

# Google PageRank

---

**Vortrag zur Seminararbeit von Tobias Kiessling**

# 1. Die Grundidee des PageRanks beruht auf der Berücksichtigung der Linkstruktur des Webs zur Sortierung von Suchergebnissen

- **1. Einleitung**
  - Hintergrund
  - Struktur des Webs
  - Grundidee PageRank
  - Traditionelles Information Retrieval

## Von der Idee die Suche im Web zu verbessern ...

- **Die nun folgende Präsentation basiert größtenteils auf einem Paper von Sergey Brin und Larry Page aus dem Jahr 1998**
- **Damals sind beide Doktorandenstudenten an der Stanford University**
- **Ihr Projekt in dessen Rahmen „PageRank“ beschrieben wird heißt Google und soll die Suche im Internet verbessern**

# 1. Idee – Entscheidend für den Stellenwert einer Seite ist, wie häufig andere Seiten auf sie verlinken *und* wer auf sie verlinkt

## ■ Einstiegsbeispiel:

- 20 Studenten schreiben eine Seminararbeit
- 14 Studenten verweisen auf Wikipedia  
=> Wikipedia muss wichtig sein, **da häufig zitiert**
- 2 Studenten verweisen auf Prof. Studer, diese 2 Studenten haben die beste Note  
=> Prof. Studer muss wichtig sein, **da von hochqualifizierten Arbeiten zitiert**

## ■ Ergebnis:

- **Es ist wichtig wie häufig eine Seite zitiert wird**
- **ABER AUCH von wem sie zitiert wird! (PageRank-Idee)**

# 1. Idee – Bisherige Algorithmen basieren nur auf der Anzahl der Backlinks, aber berücksichtigen nicht von wem sie kommen

## ■ Grundstruktur des Webs

- Viele Seiten
- Seiten variieren extrem stark (Kommerzielle Seiten, Foren, ...)
- Problem: Suchergebnisse nicht pauschal bewertbar; für A sehr gut, für B evtl. sehr schlecht

## ■ Link Struktur des Webs:

- Ca. 150 Mio. Seiten & 1,7 Mrd. Links (Stand: 98, heute > 10 Mrd. Pages)
- (Fast) Jede Seite hat ausgehende und eingehende Links
- Ausgehende Links sind leicht zu ermitteln,
- Alle eingehenden Links einer Seite zu finden, ist praktisch unmöglich

# 1. Idee – Die Basis des PageRanks ist die Linkstruktur des Webs

- **Problem „herkömmlicher“ Suchmaschinen:**
  - Anzahl der eingehenden Links (*Backlinks*) ist manipulierbar
  - Ein eingehender Link von einer *wichtigen* Seite hingegen nur schwer
  - Broken links
  
- **PageRank**
  - Beruht (angeblich) nur auf der Link-Struktur des Webs
  - Je mehr eingehende Links eine Seite hat, desto höher deren PR
  - Je höher der PR der verlinkenden Seiten, desto höher der PR der Seite, auf die verlinkt wird

# 1. Idee – Traditionelles Information Retrieval scheitert an der Größe und Heterogenität des Webs

## ■ Traditionelles Information Retrieval

- Forschung basiert primär auf „kleinen“, homogenen Datenmengen
- Das Web ist aber alles andere als klein und homogen
- => Herkömmliches IR funktioniert im Web nicht gut

## ■ Vector Space Modell

- Betrachtung von Dokument und Suchanfrage als Punkten im hochdimensionalen Vektorraum (=> Dokumentenvektoren)
- Findet Dokumente mit geringster Distanz zur Suchanfrage, also die mit größter Übereinstimmung
- Entscheidend ist die Häufigkeit der Wortvorkommnisse
- => Im Web sind dies meist sehr kurze (einige Worte), aber nicht zwangsläufige „gute“ Dokumente, da in langen Dokumenten die Dichte der Wortvorkommen normalerweise geringer wird

## 2. Der PageRank bildet einen „Random Surfer“ nach und wird iterativ berechnet

### ■ 2. Der PageRank-Algorithmus

- Formel
- Random Surfer Modell als intuitive Erklärung
- Probleme
- Modifikationen
- Berechnung

## 2.1 Algorithmus – Die Umsetzung der Idee in eine berechenbare Formel

- **Die Formel:**

$$PR(A) = (1 - d) + d * [PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn)]$$

- **Variablen:**

- PR(A): PageRank der Seite A
- PR(Ti): PageRank der Seite Ti, von der ein Link auf Seite A zeigt
- **C(Ti):** Gesamtanzahl der Links auf Seite Ti und
- d: Dämpfungsfaktor ( $0 < d < 1$  ist, meist 0,85), bestimmt den Grad der Weitergabe des PR an verlinkte Seiten
- Ti: T1 bis Tn mit n = Anzahl der Seiten im Web (> 10 Mrd.)

## 2.1 Algorithmus – Eingehende Links erhöhen den PR und unter allen Ausgehenden wird der PR aufgeteilt

- **Wiederholung der Formel:**

$$PR(A) = (1 - d) + d * [PR(T1) / C(T1) + \dots + PR(Tn) / C(Tn)]$$

- ➡ **Jeder eingehende Link erhöht den PageRank von Seite A**
- ➡ **Eingehende Links von Seiten mit hohem PageRank erhöhen eigenen PageRank stärker**
- ➡ **Je mehr ausgehende Links eine Seite hat, umso weniger „PageRank“ gibt sie an Seite A weiter**

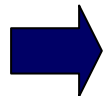
## 2.2 Random Surfer - Das „Random Surfer“ Modell liefert eine intuitive Erklärung der Formel

### ■ Random Surfer Modell:

- User startet auf einer beliebigen Seite im Web (z.B. Yahoo)
- User verfolgt beliebige Links (ohne Inhalte zu beachten)
- Irgendwann hört er die Linkverfolgung auf und beginnt auf neuer Startseite

### ■ Überlegungen:

- Wenn auf unsere Seite ein Link von Yahoo kommt, dann kommen mehr Leute zu uns, als bei einem Link von tante-emma.de
- Leute befinden sich unterm Strich häufiger auf wichtigen Seiten, als auf unwichtigen



Wahrscheinlichkeit auf Seite A zu landen ist:

Summe der Wahrscheinlichkeiten mit denen man von verlinkenden Seiten den Link zu A verfolgen kann (unter Beachtung von d)

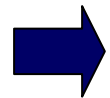
## 2.2 Random Surfer – Der PageRank gibt an mit welcher Wahrscheinlichkeit ein Random Surfer sich auf einer Seite aufhält

- **Wir erklären die Formel mit der Random Surfer Theorie:**

$$PR(A) = (1 - d) + d * [PR(T1) / C(T1) + ... + PR(Tn) / C(Tn)]$$

- **Erklärung für „PageRank“:**

- Je höher der PageRank einer Seite, desto wahrscheinlicher ist es, dass sich der User gerade auf dieser Seite befindet



PageRank ist Erwartungswert für Besuch einer Seite bei N (= Anzahl Seiten im Web) Anläufen

- **Erklärung für „ / C(Ti)“:**

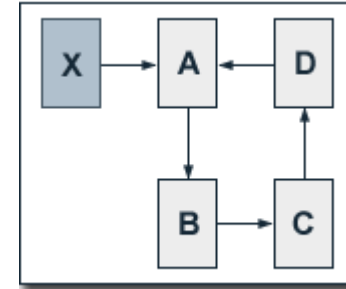
- Je mehr Links eine Seite hat, desto geringer wird die Wahrscheinlichkeit, dass der User genau den Link zu unserer Seite anklickt



## 2.2 Random Surfer – Der Dämpfungsfaktor gibt an wie geduldig ein Surfer Links verfolgt

- **Erklärung für „d“:**
  - Je länger die Linkverfolgung andauert, desto geringer wird die Wahrscheinlichkeit für weiteres Anklicken eines Links. User gelangweilt.
  - Hohes  $d$  => User verfolgt viele Links am Stück
  - kleines  $d$  => Surfer sehr ungeduldig, bricht schnell ab
- **Erklärung für „(1-d)“:**
  - Nachdem der User die Link-Verfolgung für die aktuelle Seite abbricht, ruft er eine beliebige Seite auf;  $(1-d)$  ist die Wahrscheinlichkeit dafür
  - Erweiterung: **Personalisierter PageRank** durch Multiplikation von  $(1-d)$  mit Vektor  $E(A)$  (enthält Gewichte über gesamten Webseiten)
- Länge eines Surf-Vorgangs ist **Exp-Verteilung** mit Mittel von  $d/(1-d)$

## 2.3 Probleme – Seiten, die nur untereinander verlinken, führen zu einer Endlosschleife; rank sink



### ■ Rank sink:

- Mehrere Seiten verlinken sich „in Reihe“ ohne eine Seite außerhalb zu verlinken
- Seiten würden PageRank von außen bekommen, aber niemals wieder nach außen abgeben
- Es würde eine Endlosschleife entstehen, da jede Seite ihren PageRank an die nächste überträgt, u.s.w (Entropiegedanke gestört)

### ■ Lösung

- Einführung von Quellen („*Rank source*“), also Seiten an denen dann wieder neu eingestiegen wird
- In der Formel realisiert durch *Dämpfungsfaktor*

## 2.3 Probleme – Dangling Links führen den Random Surfer in eine Sackgasse

- **Dangling Links**
  - Link der auf eine Seite zeigt, die keinen ausgehenden Link hat
  - Z.B. Seite des ausgehenden Links noch nicht indiziert, Indizierung verboten, Dateien (PDFs, Bilder, ...), E-Mails, ...
  - Random Surfer befindet sich in *Sackgasse*
- **Problem:**
  - Wohin soll das Gewicht dieser Seite gehen? (Entropiegrundsatz!)
- **Lösung:**
  - Löschung der Links vor PageRank-Berechnung, danach wieder hinzugefügt und PageRank-Weitergabe der ausgehenden Links angepasst

## 2.3 Probleme – Der PageRank liefert keine Aussage über die Qualität des Inhalts

- **Allgemeine Probleme:**

- Grundsätzlich keine Aussage über Qualität einer Seite und deren Inhalt
- Finanziell starke Seiten können eingehende Links kaufen (wird praktiziert)

## 2.4 Modifikationen – Die Summe aller PageRanks ist gleich der Anzahl an Seiten im Web

- **Ausgangsformel:**  $PR(A) = (1 - d) + d * [PR(T1)/C(T1) + ...]$ 
    - Gewichtung der Wahrscheinlichkeit des Besuchs einer Seite nach der Anzahl der Seiten des Webs
  - ➡ **PageRank = Erwartungswert** für den Besuch einer Seite (bei N Anläufen)
    - Summe aller PageRanks = Anzahl Seiten im Web
  - **Alternative Formel:**  $PR(A) = (1 - d) / N + d * [PR(T1)/C(T1) + ...]$
  - ➡ **PageRank = Tatsächliche Wahrscheinlichkeit** für Besuch einer Seite im Web
    - Summe aller PageRanks im Web wird 1 (d.h. irgendwo muss Surfer sein)
  - **Beispiel:**
    - PR2 Surfer landet 2x auf der Seite, bei 100 Versuchen bzw.  $P(A) = 2/100$
-

## 2.4 Modifikationen – Durch die Einführung eines Gewichtsvektors können spezielle Seiten einen Bonus bekommen

- **Bonus für bekannte Seiten**

- Z.B.: Yahoo, dmoz.org, Wikipedia
- Erwähnt in Patentschrift von google

- **PageRank Algorithmus wird um Vektor  $E(A)$  erweitert:**

$$PR(A) = (1 - d) * E(A) + d * [PR(T1)/C(T1) + ...]$$

$E(A)$  enthält Wahrscheinlichkeit mit der Seite A aufgerufen wird

- **Begründung:**

- Random Surfer wählt neue Startseite nicht beliebig, sondern greift auf Autoritäten o.ä. zurück

- **Welche Seiten sind diese Autoritäten?**

- Ermittlung evtl. über google-Toolbar Datensammlung

## 2.4 Modifikationen – Flexiblere Gewichtung einzelner PageRanks

- **Modifizierte Formel:**

$$PR(A) = (1 - d) + d * [PR(T1) / L(T1, A) + \dots]$$

- **Neue Variablen:**

- L: Funktion zur Bewertung des Links von Ti nach A

- **Fazit:**

- Gewichtung wird flexibler; vorher pauschal  $1/C(Ti)$
- Problem: Berechnungsaufwand steigt

## 2.5 Berechnung – Beispielhafte Berechnung des PageRanks eines Mini-Webs mit 3 Seiten

- **Annahmen:**

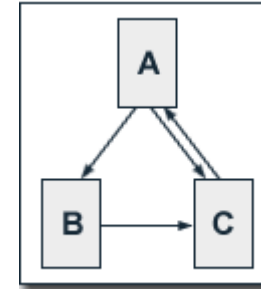
- 3 Seiten mit nebenstehender Linkstruktur
- Dämpfungsfaktor  $d = 0,5$

- **Gleichungssystem:**

- $PR(A) = 0.5 + 0.5 PR(C)$
- $PR(B) = 0.5 + 0.5 (PR(A) / 2)$
- $PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))$

- **Ergebnis:**

- $PR(A) = 14/13 = 1.07692308$
- $PR(B) = 10/13 = 0.76923077$
- $PR(C) = 15/13 = 1.15384615$
- **Summe aller PageRanks = 3**



## 2.5 Berechnung - Der PageRank wird iterativ berechnet

- **Iterative Berechnung in der Praxis:**
    1. Jeder Seite wird PR zugewiesen (z.B. PR1 für alle)
    2. PR wird in mehreren Runden ermitteltAbbruchkriterium: Konvergiert bzw. Summe aller PR  $\sim$  Anzahl Seiten im Web
  
  - **Theoretische Werte:**
    - Durchschnittlicher PR aller Seiten = ca. 1
    - Min. PR =  $(1-d)$
    - Max PR =  $(1-d) + d \cdot N$
  
  - **Laufzeit:**
    - Gutes Laufzeitverhalten, da sich Links im Web-Graphen schnell ausdehnen
    - Konvergiert in logarithmischer Zeit
    - Ca. 6 Minuten pro Schritt auf durchschnittlicher Workstation (Stand: 1998)
    - Ca. 100 Iterationen zum Berechnen des gesamten Webs mit ca. 150 Mio Seiten notwendig (Stand: 1998)
-

## 2.5 Berechnung – Zum iterativen Berechnen eines Webs mit 150 Mio Seiten, werden etwas über 100 Schritte benötigt

### ■ Iterative Berechnung des Beispiels:

- Iteration	PR(A)	PR(B)	PR(C)	Sum
- 0	1	1	1	3
- 1	1	0.75	1.125	2,875
- 2	1.0625	0.7656	1.14844	2,977
- ...				
- 11	1.07692307	0.76923077	1.15384615	~3
- 12	1.07692308	0.76923077	1.15384615	~3

## 2.5 Berechnung – Die komplette PageRank Berechnung dauert ca. 5h (Stand: 1998)

- **Kompletter Ablauf der Berechnung innerhalb Googles:**
  - Jede URL bekommt eindeutige ID
  - Link Struktur wird anhand von Parent ID sortiert
  - Dangling links entfernt
  - Initial assignment der PageRanks (Einfluss auf Geschw, nicht Ergebnis)
  - Konvergenz tritt ein
  - Dangling links hinzu, PageRanks neu berechnen
  - Dauer insgesamt ca. 5h, bei 75 Mio Links (550 / sec)
  
- **Im Vergleich zum Aufbau eines Volltext-Index sehr geringer Aufwand**

## 3. Der PageRank sorgt für eine verbesserte Sortierung der Suchergebnisse bei Google

- **3. Integration in Google**
  - Faktoren für die Bewertung einer Seite
  - Weitere, nicht öffentlich genannte Faktoren
  - Manipulationsmöglichkeiten

## 3.1 Faktoren – Neben dem PageRank sind auch Ankertext und IR-Kennzahlen wichtig

- **Offizielle Faktoren zur Sortierung der Suchergebnisse:**
  1. IR-Kennzahlen / seitenspez. Faktoren (Inhalt, Term-Frequenz / Keyword-Dichte, Formatierung, ...)
  2. Ankertext eingehender Links (Name des Links auf externen Seite)
  3. PageRank
  4. „Proximity“: Nähe der Wortvorkommen bei Suchanfragen mit mehreren Worten („Bill Clinton“ liefert nicht „Bill Smith mag John Clinton“)

## 3.1 Faktoren – Algorithmus zur Ermittlung und Sortierung der Suchergebnisse

### 1. Ermittlung einer sog. „IR-Score“ aus IR-Maßen und Ankertext

1. Vorkommen des Suchworts im Dokument werden bestimmten Typen zugeordnet (Titel, Anker, URL, Überschrift-Tag, ...)
2. Treffer in Bezug auf Typen werden linear durchgezählt („*hit counts*“), jedoch mit oberer Schranke und in Vektor gespeichert (1, 0, 1, 4, ...); „*count-weights*“
3. Dieser Vektor wird mit einem Gewichtsvektor der Typen („*type-weights*“) skalarmultipliziert und ergibt die *IR-Score*
4. Bei Suchanfragen mit mehreren Worten wird Nähe (Proximity) berücksichtigt

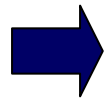
### 2. *IR-Score* wird nun mit *PageRank* multiplikativ verknüpft

- => finaler Rang der Seite im Ergebnis

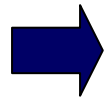
- Reihenfolge durch Kombination von Faktoren allgemeiner Bedeutung und suchanfragespezifischer Bedeutung

### 3.1 Faktoren – Zur Sortierung der Suchergebnisse wird der PageRank mit suchanfragespez. Kriterien multiplikativ verknüpft

- **Bedeutung des PageRanks für Suchergebnisse:**
  - PR hat Einfluss auf Reihenfolge der Ergebnisse
  - **Aber es wird nicht anhand des PageRanks sortiert!**



Je spezifischer eine Suchanfrage ist, desto weniger Einfluss hat der PageRank



Desto unspezifischer eine Anfrage ist (z.B. „Film“), desto höher der Benefit der PageRank-Nutzung („*common case*“)

### 3.1 Faktoren – PageRank versucht neben detaillierten Expertenseiten auch common case Ergebnisse zu liefern

- **„Common case“ Ergebnisse:**
  - Ausgabe von ganz allgemeinen Seiten zu einem Thema
  - Wichtiges Ziel bei Entwicklung des PageRanks
  - Klassische IR Verfahren liefern tendenziell Expertenseiten mit viel detaillierten Informationen
  - Kunst liegt darin keinen Ergebnistyp zu „bevorzugen“
  
- **Beispiel-Suchanfrage: „Film“**
  - Anfrage liefert kommerzielle Seite, lediglich mit Inhalt „Filme kaufen“
  - Aber auch ein detailliertes Nachschlagewerk und Portal
  - Es gibt für beide Extreme Bedarf

## 3.2 Weitere Faktoren – Empirisch ermittelte Faktoren auf Basis eines Wettbewerbs von heise online

### ■ Off-the-Page-Faktoren

- URL sehr wichtig (Keywords in URL!)
- Autoritäten bevorzugt (Yahoo, WikiPedia, ...)
- Anzahl Backlinks von untersch. Domains / IPs / Class C Netzen
- Regionale Faktoren; Herkunft des Backlinks entscheidend, unabh. vom PageRank (Sprache!)
- Alte, eingesessene Sites bevorzugt
- Aktualität und Häufigkeit der Aktualisierung
- Traffic wichtig (über Toolbar)

## 3.2 Weitere Faktoren – Empirisch ermittelte Faktoren auf Basis eines Wettbewerbs von heise online

- **On the page**
  - Ca. 5-10% Keyworddichte im Text
  - Qualität / Sinn der Inhalte unwichtig
  - Valides HTML
  - Inhalt einzigartig
  
- **Gewichtungen zwischen on- und off-the-page-Faktoren:**
  - Verlinkungen wichtiger als UI und seitenspezifische Aktualisierungen

## 3.2 Weitere Faktoren – Faktoren, die von Abakus Online Marketing vermutet werden

- **Off-the-Page-Faktoren**
  - **Viele Seiten!**
  - **Viele Links! - Am besten von Expertenseiten**
  - Interne Linkstruktur; viele interne Links! (SiteMaps, FooterLinks, ...)
  - Thematische Links bevorzugt (Themenbasierter PageRank ?)
  - *PageRank immer unwichtiger*
  - On-the-page-Faktoren unwichtig (z.B. Meta-Tags)
  
- **Vermutete neue Implementierungen:**
  - Hilltop
  - Theming
  - LocalRank

### 3.3. Manipulationsmöglichkeiten – Linkfarmen versuchen durch viele eingehende Links den PR zu verbessern

- **Manipulationsmöglichkeiten:**

- Linkfarmen
- Cloaking
- Vererbung
- Gästebücher / Foren
- 302 Hijacking
- ...

- **Populär: Linkfarmen**

- Idee: Jeder eingehende Link erhöht den PageRank
- => Man produziert selbst tausende Links auf die eigene Seite
- Google hat PageRank diesbzgl. angepasst

## 4. Eingehende Links haben starke positive Effekte auf den PageRank

- **4. Effekte von Links und Seiten auf de PageRank**
  - Effekte ausgehender Links
  - Effekte eingehender Links
  - Weitere Effekte zwischen verlinkten Seiten

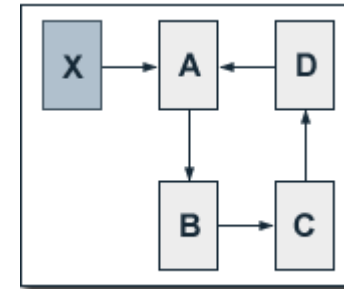
4

## 4.1 Eingehende Links - Erhöhen den PageRank

- **Wiederholung:**
  - Jeder eingehende Link erhöht PageRank
- **Vermutung:**
  - Erhöhung des PageRanks um „ $d \times PR(X) / C(X)$ “
- **Aber:**
  - Seiten ausgehender Links profitieren auch vom höheren PageRank
  - => PageRank-Erhöhung muss teilweise an sie weitergegeben werden

## 4.1 Eingehende Links – Je höher der Dämpfungsfaktor, desto stärker überträgt sich der PR auf weitere Seiten

- Ohne Link von X: PR1 für alle Seiten
- Nun eingehender Link von X (PR10) nach A:
- Rechnung:
  - $PR(A) = 0.5 + 0.5 (PR(X) + PR(D)) = 5.5 + 0.5 PR(D)$
  - $PR(B) = 0.5 + 0.5 PR(A)$
  - $PR(C) = 0.5 + 0.5 PR(B)$
  - $PR(D) = 0.5 + 0.5 PR(C)$

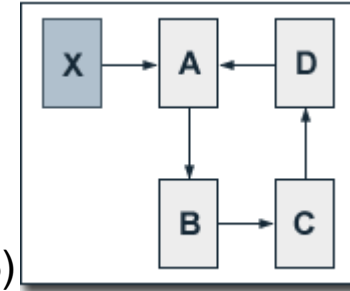


- | Ergebnis:               | Vermutung | $d = 0,75$ | Vermutung |
|-------------------------|-----------|------------|-----------|
| - $PR(A) = 19/3 = 6.33$ | = 6       | = 11,97    | = 7,5     |
| - $PR(B) = 11/3 = 3.67$ | = 1       | = 9,23     | = 1       |
| - $PR(C) = 7/3 = 2.33$  | = 1       | = 7,17     | = 1       |
| - $PR(D) = 5/3 = 1.67$  | = 1       | = 5,63     | = 1       |
- => Effekt von  $d \times PR(X) / C(X) = 5$  setzt sich unter den Seiten fort!
  - => Je höher  $d$ , desto höher der Effekt und desto gleichmäßiger!

## 4.1 Eingehende Links – Der aufaddierte PageRank im verlinkten System erhöht sich um $(d/(1-d)) * (PR(X)/C(X))$

- **Betrachtung der Summe der PageRanks:**

- 4 bei Ausgangssituation, ohne Link von X
- 14 bei  $D=0,5$
- => Der PageRank10 der Seite X wird aufgeteilt (  $0,5 / 0,5$  )
  
- 34 bei  $D=0,75$
- => Der PageRank wird um 30 erhöht (  $(0,75 / 0,25) * 10$  )



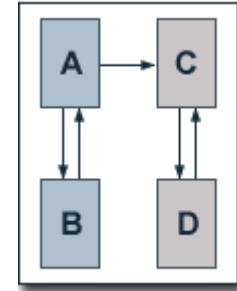
- **Ergebnis:**

- Erhöhung des aufaddierten PageRanks um:  $(d / (1-d)) \times (PR(X) / C(X))$
- Bedingung: Verlinkung in geschlossenes System (also nur theoretisch)

## 4.1 Eingehende Links - Erhöhen den PageRank stärker als zu vermuten

- **Herleitung:**
  - Wiederholung: Länge eines Random Surf Vorgangs ist *exponentialverteilt*, mit einem Mittel von  $d/(1-d)$
  - => Surfer besucht im Schnitt genau  $(d/(1-d))$ -Seiten
  - Soviel mehr PageRank muss dann auch übertragen werden
- **Fazit: Eingehende Links mit sehr starkem Einfluss**

## 4.2 Ausgehende Links - Verringern den PageRank



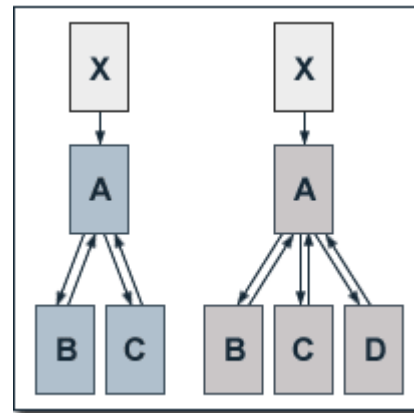
- **Beachte: Summe des PageRanks im Web konstant 1**
  - => Wenn Seiten gewinnen, dann müssen andere Seiten verlieren!
  - => Wenn eingehende Links Effekte haben, dann auch ausgehende!
- **Ausgehende Links verringern den eigenen PageRank**
  - Verlust ist der Gewinn des anderen Systems, wo Link hinführt
- **Erklärung im Random Surfer Modell:**
  - Externe Links erhöhen die Wahrscheinlichkeit, dass Random Surfer eigene Seite verlässt

## 4.2 Ausgehende Links – Der Effekt ausgehender Links ist diskussionswürdig

- **Macht das Sinn?:**
  - Mathematisch gesehen notwendig
  - Praktisch gesehen wird jedoch eine WebSite durch externe Links aufgewertet,
  
- **Annahme:**
  - Google belohnt in anderer Form ausgehende Links
  - Sonst wären Seiten, die pauschal alle ausgehenden Links löschen bevorteilt, wodurch sich das PR-Verfahren eigenen Ast abschneiden würde

## 4.3 Weitere Effekte – Eine Erhöhung der Anzahl an Unterseiten lässt den PageRank der Hauptseite steigen

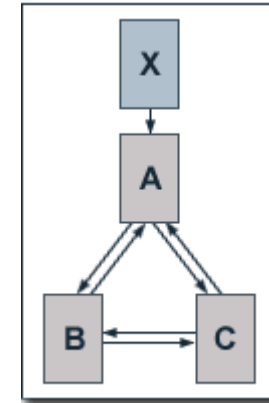
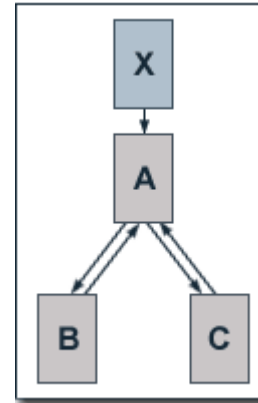
- **Wiederholung:**
  - Aufaddierter PageRank aller Seiten im Web = Anzahl Seiten im Web
- => **Jede zusätzliche Seite erhöht PageRank um 1**
- **Beispiel:**
  - PR(A) steigt
  - PR(B, C) fällt



## 4.3 Weitere Effekte – Konzentration ausgehender Links auf eine Seite erhöht den PageRank aller Seiten

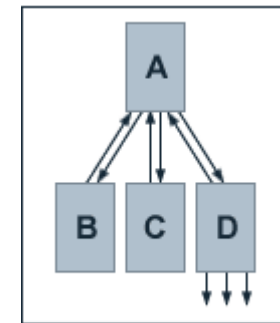
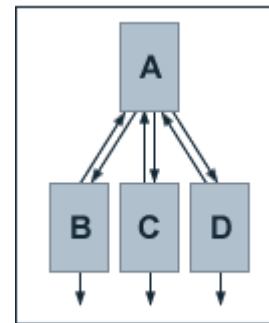
- **Verlinkung untereinander**

- PR(A) fällt
- PR(B) und PR(C) steigen



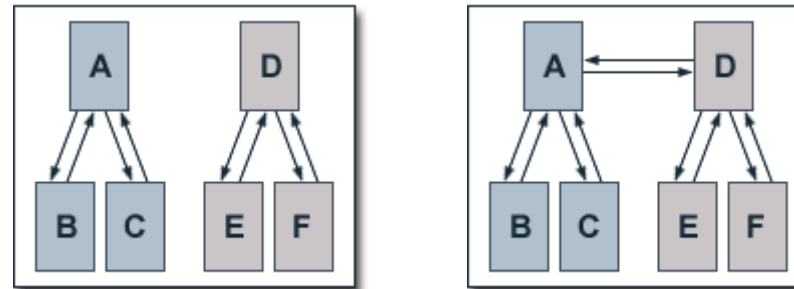
- **Konzentration ausgehender Links:**

- Alle PR steigen



## 4.3 Weitere Effekte – Ein Linkaustausch verbessert alle beteiligten Seiten

- **Linkaustausch**
  - Alle PR steigen



## 5. Der PageRank von morgen basiert auf einem personenbezogenen PageRank

- **5. Ausblick [5-10 Min]**
  - Personalisierter PageRank
  - Themenbasierter PageRank („topic sensitive PageRank“)
  - Modularer PageRank
  - Hilltop-Algorithmus
  - Anwendungsmöglichkeiten

# 5.1 Personalisierter PageRank – Berücksichtigt individuelle Interessen

- **Wiederholung:**
  - Random Surfer springt nach gewisser Zeit auf eine *zufällige* Startseite
  
- **Idee:**
  - Diese Seite ist nicht zufällig, sondern personenbezogen
  - Gewicht jeder Seite des Webs kann über E-Vektor gesteuert werden
  - Extremfall: Es wird nur eine einzige Startseite zugelassen
  
- **Mögliche Kriterien zur Ausfall der Startseiten:**
  - Nur Startseiten aus Bookmarks des Users zugelassen
  - Startseitengewinnung aus früheren Suchanfragen des Users
  - Vom User ausgewählte Ergebnisse
  - Vom User gegebene „persönlichen Informationen“
  
- => Diese Seiten haben nun **für den jeweiligen User** einen viel höheren PR

## 5.1 Personalisierter PageRank ist bereits als Beta-Version verfügbar

- **Vorteile:**

- Treffer aus persönlichen Interessen eines Users können übergewichtet werden
- Insbesondere bei doppeldeutigen Suchbegriffen wirkungsvoll
- Kaum anfällig gegen Manipulationen

- **Problem:**

- Hohe Berechnungsdauer, denn eine Seite hat nun nicht mehr nur einen einzigen PageRank, sondern für jeden Surfer einen individuellen PageRank

- **Personalisierter PageRank in der Praxis**

- Patent (u.a.) im Juli 2004 von google eingereicht
- Betaversion verfügbar: [www.google.com/psearch](http://www.google.com/psearch)
- Demnächst wohl in google integriert

## 5.2 Themenbasierter PageRank – Berücksichtigt den thematischen Bezug eines Surf-Vorgangs

- **Themenbasierter PageRank**
  - Themen der Seiten eingehender Links bzw. ausgehender Links haben Einfluss auf PageRank Berechnung
  - Random Surfer wird eher eine *themenverwandte* neue Startseite aufrufen, als eine *zufällige* neue Startseite
- **Einsatz in der Praxis:**
  - Denkbare Einteilung in 16 größte Themen (Gesundheit, Sport, Shopping, ... wie bei dmoz.org) => „nur“ 16 PageRanks pro Seite
  - Bei der PR-Berechnung für Gesundheit erhalten Seiten, die von dmoz.org: Gesundheit verlinkt sind ein hohes E-Vektor-Gewicht
  - => Vereinfachte Form des personenebezogenen PageRanks
  - Vermutlich bereits teilweise implementiert

## 5.3 Modularer PageRank - Lässt nur die wichtigsten Seiten des Webs als Startseite zu

- **Modularer PageRank**
  - Es werden als mögliche Startseiten nur die wichtigsten Seiten des Webs zugelassen
  - Einfache Form des personalisierten PageRanks realisierbar durch Gewichtung dieser Seiten anhand persönlicher Informationen

## 5.4 Hilltop-Algorithmus – Berücksichtigt nur eingehende Links, die von Expertenseiten kommen

### ■ Hilltop Algorithmus

- 1999 von Bharat und Mihaila an der Universität Toronto entwickelt
- 2003 erwarb Google das Patent am Algorithmus
- Google nutzt vermutlich seit Jan 2004 den Hilltop-Algorithmus

### ■ Prinzip:

- Bewertung eingehender Links nur von sog. „**Expertenseiten**“
- Expertenseiten sind Seiten mit Links *eines Themas* zu *vielen unabhängigen* Seiten (z.B. redaktionell gepflegte Verzeichnisse)
- Hat eine Seite *mehrere* eingehende Links von den *besten* Expertenseiten, so wird die Seite zur „**Autorität**“ (z.B. Wikipedia)
- Google listet solche Autoritäten sehr weit oben
- Problem: Einige große Seiten werden übergewichtet

## 5.5 Anwendungen - Die Google Toolbar zeigt den PageRank von Seiten im Browser und sammelt personenbezogene Daten

- **Toolbar zeigt PageRank Werte von 0 bis 10**
  - **Dazu werden die berechneten Werte logarithmisch skaliert**
  - **Vermutungen:**
    - Seite mit höchstem PageRank bekommt PR11
    - Alle anderen Seiten werden entsprechend skaliert
    - logarithmische Skalierung zur Basis 6,x
  
  - **Populäre Beispiele:**
    - PR10: apple.com
    - PR9: [microsoft.com](http://microsoft.com)
    - PR8: Uni-karlsruhe.de
    - PR7: Uni-mannheim.de
    - PR6: AIFB
  
  - **Download: [toolbar.google.com](http://toolbar.google.com)**
-

## 5.5 Anwendungen – Der PageRank eignet sich zum Schätzen von Traffic und zukünftigen Backlinks

- **Traffic schätzen**
  - Traffic korreliert mit PageRank
- **Backlink Predictor**
  - Zählen der Links auf einen Physiker lässt auf Wahrscheinlichkeit für Nobelpreisgewinn schließen
- **Navigationshilfe**
  - Anzeige des PR neben jeden Link
  - User kennt Wichtigkeit der verlinkten Seiten schon vor dem Klick

**... wurden Brin und Page in nur 8 Jahren zu den reichsten Menschen der Welt.**

- **Seit dem Erscheinen des Papers sind nun 8 Jahre vergangen**
- **Sergey Brin und Larry Page sind Hauptaktionäre der Google Inc. mit Sitz in Mountain View, California, USA**
- **Googles Wert an der Börse beträgt 123,07 Mrd USD, das ist ein höherer Wert als BMW (27), DaimlerChrysler (45,2), Porsche (7) und VW (19,6) zusammen haben**