Advanced Topics in Databases

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Hasso-Plattner-Institut Potsdam
Fachgebiet Informationssysteme
Markus Güntert

WS 2008/2009

The Anatomy of a Large-Scale Hypertextual Web Search Engine

- Sergey Brin, Lawrence Page
- Computer Science Department, Stanford University
- 1998 veröffentlicht
- *“an in-depth description […] – the first such detailed public description we know of to date”*

Agenda

- Autoren
- Ziele von Google
- Standpunkt 1998
- PageRank
- weitere Hypertext-Informationen
- Architektur (vereinfacht)
- Suchanfragen
- Future Work
- Fazit

Autoren

- **Sergey Brin**

  - * 1973 in Moskau, 1979 Migration in die USA
  - Bachelor of Science in Mathematik u. Informatik (University of Maryland, College Park), Masterstudium in Stanford
  - Promotion bis heute nicht fertig gestellt

- **Lawrence Page**

  - * 1973 in Michigan
  - Bachelor of Science in Computer Engineering (University of Michigan), Masterstudium in Stanford
  - *“The ultimate search engine would understand exactly what you mean and give back exactly what you want.”*

Ziele von Google

- Google ist Prototyp einer “large-scale search engine”
  - Crawling, Indexing, Sorting
  - Konzentration auf Qualität von Suchergebnissen
  - Berücksichtigung des Wachsens des Webs sowie des technologischen Fortschrittes
  - wissenschaftliche Arbeit im Bereich der Suchmaschinen
  - Zugänglichkeit für Jedermann

  - ~ 100 000 000 indizierte Webseiten
    - 2008: ~ 1 000 000 000 000 [1]
  - ~ 10 000 000 Anfragen pro Tag
    - 2006: ~ 91 000 000 [2]

[1] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
[2] http://searchenginewatch.com/2156461

Standpunkt 1998

- “human-maintained” Seiten wie *Yahoo!*
  - subjektiv, teuer, unvollständig, langsam
- reines Keyword-Matching liefert viele schlechte Ergebnisse
  - kein (ausgereiftes) Ranking der Ergebnisse
  - Indizes werden größer
  - Aufnahmevermögen des Nutzers bleibt gleich
- Manipulation von Suchmaschinen
  - “Junk results”


PageRank

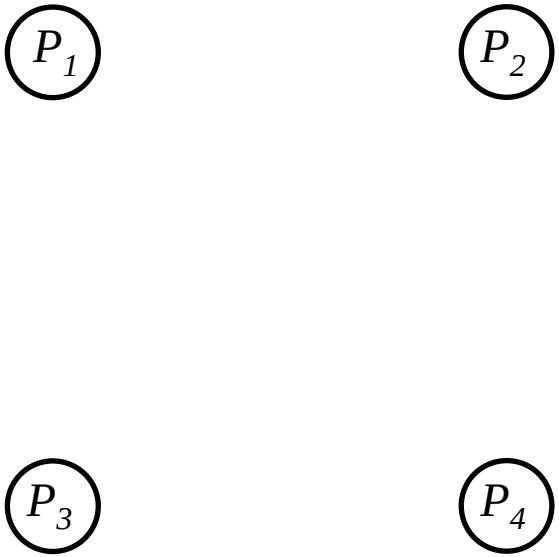
PageRank

- Ziel: Priorisierung von Suchergebnissen

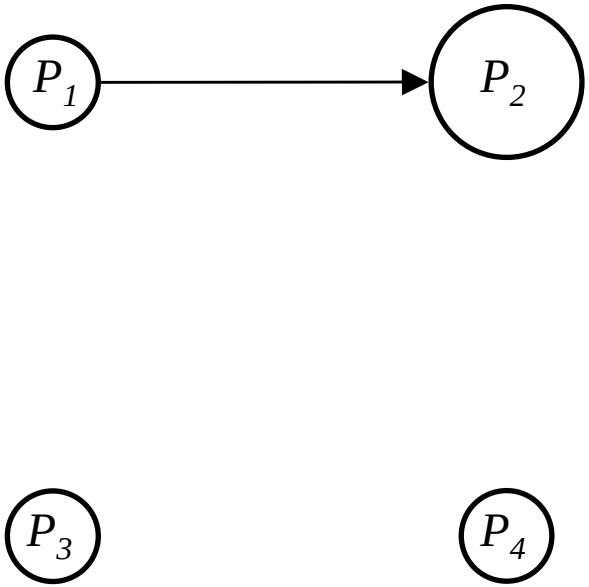
- Menge verlinkter Dokumente anhand Struktur bewerten und gewichten
  - Seite gilt als wichtig, wenn viele Seiten auf sie verlinken
  - Seite gilt als wichtig, wenn wichtige Seiten auf sie verlinken

- Prinzip des Zitats für Web adaptiert
  - subjektives Verständnis von Wichtigkeit kommt dem nahe

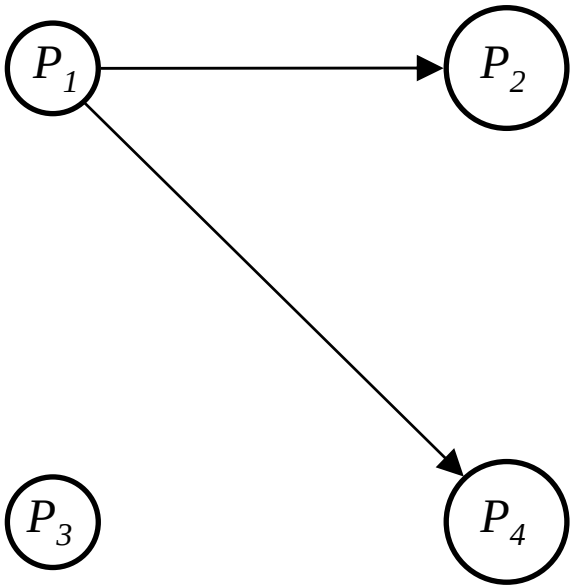
PageRank – Beispiel



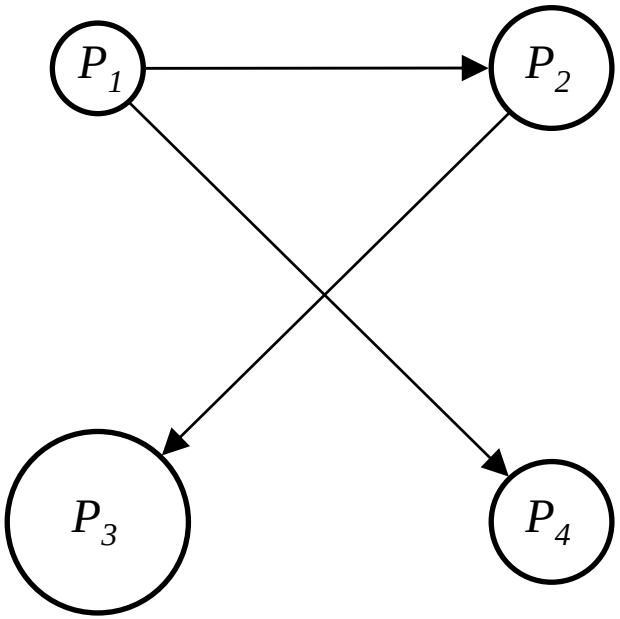
PageRank – Beispiel



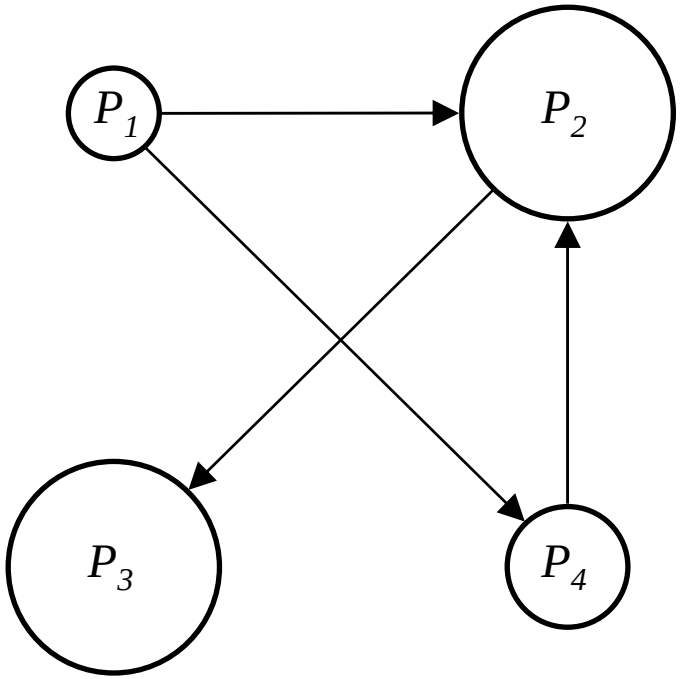
PageRank – Beispiel



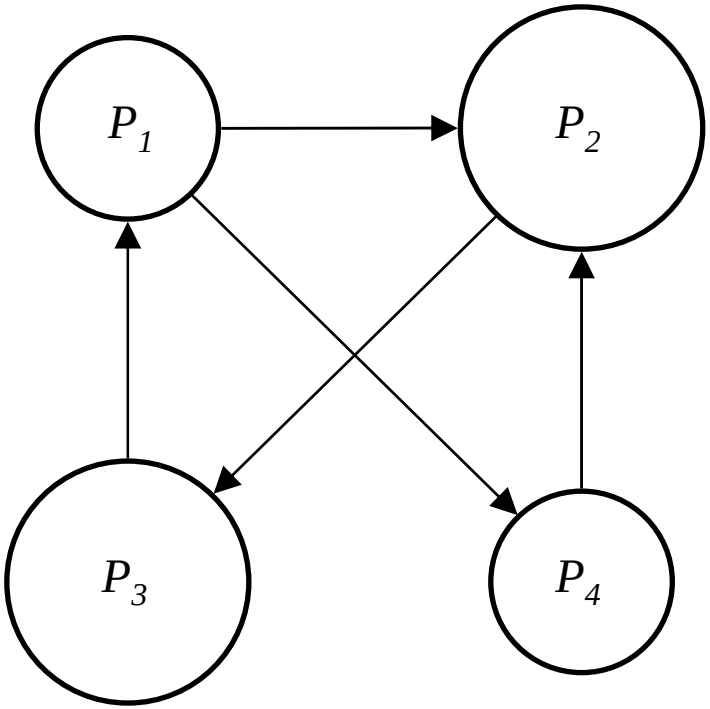
PageRank – Beispiel



PageRank – Beispiel



PageRank – Beispiel

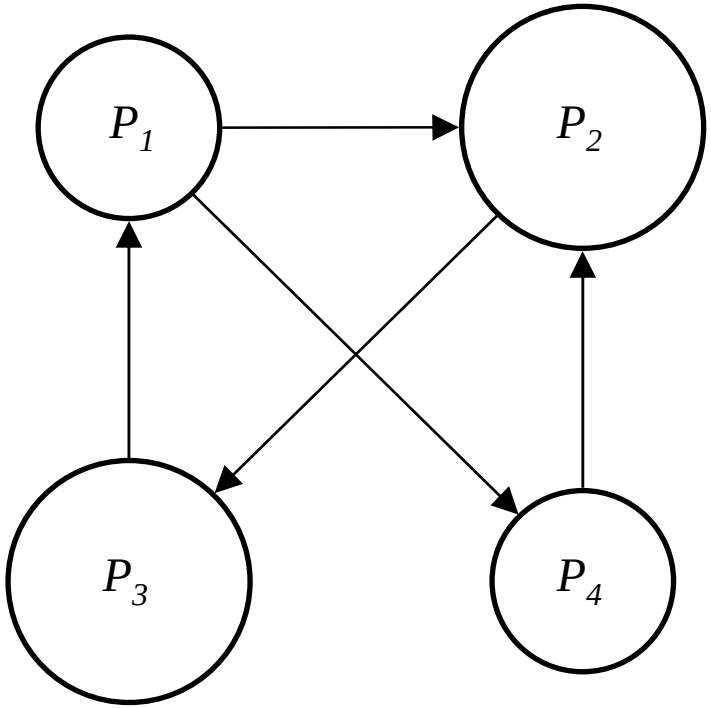


PageRank – Berechnung

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

- Summe aller PageRanks von verlinkenden Seiten P_j
  - normalisiert
- rekursiv!

PageRank – Berechnung



- iterativer Ansatz

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

Iteration 0	Iteration 1	Iteration 2	Rank nach 100 Iterationen
$r_0(P_1) = \frac{1}{4}$	$r_1(P_1) = \frac{1}{4}$	$r_2(P_1) = \frac{1}{4}$	3
$r_0(P_2) = \frac{1}{4}$	$r_1(P_2) = \frac{3}{8}$	$r_2(P_2) = \frac{1}{4}$	1
$r_0(P_3) = \frac{1}{4}$	$r_1(P_3) = \frac{1}{4}$	$r_2(P_3) = \frac{3}{8}$	4
$r_0(P_4) = \frac{1}{4}$	$r_1(P_4) = \frac{1}{8}$	$r_2(P_4) = \frac{1}{8}$	2

PageRank – intuitiv

- Nachahmung eines zufällig durch das Netz surfenden Users
  - zufällige Startseite
  - Klicken von Links (ohne jemals zurück zu gehen)
  - Anfordern einer neuen zufälligen Startseite möglich
  - Wahrscheinlichkeit für Finden einer Seite entspricht deren PageRank

weitere Hypertext-Informationen

- Anchor Text `<a href="http://www.google.de">Large-Scale Hypertextual Web Search Engine</a>`
  - teilweise präzisere Beschreibungen von Webseiten
  - nicht-indizierbare Inhalte (Bilder, Audio, Video,…)

- Visual Presentation Details
  - (relative) Schriftgröße
  - fett, kursiv etc.

DAS ist wichtiger als das.

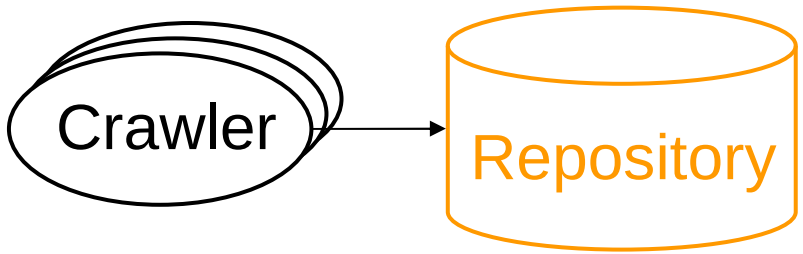
- Location Information
  - URL, <title>, <meta>
  - Proximität

Architektur
(vereinfacht)

Architektur

- automatisiertes Durchsuchen und Analysieren von Webseiten

Architektur



- Speicherung der gecrawlten Webseiten
  - in komprimierter Form
- jede Webseite erhält eindeutige *docID*

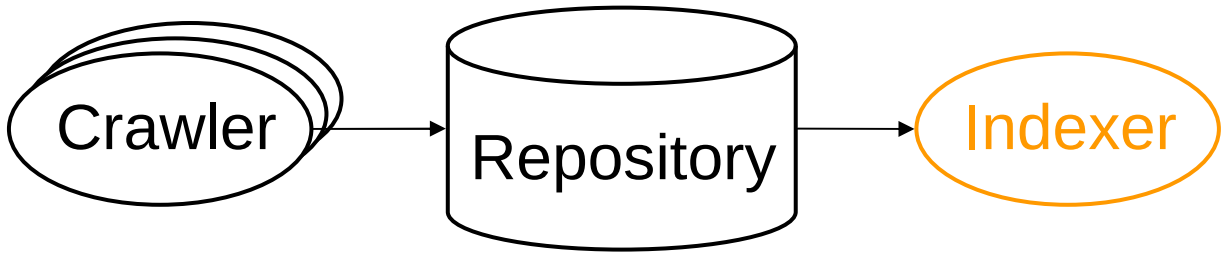
Repository

docID	
1234	**The Anatomy of a Search Engine** In this paper we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying results that existing systems …
.	⋮
98765	**Google Search – Wikipedia** Google search is a Web search engine owned by Google Inc. and is the most used search engine on the Web. Google receives several hundred million queries each day through its various services …

http://infolab.stanford.edu/~backrub/google.html

http://en.wikipedia.org/wiki/Google_search

Architektur



- Parsen der Seiten im Repository
- für jedes Dokument wird eine 'Hit List' von Wörtern angelegt, …
  - *wordID*

Hits

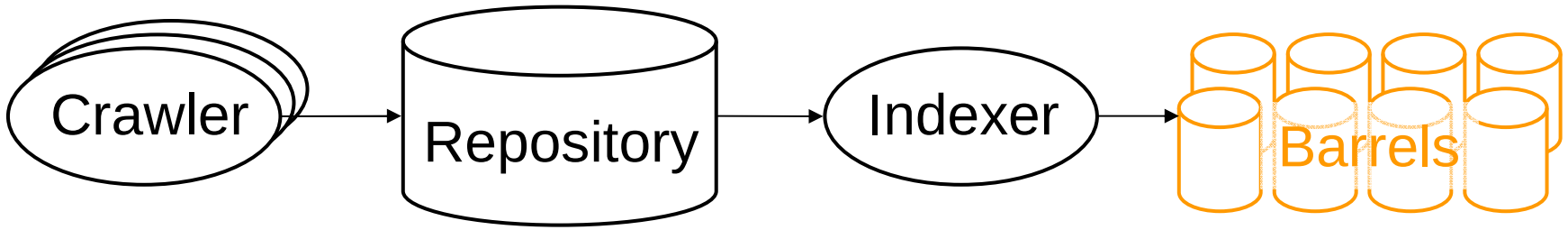
- jedes Wort, das in einem Dokument vorkommt, ist ein Hit

- plain Hits
- fancy Hits
  - URL, <title>, <meta>, Anchor Text

Hit: 2 bytes

plain:	cap:1	imp:3	position: 12	
fancy:	cap:1	imp = 7	type: 4	position: 8

Architektur

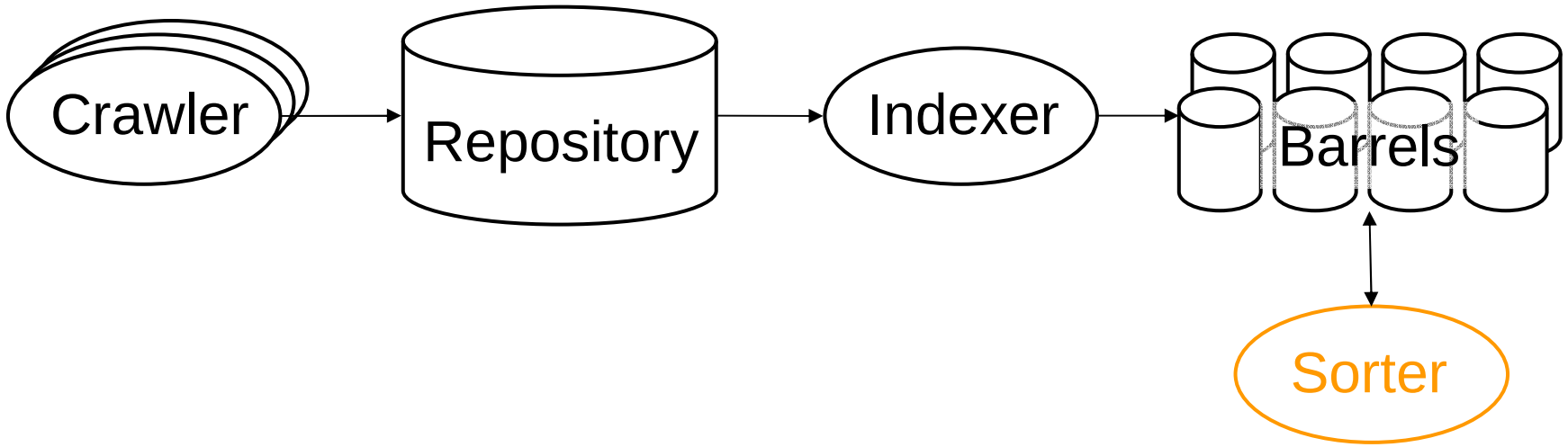


- … diese werden als *Forward Index* gespeichert
  - sortiert nach *docID*

Forward Index

docID	*wordID*	Hit
1234	anatomy	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	search	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	engine	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	Google	plain: cap:1 | imp:3 | position: 12
	Web	plain: cap:1 | imp:3 | position: 12
⋮	⋮	⋮
98765	Google	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	search	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	Google	plain: cap:1 | imp:3 | position: 12
	Web	plain: cap:1 | imp:3 | position: 12
	engine	plain: cap:1 | imp:3 | position: 12

Architektur

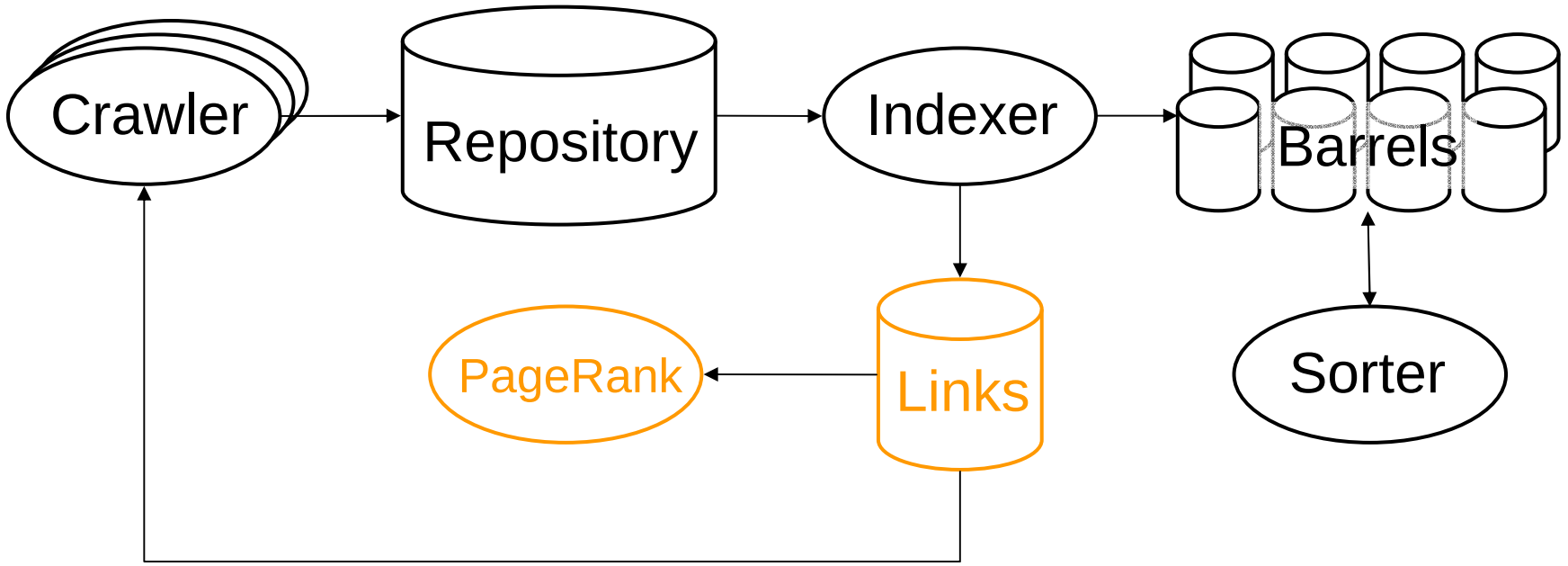


- Reorganisation zu *Inverted Index*
  - sortiert nach *wordID*

Inverted Index

wordID	*docID*	Hit
anatomy	1234	fancy: cap:1 | imp = 7 | type: 4 | position: 8
search	1234	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	98765	fancy: cap:1 | imp = 7 | type: 4 | position: 8
engine	1234	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	98765	plain: cap:1 | imp:3 | position: 12
Google	98765	fancy: cap:1 | imp = 7 | type: 4 | position: 8
	98765	plain: cap:1 | imp:3 | position: 12
	1234	plain: cap:1 | imp:3 | position: 12
Web	1234	plain: cap:1 | imp:3 | position: 12
	98765	plain: cap:1 | imp:3 | position: 12

Architektur

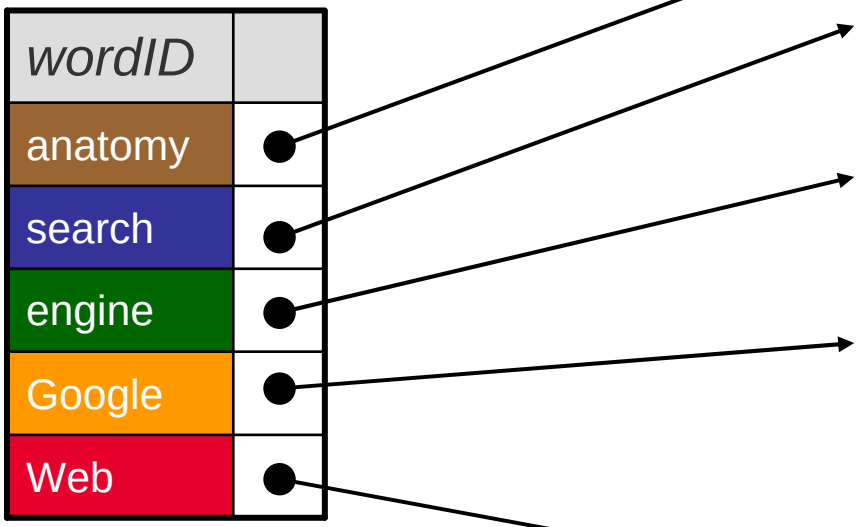


- Auswertung der Link-Struktur
  - Berechnung des PageRanks aller Webseiten
  - neue Links crawlen

Suchanfragen

- *One-Word-Query*
  - Suchen von Hit List im Inverted Index
    - Lexicon (in-memory)

Lexicon



wordID	*docID*	Hit
anatomy	1234	fancy: ...
search	1234	fancy: ...
	98765	fancy: ...
engine	1234	fancy: ...
	98765	plain: ...
Google	98765	fancy: ...
	98765	plain: ...
	1234	plain: ...
Web	1234	plain: ...
	98765	plain: ...

Lexicon (in-memory)

Inverted Index

Suchanfragen

- *One-Word-Query*
  - Suchen von Hit List im Inverted Index
    - Lexicon (in-memory)
  - Gewichtung
    - Informationen in den Hit Listen
    - Anzahl der Hits pro Dokument
    - PageRank

- *Multi-Word-Query*
  - zusätzlich Betrachtung von Proximität
    - besser, wenn Wörter nahe beieinander sind

Future Work

- Query Caching
- Kontrolliertes Re-Crawling
  - statistische Informationen über Updates von Webseiten
- Stemming
- Boolsche Operatoren in Suchanfragen
- Zusammenfassung der Ergebnisse

- 2000
  - 1 000 000 000 Webseiten indiziert
  - Internationalisierung
  - Ads
- 2004
  - Börsengang

Fazit

- Primärziel sind hochqualitative Suchergebnisse im immer größer werdenden Web
- Priorisierung der Suchergebnisse durch Auswertung von Hypertext-Informationen
  - PageRank
  - Anchor Text
  - etc.
- Architektur von vornherein auf Skalierbarkeit ausgelegt