights to the data above are restricted by copyright.

ronny@asgard=>
What is the geography of a page?

asgard=> whois -h whois.isoc.org.il iguide.co.il

ronny@asgard> nslookup www.iguide.co.il

Name: stalker.iguide.co.il
Address: 194.90.246.9
Aliases: www.iguide.co.il

donny@asgard> whois -h whois.ripe.net 194.90.246.9

descr: NetVision
country: IL

descr: Netvision
descr: Omega Bldg.
descr: MATAM industrial park
descr: Haifa 31905
descr: Israel

ronny@asgard>

% Rights to the data above are restricted by copyright.
What is the geography of a page?
What is the geography of a page?
What is the geography of a page?