Informatik der digitalen Medien

Ergänzungs-Studienangebot der Mediendidaktik für Lehramtstudenten

Dr. Harald Sack
Institut für Informatik
FSU Jena

Sommersemester 2007

http://www.informatik.uni-jena.de/~sack/SS07/infod.htm

Informatik der digitalen Medien

1 2 3 4 5 6 7 8 9 10 11 **16.07.2007 – Vorlesung Nr. 12**

3. Internet und WWW (Teil 5)

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele

Cast……..

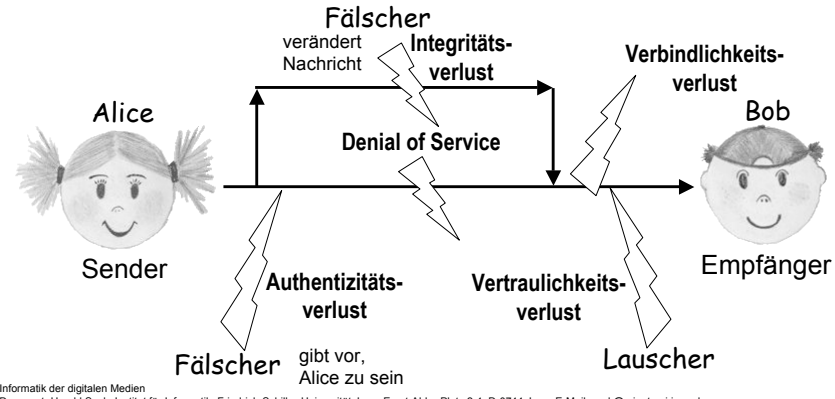
Trudy
(Eindringling,Lauscherin,
Fälscherin, etc..)

Alice (Sender)

Bob (Empfänger)

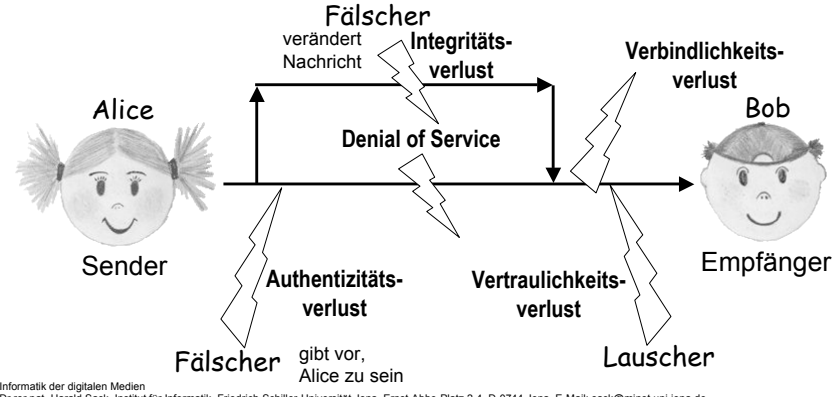
Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - „sicher ist, dass nichts sicher ist….und selbst das nicht."



Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - „sicher ist, dass nichts sicher ist….und selbst das nicht."

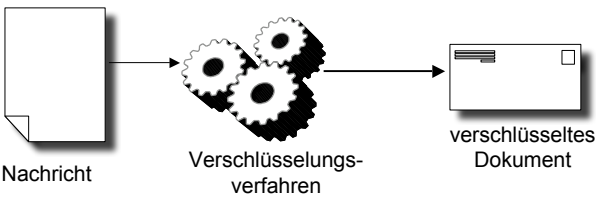


Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - **Verfügbarkeit**
    Die zuverlässige Funktionstüchtigkeit der zur Kommunikation verwendeten Medien darf nicht gestört werden können
  - **Datenintegrität**
    Die übertragene Nachricht muss den Empfänger im Originalzustand erreichen und darf nicht verändert werden
  - **Vertraulichkeit**
    Der Inhalt der übermittelten Nachricht darf nur für Sender und Empfänger, nicht für unbefugte Dritte lesbar sein.
  - **Authentifikation**
    Der Empfänger muss sich darauf verlassen können, dass der Absender der Nachricht diese auch tatsächlich verfasst hat
  - **Autorisation**
    Es muss sichergestellt werden, dass niemand anderes als der designierte Empfänger einer Nachricht die Berechtigung hat, diese zu lesen.

Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - Geheimhaltung durch Verschlüsselung
    - Um eine Nachricht zu verschlüsseln benötigt man dazu ein geeignetes Verfahren



  - **Problem**:
    Wird das Verfahren bekannt, muss man sich ein neues ausdenken (kompliziert)

Internet und WWW (5) - Grundlagen der Kryptografie

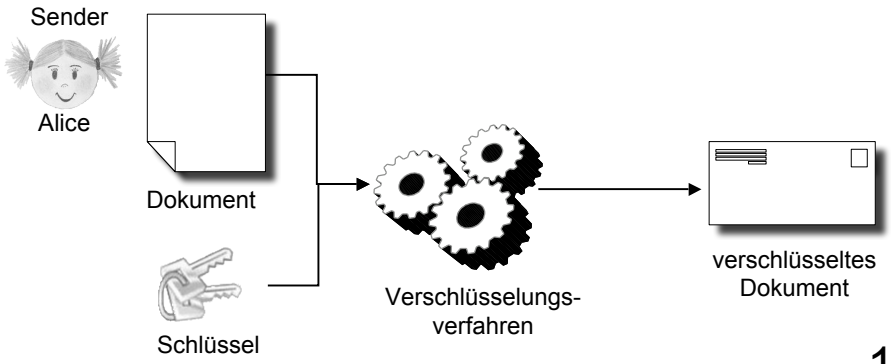
- Sicherheitsziele
  - Geheimhaltung durch Verschlüsselung
    - Besser ist ein Verfahren, das auf einfache Weise, **Variationsmöglichkeiten** der durchzuführenden Verschlüsselung bietet
    - Die Parameter zur Einstellung der Variationsmöglichkeiten werden als **Schlüssel** bezeichnet

Schlüssel

    - **Vorteil:**
      Verfahren kann bekannt sein, nur der jeweilige **Schlüssel muss geheim gehalten werden**

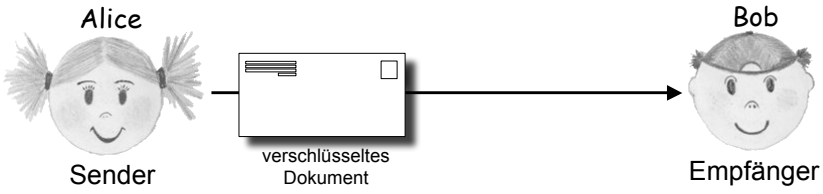
Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - Geheimhaltung durch Verschlüsselung



Internet und WWW (5) - Grundlagen der Kryptografie

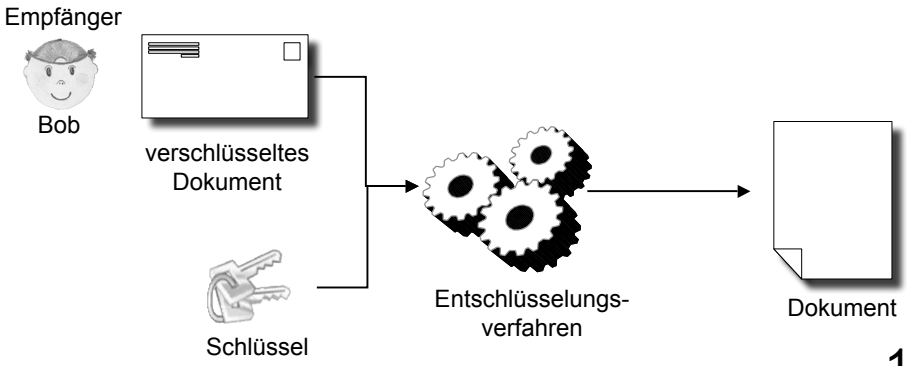
- Sicherheitsziele
  - Übermittlung verschlüsselter Nachrichten



Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

- Sicherheitsziele
  - Geheimhaltung durch Verschlüsselung



Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

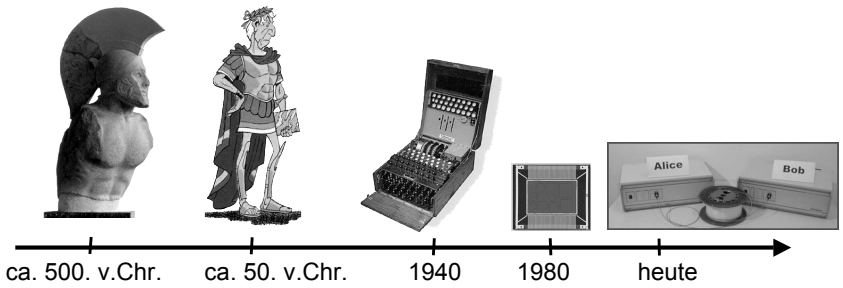
Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - Überblick



Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - **Die griechische Skytale**
    - **Transpositionschiffre**
      - Im 5. Jhd. v. Chr. verschlüsselten die Spartaner Nachrichten mit Hilfe der Skytale
      - **Position** der Einzelzeichen wird nach einem festen Schema **vertauscht**

Klartext	DAS IST EIN GEHEIMNIS
Leerzeichen löschen	DASISTEINGEHEIMNIS
verschlüsseln	DEEAIISNMIGNSEITHS

Entschlüsselung

DEE**A**II**S**NM**I**GN**S**EI**T**HS
D**E**EA**I**IS**N**MI**G**NS**E**IT**H**S
DE**E**AI**I**SN**M**IG**N**SE**I**TH**S**

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

15

Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - **Cäsar's Verschlüsselung**
    - **Substitutionschiffre**
      - im 1. Jhd. v. Chr. nutzte Gaius Julius Cäsar ein einfaches Ersetzungsverfahren als Verschlüsselung
      - jedes **einzelne Zeichen** wird nach festem Schema durch ein anderes Zeichen **ersetzt**

Gaius Julius Cäsar
(100—44 v. Chr.)

Klartext	ROMANI ITE DOMUM
Verschlüsselung	verschiebe alle Buchstaben um drei Buchstabenwerte weiter
Chiffrat	URPDQL LWH GRPXP

A→D
B→E
C→F
D→G
E→H
…

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

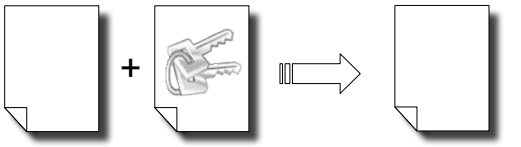
16

Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - **One Time Pad**
    - wähle einen Schlüssel, der
      - nur **ein einziges Mal** zum Verschlüsseln einer einzigen Nachricht genutzt wird und
      - **genauso lang** ist, wie die Nachricht selbst
    - verknüpfe jedes einzelne Zeichen der Nachricht mit einem Zeichen des Schlüssels

Blaise de Vigenère
(1523-1596)

Gilbert Vernam, 1917



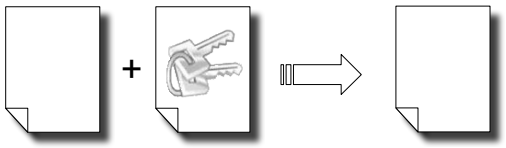
Nachricht Schlüssel verschlüsselte Nachricht

Einfachstes und nachweislich sicheres Verfahren

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - **One Time Pad**



Nachricht Schlüssel verschlüsselte Nachricht

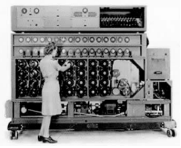
- **Merke**: je **länger** und je **zufälliger** der gewählte **Schlüssel**, desto **schwieriger** ist das Verfahren zu „knacken“!

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

- Kurze Geschichte der Kryptografie
  - **Verschlüsselungsmaschinen**
    - Kombination von Transpositionen und Substitutionen mit dynamisch wechselndem Schlüssel
    - Abfolge und Parameter werden durch **geheimen Schlüssel** bestimmt
    - Berühmtestes Beispiel:
    - **Enigma**
      verschlüsselte Funksprüche der deutschen Wehrmacht im 2. Weltkrieg

Alan Turing
(1916-1954)



„Bombe“ – zur automatischen Entschlüsselung

Enigma - Maschine

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

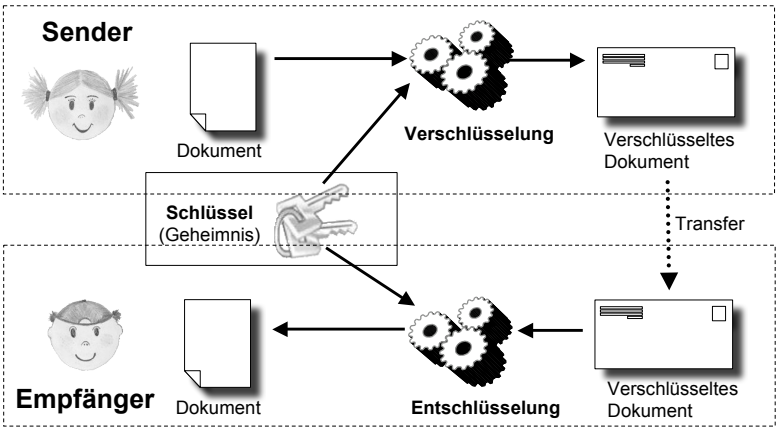
- Kurze Geschichte der Kryptografie
  - **Offene Geheimnisse – öffentliche Schlüssel**
    - Wie komplex die Verschlüsselungsverfahren auch sind, alle hängen bislang von einem **sicheren Austausch der** verwendeten **Schlüssel** ab
    - **Idee:**
      - Gibt es ein Verfahren zur Verschlüsselung, das ohne Austausch eines geheimen Schlüssels auskommt?
    - Kommunikation mit **öffentlichen Schlüsseln**
      - **öffentlicher Schlüssel** zum Verschlüsseln (kann von jedem genutzt werden)
      - **geheimer Schlüssel** zum Entschlüsseln (bleibt beim Besitzer)

Whitfield Diffie
Martin Hellmann
Ralph Merkle
(1976)

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Symmetrische Verschlüsselungsverfahren**



Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Symmetrische Verschlüsselungsverfahren**
    - Sender und Empfänger verwenden einen **identischen Schlüssel**, der nur jeweils den beiden bekannt ist
    - Verschlüsselungsverfahren kann allgemein bekannt sein
    - **Problem:**
      - Sender und Empfänger müssen den jeweils verwendeten Schlüssel zuvor austauschen
      - der Schlüsselaustausch muss geheim gehalten werden!

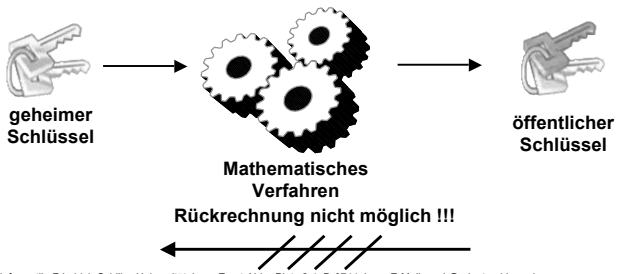
Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

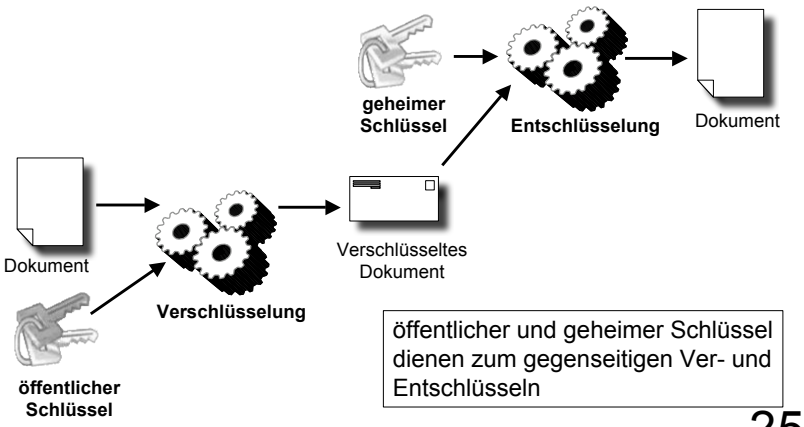
Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**
    - Problem bei symmetrischen Verfahren ➔ Schlüsselaustausch
    - Ist es möglich, **ohne** einen geheimen **Schlüsselaustausch** auszukommen?
    - Voraussetzung dazu sind **mathematische Einwegfunktionen**



Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**

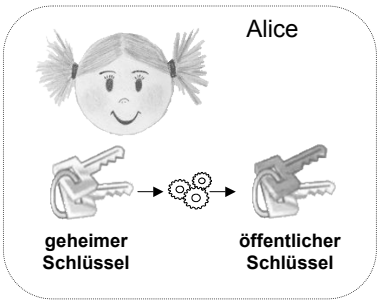


öffentlicher und geheimer Schlüssel dienen zum gegenseitigen Ver- und Entschlüsseln

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**
    - Sender **behält den geheimen Schlüssel** für sich
    - nur der **öffentliche Schlüssel wird an alle weitergegeben**, die mit dem Sender kommunizieren wollen

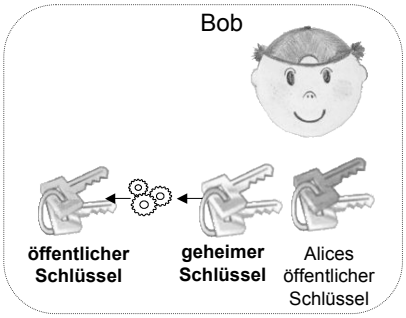


Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Grundlagen der Kryptografie

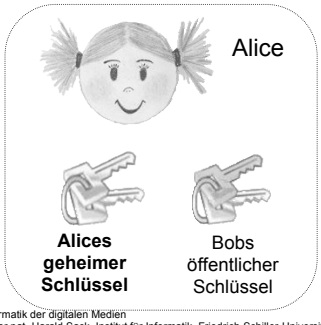
- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**
    - Sender **behält den geheimen Schlüssel** für sich
    - nur der **öffentliche Schlüssel wird an alle weitergegeben**, die mit dem Sender kommunizieren wollen

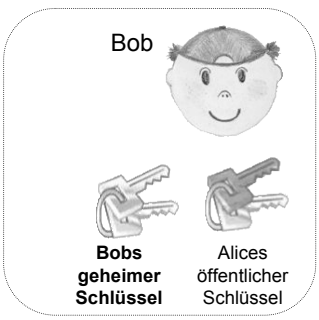
Alice



Internet und WWW (5) - Grundlagen der Kryptografie

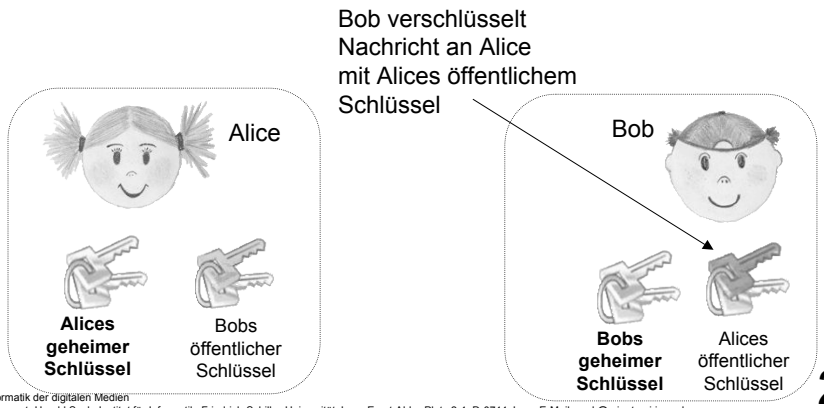
- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**
    - Sender **behält den geheimen Schlüssel** für sich
    - nur der **öffentliche Schlüssel wird an alle weitergegeben**, die mit dem Sender kommunizieren wollen





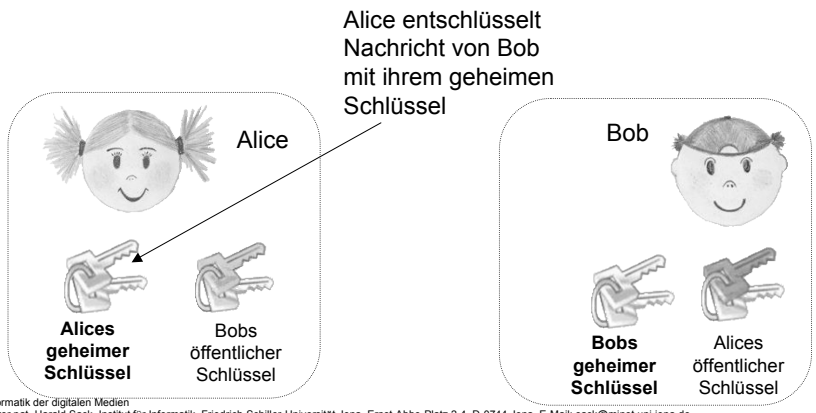
Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**



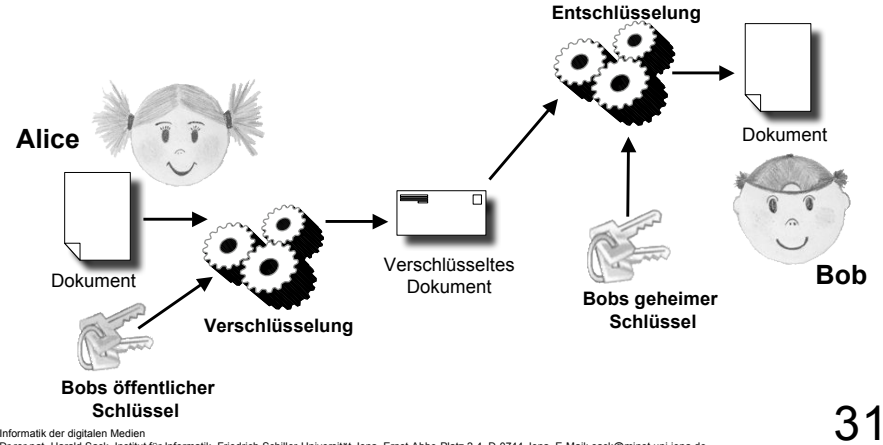
Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**



Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**



Internet und WWW (5) - Grundlagen der Kryptografie

- Grundlagen der Kryptografie
  - **Verfahren mit öffentlichem Schlüssel**
    - Niemand außer Alice kann eine Nachricht entschlüsseln, die mit ihrem öffentlichen Schlüssel verschlüsselt wurde
    - Bob kann also sicher sein, dass seine Nachricht von niemandem sonst als Alice gelesen werden kann
    - Verfahren garantiert
      - **Vertraulichkeit der Nachricht**
      - **Authentizität des Empfängers**
    - Aber wer garantiert, dass die empfangene Nachricht nicht doch verfälscht wurde… (**Integrität**)...??

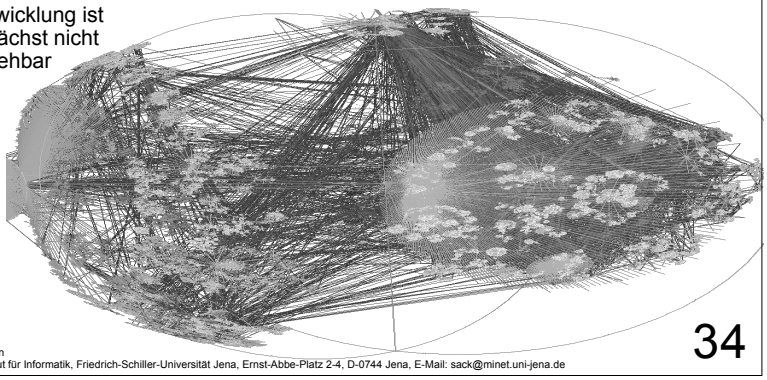
Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - WWW bietet Zugriff auf eine gigantische Informationsfülle
  - Schätzungen gehen von über 55 Milliarden Dokumenten im WWW aus
    - davon derzeit (11/2005) etwa 19,5 Mrd. in **Google** indiziert
  - Dokumentenbestand im WWW verdoppelt sich etwa alle 6 Monate
  - Ein Ende dieser Entwicklung ist zunächst nicht absehbar



Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - um gezielt auf Informationen im WWW zugreifen zu können, muss der Nutzer durch geeignete Werkzeuge unterstützt werden

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **allgemeine Aufgaben von Suchmaschinen**:
    - Unterstützung des Nutzers bei der Informationsbeschaffung im WWW
    - Erschließung eines **möglichst vollständigen** Datenbestands aller Dokumente des WWW
    - **Zuordnung** der einzelnen **Dokumente** des WWW zu bestimmten **Schlüsselbegriffen**
    - Wichtigste Kriterien für Nutzer und Anbieter:
      - **Vollständigkeit** und **Genauigkeit**
    - Jeder Informationsanbieter „möchte gefunden werden“

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Grundtypen der WWW-Suchdienste**:
    - Webkataloge (katalogbasierte Suchmaschinen)
    - (indexbasierte) Suchmaschinen
    - Meta-Suchmaschinen
    - Payed Placement-Suchmaschinen

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - Webkataloge (katalogbasierte Suchmaschinen)
    - Suchdienst, dessen Datenbestand von **menschlichen Redakteuren** zusammengestellt wird
    - Redakteure stellen einen thematisch gegliederten Suchkatalog zusammen
    - Web-Seiten werden dazu
      - manuell geprüft
      - redaktionell bewertet
      - verworfen oder für Aufnahme in den Katalog akzeptiert
    - Suche erfolgt durch Blättern des Suchkataloges

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - Webkataloge (katalogbasierte Suchmaschinen)
    - **Pro**:
      - Intellektuelle Bewertung der Web-Seiten durch Menschen erhöht die **Qualität** (Präzision) der Suchergebnisse
    - **Contra**:
      - Relativ kleiner Datenbestand
      - nur wenige Informationsangebote können berücksichtigt werden
      - Problem: Aktualität
      - Neue WebSites/Dokumente müssen stets angemeldet werden

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Webkataloge (katalogbasierte Suchmaschinen)**
    - **www.dmoz.org**
      - seit 1998
      - "Directory Mozilla" → Open Directory Project
      - Nichtkommerziell, jeder kann sich beteiligen
      - > 5,2 Mio Sites im Katalog verzeichnet
      - 590.000 unterschiedliche Kategorien
      - > 71.000 freiwillige Editoren

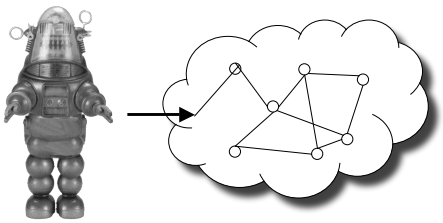
Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - Indexbasierte Suchmaschinen
    - Index-Datenbestand wird vollautomatisch gewonnen und verarbeitet
    - **Basisfunktionen:**
      1. Datenbeschaffung
      2. Dokumentenanalyse und Dokumentenbewertung
      3. Aufbau und Verwaltung von Index-Datenstrukturen
      4. Beantwortung von Suchanfragen unter Einbeziehung von Relevanzwerten

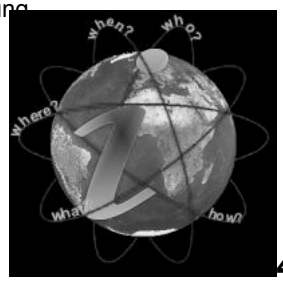
Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Indexbasierte Suchmaschinen**
    1. **Datenbeschaffung:**
       - Einsatz von speziellen, autonom arbeitenden Software-Werkzeugen (**Robots**, **WebRobots**)
       - Robots können automatisch **neue WebSites** und Dokumente im WWW ausfindig machen
       - bereits im Datenbestand vorhandene Dokumente müssen periodisch auf **Konsistenz/Veränderungen** überprüft werden



Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Indexbasierte Suchmaschinen**
    2. **Dokumentenananalyse und -bewertung:**
       - Einsatz von Software-Werkzeugen zur vollständig **automatisierten Analyse und inhaltlichen Bewertung** von Dokumenten (**Information Retrieval Systeme**)
       - Manuelle Eingriffe erfolgen in der Regel nur bei Verstößen gegen die jeweilige Nutzungsordnung



Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

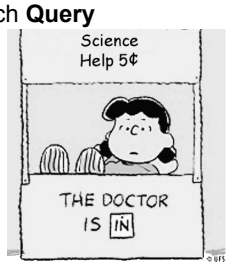
Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Indexbasierte Suchmaschinen**
    3. **Aufbau und Verwaltung von Index-Datenstrukturen:**
       - **Information Retrieval Systeme** ermitteln inhaltliche Schwerpunkte der untersuchten Dokumente und
       - legen die analysierten Dokumente entsprechend der relevanten Kategorien (**Schlüsselworte**) innerhalb einer Datenbank ab
       - Einzelnen Dokumenten wird entsprechend ihrer Relevanz bzgl. der darin behandelten Themen eine Gewichtung zugewiesen
       - Verfahren zur Erstellung eines durchsuchbaren Datenbestandes werden als **Indexierung** bezeichnet

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Indexbasierte Suchmaschinen**
    4. **Beantwortung von Suchabfragen :**
       - Suche erfolgt durch Eingabe eines/mehrerer Suchbegriffe
       - **Automatische Relevanzbewertung** der einzelnen Dokumente des Datenbestands führt zur Auswahl eines Ergebnisses entsprechend dem eingegebenen Suchbegriff
       - Auswahl der Ergebnis-Dokumente erfolgt durch **Query Processor** (eigentliche „Suchmaschine“)
       - Anzeige der gefundenen Dokumente erfolgt entsprechend der Relevanzgewichtung



Internet und WWW (5) - Suchmaschinen im WWW

- Suchmaschinentechnologie
  - **Indexbasierte Suchmaschinen**
    - **Pro**:
      - Automatische Datenbeschaffung ermöglicht möglichst **aktuellen und vollständigen** Datenbestand
    - **Contra**:
      - Zielgenauigkeit ist abhängig von den zur Relevanzbewertung eingesetzten Algorithmen
      - Automatische Relevanzbewertung führt zu qualitativ minderwertigeren Ergebnissen

Google

Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Google-Datenbeschaffung – Was?**
    - **Problem 1**: **Datenvielfalt des WWW**
      - statische HTML-Dokumente
      - Dynamisch erzeugte HTML-Dokumente
      - Bilder (JPG/GIF/PNG/…)
      - Postscript/PDF-Dokumente
      - Word/Powerpoint-Dokumente
      - etc…

⇒ Festlegung, **welche Datentypen** archiviert werden

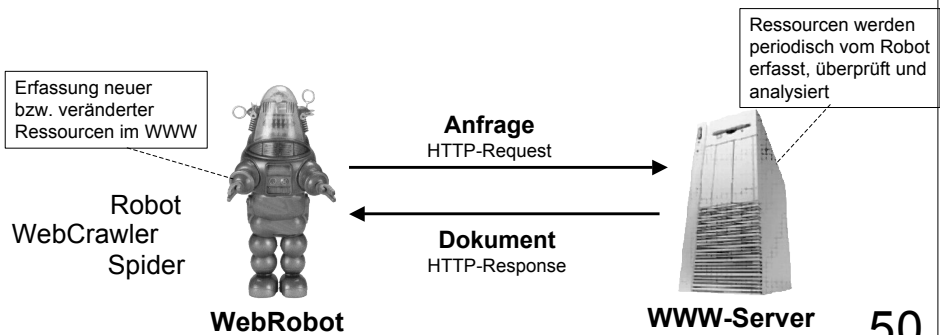
Internet und WWW (5) - Suchmaschinen im WWW

- **Wie funktioniert eigentlich Google?**
  - **Google-Datenbeschaffung – Wann?**
    - **Problem 2: das WWW lebt…**
    - Daten und Dokumente im Wissensraum WWW
      - haben nur eine „kurze“ Lebenszeit
      - unterliegen ständigen Veränderungen
      - sind von anderen Dokumenten abhängig (Links)

⇒ Erfasster Datenbestand muss **periodisch gewartet** werden

Internet und WWW (5) - Suchmaschinen im WWW

- **Wie funktioniert eigentlich Google?**
  - **Google-Datenbeschaffung**
    - Google verwendet WebRobots (Crawler) zur Erschließung des Datenbestands im WWW
    - WebRobot arbeitet verteilt nach dem Client-/Server-Prinzip



Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - Arbeitsweise eines WebCrawlers (vereinfacht)

1. Initialisiere Warteschlage mit zufällig gewählten URLs
2. Lade Dokument zur ersten URL in der Warteschlange
3. Finde alle Hyperlinks im untersuchten Dokument und hänge diese in die Warteschlange
4. Speichere das untersuchte Dokument
5. GOTO 2

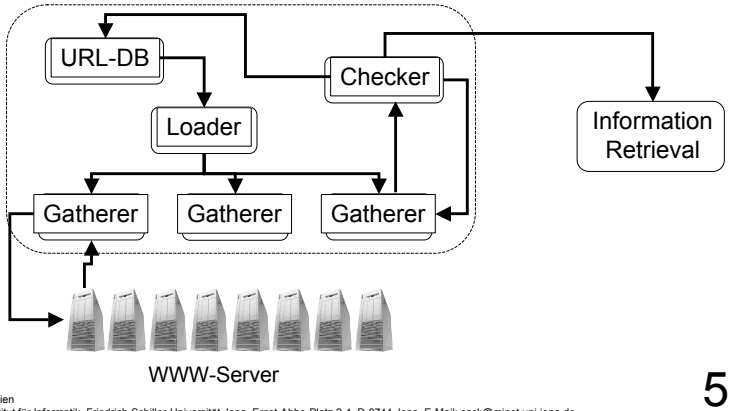
Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Komponenten eines WebRobot-Systems (vereinfacht)**

- **Gatherer**
  Dokumentensammlung aus dem WWW
- **Loader**
  Organisation der Beschaffungsaufträge
- **URL-Datenbank**
  Verwaltung des gesammelten Datenbestands
- **Checker**
  Filterung der gesammelten Daten

Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Komponenten eines WebRobot-Systems (vereinfacht)**



Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Komponenten eines WebRobot-Systems**
    - **Checker**

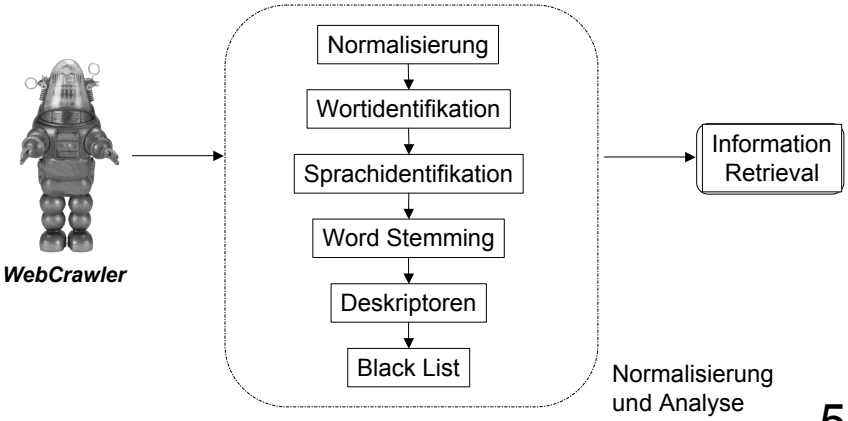
- entscheidet, welche Dokumente vom Gatherer an das **Information Retrieval-System** weitergegeben werden z.B. Auswahl nach
  - Dokumententyp,
  - syntaktischer Korrektheit,
  - Verfügbarkeit…
- entscheidet, nach welchen Links **weiter gesucht** werden soll
- Vermeidung von SPAM, defekten Links, Redirects, etc.
- eliminiert Duplikate

Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Google Datenaufbereitung und -analyse**
    - Umwandlung der Dokumente in **einheitlichen Dokumententyp** (HTML, Postscript, PDF, DOC, PPT in Text umwandeln)
      → effizient durchsuchbarer Datenbestand
    - Auffinden **relevanter Zeichenfolgen** durch semantische Analyse der Textdatei
      - Schlüsselworte,
      - Überschriften,
      - Aufzählungspunkte,… analysieren)
    - Zuordnung **Schlüsselworte** (Suchbegriffe) zu Dokumenten
    - unter Berücksichtigung von Bewertungskriterien **Rangfolge** bilden

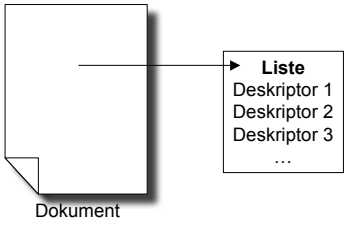
Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Google Datenaufbereitung und –analyse**



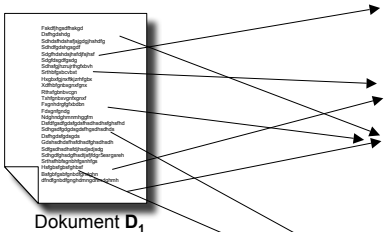
Internet und WWW (5) - Suchmaschinen im WWW

- Wie funktioniert eigentlich Google?
  - **Google Datenstrukturen im Information Retrieval**
    - Erfordernis der schnellen Beantwortung von Suchabfragen macht spezielle Datenstrukturen erforderlich
    - **Reguläres (direktes) Dateisystem**
      - Speicherung von Dokument und Liste der daraus extrahierten Deskriptoren (Schlüsselworte)



Internet und WWW (5) - Suchmaschinen im WWW

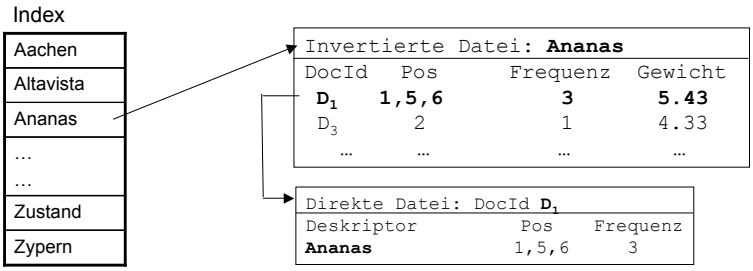
- Wie funktioniert eigentlich Google?
  - **Google Datenstrukturen im Information Retrieval**



D$_1$	*D*$_3$	…	*D*$_n$	
1	*0*	…	*0*	Aachen
0	*0*	…	*1*	Altavista
3	*1*	…	*0*	Ananas
2	*3*	…	*2*	Buttermilch
0	*2*	…	*3*	Bratwurst
0	*0*	…	*0*	Chromatograph
0	*0*	…	*0*	Dynamik
…	…	…	…	…
0	*0*	…	*2*	Zustand
0	*0*	…	*0*	Zypern

Internet und WWW (5) - Suchmaschinen im WWW

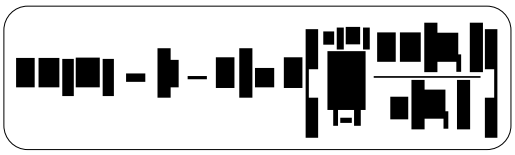
- Wie funktioniert eigentlich Google?
  - **Google Datenstrukturen im Information Retrieval**
    - Erfordernis der schnellen Beantwortung von Suchabfragen macht spezielle Datenstrukturen erforderlich
    - **Invertiertes Dateisystem**
      - Umgekehrt wird zu einem Deskriptor eine Reihe von relevanten Dokumenten zugeordnet



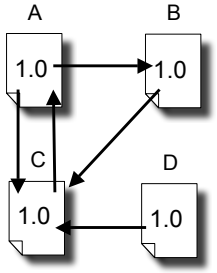
Internet und WWW (5) - Suchmaschinen im WWW

- **Wie funktioniert eigentlich Google?**
  - **Google Relevanz- und Gewichtungsmodelle**
    - Um qualitativ hochwertige Suchergebnisse zu erzielen, müssen die aus dem invertierten Index gewonnenen Dokumente entsprechend ihrer Relevanz gewichtet werden
    - **Wie funktioniert das?**
    - Google unterscheidet „wichtige“ von „unwichtigen“ Dokumenten
    - **„Wichtig“:**
      1. ein Dokument ist um so „wichtiger“, je mehr andere Dokumente auf dieses Dokument via Links verweisen
      2. ein Dokument, auf das ein „wichtiges“ Dokument via Link verweist, ist selbst „wichtig“
      3. je mehr Links ein Dokument auf andere Dokumente enthält, desto „unwichtiger“ ist ein einzelner Link

- **Wie funktioniert eigentlich Google?**
  - **Google Relevanz- und Gewichtungsmodelle**
    - aus 1-3 läßt sich eine Formel zur Berechnung der „Wichtigkeit“ (**PageRank**, **PR**) eines Dokuments gewinnen
      - sei **PR(A)** der zu ermittelnde PageRank des Dokuments **A**
      - seien $\mathbf{T_1 \ldots T_n}$ Dokumente, die einen Link auf A enthalten
      - seien $\mathbf{PR(T_1)}$ … $\mathbf{PR(T_n)}$ die PageRanks der Dokumente $\mathbf{T_1} \ldots \mathbf{T_n}$
      - sei $\mathbf{c(T_i)}$ die Anzahl der ausgehenden Links in Dokument $\mathbf{T_i}$
      - sei d ein Dämpfungsfaktor (0<d<1)



- **Wie funktioniert eigentlich Google?**
  - **Google Relevanz- und Gewichtungsmodelle**
    - Beispiel für dir PageRank-Berechnung

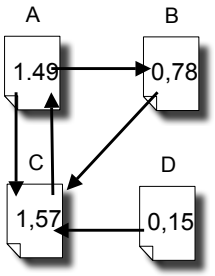


Berechnung wird iterativ durchgeführt, bis sich ein stabiler Zustand (**Fixpunkt**) ergibt.

Nr.	PR(A)	PR(B)	PR(C)	PR(D)
1	1,0	1,0	1,0	1,0
2	1,0	0,575	2,275	0,15
3	2,083	0,575	1,1912	0,15
…	…	…	…	…
n	1,49	0,7833	1,577	0,15

Internet und WWW (5) - Suchmaschinen im WWW

- **Wie funktioniert eigentlich Google?**
  - **Google Relevanz- und Gewichtungsmodelle**
    - Beispiel für dir PageRank-Berechnung



Berechnung wird iterativ durchgeführt, bis sich ein stabiler Zustand (**Fixpunkt**) ergibt.

Nr.	PR(A)	PR(B)	PR(C)	PR(D)
1	1,0	1,0	1,0	1,0
2	1,0	0,575	2,275	0,15
3	2,083	0,575	1,1912	0,15
…	…	…	…	…
n	1,49	0,7833	1,577	0,15

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Internet und WWW (5) - Suchmaschinen im WWW

- Alternativen und bessere der Suchergebnisse
  - **PageRank und darüber hinaus…**
    - bei der Darstellung des Suchergebnisses erscheinen Dokumente mit hohem PageRank vor Dokumenten mit niedrigem PageRank
    - **Welche Probleme gibt es darüber hinaus?**
      - Synonyme und Homonyme (z.B. „**Golf**“…)

      - Text in Grafiken
      - PageRank Manipulation

Informatik der digitalen Medien
Dr.rer.nat. Harald Sack, Institut für Informatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2-4, D-0744 Jena, E-Mail: sack@minet.uni-jena.de

Informatik der digitalen Medien

3. Internet und WWW (5)

- Grundlagen der Kryptografie
  - Sicherheitsziele
  - Kurze Geschichte der Kryptografie
  - Verfahren mit öffentlichem Schlüssel
  - #
- Suchmaschinen im WWW
  - Suchmaschinentechnologie
  - Wie funktioniert eigentlich Google?

Informatik der digitalen Medien

Internet und WWW (5) - Suchmaschinen im WWW

- **Literatur**

- Ch. Meinel, H. Sack:
  WWW – Kommunikation, Internetworking, Web Technologien,
  Springer, 2004.

- M. Glöggler:
  Suchmaschinen im Internet,
  Springer, 2003..