Query GeoParser: A Spatial-Keyword Query Parser Using Regular Expressions

Jason Hines and Tony Abou-Assaleh
GenieKnows.com

GenieKnows.com

Outline

- Domain
- Our Approach
- Advantages
- Disadvantages
- Conclusion

Local Search

- Many local search engines exist to fulfill spatial-keyword (SK) queries
  - *Hotels near 1567 Argyle St, Halifax, NS*

GenieKnows.com

Local Search

- Many local search engines exist to fulfill spatial-keyword (SK) queries
  - *Hotels near 1567 Argyle St, Halifax, NS*

GenieKnows.com

Local Search

- Many local search engines exist to fulfill spatial-keyword (SK) queries
  - *Hotels near 1567 Argyle St, Halifax, NS*

GenieKnows.com

Local Search

- Many local search engines exist to fulfill spatial-keyword (SK) queries
  - *Hotels near 1567 Argyle St, Halifax, NS*

- Many use two text fields
  - e.g., what & where
  - Popular, but somewhat restricting

GenieKnows.com

Local Search

- Many local search engines exist to fulfill spatial-keyword (SK) queries
  - *Hotels near 1567 Argyle St, Halifax, NS*

- Many use two text fields
  - e.g., what & where
  - Popular, but somewhat restricting
    - And ... not that interesting from a research perspective

GenieKnows.com

Local Search

- Single text fields
  - require a more intelligent parser

Local Search

- Single text fields
  - require a more intelligent parser
    - Part-of-speech tagging with Hidden Markov Models

Local Search

- Single text fields
  - require a more intelligent parser
    - Part-of-speech tagging with Hidden Markov Models
    - Statistical approaches

Local Search

- Single text fields
  - require a more intelligent parser
    - Part-of-speech tagging with Hidden Markov Models
    - Statistical approaches
    - Regular expressions

Local Search

- Single text fields
  - require a more intelligent parser
    - Part-of-speech tagging with Hidden Markov Models
    - Statistical approaches
    - Regular expressions
    - Hybrid models

Query GeoParser

- Perl / C++
  - Extensive regular expression capabilities
    - Wildcards
    - Grouping
    - Pre-compiling
    - Look-ahead
    - Building blocks

GenieKnows.com

Approach

Approach

Clean the query

Approach

Clean the query

Mark the query

Approach

Clean the query

Mark the query

Match query tokens against templates

GenieKnows.com

Approach

Clean the query

Mark the query

Match query tokens against templates

Decode and disambiguate matches

GenieKnows.com

Approach

Clean the query

Mark the query

Match query tokens against templates

Decode and disambiguate matches

GenieKnows.com

Cleaning

- Consider the following query
  - *Hotels near 1567 Argyle St, Halifax, NS*

- *Convert to lowercase, pad commas*

GenieKnows.com

Cleaning

- Consider the following query
  - *Hotels near 1567 Argyle St, Halifax, NS*
- *Convert to lowercase, pad commas*
- *{‘hotels’, ‘near’, ‘1567’, ‘argyle’, ‘st’, ‘,’, ‘halifax’, ‘,’, ‘ns’}*

GenieKnows.com

Marking

- Simple back-tracking
  - First and longest combinations of consecutive tokens
    - Prepositions, zip codes, popular names, etc.
- Encode to make efficient use of wildcards

Marking

- Simple back-tracking
  - First and longest combinations of consecutive tokens
    - Prepositions, zip codes, popular names, etc.
- Encode to make efficient use of wildcards
- ‘hotels near|prep 1567|number argyle|city|popb st|type , halifax78|city|popb , ns|state’

GenieKnows.com

Marking

- Simple back-tracking
  - First and longest combinations of consecutive tokens
    - Prepositions, zip codes, popular names, etc.
- Encode to make efficient use of wildcards
- ‘hotels near|prep 1567|number argyle|city|popb st|type , **halifax78**|city|popb , ns|state’

GenieKnows.com

Matching

- Match marked query against templates
  - Order matters!
    - In general, stricter, or specific, templates first
- Over 400 templates

Matching

- Match marked query against templates
  - Order matters!
    - In general, stricter, or specific, templates first
- Over 400 templates
- <keywords> <preposition> <street>, <city>, <state>

GenieKnows.com

Decoding

- If template matches, remaining parts are known
- Decode, disambiguate, and print matches

Decoding

- If template matches, remaining parts are known
- Decode and print matches
- {‘keywords’=>’hotels’, ‘prep’=>’near’, ‘number’=>’1567’, ‘street’=>’argyle’, ‘type’=>’st’, ‘city’=>’halifax’, ‘state’=>’ns’}

GenieKnows.com

Advantages

- Simplicity and readability of source code
- Mark and match improves efficiency
- Over 400 templates

Disadvantages

- Must build template set
- Not hierarchical
- Strict matches are required
  - How do you deal with misspellings?

Disadvantages

- Piza near Pizza st, Lawrence, NY
- Kentucky Fried Chicken
- Janes on the Common
- Street street Street MD => Street rd, Street, MD

Conclusion

- Not the be all end all solution
  - But is a viable approach
- Over 400 successful templates
- Code simplicity

GenieKnows.com

Thank You!

- http://www.genieknows.com
- research@genieknows.com