Suchen im WWW

Einführung

Das World Wide Web

- 1990 von Tim Berners-Lee in CERN entwickelt, um im Internet verfügbare Forschungsdokumente zu organisieren.
- Verbindet zur Verlinkung von Dokumenten die Idee von durch FTP verfügbaren Dokumenten mit der Idee von *Hypertext*.
- Entwickelte das ursprüngliche HTTP-Netzwerk-Protokoll, URLs, HTML und den ersten "Webserver".

Web-Vorgeschichte

- Ted Nelson entwickelte 1965 die Idee des Hypertexts.
- Doug Engelbart erfand die Maus und bildete die erste Implementierung von Hypertext in den späten 60igern bei SRI.
- ARPANET wurde Anfang der 70iger Jahre entwickelt.
- Die grundlegende Technologie war in den 70igern etabliert; aber die PC-Revolution und verbreitete Vernetzung waren notwendig, um das Web zu inspirieren und es nutzbar zu machen.

Web-Browser-Geschichte

- Frühe Browser wurden 1992 (Erwise, ViolaWWW) entwickelt.
- In 1993 entwickelten Marc Andreessen und Eric Bina in UIUC NCSA den Mosaic Browser und sorgten für seine weite Verbreitung.
- Andreessen verband sich mit James Clark (Stanford Prof. und Silicon Graphics Gründer), um die Mosaic Communications Inc. in 1994 zu gründen (die Netscape wurde, um einen Konflikt mit UIUC zu vermeiden).
- Microsoft lizensierte das ursprüngliche Mosaic von UIUC und verwendete es 1995 für die Erstellung des Internet Explorers.

Suchmaschinen Frühgeschichte

- Ende 1980 waren viele Dateien durch anonymes FTP verfügbar.
- In 1990 entwickelte Alan Emtage von McGill Univ. Archie (Kurzform für “Archive”)
  - zusammengestellte Listen von auf vielen FTP-Servern verfügbaren Dateien.
  - Erlaubte regex-Suche dieser Dateinamen.
- In 1993 wurden Veronica und Jughead entwickelt, um die Namen von durch Gopher Server verfügbaren Textdateien zu suchen.

Geschichte der Websuche

- 1993 wurden die ersten Webrobots (spiders) gebaut, um URLs zu sammeln:
  - Wanderer
  - ALIWEB (Archie-ähnlicher Index des WEB)
  - WWW Worm (indiziert URL's und Titel für regex Suche)

- In 1994 begannen die Stanford-Studenten David Filo und Jerry Yang mit der manuellen Sammlung beliebter Webseiten in einer thematischen Hierachie, genannt Yahoo.

Forts. Geschichte Websuche

- Anfang 1994 entwickelte Brian Pinkerton WebCrawler als ein Lehrprojekt an der U Wash. (wurde letzendlich Teil von Excite und AOL).

- Einige Monate später entwickelte Fuzzy Maudlin, ein Student an der CMU, Lycos - das erste Standard-IR-System, entwickelt im DARPA Tipster-Projekt. Das erste System mit einer großen Menge von indizierten Seiten.

- Ende 1995 entwickelte DEC Altavista. Sie benutzen eine große Serverfarm von Alpha-Maschinen, um schnell große Mengen von Daten zu verarbeiten. Sie unterstützten Boolesche Opertoren, Phrasen und “reverse pointer”-Anfragen.

Websuche - jüngste Geschichte

- In 1998 starteten Larry Page und Sergey Brin, Doktoranden in Stanford, mit Google. Der wesentliche Fortschritt ist die Nutzung einer *Link-Analyse* zum Ranking der Ergebnisse, die auf der Weitergabe von Relevanz entlang von Hyperlinks basiert.

Web-Herausforderungen für IR

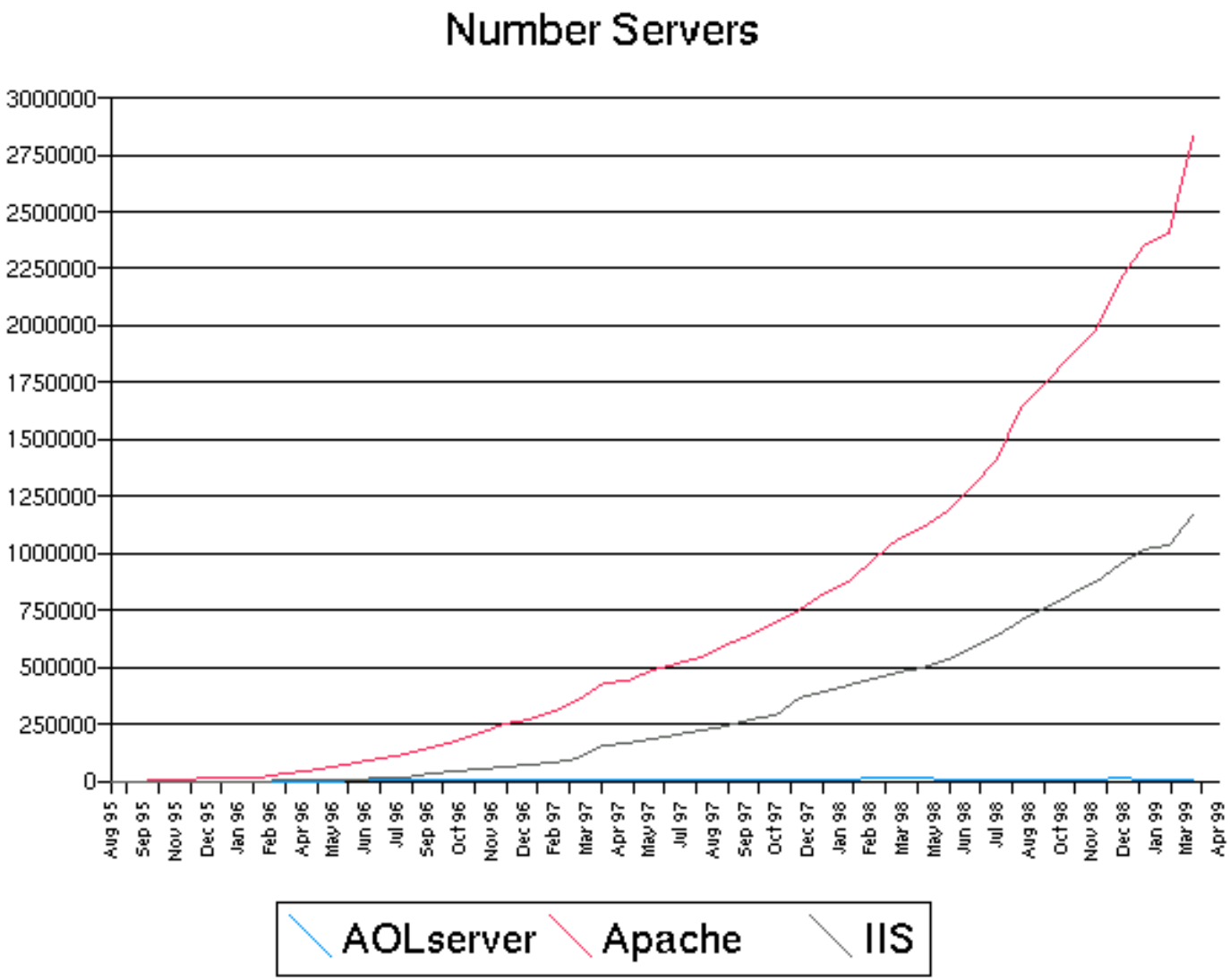
- Verteilte Daten: Dokumente sind über Millionen verschiedener Webserver verteilt.
- Flüchtige Daten: Viele Dokumente ändern sich oder verschwinden schnell (z.B. tote Links).
- Großes Volumen: Billionen von separaten Dokumenten.
- Unstrukturierte und redundante Daten: Keine einheitliche Struktur, HTML Fehler, bis zu 30% (nahezu) doppelte Dokumente.

Web-Herausforderungen für IR

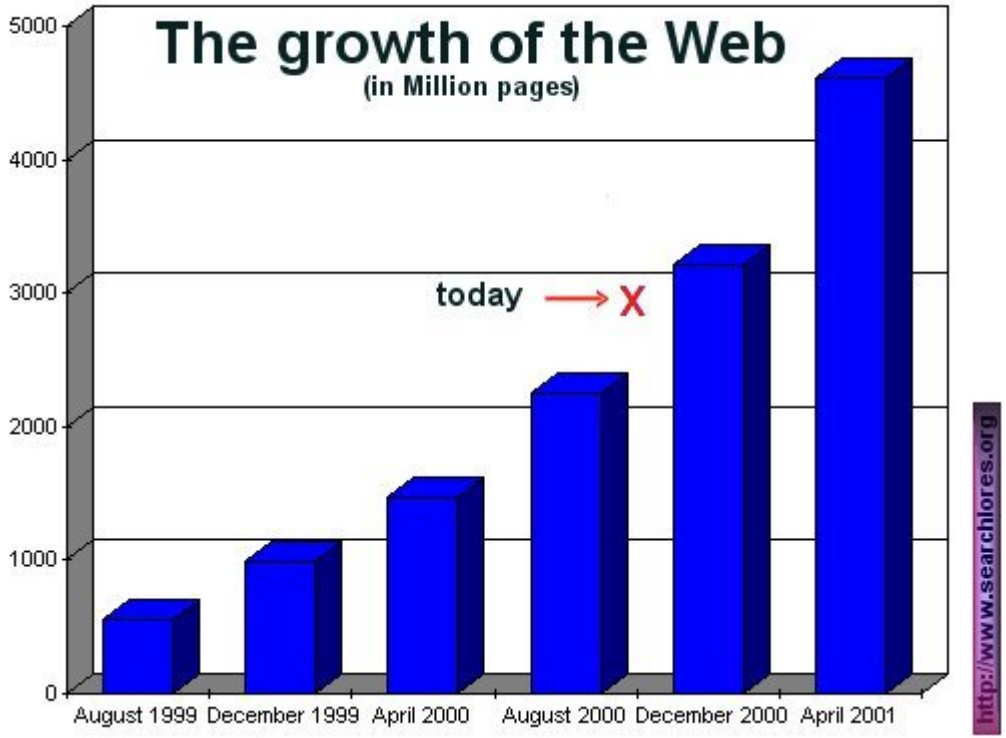
- Qualität der Daten: Keine redaktionelle Kontrolle, falsche Informationen, schlechte Schreibweise, Tippfehler, etc.

- Heterogene Daten: Multiple Medien-Typen (Bilder, Video, VRML), Sprachen, Zeichensätze, etc.

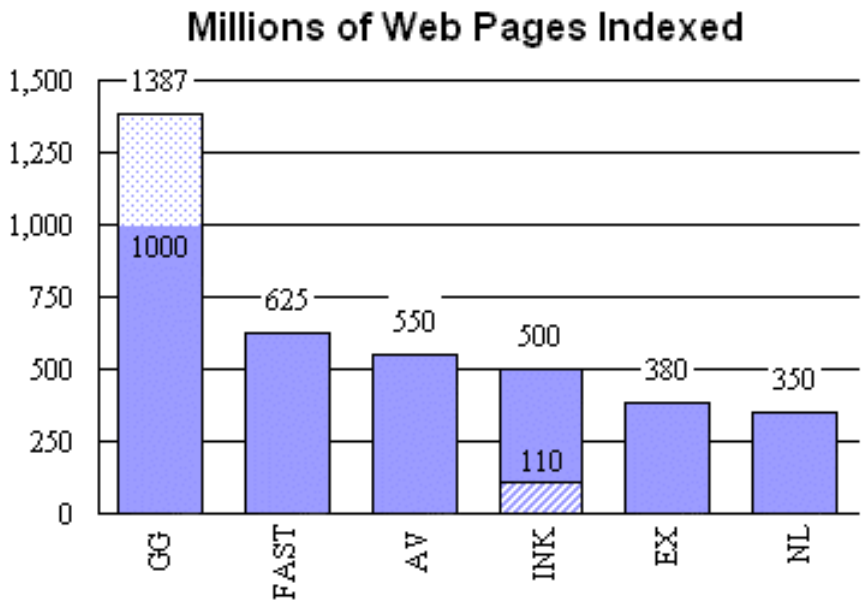
Anzahl der Webserver



Anzahl der Webseiten



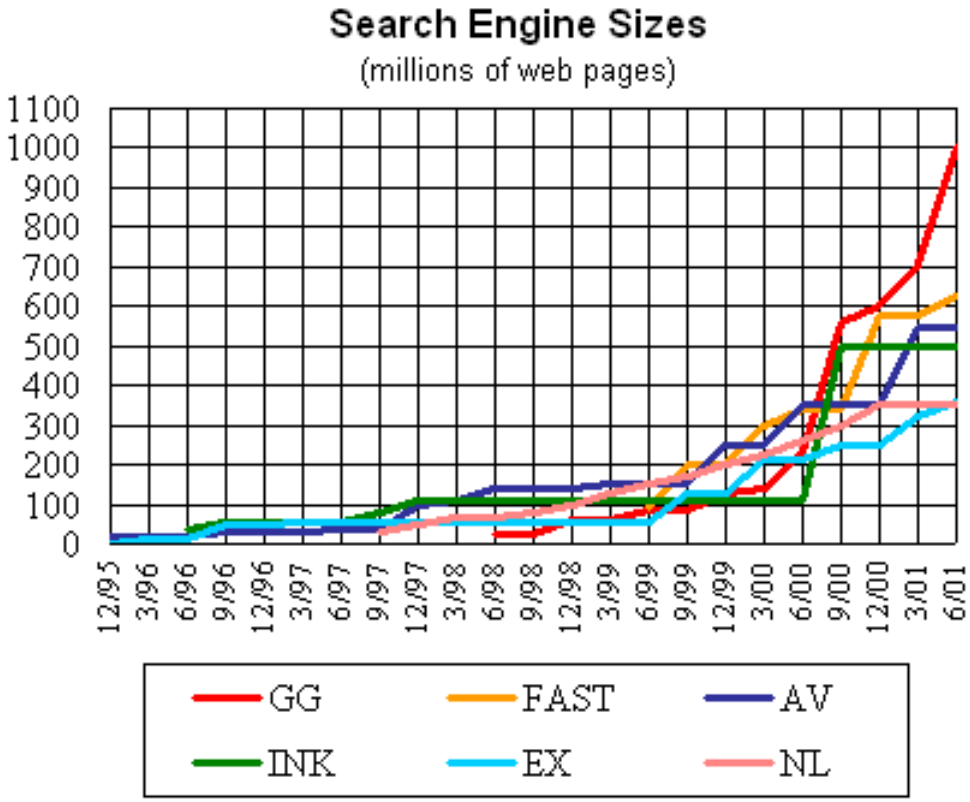
Anzahl der indizierten Webseiten



SearchEngineWatch, Aug. 15, 2001

Unter der Annahme von ca. 20KB pro Seite entsprechen 1 Milliarde Seiten ca. 20 Terabyte Daten.

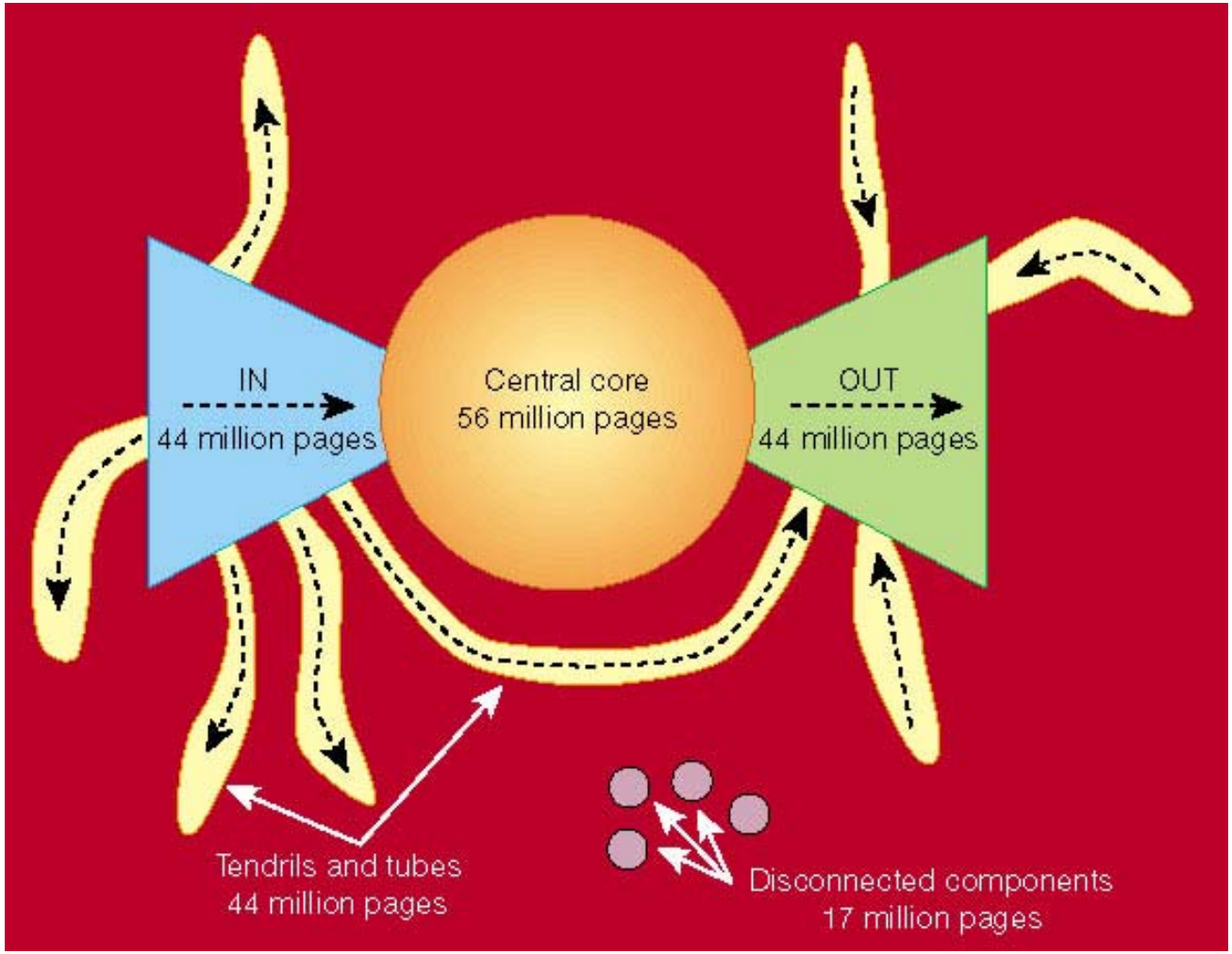
Wachstum der indizierten Webseiten



SearchEngineWatch, Aug. 15, 2001

Google listet die aktuelle Anzahl der besuchten Daten.

Graph-Struktur des Webs



http://www9.org/w9cdrom/160/160.html

Zipf's Gesetz im Web

- Die Anzahl der in-links/out-links zu/von einer Seite hat eine Zipfsche Verteilung.
- Die Länge der Webseiten hat eine Zipfsche Verteilung.
- Die Anzahl der Treffer auf einer Webseite hat eine Zipfsche Verteilung.

Manuelle hierarchische Web-Taxonomie

- Der Ansatz von Yahoo ist es, durch die Unterstützung menschlicher Herausgeber ein großes hierarchisch strukturiertes Verzeichnis von Webseiten zusammenzustellen.
  - http://www.yahoo.com/

- Open Directory Project ist ein ähnlicher Ansatz, der auf der Verteilung der Arbeit auf freiwillige Herausgeber basiert ("net-citizens provide the collective brain"). Dies wird auch von vielen meisten anderen Suchmaschinen verwendet. Angefangen bei Netscape.
  - http://www.dmoz.org/

Automatische Dokument-Klassifizierung

- Die manuelle Klassifizierung in eine gegebene Hierarchie ist arbeitsintensiv, subjektiv und fehleranfällig.
- Methoden zur Text-Kategorisierung bieten einen Weg, um Dokumente automatisch zu klassifizieren.
- Die besten Methoden basieren auf dem Trainieren eines *Maschinellen-Lern-* (*Mustererkennung*-)Systems mit einer vorklassifizierten Menge von Beispielen (*supervised learning*).
- Textkategorisierung ist ein Thema, das wir zu einem späteren Zeitpunkt im Kurs diskutieren werden, sowie insbesondere in der Vorlesung Knowledge Discovery.

Automatische Dokumenten-Hierarchie

- Manuelle Hierarchieentwicklung ist arbeitsintensiv, subjektiv und fehleranfällig.
- Es wäre schön, automatisch eine bedeutungsvolle hierarchische Taxonomie aus einem Dokumenten-Korpus zu konstruieren.
- Dies ist mit hierarchischem Textclustering (unsupervised learning) möglich.
  - Hierarchical Agglomerative Clustering (HAC)
- Textclustering ist ein weiteres Thema, das wir zu einem späteren Zeitpunkt und in der Vorlesung Knowledge Discovery diskutieren werden.

Websuche unter Verwendung von IR