Google

Die Suche nach der Stecknadel im weltweiten Heuhaufen

Heinz Gnehm, 18. Januar 2006

Klassifikation von Information

- Präkoordination
  - Dezimalklassifikation
  - Verschlagwortung (Thesaurus)
- Postkoordination
  - Indexierung (manuell, automatisch)
  - Volltextsuche

Präkoordination
Dezimalklassifikation

- Dewey-Schema (DDC)
  - von Melvil Dewey 1876 erfunden
  - allumfassende Wissensklassifikation
  - bei öffentlichen und Schulbibliotheken im Einsatz
- Patentklassifikation (IPC)
- Eidgenössischer Zolltarif
- Registraturplan (-> Aktenzeichen)
  - Akten- und Dossierverwaltung in öffentlichen Ämtern und Archiven

Präkoordination
Verschlagwortung

- Thesaurus
  - griechisch für Schatzkästchen
  - Sammlung von normierten Schlagwörtern eines bestimmten Fachgebiets
  - stellt Relationen von Begriffen dar
    - Related Term (RT) - verwandter Begriff
    - Broader Term (BT) - übergeordneter Begriff
    - Narrower Term (NT) - untergeordneter Begriff
    - Used For (UF) - gebraucht anstelle von ...
    - Use (USE) - gebrauche ...

Postkoordination
Indexierung

- manuell
  - Indexierung von Büchern (Register)
  - wird von menschlichen Indexierern durchgeführt
- automatisch
  - generiert aus den Wörtern einer Datei einen Index
  - kann Schlagworte vereinheitlichen und Schlagwortlisten wie einen Thesaurus zu Hilfe nehmen

Postkoordination
Volltextsuche

- gab es bereits vor dem Computer
- hat mit Computern und der Informationsflut stark an Bedeutung gewonnen
- wird zur Suche in grossen kommerziellen Datenbanken verwendet, häufig kombiniert mit einer speziellen Abfragesprache

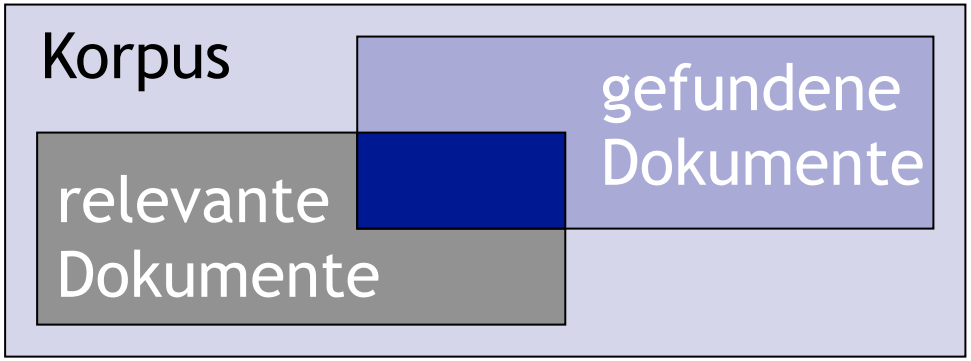
Klassifikation im Internet

- Präkoordination
  - Yahoo!
  - dmoz (Open Directory Project)
  - META-Tags in Webseiten
- Postkoordination
  - Suchmaschinen (Google, MSN Search, A9 etc.)

Information Retrieval (IR)

- Formatkonversion
- Sprachidentifizierung
- Lemmatisierung (Grammatik, Normalformen)
- Stoppworterkennung
- Stammformerkennung (Stemming)
- Kategorisierung
- Indexierung

Messung der Effizienz



- Präzision $= \frac{relevante\_gefundene\_Dokumente}{gefundende\_Dokumente}$

- Recall $= \frac{relevante\_gefundene\_Dokumente}{relevante\_Dokumente}$

Google - Geschichte [i]

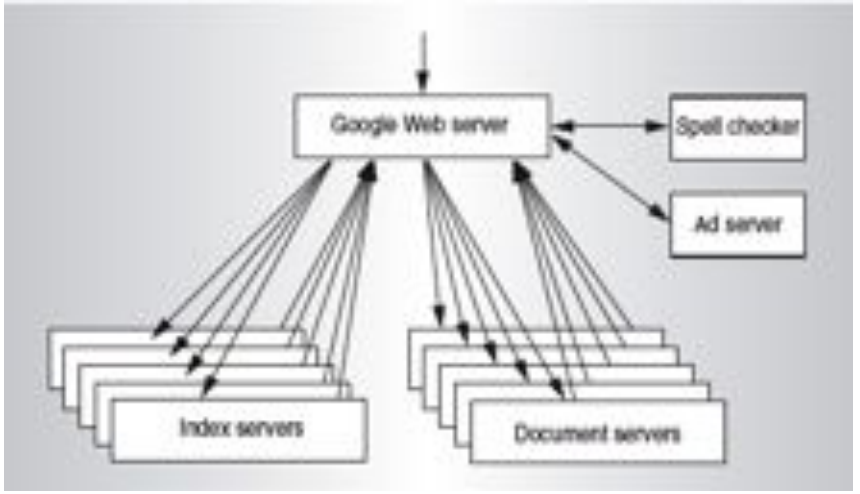
- begann 1996 als Forschungsprojekt der beiden Studenten Larry Page und Sergey Brin an der Stanford University in Palo Alto, Kalifornien (Projekt BackRub)
- im September 1998 wurde Google Inc. in einer Garage in Menlo Park gegründet
- im September 2004 ging Google an die New Yorker Technologie-Börse NASDAQ

Google - Geschichte [ii]

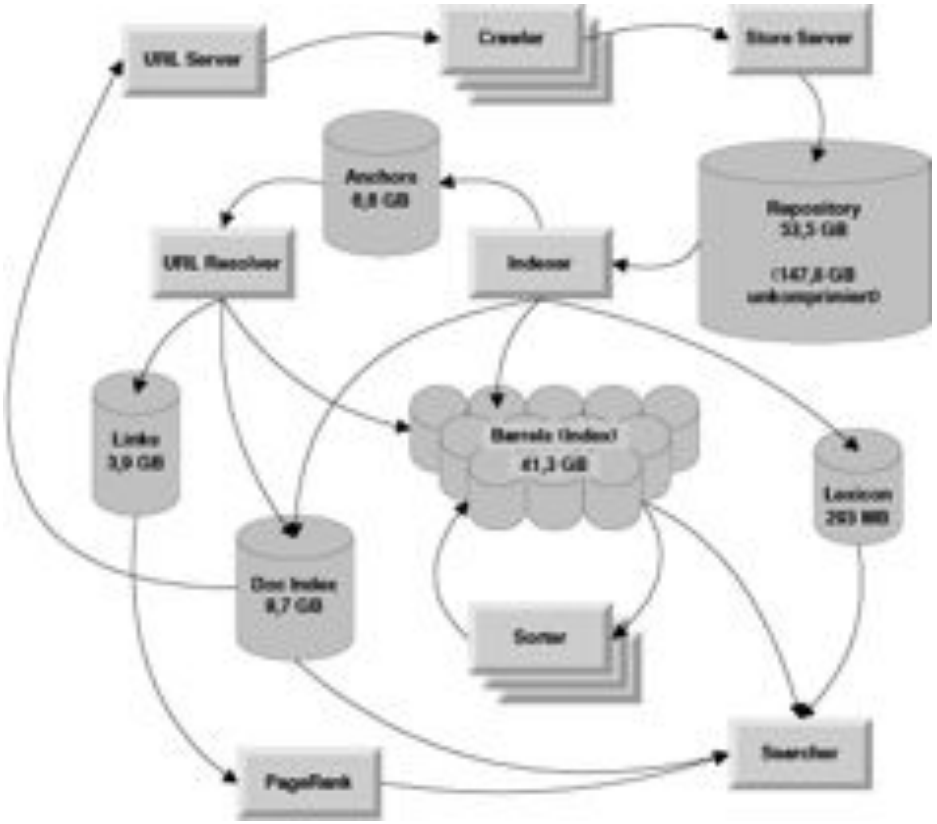
- Von Januar bis September 2005 hat Google einen Umsatz von über 4 Mia US$ und einen Gewinn von über 1 Mia US$ erzielt
- Google beschäftigt heute weltweit über 4 000 Mitarbeiter und hat Niederlassungen in 18 Ländern
- Pro Monat benutzen über 380 Millionen Menschen die Suchmaschine Google, pro Tag sind es über 250 Millionen Suchanfragen

Google - Infrastruktur

- Google betreibt weltweit über zehn Rechenzentren mit 100 000 billigen Linux-Servern
- DNS wird zum Load Balancing eingesetzt
- Google indexiert etwa 4 Milliarden Webseiten mit einer Grösse von ca. 40 TByte



Google - Architektur



Google - Ablauf

- Crawler
  - sucht das Internet nach Webseiten ab und speichert sie komprimiert ab
- Indexer
  - erstellt aus den Webseiten einen Dokumentindex und daraus den invertierten Wortindex
- Query
  - beantwortet Suchanfragen und liefert die Resultate zurück

Google - Crawler

- Erkennt folgende Dateiformate
  - HyperText Markup Language (html)
  - Adobe Portable Document Format (pdf), PostScript
  - Text (ans, txt), Rich Text Format (rtf)
  - Microsoft Word (doc), Excel (xls), PowerPoint (ppt), Write (wri)
  - Lotus 1-2-3 (wk1, wk2, wk3, wk4, wk5, wki, wks), WordPro (lwp)
  - MacWrite (mw)
  - Shockwave Flash (swf)

Google - Indexer [i]

- Extrahiert die Links und speichert sie in die Anchors-Datei
- Anhand der URL wird die docID festgestellt oder eine neue docID erzeugt
- Anhand des Lexikons wird die wordID festgestellt oder eine neue wordID erzeugt
- Das Auftreten jeder wordID wird mit einem Hit markiert

Google - Indexer [ii]

- Hits werden nach docID sortiert in einem Forward Index in Barrels gespeichert
- Der Forward Index wird nach der wordID sortiert und als invertierter Index in Barrels abgespeichert
- Aus der Anchors-Datei werden docIDs erzeugt und in der Links-Datei gespeichert
- Der PageRank-Algorithmus berechnet den PageRank jeder docID

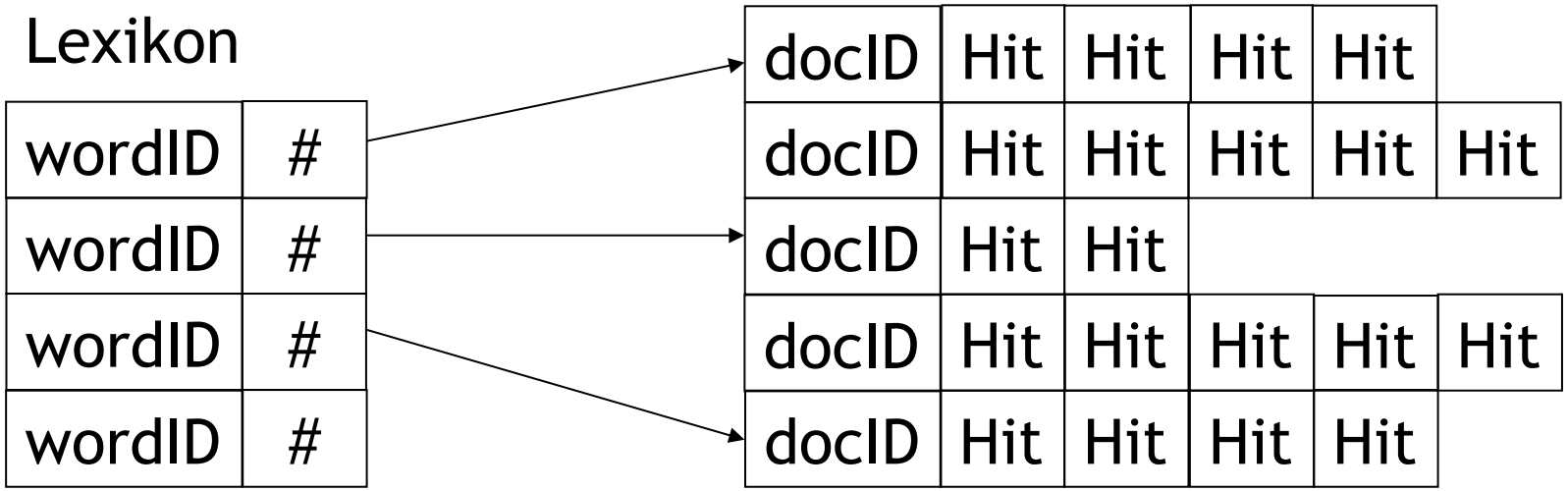
Google - Forward Index

- Enthält die docID mit den Hits pro wordID und wird nach wordID geordnet in mehreren Barrels abgespeichert

docID	wordID	Hit	Hit	Hit	Hit	
	wordID	Hit	Hit	Hit	Hit	Hit
docID	wordID	Hit	Hit			
	wordID	Hit	Hit	Hit	Hit	Hit
	wordID	Hit	Hit	Hit	Hit	

Google - Inverted Index

- Enthält die Hits pro docID und wird nach wordID geordnet in mehreren Barrels abgespeichert



Google - Hashing

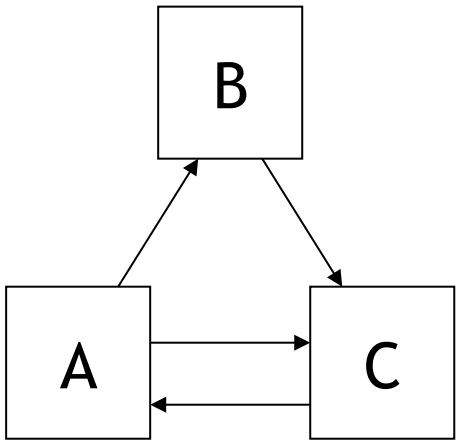
- Sowohl die docID und die wordID werden durch eine Hash-Funktion bestimmt
- ermöglicht einen schnellen Zugriff
- braucht vergleichsweise viel Speicherplatz
- je nach gewählter Hash-Funktion kann die Zahl der nötigen Zugriffe oder der verwendete Speicherplatz optimiert werden

Google - PageRank [i]

- patentgeschütztes Ranking von Google
- basiert auf dem Modell eines sogenanten «Random Surfers»
- Ist ein iteratives Verfahren, das jede Webseite nach den ein- und ausgehenden Links bewertet
- je häufiger auf eine Seite verwiesen wird, desto relevanter ist wahrscheinlich ihr Inhalt

Google - PageRank [ii]

$$PageRank(A) = (1-d) + d \cdot \left(\frac{PageRank(B)}{Links(B)} + \frac{PageRank(C)}{Links(C)} + \cdots \right)$$



	PageRank(A)	PageRank(B)	PageRank(C)
1	1	1	1
2	1	0.75	1.25
3	1.125	0.75	1.125
4	1.0625	0.78125	1.15625
5	1.078125	0.765625	1.15625
6	1.078125	0.76953125	1.15234375
7	1.076171875	0.76953125	1.154296875
8	1.077148438	0.769042969	1.153808594
9	1.076904297	0.769287109	1.153808594
10	1.076904297	0.769226074	1.153869629

Google - Relevanz (Ranking)

- Neben dem PageRank werden auch noch andere Merkmale zum Ranking verwendet
  - Vorkommen des Suchbegriffs im Titel oder der URL
  - Vorkommen des Suchbegriffs in eingehenden Links (Ankertext)
  - Textgrösse und Position des Suchbegriffs
  - Häufigkeit des Suchbegriffs auf der Webseite

Google - Query

- Die Suchwörter werden in wordIDs umgewandelt (-> Hash-Funktion)
- Für jedes Suchwort wird im invertierten Index die Dokumentenliste geholt und nach Übereinstimmung mit den anderen Suchwörtern durchsucht (-> Merge-Funktion)
- Für jedes gefundene Dokument wird der PageRank berechnet