Die Bearbeitung von Concept@Location Queries mit vager Spezifikation der Location

Andreas Henrich, Volker Lüdecke

andreas.henrich@uni-bamberg.de
volker.luedecke@uni-bamberg.de

Lehrstuhl für Medieninformatik
Universität Bamberg

Bedeutung geografischer Anfragen

- Analyse der „AOL 500k User Session Collection“
- Anfragen an eine „normale“ Suchmaschine!
- Genutztes Suchpattern:

```
<concept> <preposition> <toponym t> <statequalifier t is in>
```

  - part time jobs near lansing illinois
  - motels in santa monica ca

- Toponym:
  - Jeder der 26000 Ortsnamen in den USA aus dem WorldGazetteer

Analyseprozess

10 mio queries

hotels in houston texas
john f kennedy
hotels in austin tx
dan brown books
hotels in santa monica ca

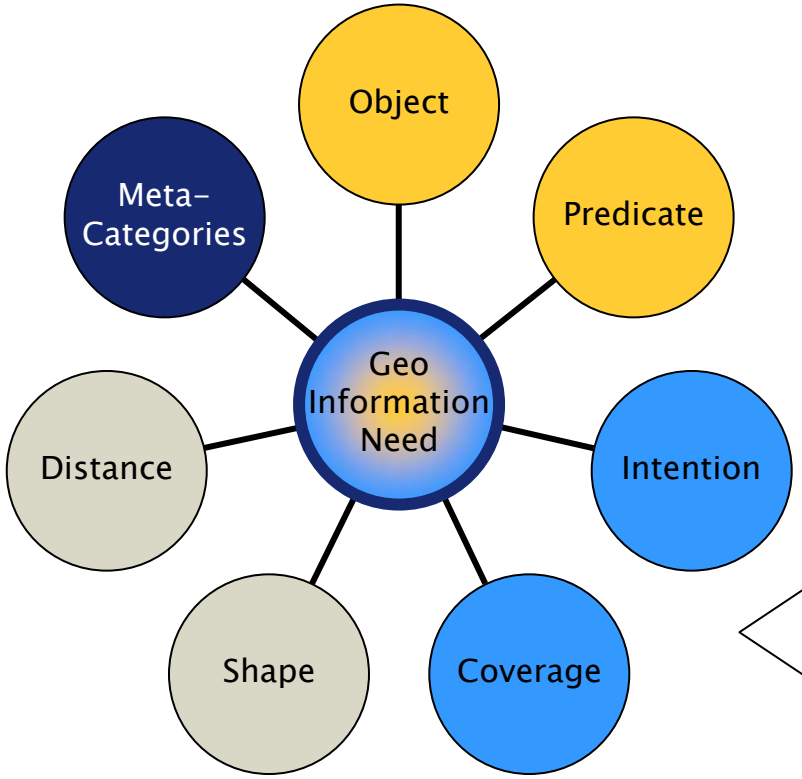
Pattern-Filter

hotels in houston texas
hotels in austin tx
hotels in santa monica ca

225421 geographic queries (pattern 1)

stripping geographic reference

aggregating



Analysis

hotels (3)

225421 concept parts of geographic queries (pattern 1)

Beispielergebnisse

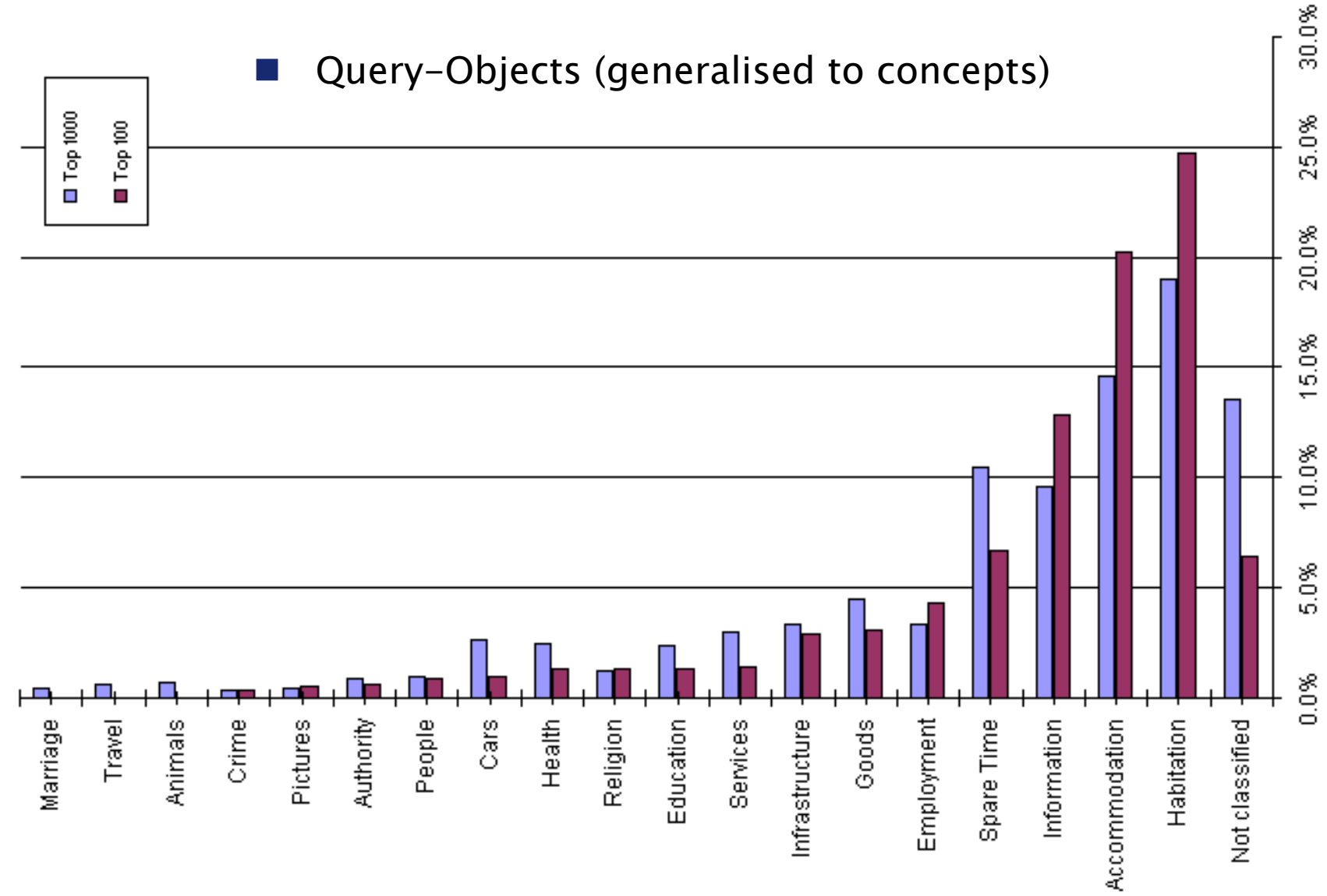
concept query	#
hotels	2115
city	2053
homes for sale	1062
jobs	981
real estate	947

Most frequent concept parts of queries

Habitation	Accommodation	Spare Time	Information
homes for sale	hotels	restaurants	map
real estate	motels	things to do	weather
apartments	holiday inn	ymca	zip code
houses for sale	hampton inn	movie theaters	newspaper
houses for rent	comfort inn	campgrounds	population

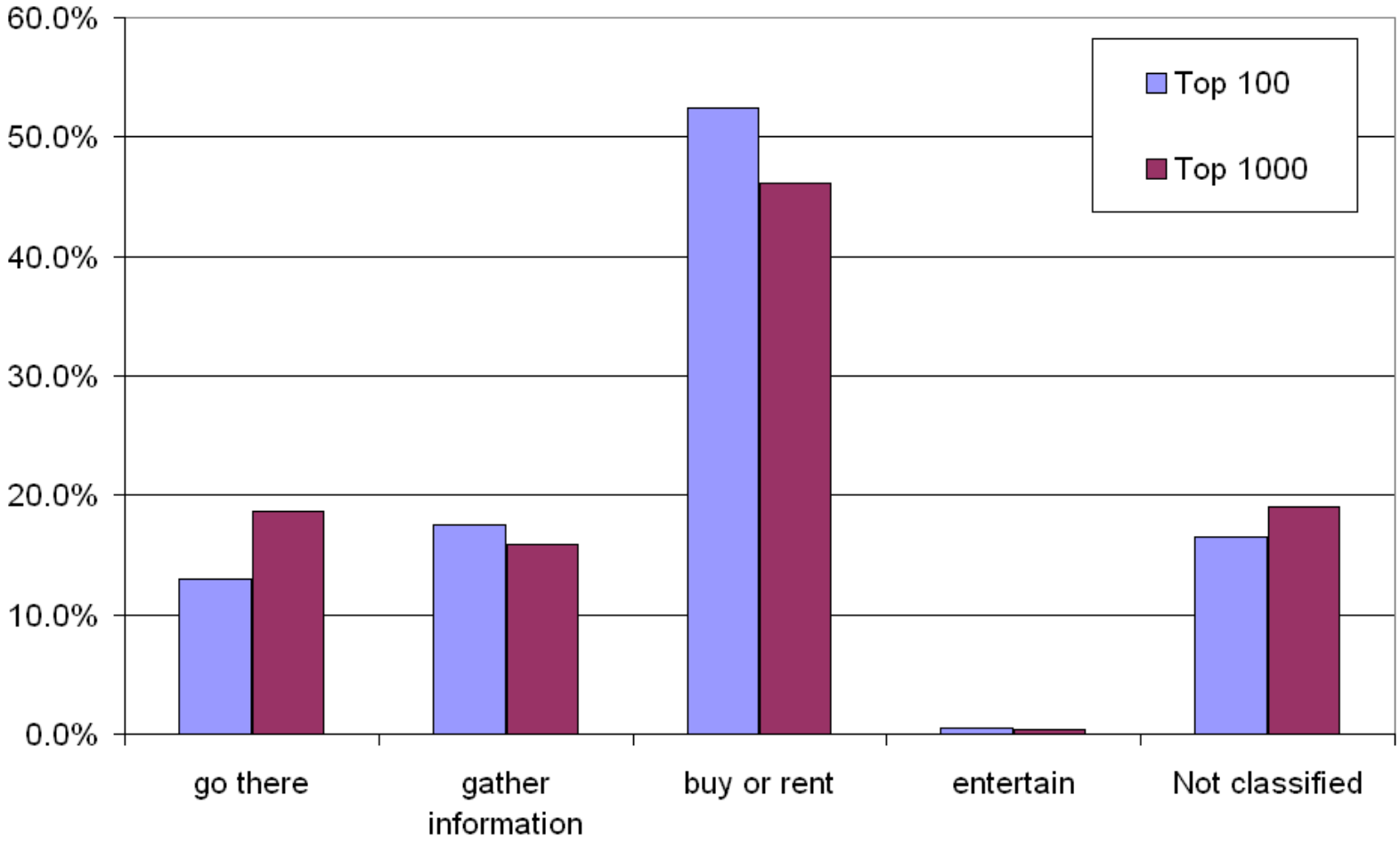
Top 5 queries for the 4 most frequent Object-concepts

Aspects of real Geographic Information Needs

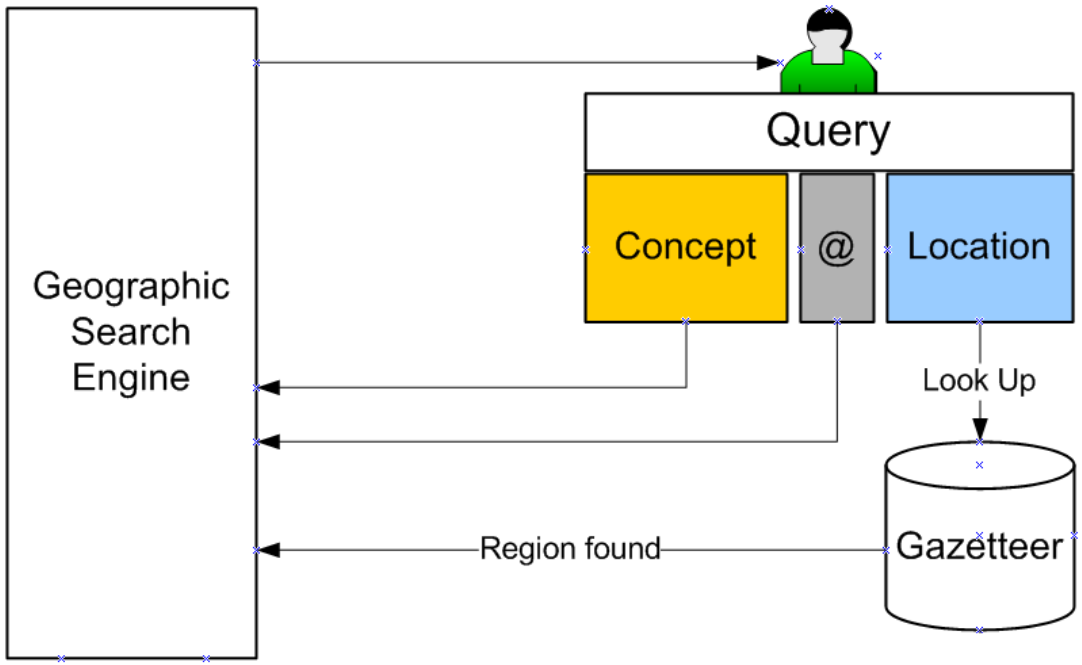


Aspects of real Geographic Information Needs

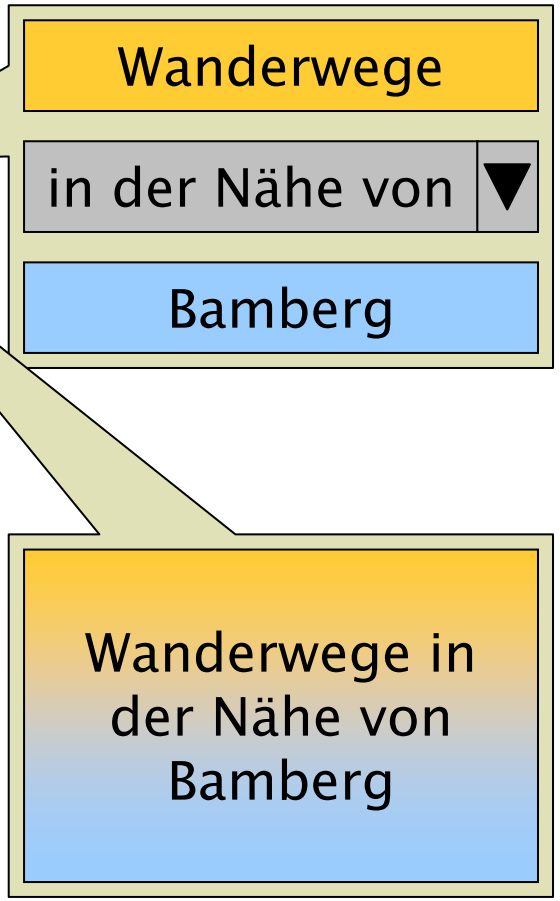
- Predicate (Intention of query)



Szenario



strukturierte Eingabe



Analyse per NLP

Gazetteer: Beispiel OpenGeoDB

OpenGeoDB - freie Geokoordinaten-Dat

Ortsdatenbank

Bamberg | Suche starten

Ihre Suche nach "Bamberg" ergab 3 Treffer:

1. **Altenbamberg**
   Deutschland > Rheinland-Pfalz > Landkreis Bad Kreuznach
2. **Bamberg**
   Deutschland > Bayern > Kreisfreie Stadt Bamberg
3. **Neu-Bamberg**
   Deutschland > Rheinland-Pfalz > Landkreis Bad Kreuznach

OpenGeoDB - freie Geokoordinaten-Datenbank

Ortsdatenbank

Bamberg | Suche starten

Bamberg

Verwaltungsgliederung

Staat	Deutschland
Bundesland	Bayern
Regierungsbezirk	Oberfranken
Landkreis	Kreisfreie Stadt Bamberg

Koordinaten

Länge	10.8667 / E 10° 52' 0"
Breite	49.8667 / N 49° 52' 0"

Zusatzinformationen

Gemeindekennzahl	09461000
Kfz-Kennzeichen	BA
Postleitzahl(en)	96047, 96049, 96050 ...

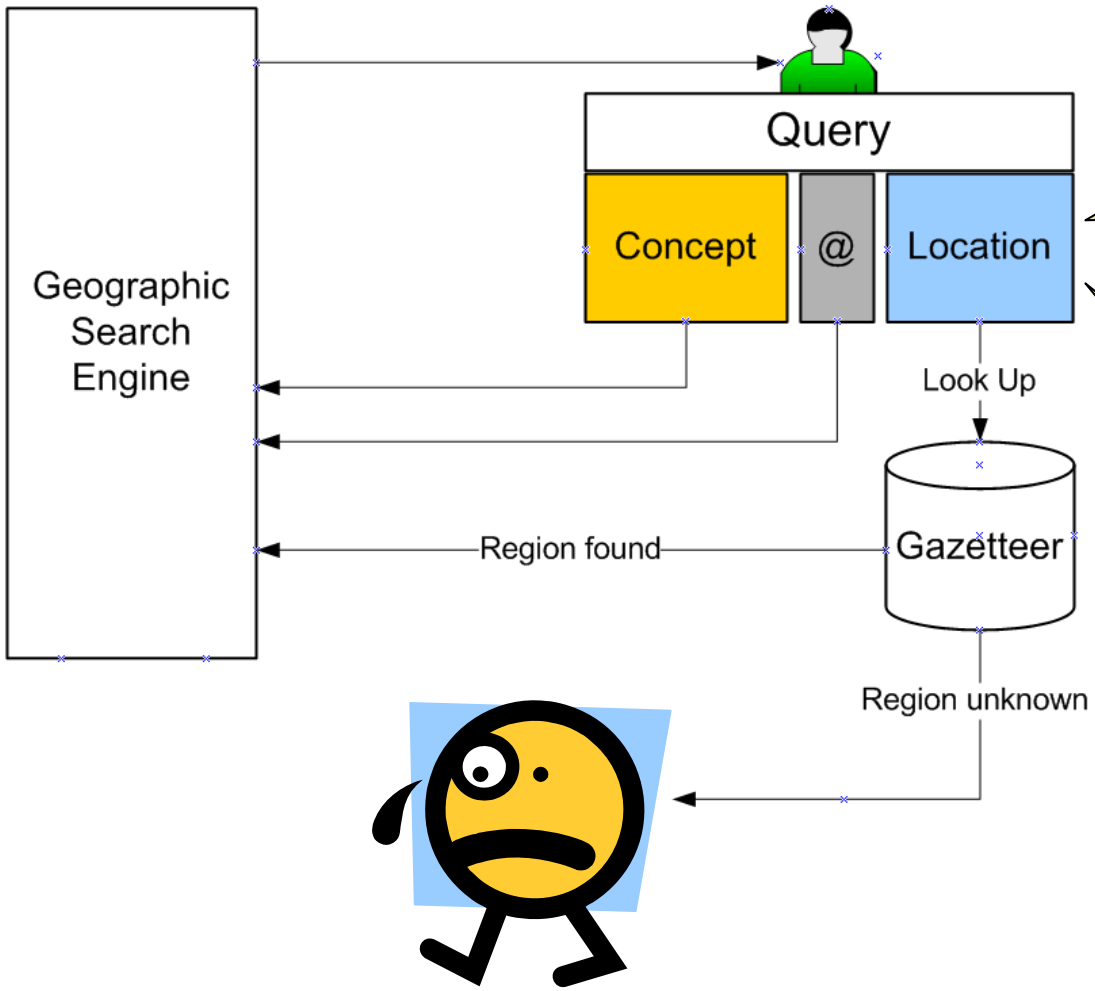
Externe Informationsquellen

- **Bamberg** suchen in *Wikipedia - Die freie Enzyklopädie*
- **Bamberg** suchen in *Google*

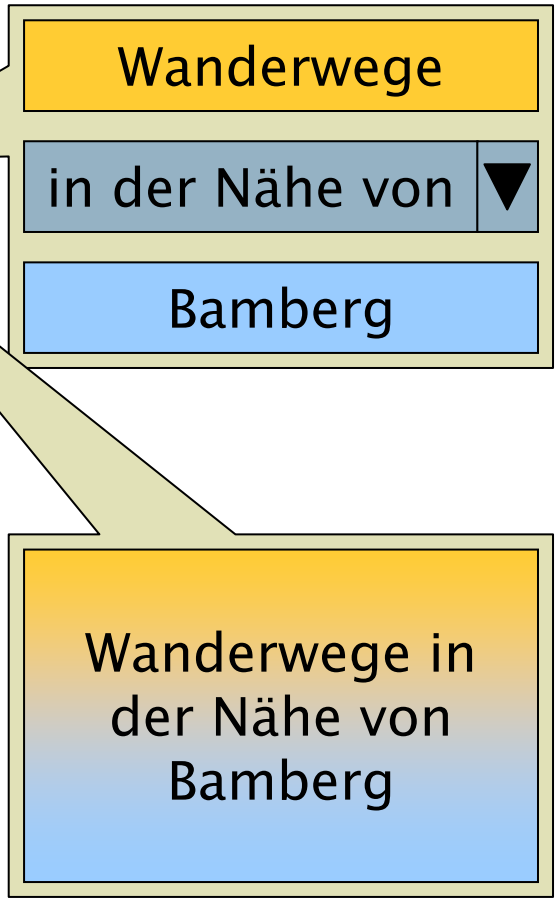
Orte im Umkreis von 10 km

1. **Stegaurach**
   Deutschland > Bayern > Landkreis Bamberg
   Entfernung: 2.39 km
2. **Frensdorf**

Szenario



strukturierte Eingabe



Analyse per NLP

Konsequenzen

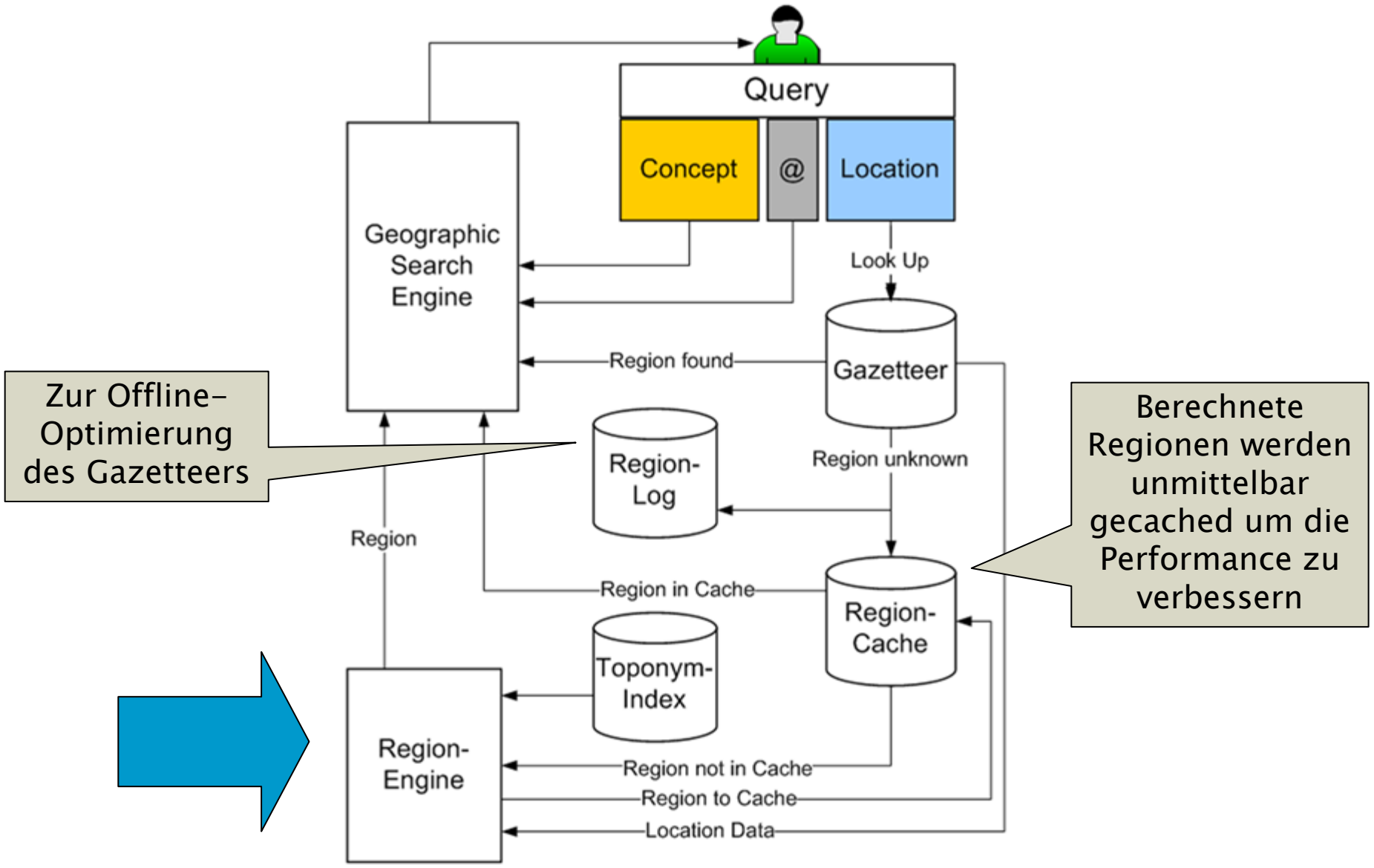
- Was ist zu tun, wenn die „Location“ …
  - ein Ort / eine Region ist, die der Gazetteer nicht kennt
    - z.B. bei umgangssprachlichen Bezeichnungen
  - gar keine geografische Region ist: z.B.
    - „Campingplatz in der Nähe von Freizeitpark“
    - „Radwanderweg in der Nähe von Brauerei“

➢ Bestimme die Grenzen der „Location“ zur Anfragezeit!

Resultierende Anforderungen

- Schnelle Berechnung der Regionen
- Bestimmung einer geografischen Repräsentation für beliebige Begriffe, die als „Location“ angeben werden
- Vollständig automatischer Prozess
- Integration in die Anfragebearbeitung der Geo-Suchmaschine
- Angemessene Ergebnisqualität als Basis für ein Ranking

Systemarchitektur



Region Engine

- Bestimmt die geografische Repräsentation für einen oder mehrere Begriffe, die als „Location“ gegeben sind

1. Verwende die „Location“ als „normale“ Anfrage
   - (wir nutzen aktuell die Google-API um gecrawlte Daten zu simulieren)
   - die ersten 500 Treffer
   - nur HTML-Seiten
   - die ersten *k* dieser Treffer werden betrachtet (Parameter)

Finales System kann den Index der Geo-Suchmaschine selbst nutzen. So können Dokumente mit bereits zugeordneten und aufgelösten Toponymen verwendet werden.

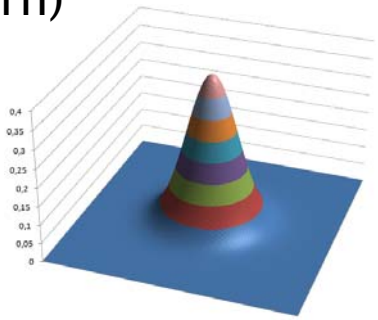
2. Extraktion und Auflösung der Toponyme (inkl. Disambiguierung, …)
   - Aufbau eines Toponym-Index für eine bessere Performance
   - Auflösung von Toponymen wird so zum einfachen Index-Lookup

3. Bestimmung der geografischen Repräsentation auf Basis der in den Dokumenten gefundenen Toponyme

Region Engine

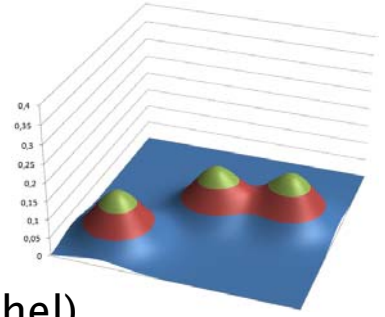
- Density Surfaces als Repräsentation
  (analog zu Purves, Clough, Joho 2005)
  - kernel density estimation, Gaussian kernel function
  - Jedes Toponym hat eine „Einflussregion“ (Gaußkern)

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$



  - Dokumente: Gew. Überlagerung der „Gaußkerne“

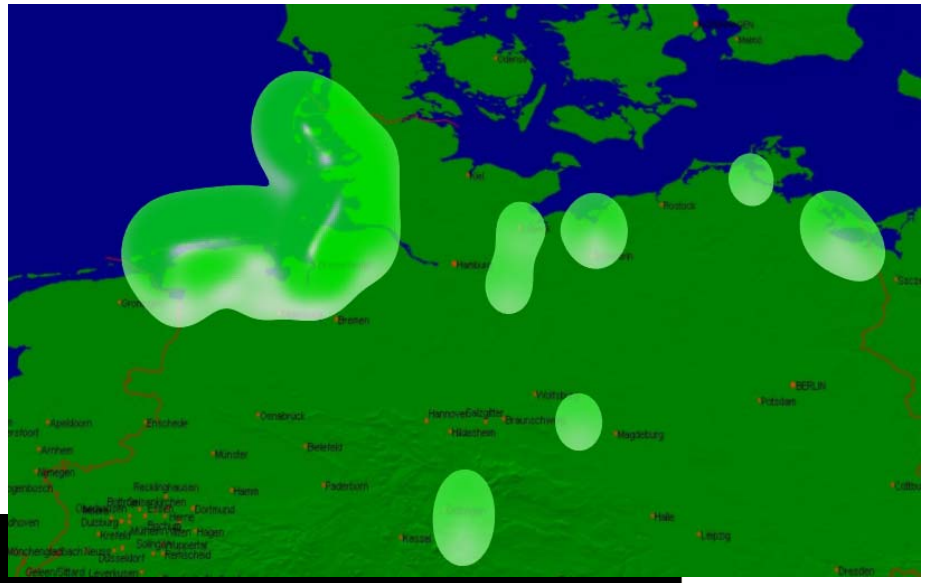
$$\widehat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

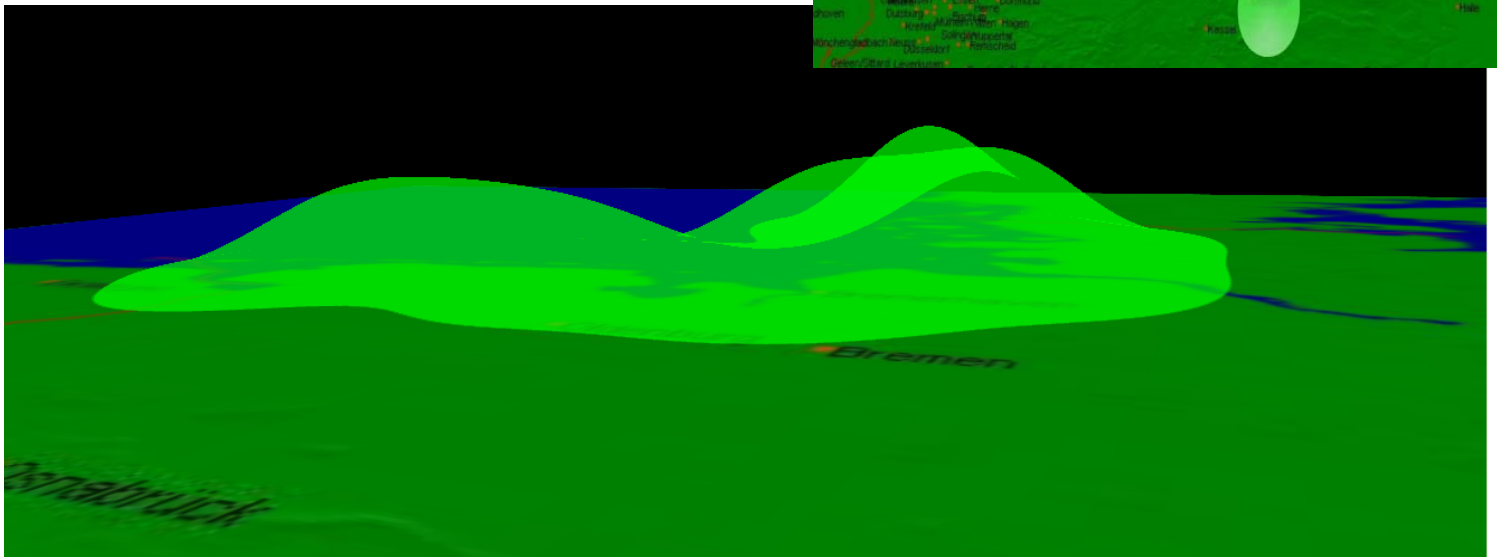


  - Die ganze Region wird in Kacheln eingeteilt
    - Kachelgöße → Designparameter (z.B. 1 km² pro Kachel)
  - Bestimmung eines Schwellenwertes *T*
    - Alle Kacheln mit Werten > *T* repräsentieren „Location“

Region Engine

- Illustration: (Nordsee)

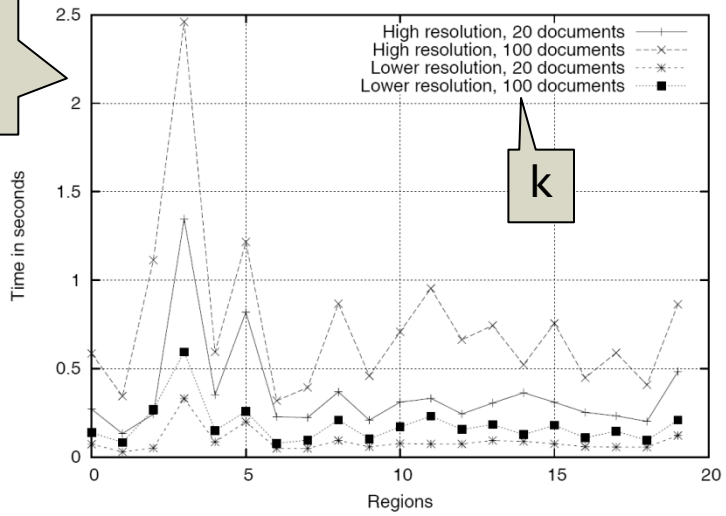


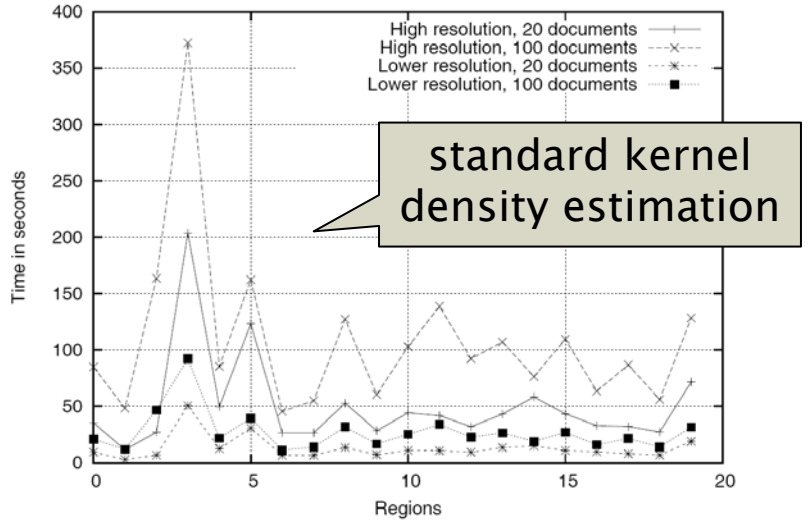


Region Engine: Performance Verbesserung

- Inkrementelle Berechnung für bessere Performance
  - In der Praxis haben viele Kacheln einen Wert nur knapp über 0
  - ➢ Idee: berechne die Werte nur für Kacheln mit Werten deutlich > 0
- Unser Ansatz:
  - Iteriere über Toponyme, nicht über Kacheln
  - Toponyme, die mehrfach vorkommen, werden entspr. gewichtet
  - Je Toponym wird der Einfluss auf alle benachbarten Kacheln bestimmt
  - Ein Grenzwert wird genutzt um dieses „Ausbreiten“ zu begrenzen

tuned kernel density estimation





Näherungsweise Bestimmung des Schwellenwertes *T*

- Ziel: Bestimme den optimalen Schwellenwert *T* automatisch
- Idee:
  - Berechne Näherung von *T* auf Basis einfach bestimmbarer Kenngrößen
- Mögliche bestimmbare Kenngrößen:
  - Anzahl der Toponyme in den Dokumenten
  - Größe der Region (Anzahl Kacheln mit Werten > des Schwellenwertes)
  - Maximaler Wert einer Kachel (V_{max})
  - Summe der Werte aller Kacheln oberhalb des Schwellenwertes
  - Summe der Werte aller Kacheln
- Problem:
  - Wie bestimmt man, was eine gute Näherung ist?
  - Testdaten („ground truth“) werden benötigt

Näherungsweise Bestimmung des Schwellenwertes *T*

- Manuelle Bestimmung der Grenzen von Beispielregionen
  - 2 Experten haben die Regionen auf Basis einer Literaturrecherche bestimmt
  - Grenzen wurden per Google Maps gezeichnet
  - Umwandlung der Polygone in Koordinaten
  - Problem:
    - Eine gewisse Vagheit der Regionen bleibt
  - Trotzdem:
    - Für unsere Zwecke gute Ausgangsbasis
  - Beispiel: Havelland

Wann „passt“ die automatisch bestimmte Region?

- Maße für die Übereinstimmung *sim*:
  - n: Anzahl der Kacheln innerhalb der „tatsächlichen Region“
  - m: Anzahl der Kacheln außerhalb der „tatsächlichen Region“
  - Val: Wert einer Kachel $\widehat{f}_h(x)$

$$sim_{bin} = 2 * n - m$$

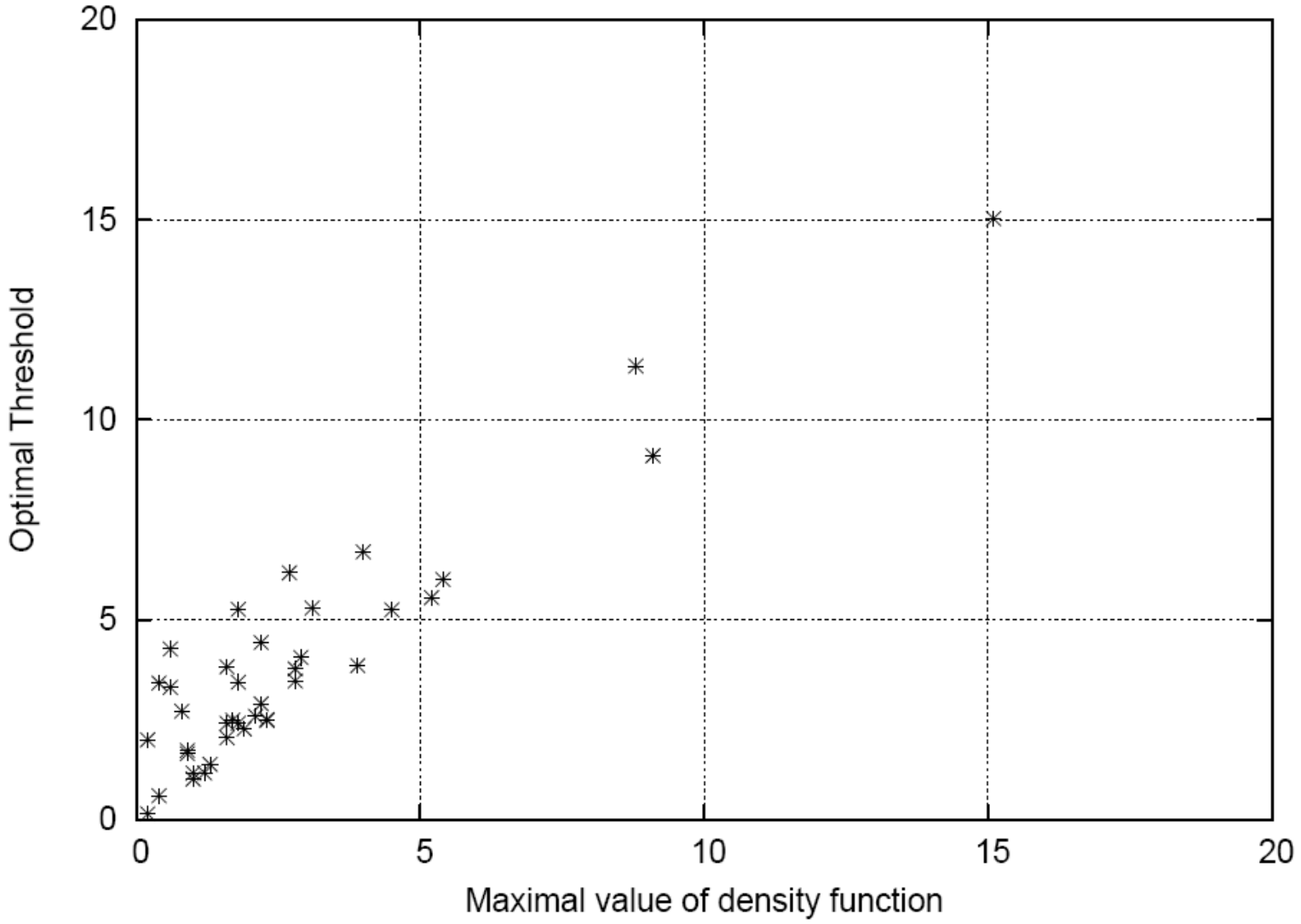
$$sim_{val} = 2 * \sum_{i=1}^{n} Val_i - \sum_{j=1}^{m} Val_j$$

Bestimmung von $T_{predicted}$

- Aus welcher Kenngröße lässt sich der beste Schätzer für T bestimmen?
- Ermittle die Schwellenwerte, für die *sim* am höchsten ist
- Speichere hierzu je auch die Werte für die 5 Kenngrößen
- Bestimme die Korrelationen zwischen Kenngrößen und optimalem Schwellenwert

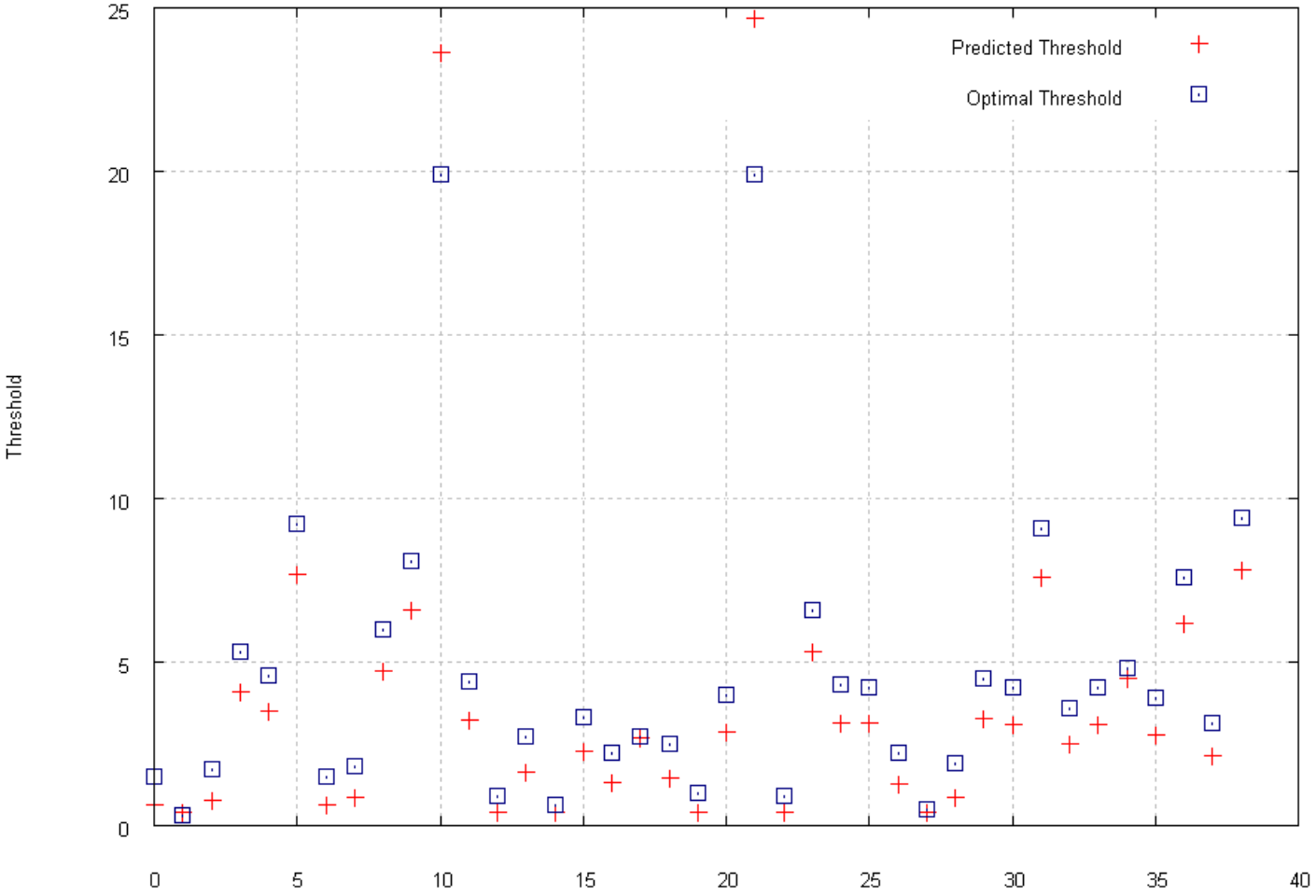
- Ergebnis: Nur V_{max} hat eine gute Korrelation

Korrelation



Evaluation

- Qualität des geschätzten $T_{predicted} = \max(0.2, 0.91 * V_{\max} - 0.79)$

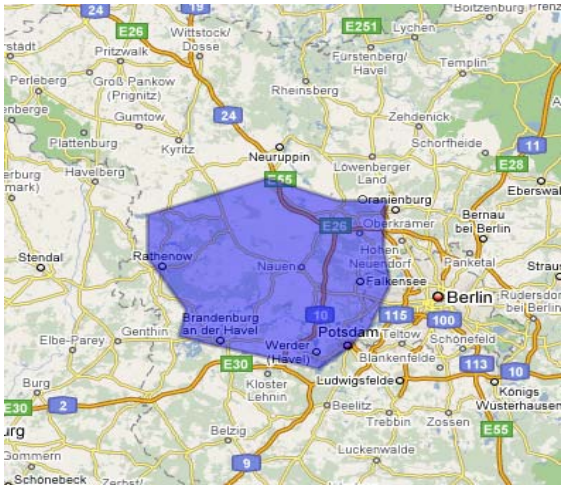


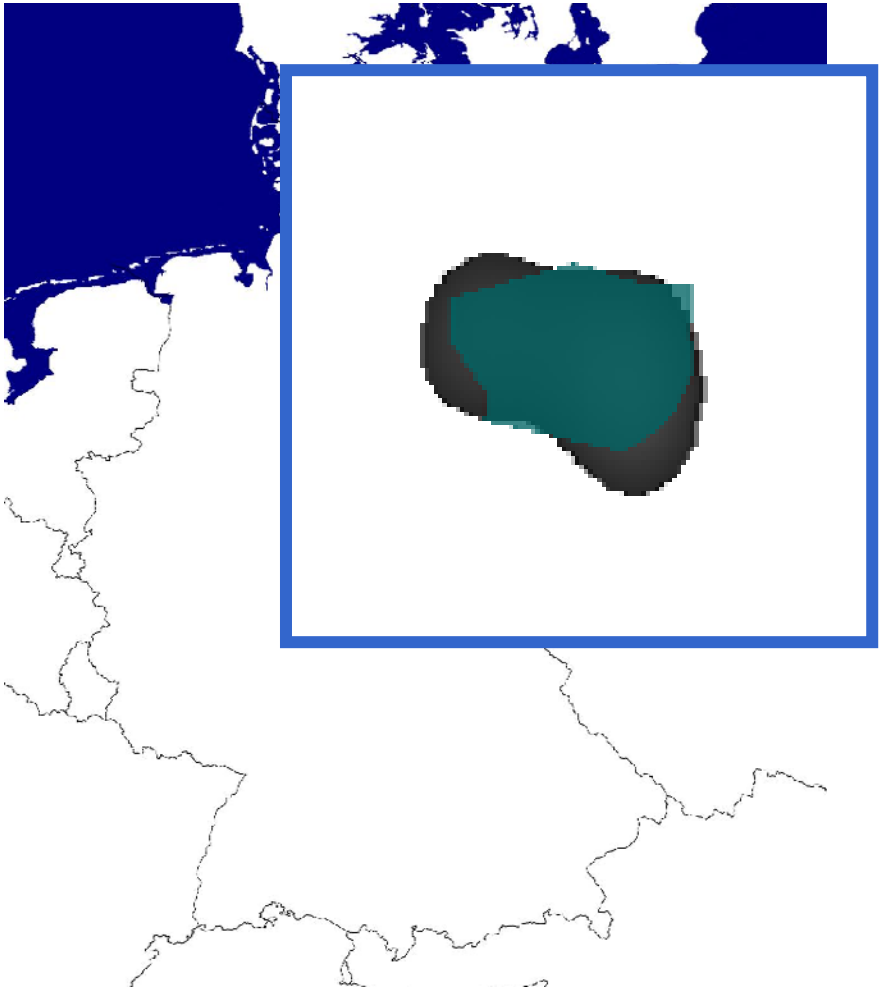
Motivation für die Evaluierung

- Anwendung des Ansatzes zur geografischen Repräsentation beliebiger Begriffe
  - Geografische Nutzung beliebiger Begriffe in Anfragen:
    - Campingplatz in der Nähe von Freizeitpark
    - Hotel in der Nähe von Brauerei
  - *wo-ist*-artige Anfrage
    - Wo waren die olympischen Spiele 1972
  - Untersuchung, ob bestimmte Begriffe eine geografische Korrelation aufweisen
    - Sprachstudien
    - Marketing, Tourismus

Evaluation: Qualität der Repräsentationen

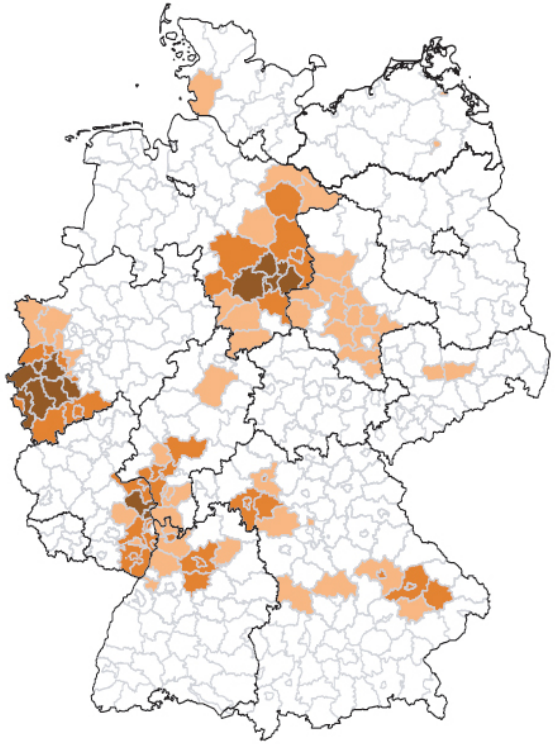
- Havelland

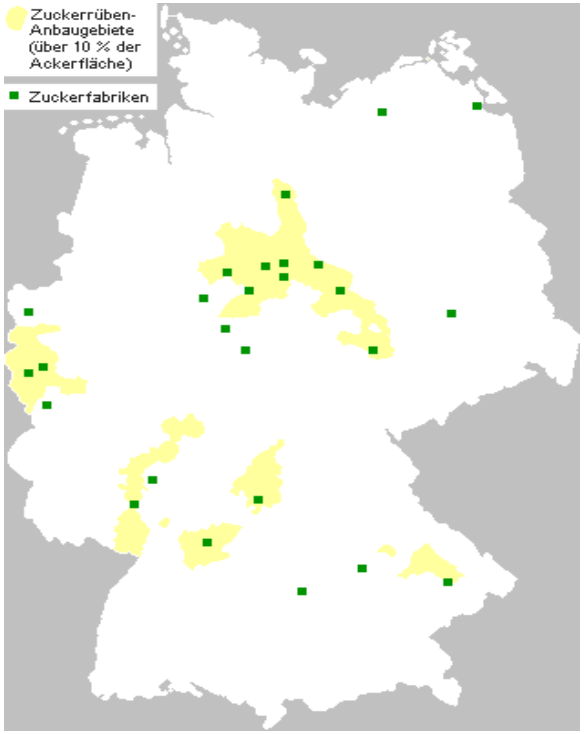

Wittstock/ Dosse
Pritzwalk
Fürstenberg/ Havel
Templin
Perleberg
Groß Pankow (Prignitz)
Rheinsberg
Gumtow
Zehdenick
Plattenburg
Kyritz
Neuruppin
Schorfheide
Havelberg
Lowenberger Land
Oranienburg
Bernau bei Berlin
Oberkrämer
Stendal
Rathenow
Nauen
Hohen Neuendorf
Panketal
Falkensee
Berlin
Brandenburg an der Havel
Potsdam
Teltow
Schönefeld
Elbe-Parey
Genthin
Werder (Havel)
Blankenfelde
Ludwigsfelde
Königs Wusterhausen
Kloster Lehnin
Beelitz
Trebbin
Zossen
Belzig
Luckenwalde
Schönebeck



Evaluation: Qualität der Repräsentationen

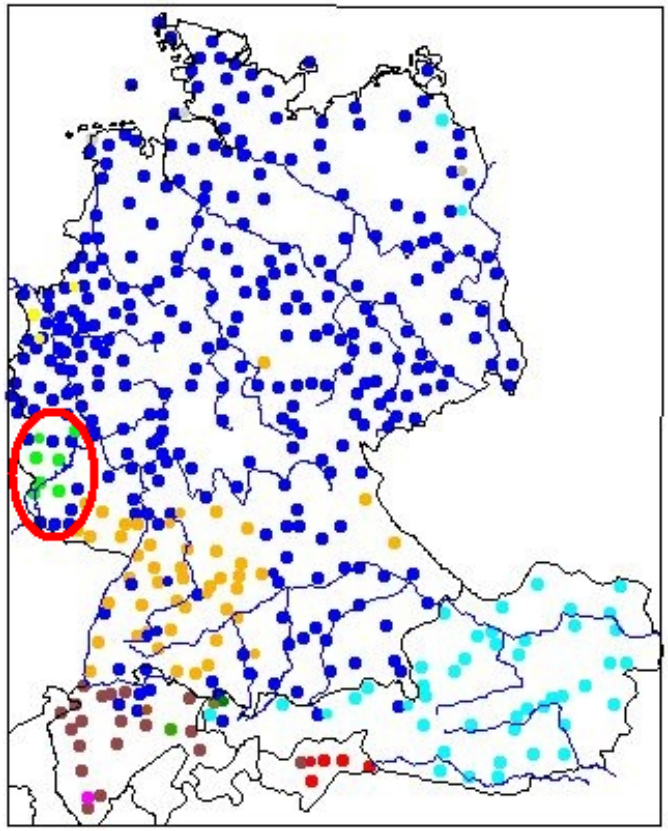
- Anbau von Zuckerrüben in Deutschland





Evaluation: Qualität der Repräsentationen

- „Knuppauto“, umgangssprachlich für „Autoscooter“

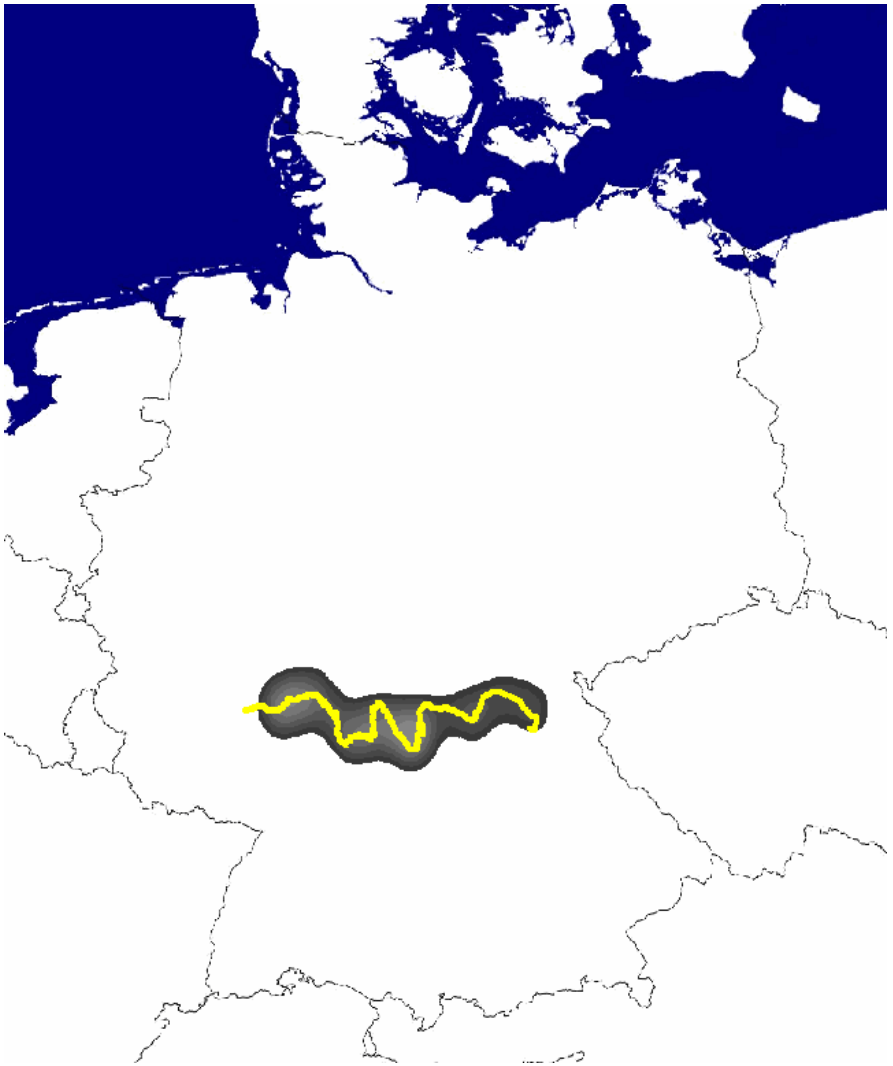


Resultat einer Studie

Resultat der Suchmaschine

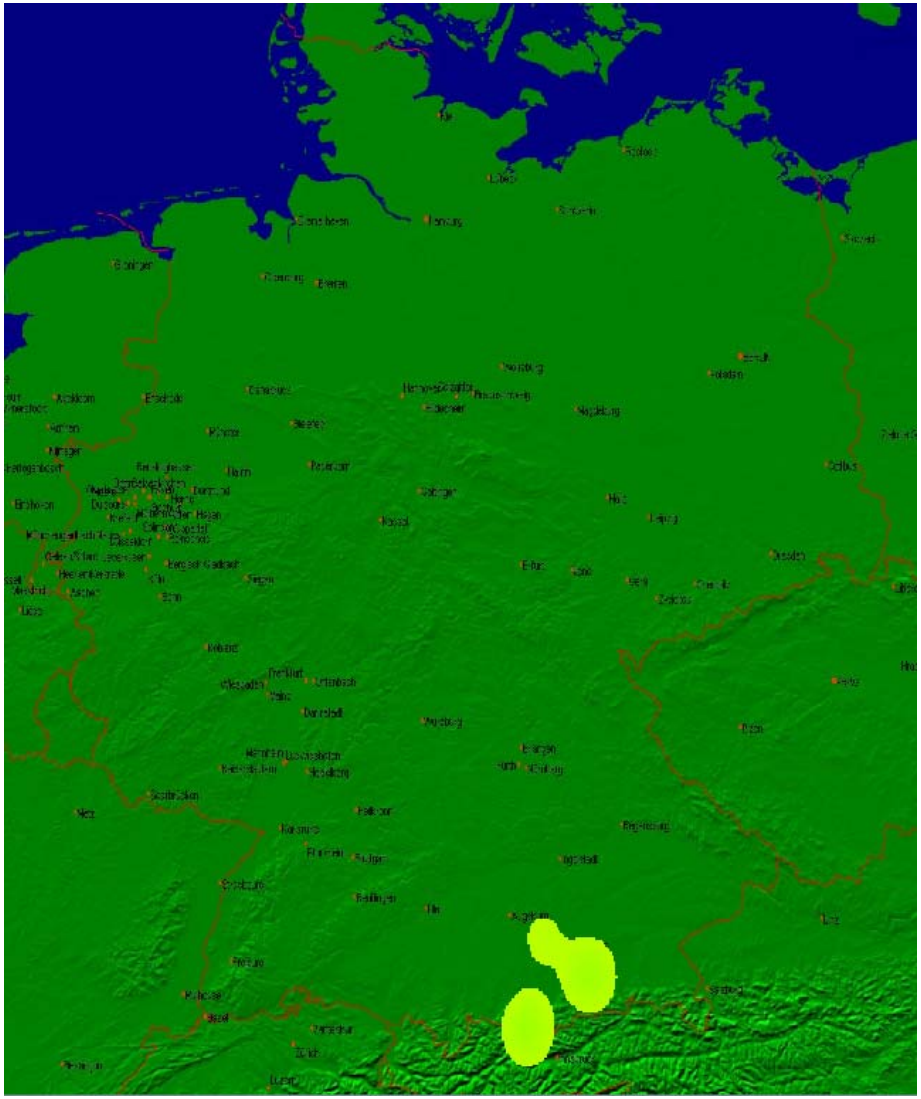
Evaluation: Qualität der Repräsentationen

- „Weißwurstäquator“

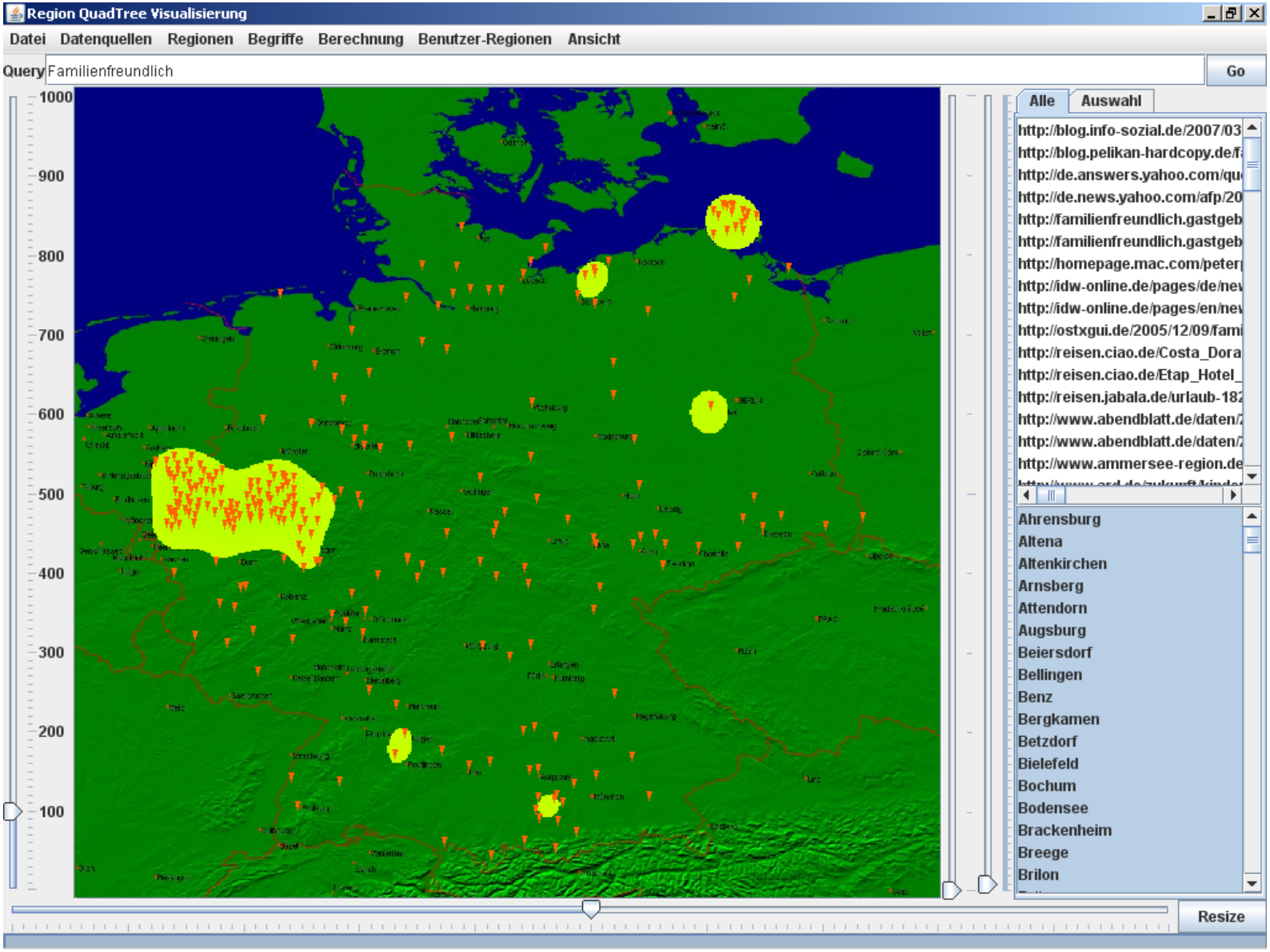


Evaluation: Qualität der Repräsentationen

- Olympische Spiele 1972



Demo


Region QuadTree Visualisierung
Datei
Datenquellen
Regionen
Begriffe
Berechnung
Benutzer-Regionen
Ansicht
Query
Familienfreundlich
Go
1000
900
800
700
600
500
400
300
200
100
Alle
Auswahl
http://blog.info-sozial.de/2007/03
http://blog.pelikan-hardcopy.de/f
http://de.answers.yahoo.com/qu
http://de.news.yahoo.com/afp/20
http://familienfreundlich.gastgeb
http://familienfreundlich.gastgeb
http://homepage.mac.com/peter
http://idw-online.de/pages/de/ne
http://idw-online.de/pages/en/ne
http://ostxgui.de/2005/12/09/fam
http://reisen.ciao.de/Costa_Dora
http://reisen.ciao.de/Etap_Hotel_
http://reisen.jabala.de/urlaub-18
http://www.abendblatt.de/daten/
http://www.abendblatt.de/daten/
http://www.ammersee-region.de
Ahrensburg
Altena
Altenkirchen
Arnsberg
Attendorn
Augsburg
Beiersdorf
Bellingen
Benz
Bergkamen
Betzdorf
Bielefeld
Bochum
Bodensee
Brackenheim
Breege
Brilon
Resize

Zusammenfassung

- Integration eines Ansatzes zur Generierung geografischer Begriffsrepräsentationen zur Anfragezeit
- Angemessene Performance nach der Optimierung
- Brauchbare Ergebnisqualität zur Bildung eines Rankings
- Allgemein anwendbar auf beliebige Begriffe;
  nicht auf tatsächliche geografische Regionen beschränkt
- Nicht sehr nützlich bei Begriffen ohne geografische Korrelation (breit verteilt ohne klare Cluster)

Diskussion

Vielen Dank für Ihre „Aufmerksamkeit“!