

Geographisch Fokussierte Websuche

Dirk Ahlers

OFFIS, Oldenburg

Susanne Boll

Universität Oldenburg

Stadt der Wissenschaft 2009
Übermorgenstadt
OLDENBURG



08.-09.05.2008

GI-Fachgruppentreffen

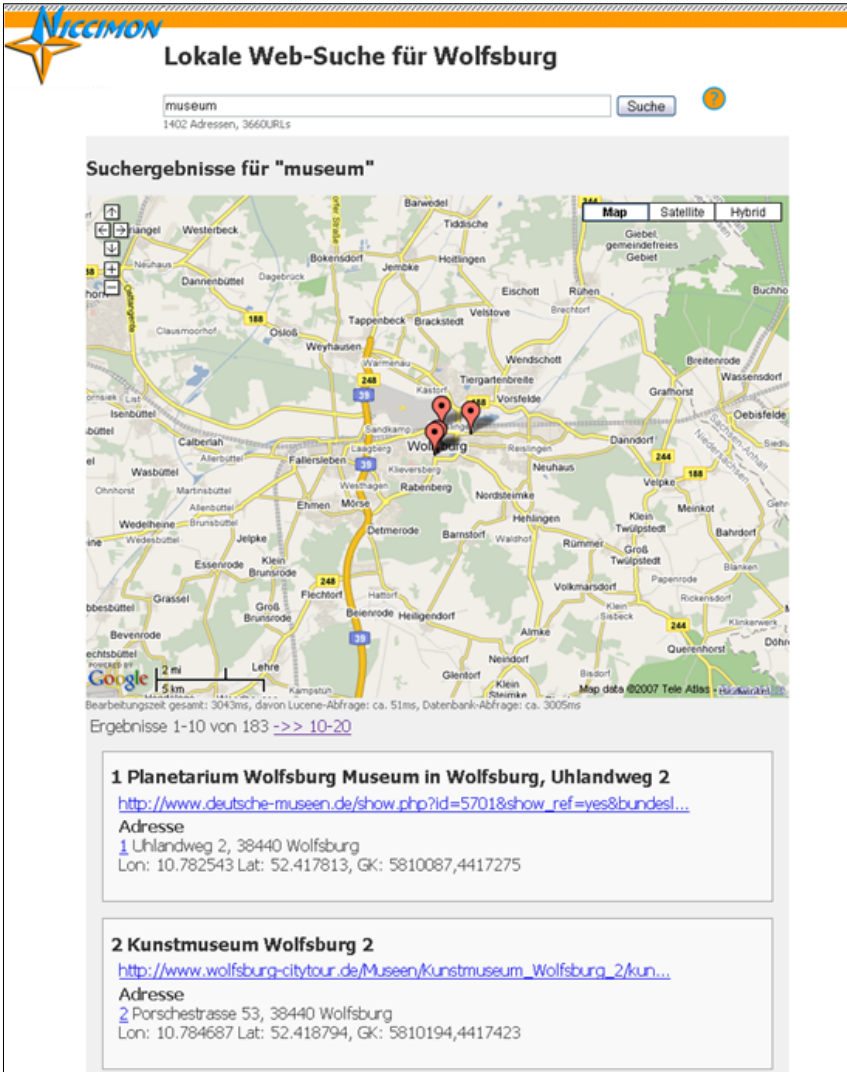
Bamberg

49° 53' 29.62" N 10° 53' 7.51" E



- ▶ Motivation
- ▶ Geographische Websuche
- ▶ Geographisch fokussiertes Crawling
- ▶ Evaluation
- ▶ Fazit

- ▶ Navigation im Web benötigt leistungsfähige Suchmaschinen
 - Stichwortsuche
- ▶ Ortsbasierte Suche
 - Einschränkung des Suchraumes
 - Semantische Erschließung des Ortsbezuges
 - Auffinden, Extraktion, Aufbereitung von Ortsinformationen
 - Anwendungsfall: Information für mobile Benutzer



VICIMON Lokale Web-Suche für Wolfsburg

museum 1402 Adressen, 3660URLs

Suchergebnisse für "museum"

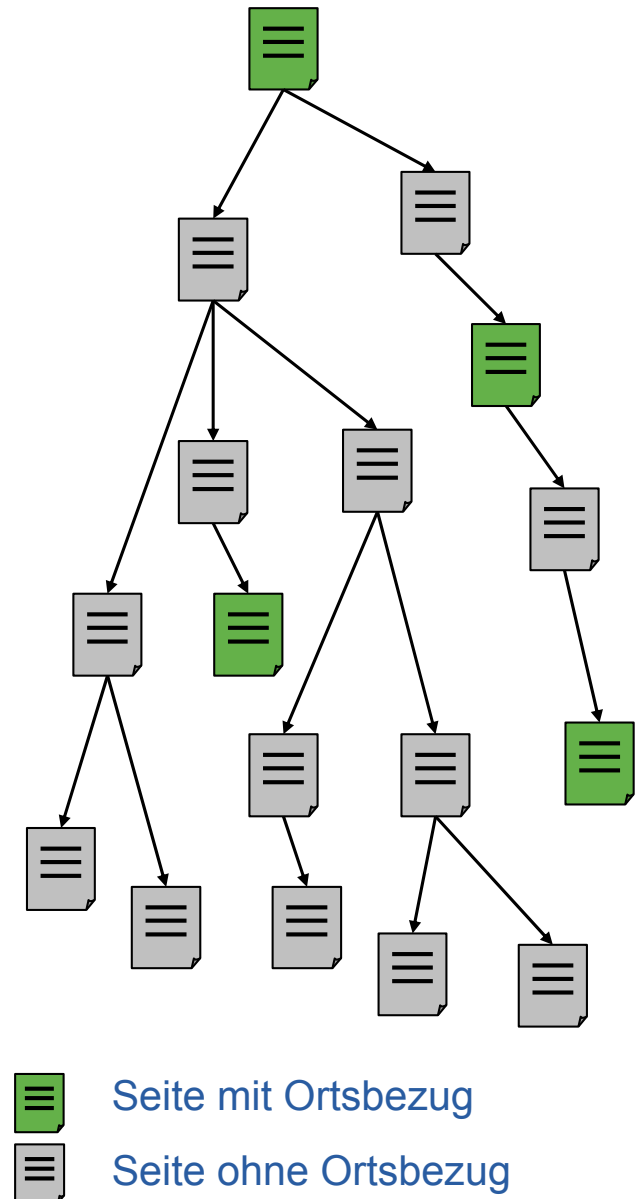
Map Satellite Hybrid

Ergebnisse 1-10 von 183 -> > 10-20

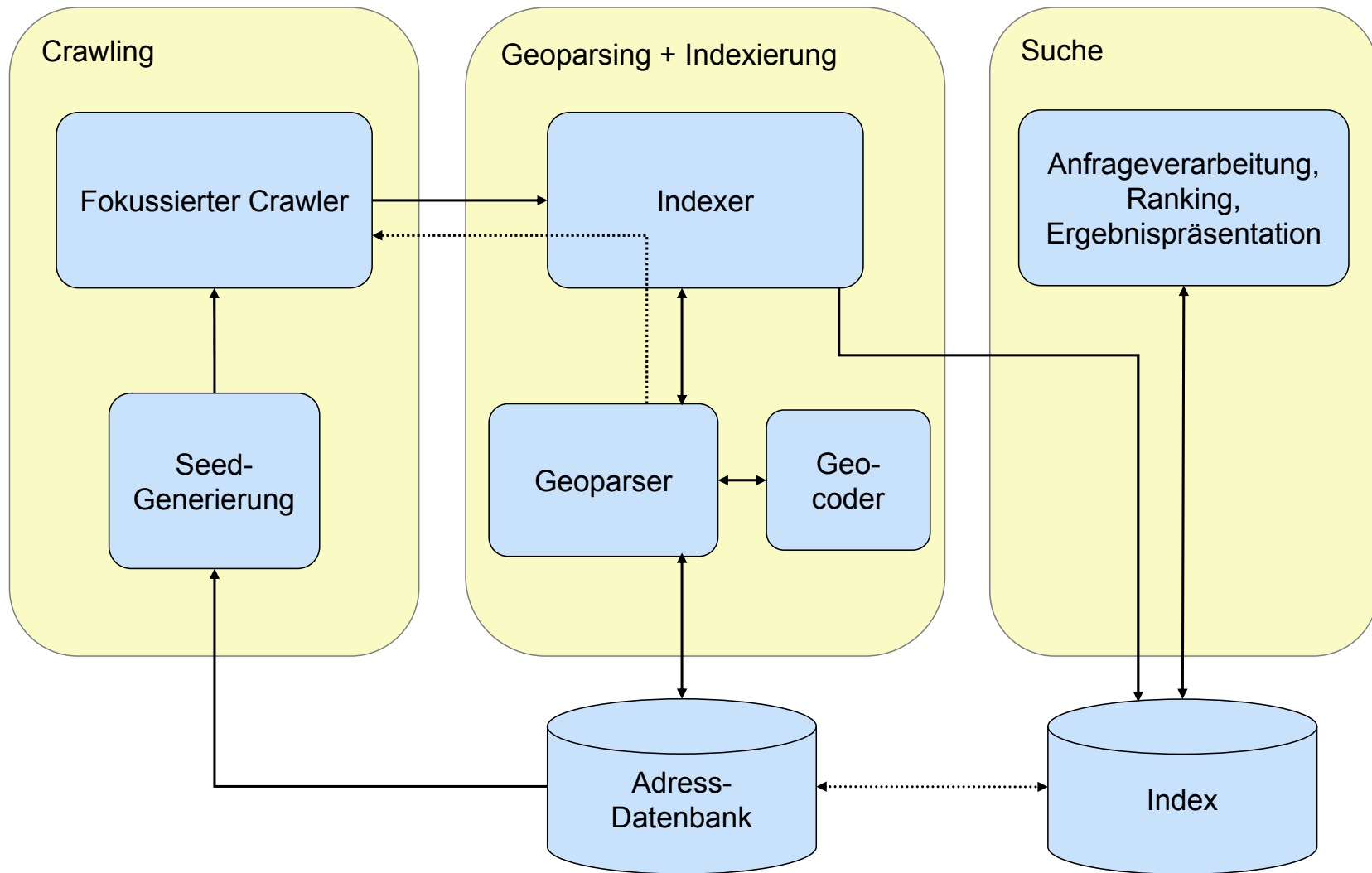
1 Planetarium Wolfsburg Museum in Wolfsburg, Umlandweg 2
http://www.deutsche-museen.de/show.php?id=5701&show_ref=yes&bundesl...
Adresse
1 Umlandweg 2, 38440 Wolfsburg
Lon: 10.782543 Lat: 52.417813, GK: 5810087,4417275

2 Kunstmuseum Wolfsburg 2
http://www.wolfsburg-citytour.de/Museen/Kunstmuseum_Wolfsburg_2_kun...
Adresse
2 Porschestraße 53, 38440 Wolfsburg
Lon: 10.784687 Lat: 52.418794, GK: 5810194,4417423

- ▶ Das Web beinhaltet eine Vielzahl von ortsbezogenen Informationen
- ▶ Ortsinformationen im Web sind ...
 - Verstreut
 - Nicht indiziert
 - Versteckt in Webseiten
 - Nicht explizit strukturiert
- ▶ Effizientes Auffinden und Aufbereitung dieser Informationen ist notwendig
- ➔ Anwendbarkeit von Fokussiertem Crawling

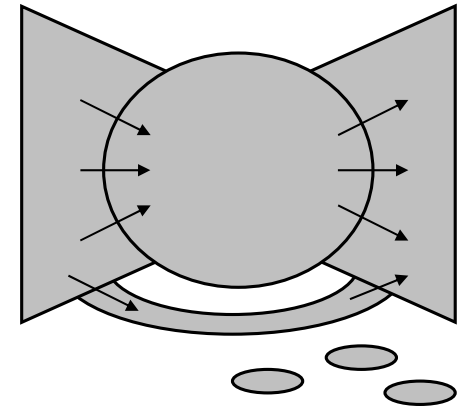


Komponenten einer Geo-Suchmaschine

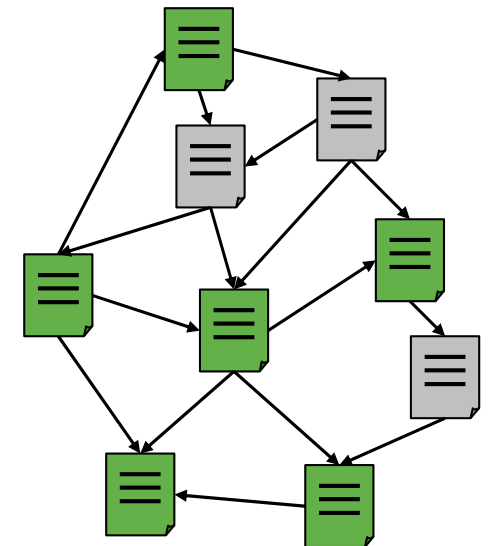


Datengrundlage: Struktur des Web

- ▶ Webcrawler erschließen das Web entlang von Hyperlinks
- ▶ Analyse des Web-Linkgraphen [BKM+00]
 - “Bow-tie structure“
 - Kein zusammenhängender Graph
 - Kleine-Welt-Phänomen
- ▶ Aber:
- ▶ *Themen* sind stärker verlinkt [CBD99]
 - Thematischer Zusammenhang kann durch einen Crawler ausgenutzt werden
 - Verbindungen teilweise indirekt
 - Nichtrelevante Seiten auf einem Linkpfad

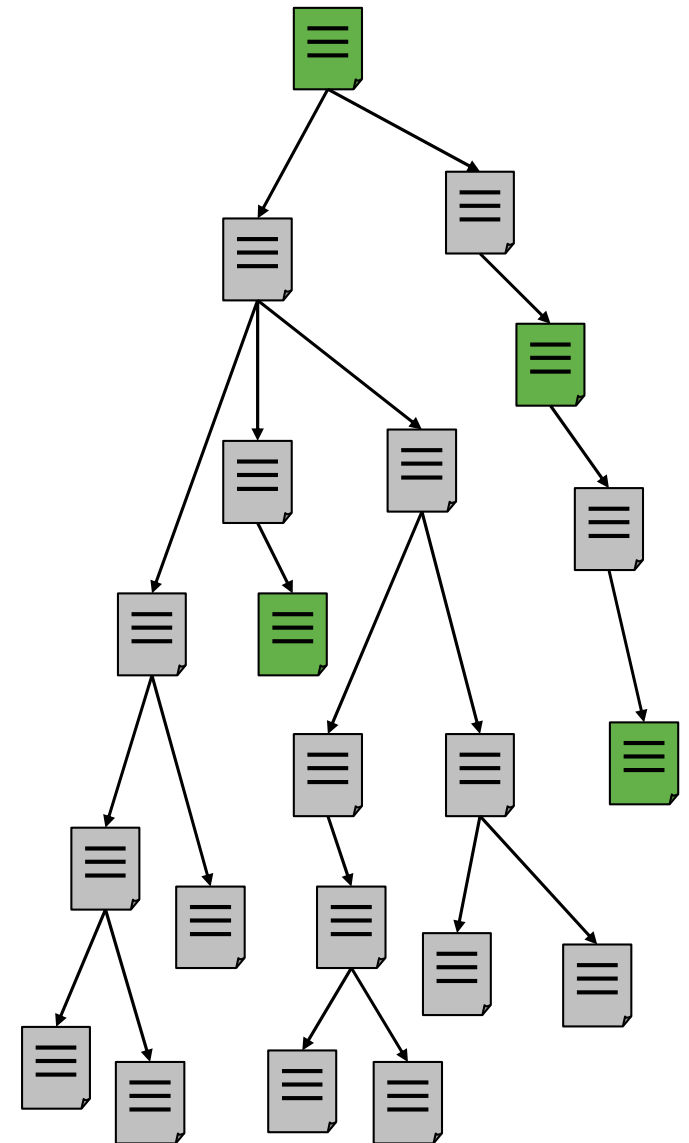


Link-Struktur des WWW



Verlinkte thematische Seiten

- ▶ Webcrawler zum effizienten Erschließen eines Themas
- ▶ Einschränkung des Crawls auf nur einen Teil des Web
- ▶ Einsatz von Heuristiken für hohe Ausbeute relevanter Seiten
- ▶ Ausnutzen des Linkgraphen
- ▶ Klassifikation besuchter Seiten
- ▶ Rückkopplung zum Crawler



Geographisch Fokussiertes Crawlen

Landesbibliothek Oldenburg

Unser Service | Online-Katalog | Datenbanken | Region Nordwest | Veranstaltungen | Über die LBO | A bis Z | Kontakt

Aktuelles:

- Veranstaltungskalender
- Ausstellungen
- Pressemitteilungen

Tipp:

- Oldenburgische Bibliographie
- Verzeichnis niederdeutscher Autorinnen und Autoren
- Neue Oldenburg-Publikationen
- Aktion "Buch in Not"
- Schulangebote
- Oldenburgische Bibliotheksgesellschaft

Suche:

- Webseiten der LBO durchsuchen

Landesbibliothek
Pferdemarkt 15
26121 Oldenburg
Tel. 0441-799-2600
Fax. 0441-799-2865
lbo@lb-oldenburg.de

Postanschrift:
Postfach 3480
26024 Oldenburg

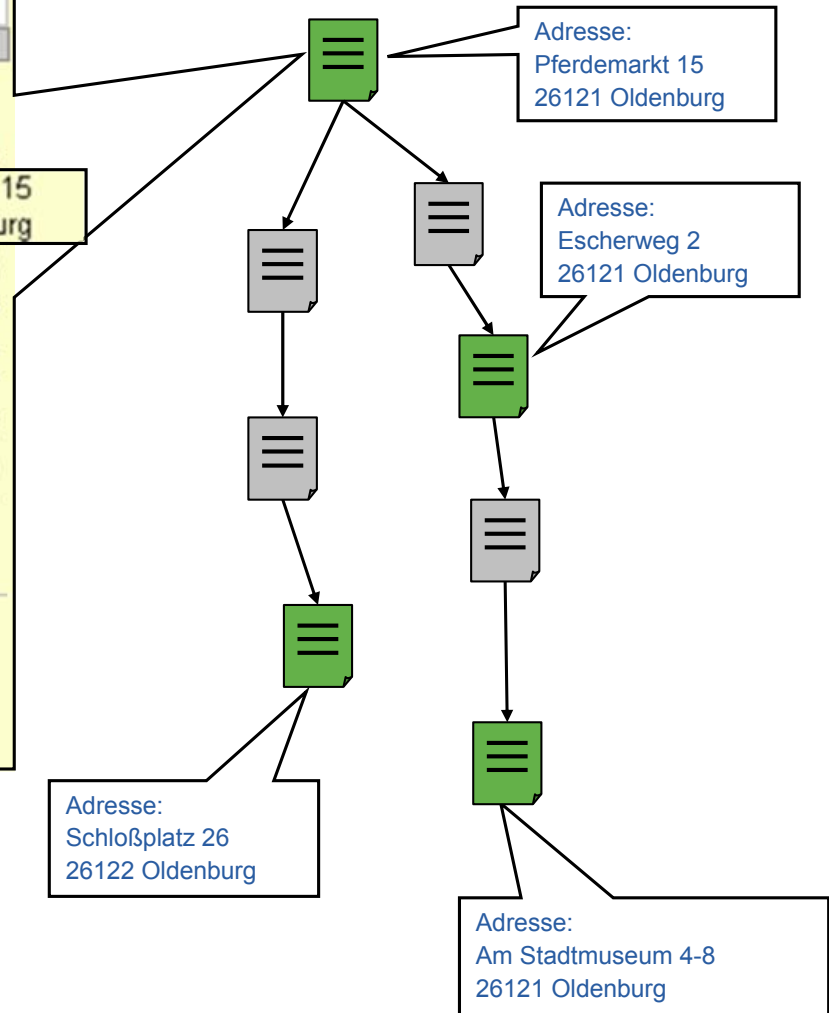
Öffnungszeiten:
Mo-Mi 10-18 Uhr
Do 10-19 Uhr
Fr 10-17 Uhr
Sa 9-12 Uhr

Willkommen in der Regionalbibliothek des Nordwestens

Impressum @ 2003

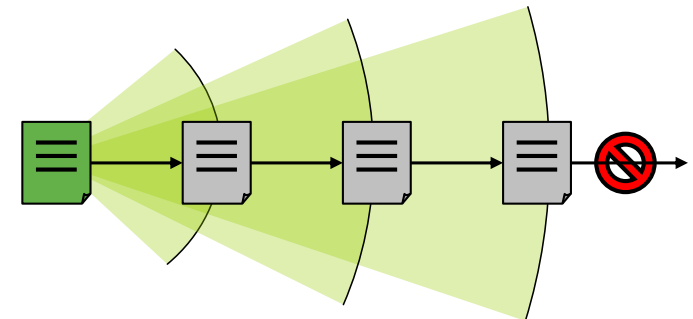
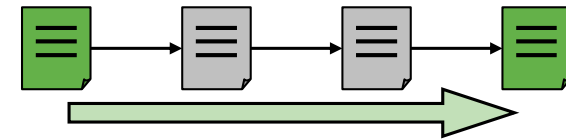
W3C HTML 4.01 ✓ W3C CSS ✓

Portal Niedersachsen

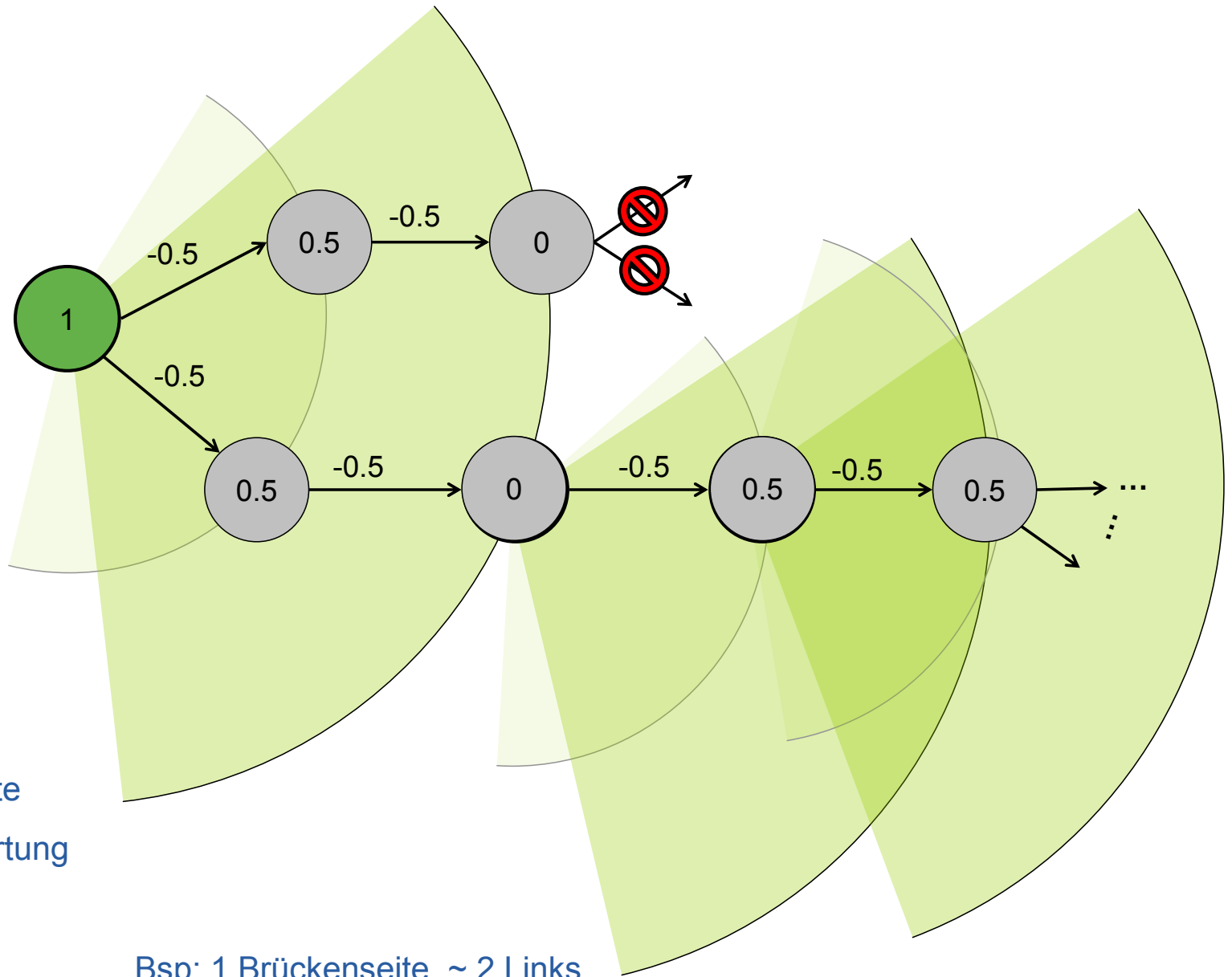


Einbeziehung von Brückenseiten

- ▶ Adaptierte Breitensuche
- ▶ Distanzbewertung von Seiten
- ▶ Erreichen indirekt verlinkter Seiten
 - Nicht-relevante Seiten auf einem Linkpfad werden erfasst
- ▶ Ausschluss von erfolglosen Zweigen
 - Definition einer Suchtiefe um relevante Seiten
 - Dynamisches Beschneiden

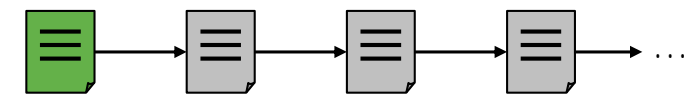


Brückenseiten: Beispiel

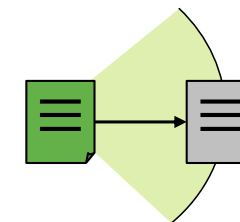


Bsp: 1 Brückenseite ~ 2 Links

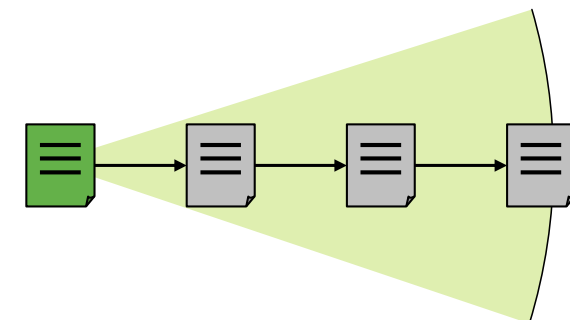
- ▶ Vergleich geographisch fokussiertes gegen unfokussiertes Crawl
 - Basiscrawl
- ▶ Vergleich verschiedener Parameter für fokussiertes Crawl
 - Zahl zulässiger Brückenseiten
- ▶ Crawls
 - Basiscrawl & 0 bis 6 Brückenseiten
 - Crawls über 24 Stunden
 - Seeds aus DMOZ für Oldenburg



Basiscrawl

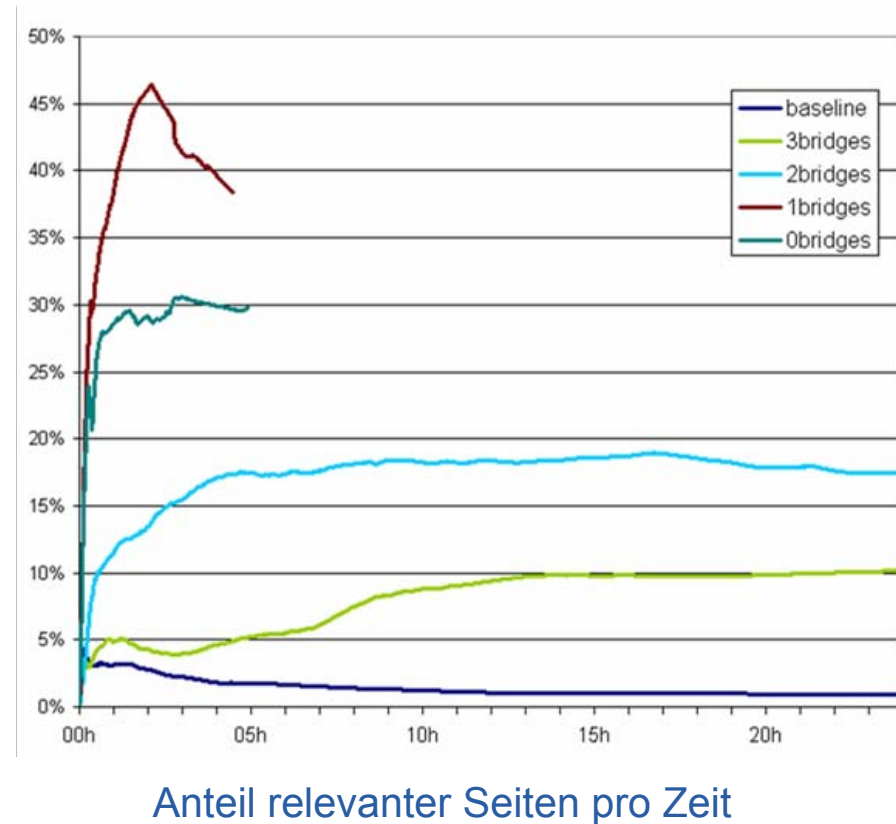


0 Brückenseiten



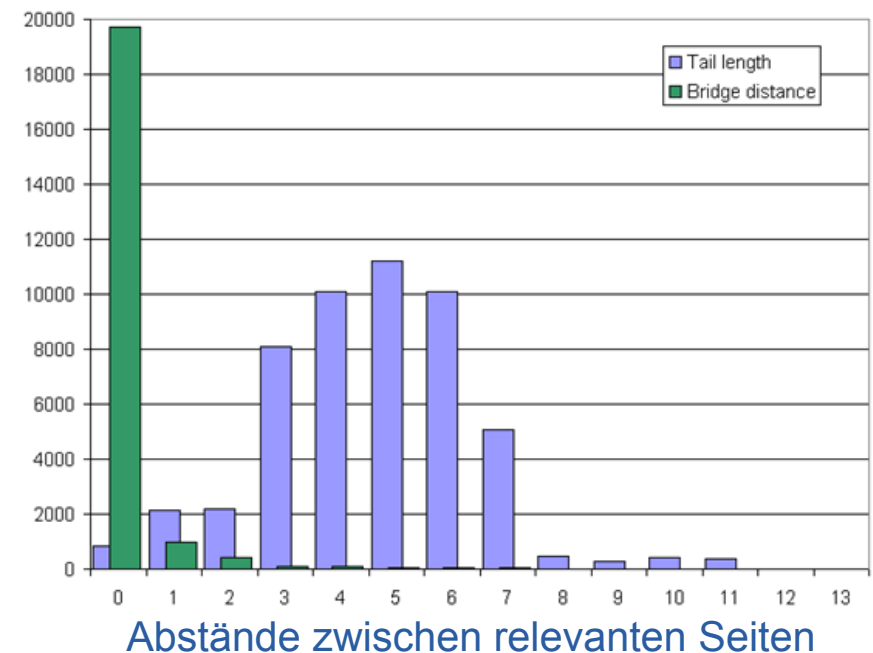
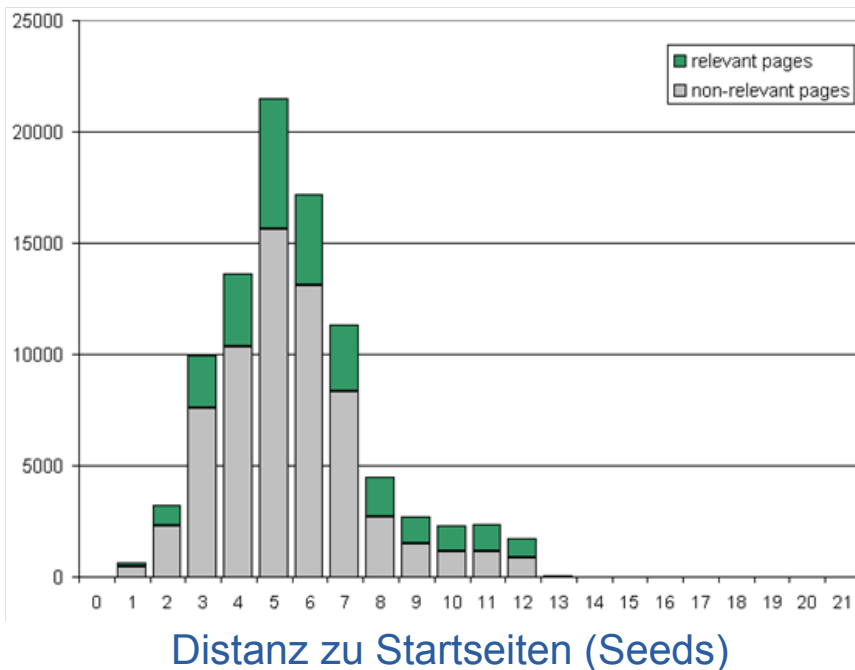
2 Brückenseiten

- ▶ Unfokussierter Crawl erfasst viele nicht-relevante Seiten, ebenso Brückenseiten > 3
- ▶ Höchstwerte > 30% nicht dauerhaft
- ▶ Ergebnisanteil ~10–20% für 2 & 3
- ▶ Stabiles Langzeit-Verhalten
- ▶ Parametrierung der Brückenseiten führt zu unterscheidbarem Verhalten und Steuerung des Ergebnisses
- ▶ In diesem Fall: Optimum bei 2 Brückenseiten

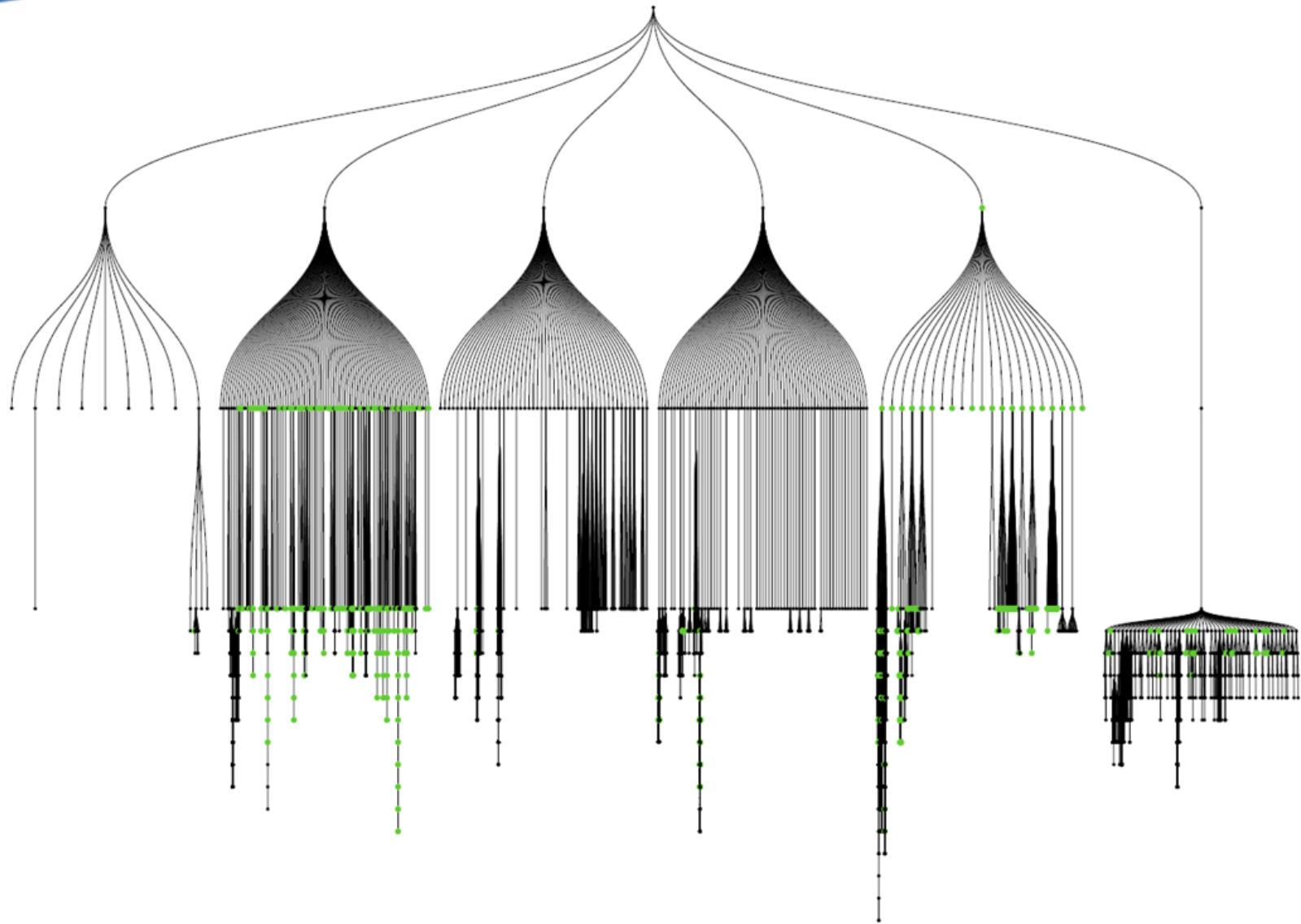


- ▶ Distanz zu Startseiten wenig aussagekräftig
- ▶ Relevante Seiten sind anteilig über alle Abstände verteilt

- ▶ Relevante Seiten sind stark verlinkt
- ▶ Deutliche Häufung bei kleinen Distanzen → enge Verlinkung

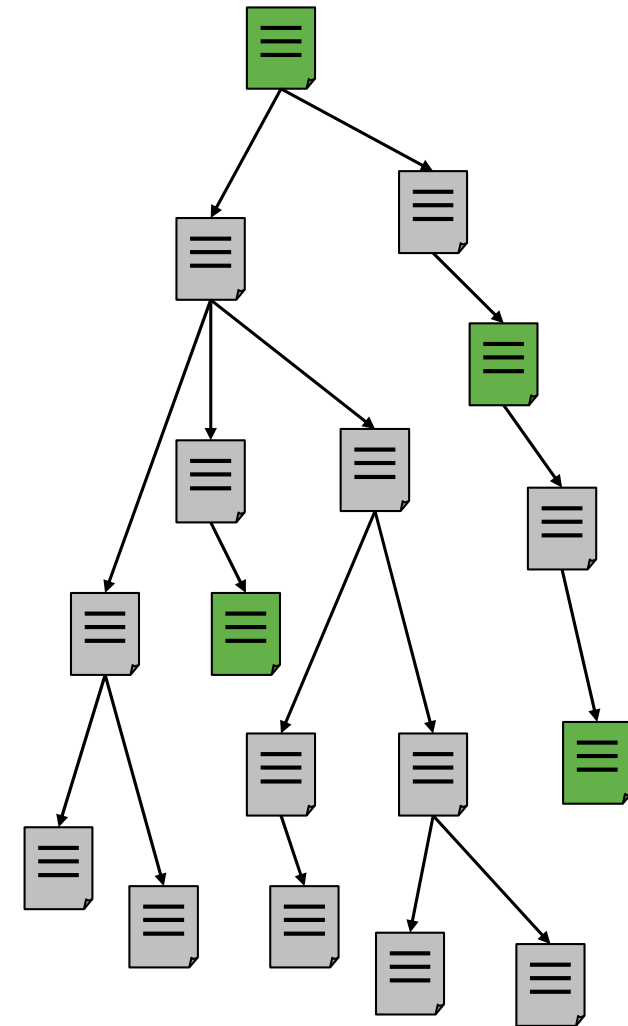


Verteilung des Ortsbezugs



Ausschnitt eines Crawlbaumes: Tiefe und Distanzen relevanter Seiten

- ▶ Geographisch fokussiertes Crawlern ermöglicht effizienten Indexaufbau
- ▶ Ortsbezug der Webseiten und Verteilung ist ausreichend
- ▶ Gutes Auffinden relevanter Seiten
 - Unfokussiert ~2%, fokussiert ~10–20%
- ▶ In Zukunft stärkere Dynamisierung des Crawlers
 - Priorisierung
 - Linkvorhersage



A satellite-style map of a coastal region, likely in the North Sea area. A river flows from the top right towards the bottom left, emptying into a bay. The land is green and brown, and the water is blue. The text 'Q&A' is overlaid in the center.

Q&A

Dirk Ahlers

OFFIS Oldenburg

53° 8' 55.9" N 8° 12' 0.43" E

ahlers@offis.de

- ▶ [AB07a] *Location-based Web Search*
- ▶ D. Ahlers, S. Boll. In: *The Geospatial Web. How Geo-Browsers, Social Software and the Web 2.0 are Shaping the Network Society*. Arno Scharl, Klaus Tochtermann (Eds.). Springer, 2007.
- ▶ [AB07b] *Geospatially Focused Web Crawling*
- ▶ D. Ahlers, S. Boll. *Datenbank-Spektrum* 7(23): 3-12, Themenheft "Fokussierte Suche", 2007.
- ▶ [AB08] *Urban Web Crawling*
- ▶ D. Ahlers, S. Boll. *Proceedings of the First International Workshop on Location and the Web (LocWeb 2008)*, Workshop auf der WWW 2008, Peking, China, 2008
- ▶ [BA08] *A Web more Geospatial: Insights into the Location Inside*
- ▶ S. Boll and D. Ahlers. *Understanding Web Evolution: A Prerequisite for Web Science*, Workshop auf der WWW 2008, Peking, China, 2008
- ▶ [BKM+00] *Graph structure in the Web*
- ▶ A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. *Computer Networks*, 33(1):309–320, 2000.
- ▶ [CBD99] *Focused Crawling: A New Approach to Topic-specific Web Resource Discovery*.
- ▶ S. Chakrabarti, M. van den Berg, B. Dom. *Computer Networks*, 31(11–16): 1623–1640. 1999
- ▶ [GHLS06] *Geografisches Information Retrieval*
- ▶ J. Gräf, A. Henrich, V. Lüdecke, C. Schlieder. *Datenbank-Spektrum*, 6(18): 48–56. 2006
- ▶ [MCS+05] *Design and Implementation of a Geographic Search Engine*
- ▶ A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, B. Seeger, *WebDB2005*