EXTRACTING CENTRAL PLACES FROM THE LINK STRUCTURE IN WIKIPEDIA

Carsten Keßler
Department of Planning, Aalborg University Copenhagen
http://carsten.io – @carstenkessler

GEOINFORMATICS
AALBORG UNIVERSITY COPENHAGEN

Are these patterns reflected on the web?

- Research Question: Do relationships between pages about places could reflect CPT patterns?
- Method:
  1. Take a large set of pages about places
  2. Extract relationships
  3. Compare to real-world patterns

Solution: Reference dataset

- Central places have to be declared in German spatial planning documents by each state
- Reference dataset of **123** upper centers and **874** middle centers

Wikipedia dataset

- German Wikipedia dump used to compare to reference dataset
- Process:
  - Pages about cities selected
  - Lat/lon extracted
  - References (links + “mentions”) counted
  - Import everything into PostGIS

Wikipedia dataset stats

- ~ 2.3 mio. pages
- ~ 73 mio. pairs of pages with links between them
- ~ 91 mio. links
- ~ 2.1 bio. mentions
- **~ 2.2 bio. References**

- Can the references be used to infer “centrality” of a place?

Bottom-up approach

- Select the most referenced city for every city in the dataset (with count)
- Add up incoming references for every city based on those counts
- Rank in descending order
- Select 125 first as upper centers, 875 next as middle centers

Results

- Within those 1000 results:
  - 110 of 123 upper centers
  - 404 of 874 middle centers

Conclusions

- Results show first indication that real-world spatial relationships between places are reflected on the web
- One attribute (# references) can predict centers with F-score of 0.51
- Needs to be confirmed with other kinds of data sources from the web