Data Acquisition through Geospatial Information Retrieval

John Farrell
L3Harris Technologies
Principle Investigator, Geospatial AI Analytics IRAD

john.farrell@l3harris.com
Wouldn’t it be nice if you could say “Alexa, find labels and annotations at 39.7499° N, 104.9917° W, and show me the results on my map”. This may be possible one day using Geospatial Information Retrieval or GIR.
Wouldn’t it be nice if you could say “Alexa, find labels and annotations at 39.7499° N, 104.9917° W, and show me the results on my map”. This may be possible one day using Geospatial Information Retrieval or GIR.

- What can I learn about a location from the Internet?
- How can I augment my geospatial understanding of a location?
- How can I do this automatically?
A note about structure and search

• Structured, Semi-structured and Unstructured data
  • Structured
  • Semi-Structured
  • Unstructured

• Search and Query
  • Query is precise, complete
  • Search is imprecise, ranked

OGC is a registered trademark of Open Geospatial Consortium.
A note about structure and search

- Structured, Semi-structured and Unstructured data
  - Structured
  - Semi-Structured
  - Unstructured

- Search and Query
  - Query is precise, complete
  - Search is imprecise, ranked

- Much of the information on the Internet contains references to places on earth.
- Traditionally Geographic Information Systems (GIS) focus on geospatial information available as structured data (OGC, KML, GML, etc.).
- Growing amounts of spatial references are available in unstructured or semi-structured data available today.

OGC is a registered trademark of Open Geospatial Consortium.
Denver

From Wikipedia, the free encyclopedia

For other uses, see Denver (disambiguation).

*Mile-High City* redirects here. For the capital of Lesotho, see Maseru.

Denver (i/ˈdɛnvr/), officially the City and County of Denver, is the capital and most populous municipality of the U.S. state of Colorado. Denver is located in the South Platte River Valley on the western edge of the High Plains just east of the Front Range of the Rocky Mountains. The Denver downtown district is immediately east of the confluence of Cherry Creek with the South Platte River, approximately 12 mi (19 km) east of the foothills of the Rocky Mountains. Denver is named after James W. Denver, a governor of the Kansas Territory, and it is nicknamed the Mile High City because its official elevation is exactly one mile (5280 feet or 1609.3 meters) above sea level. The 109th meridian west of Greenwich, the longitudinal reference for the Mountain Time Zone, passes directly through Denver Union Station.

Denver is ranked as a Beta world city by the Globalization and World Cities Research Network. With an estimated population of 716,492 in 2018, Denver is the 19th-most populous U.S. city, and with a 19.38% increase since the 2010 United States Census, it has been one of the fastest-growing major cities in the United States. The 10-county Denver-Aurora-Lakewood, CO Metropolitan Statistical Area had an estimated 2018 population of 2,932,415 and is the 19th most populous U.S. metropolitan statistical area. The 12-county Denver-Aurora, CO Combined Statistical Area had an estimated 2018 population of 3,572,798 and is the 15th most populous U.S. metropolitan area. Denver is the most populous city of the 18-county Front Range Urban Corridor, an oblong urban region stretching across two states with an estimated 2018 population of 4,976,781. Denver is the most populous city within a 500-mile (800 km) radius and the second-most populous city in the Mountain West after Phoenix, Arizona. In 2016, Denver was named the best place to live in the United States by U.S. News & World Report.
Basic approach

- Query Processing: build a query that represents location and space using natural language.
- Discovery/extraction of geographical information: find and formulate structured geographical data from the text.
- Gazetteers: catalogs of place name related information, including geographic coordinates.
- Corpus: source of our content.
- Index: the ‘card catalog’ for the extracted geographic data.
- Result Ranking: Results ranked by some priority determination.
We already use GIR

• Search engines have used local search in recent years.
• Works well with commercial directory listings and travel pages.
• Google has program to get your business listed and located.
We already use GIR

• Search engines have used local search in recent years.
• Works well with commercial directory listings and travel pages.
• Google has program to get your business listed and located.

What if I want something not in a business directory, or travel ad?

• The web provides a very large amount unstructured web content that refers to geographical information in semi-structured text.
• Easy to retrieve if there is a direct match between our query and the terms used in the web page.
• This presentation will introduce approaches to developing methods to find, index, rank and retrieve geographical data from text.
Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts:
- A theme (attractions)
- A spatial relationship (near)
- A location (Ritz-Carlton Denver).
Understanding geographical information in text

• Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts
  • A theme (attractions)
  • A spatial relationship (near)
  • A location (Ritz-Carlton Denver).

• It is likely that by Denver we mean Denver, CO, but this is an assumption.
• Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts
  • A theme (attractions)
  • A spatial relationship (near)
  • A location (Ritz-Carlton Denver).

• It is likely that by Denver we mean Denver, CO, but this is an assumption.
• What do we mean by near? Next door? Walking distance? Driving distance?
• Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts
  • A theme (attractions)
  • A spatial relationship (near)
  • A location (Ritz-Carlton Denver).

• It is likely that by Denver we mean Denver, CO, but this is an assumption.
• What do we mean by near? Next door? Walking distance? Driving distance?
• Are attractions ranked how close they are to the centroid of the Denver city limits? The centroid of ‘downtown’ Denver? The ratings of the attractions?
Understanding geographical information in text

• Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts
  • A theme (attractions)
  • A spatial relationship (near)
  • A location (Ritz-Carlton Denver).

• It is likely that by Denver we mean Denver, CO, but this is an assumption.
• What do we mean by near? Next door? Walking distance? Driving distance?
• Are attractions ranked how close they are to the centroid of the Denver city limits? The centroid of ‘downtown’ Denver? The ratings of the attractions?
• What are the attractions? A monument (point), state park (polygon) or a scenic drive (polyline)?
Understanding geographical information in text

• Consider the query “attractions near the Ritz-Carlton Denver”. It consists of three important parts
  • A theme (attractions)
  • A spatial relationship (near)
  • A location (Ritz-Carlton Denver).
• It is likely that by Denver we mean Denver, CO, but this is an assumption.
• What do we mean by near? Next door? Walking distance? Driving distance?
• Are attractions ranked how close they are to the centroid of the Denver city limits? The centroid of ‘downtown’ Denver? The ratings of the attractions?
• What are the attractions? A monument (point), state park (polygon) or a scenic drive (polyline)?
• Theme/Spatial Relationship/Location is a rule. More on that later.
We need to find geographic references in text, and unambiguously resolve these to locations on the earth’s surface.
Step 1: Identify geographical references unambiguously

The located object and the reference object can be given place names that can vary in geographical granularity.
GeoNames

Containment
Place names and spatial references are ambiguous.

- Next to
- Located in
- In the neighborhood
- Foothills of
- Near
- Down the road
- Western edge of
- East of
- Downtown
- Passes through
Automating geographical reference identification

• There are three basic approaches to identifying candidate georeferences:
  • List lookup
    • Need exact match
    • Can use capitalization, stemming, thesaurus, etc.
  • Manual

Buildings -> buildings -> building
Automating geographical reference identification

There are three basic approaches to identifying candidate georeferences:

- List lookup
  - Need exact match
  - Can use capitalization, stemming, thesaurus, etc.
  - Manual

- Knowledge or rule-based methods
  - Take context into account
  - Natural Language Processing (NLP)
  - More general than list lookup
  - Tokenization, stemming, stop words, named entity resolution (starts to use machine learning)

Geographic statement = Theme/Spatial Relationship/Location

Buildings -> buildings -> building
Automating geographical reference identification

- There are three basic approaches to identifying candidate georeferences:
  - List lookup
    - Need exact match
    - Can use capitalization, stemming, thesaurus, etc.
  - Manual
  - Knowledge or rule-based methods
    - Take context into account
    - Natural Language Processing (NLP)
    - More general than list lookup
    - Tokenization, stemming, stop words, named entity resolution (starts to use machine learning)
  - Machine learning
    - Supervised and unsupervised
    - Less manual effort (with exception of labeling data)
    - Topic modeling, deep encoding, segmentation, classification

Geographic statement = Theme + Spatial Relationship + Location

Buildings -> buildings -> building
Step 2: Assign spatial coordinates (Geocode)

• Associated the georeference with:
  • A point, line or polygon
  • A raster cell

U.S. National Grid https://www.arcgis.com/home/webmap/viewer.html?layers=dc352c5f18854d82b32bce92c0b6656b

© OpenStreetMap contributors
Conclusion

• Geospatial Information Retrieval (GIR) is still in the early stages, but promising
• Hopefully you gained some understanding of Geospatial Information Retrieval
  • Geospatial Information Retrieval is fairly easy for humans
  • Automating Geospatial Information Retrieval is hard, but there has been progress

Questions?