Information Retrieval

09 - Retrieval im Web: Crawling und Link-Analyse

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-05-27

Technology
Arts Sciences
TH Köln

Das World Wide Web

Web-Retrieval ist anders, als bisheriges IR...

- **Unklare Anzahl** an Dokumenten (Größe des Web = Größe Suchmaschinenindex?)
- Verschiedene **Formate** (html, doc, pdf, jpeg, mp3…)
- Verschiedene **Strukturen** (Artikel, Tabellen, Blogs, etc…)
- Verschiedene **Textqualitäten**
- Verschiedene **Vokabulare**
- **SPAM**
- Duplikate, **Redundanzen**
- **Hyperlinks**
- **Dynamisch** erstellte Inhalte („Deep web“, „Inivisible web“)
- Was ist ein Dokument? → **Site, page, section**?

Web-IR vs. „klassisches“ IR

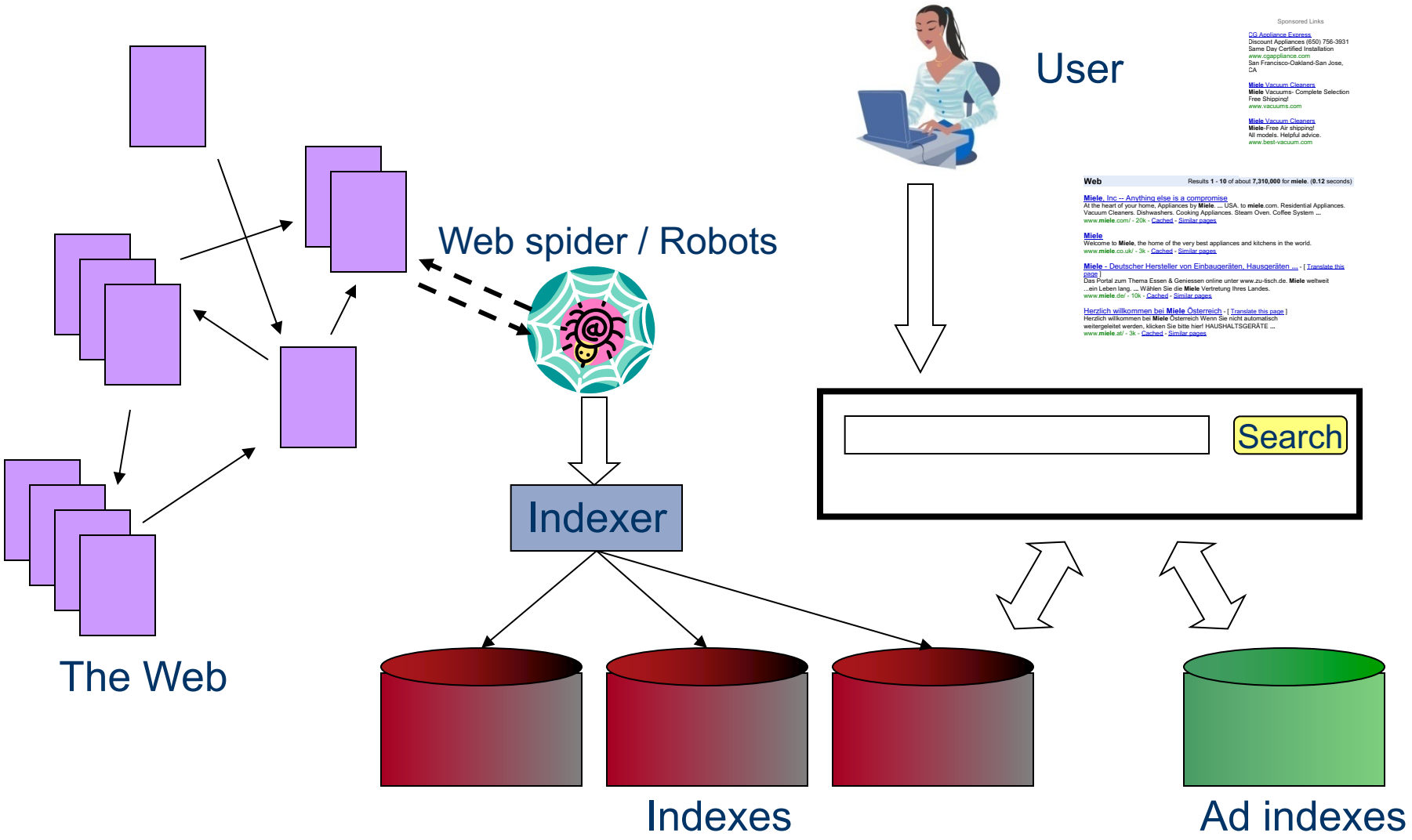
	Web-IR	„klassisches“ IR
I. Dokumente		
Sprachen	Viele	i.d.R. einheitliche Indexierungssprache
Formate	Viele	i.d.R. ein Format
Länge	Unterschiedlich	Bibliograph. DB: in etwas gleich
Teile	Unterschiedlich (Bilder, Anker…)	Genau ein Datensatz
Verlinkung	Hyperlinks	Ggf. Referenzen & Zitationen
Spam	Ja	Nein
Struktur	Schwach	Feldstruktur
Inhalt	Heterogen	homogen

Nach Lewandowski, D. (2005): Web Information Retrieval. Technologien zur Informationssuche im Internet. Frankfurt: DGI:

Web-IR vs. „klassisches“ IR

	Web-IR	„klassisches“ IR
II. Grundgesamtheit		
Größe	Unbekannt	(in etwa) bekannt
Abdeckung	Nicht messbar	(in etwa) messbar
Duplikate	Ja	Nein
III. Nutzer		
Zielgruppe	Alle Web-User	i.d.R. Fachexperten
Bedarf	Unterschiedlich	Fachbezogen
Kenntnisse	Gering	Hoch – sehr hoch
IV. IR-System		
Interface	Einfach	Oft (sehr) komplex
Funktionalität	Gering	Hoch
Ranking	Ja, basierend auf Linkanalyse	Ja, z.B. Vektorraummodell

Prinzip Web-Suchmaschinen



Allgemein: Crawling

Datenbank der Suchmaschinen-Betreiber

- **Index**, so wie Sie ihn auch als invertierte Liste kennen!

Andere Namen für **Robots**:

- **Web Spider**, **Crawler**, z.B. **GoogleBot**
- **Robots durchsuchen das Web**.

Der Robot:

- findet **neue Dokumente**
- überprüft vorhandene Dokumente auf **Existenz**
- überprüft vorhandene Dokumente auf **Veränderung**

Der Robot hangelt sich dabei **von Link zu Link auf Webseiten** und findet von sich aus keine Dokumente, die nicht mit anderen verlinkt sind.

Allgemein: Crawling

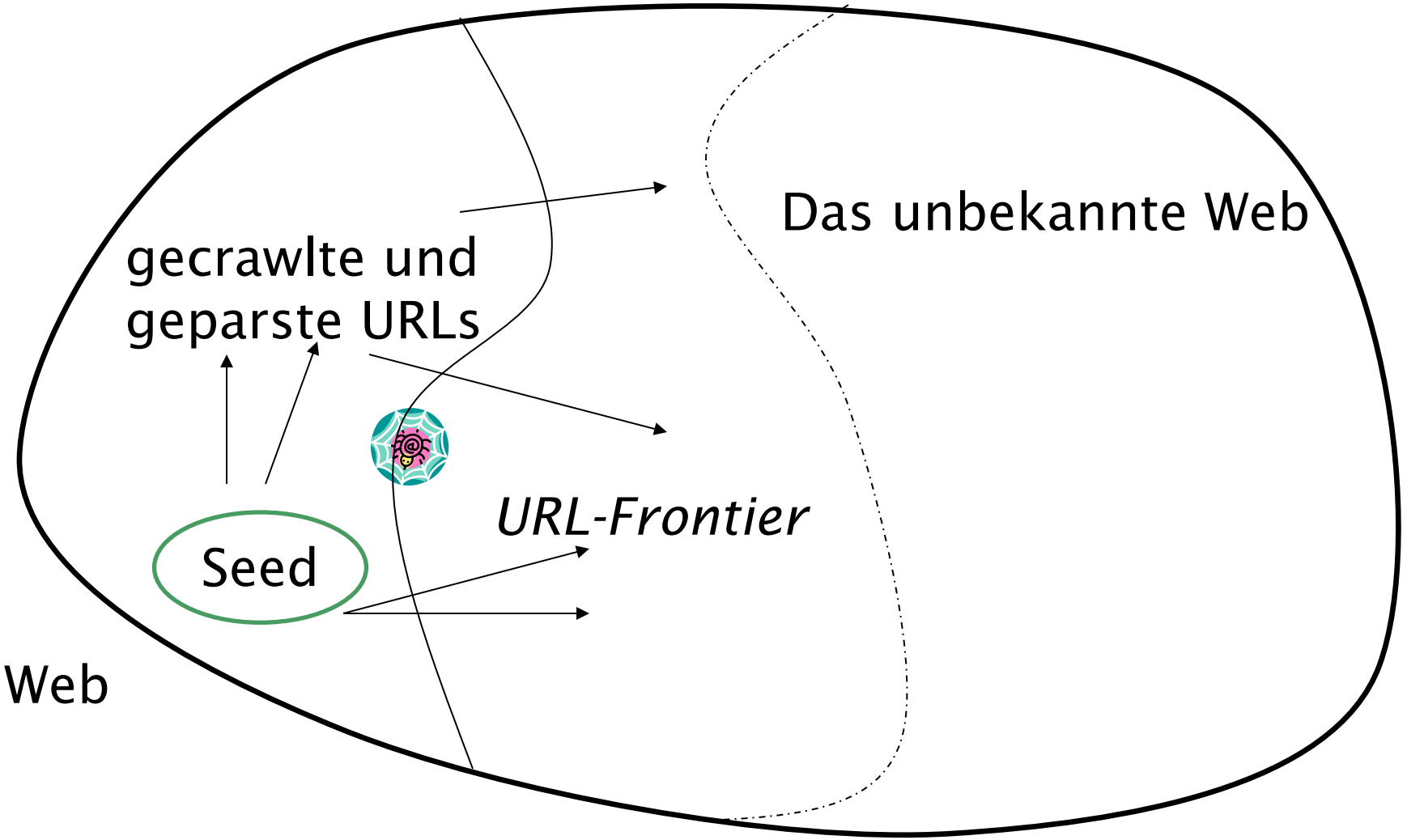
Allgemeines Vorgehen

1. Die Robots speichern eine Kopie der gefundenen Website (Startseite inklusive Unterseiten).
2. Jeder katalogisierten Seite wird eine Nummer zugeordnet.
3. Analysiert und aufgenommen werden in der Regel der Volltext, Dateinamen von Bildern, Seitentitel (Title Tag).

Grundlegende **Crawler-Operationen**

1. Beginne mit einer bekannten Menge von URLs (“seed”)
2. Lade und verarbeite die URLs
   - Extrahiere URLs auf die verwiesen wird
   - Platziere extrahierte URLs innerhalb der Warteschlange (“queue”)
3. Lade jede URL in der Warteschlange und wiederhole Ablauf

Web-Crawling verbildlicht



Komplikationen

Web-Crawling ist nicht machbar auf nur einem System

- Jeder der vorher genannten Schritte wird verteilt ausgeführt

„**Bösartige**“ Seiten

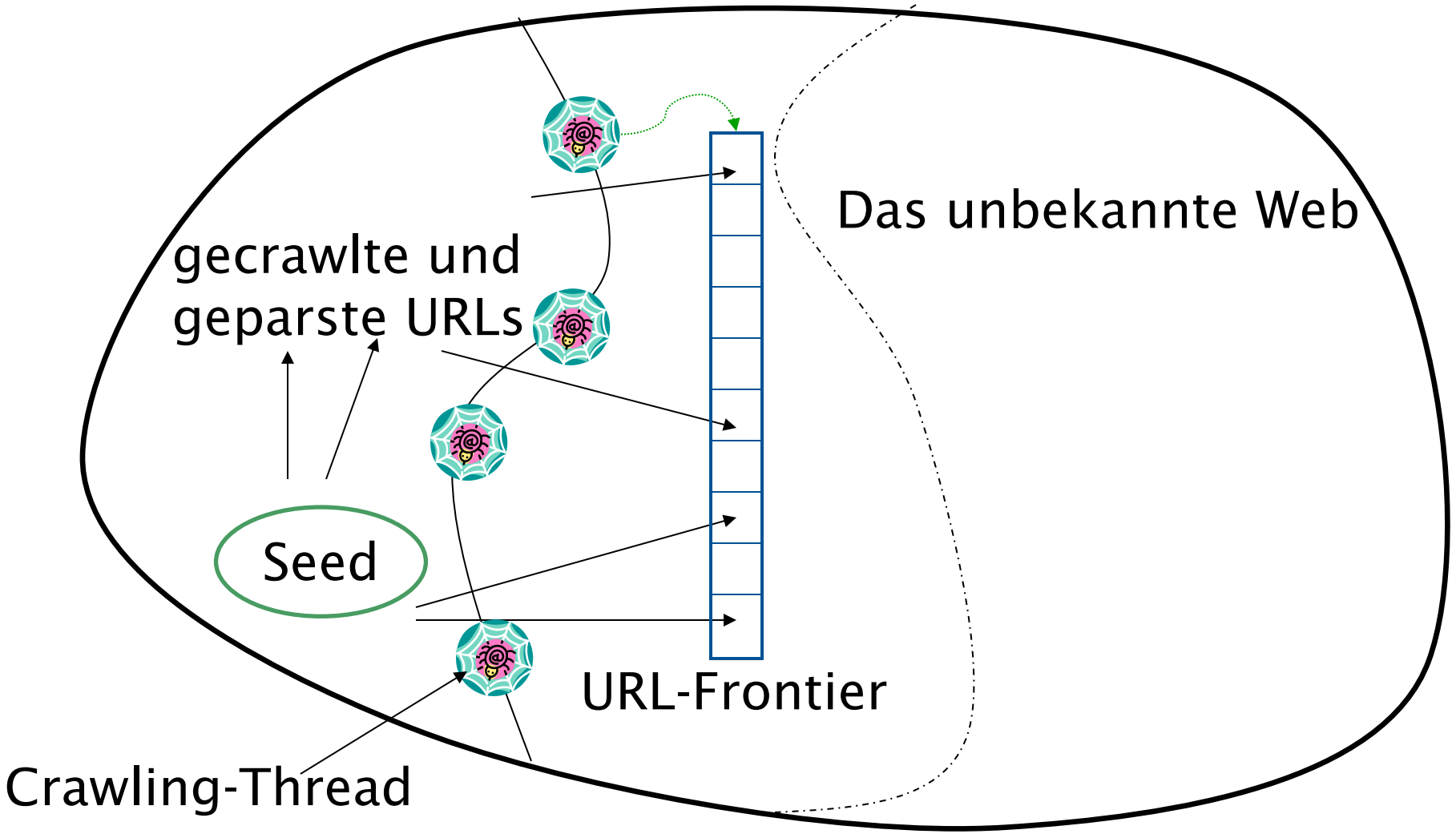
- Spam-Seiten
- „Spider-Fallen“– inkl. dynamisch generierter Inhalte

Auch „gutartige” Seiten sind **herausfordernd**

- Antwortzeiten/Bandbreite des gecrawlten Servers sind verschieden
- Bestimmungen des Webmasters
  - Wie tief soll eine Website-Hierarchie gecrawlt werden?
- Website Spiegelungen und Duplikate

Höflichkeit – ein Server darf nicht zu oft angefragt werden

Web-Crawling verbildlicht (Update!)



„Klassisches“ IR im Web?

Prinzipiell geht das... Frühe Suchmaschinen arbeiteten z.B. mit Termgewichten wie **TF-IDF im Web**.

Nachteile:

- Diese Ansätze sind sehr **spamanfällig** und
- die **Größe des Web** ist ein Problem,
- genauso wie das **Sprachproblem**!

Es brauchte **alternative Ansätze** um den besonderen Problemen im Web Herr zu werden!

PageRank als Schlüssel zum Web

The PageRank Citation Ranking: Bringing Order to the Web.

Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report. Stanford InfoLab.

http://ilpubs.stanford.edu:8090/422/

Hintergrund: Hypertext und Links

Wir schauen **hinter den Inhalt** von Web-Dokumenten

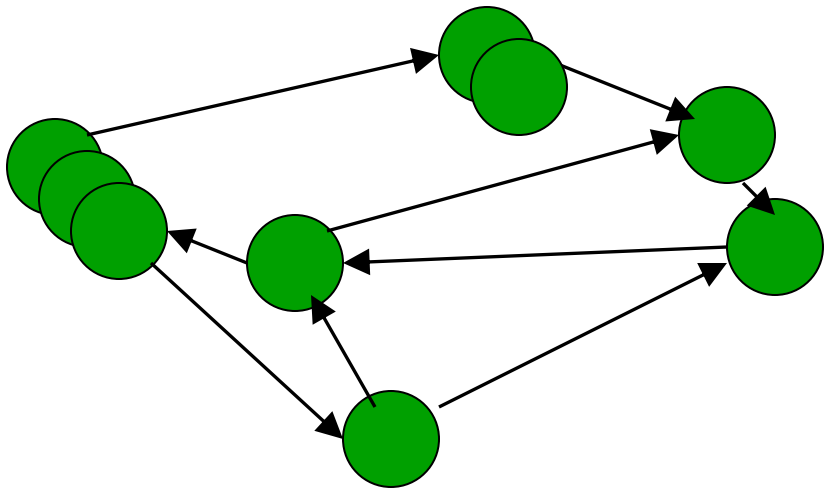
- Uns interessieren nun vielmehr die **Hyperlinks** zwischen ihnen

Dies wirft Fragen auf…

- Zeigen Links zwischen Webseiten bzw. Links auf eine bestimmte Webseite die **Wichtigkeit einer Webseite**?
- Können wir diese Information für das **Ranking** verwenden?

Diese Fragen finden Anwendung

- Im Web
- Bei Email
- In Social-Networks
- …

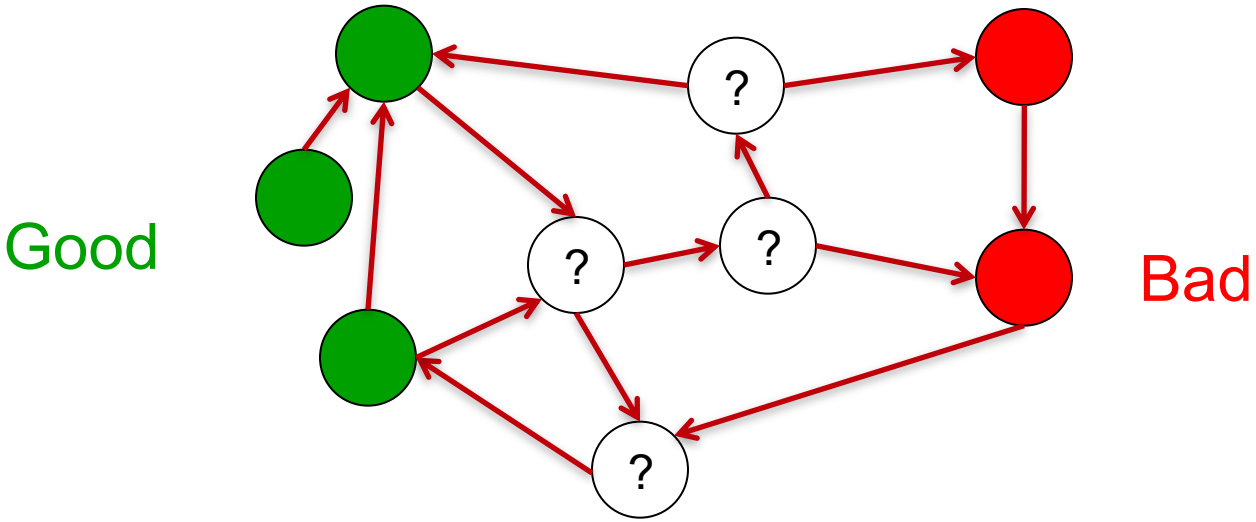


Links sind überall

Sehr mächtige Quellen für **Glaubwürdigkeit** und **Einfluss**

- Mail-Spam – Welche Email-Accounts sind Spammer?
- Anbieter-Qualität – Welche Anbieter sind „böse“?

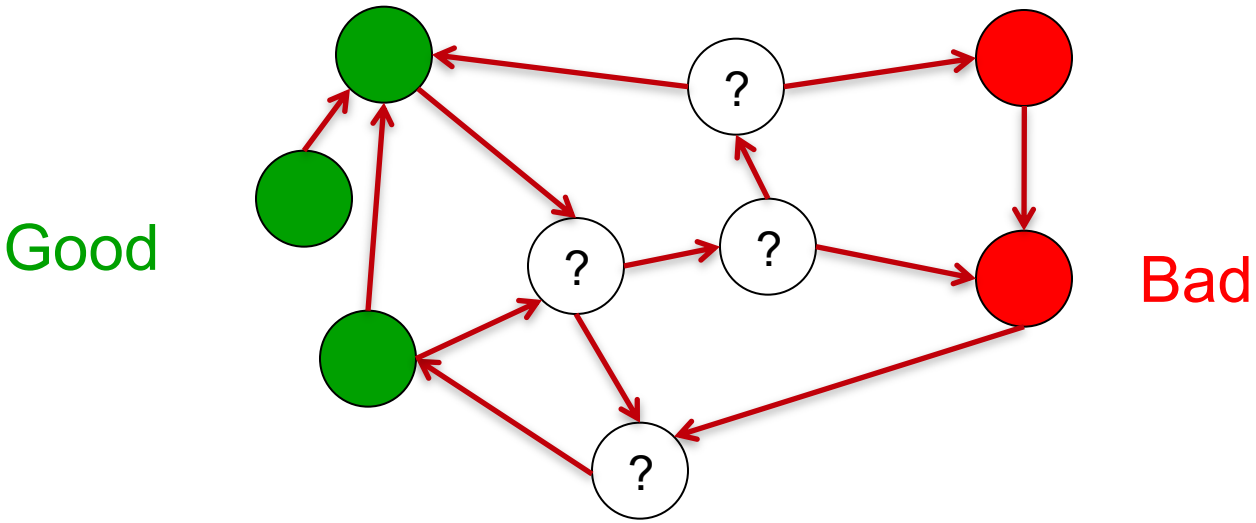
Ein Beispiel: The Good, The Bad and The Unknown



Einfache interative Logik

The Good, The Bad and The Unknown

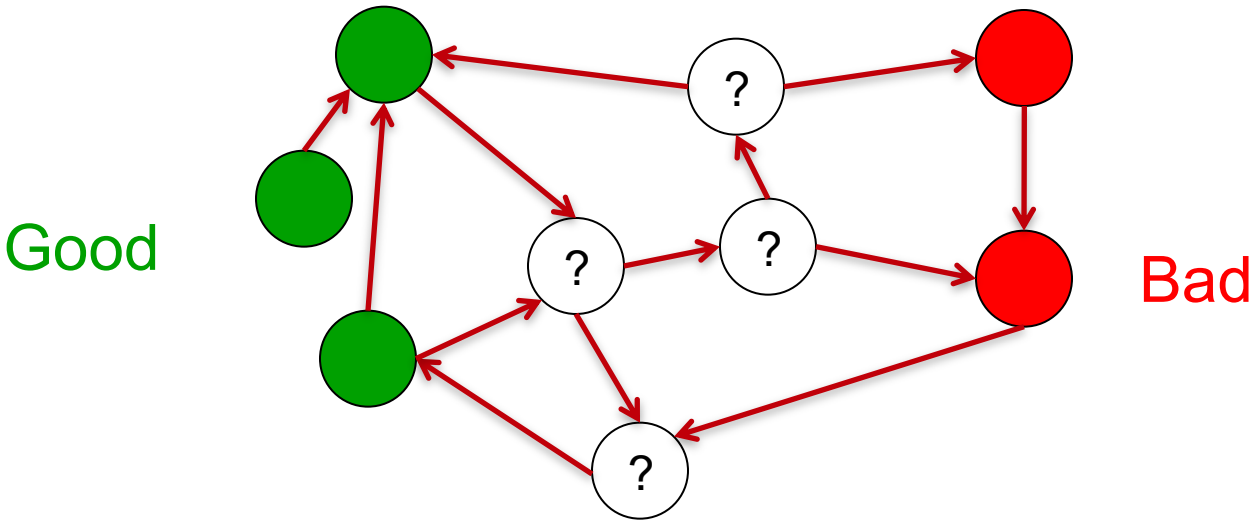
- Gute Knoten zeigen nicht auf böse Knoten
- Alle anderen Kombinationen sind möglich



Einfache interative Logik

Gute Knoten zeigen nicht auf böse Knoten

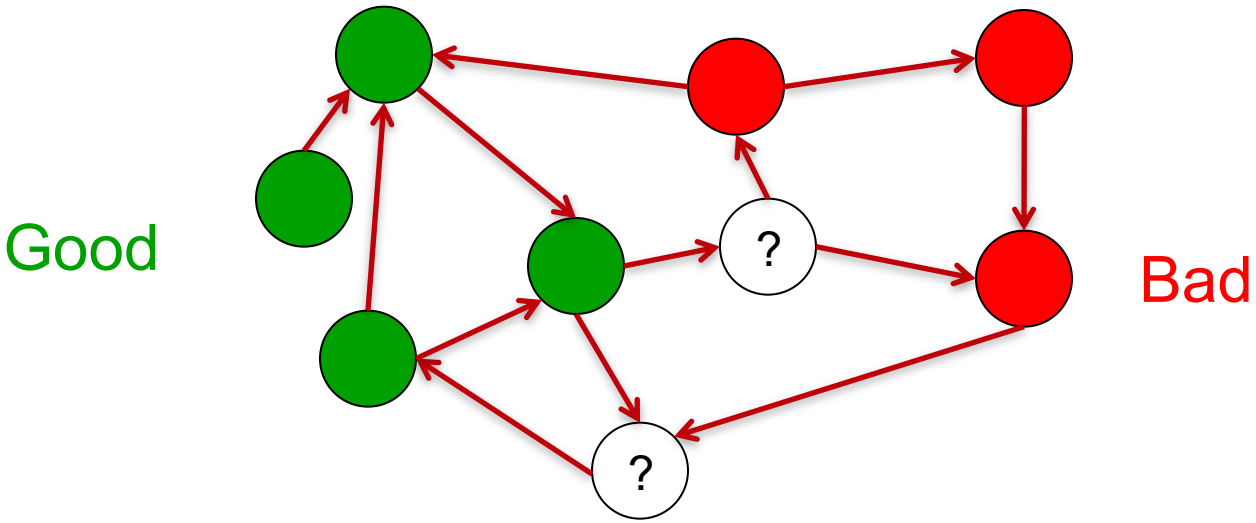
- Wenn du auf einen bösen Knoten zeigst, bist du böse
- Wenn ein guter Knoten auf dich zeigt, bist du gut



Einfache interative Logik

Gute Knoten zeigen nicht auf böse Knoten

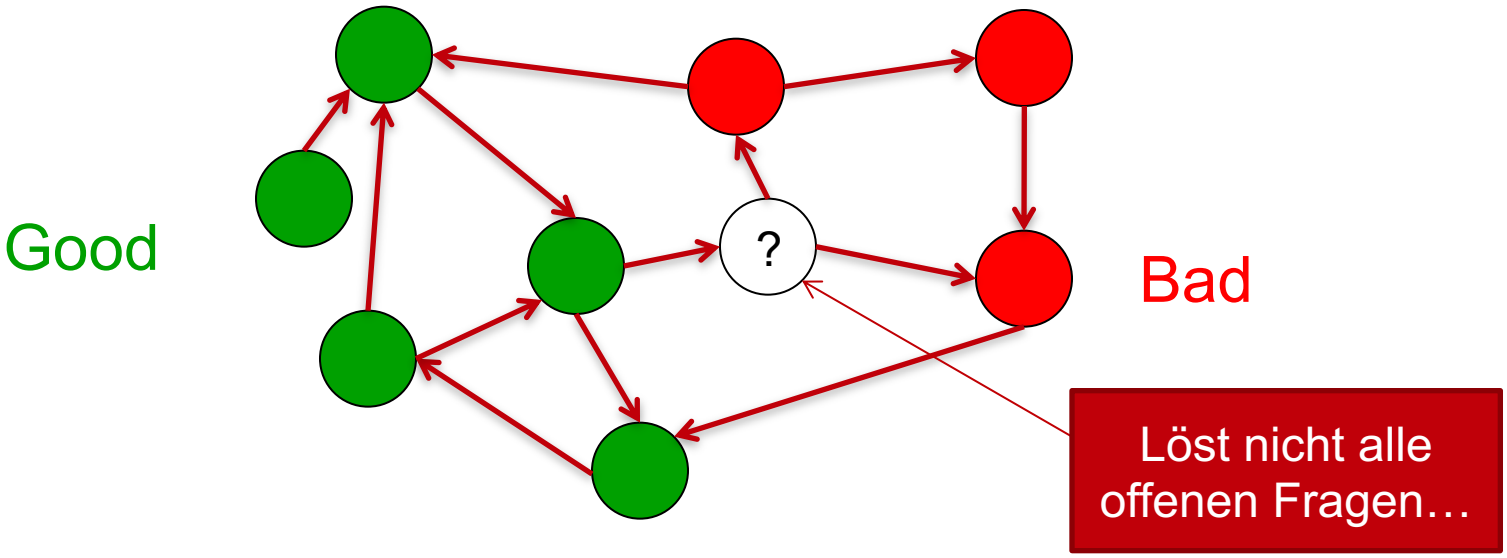
- Wenn du auf einen bösen Knoten zeigst, bist du böse
- Wenn ein guter Knoten auf dich zeigt, bist du gut



Einfache interative Logik

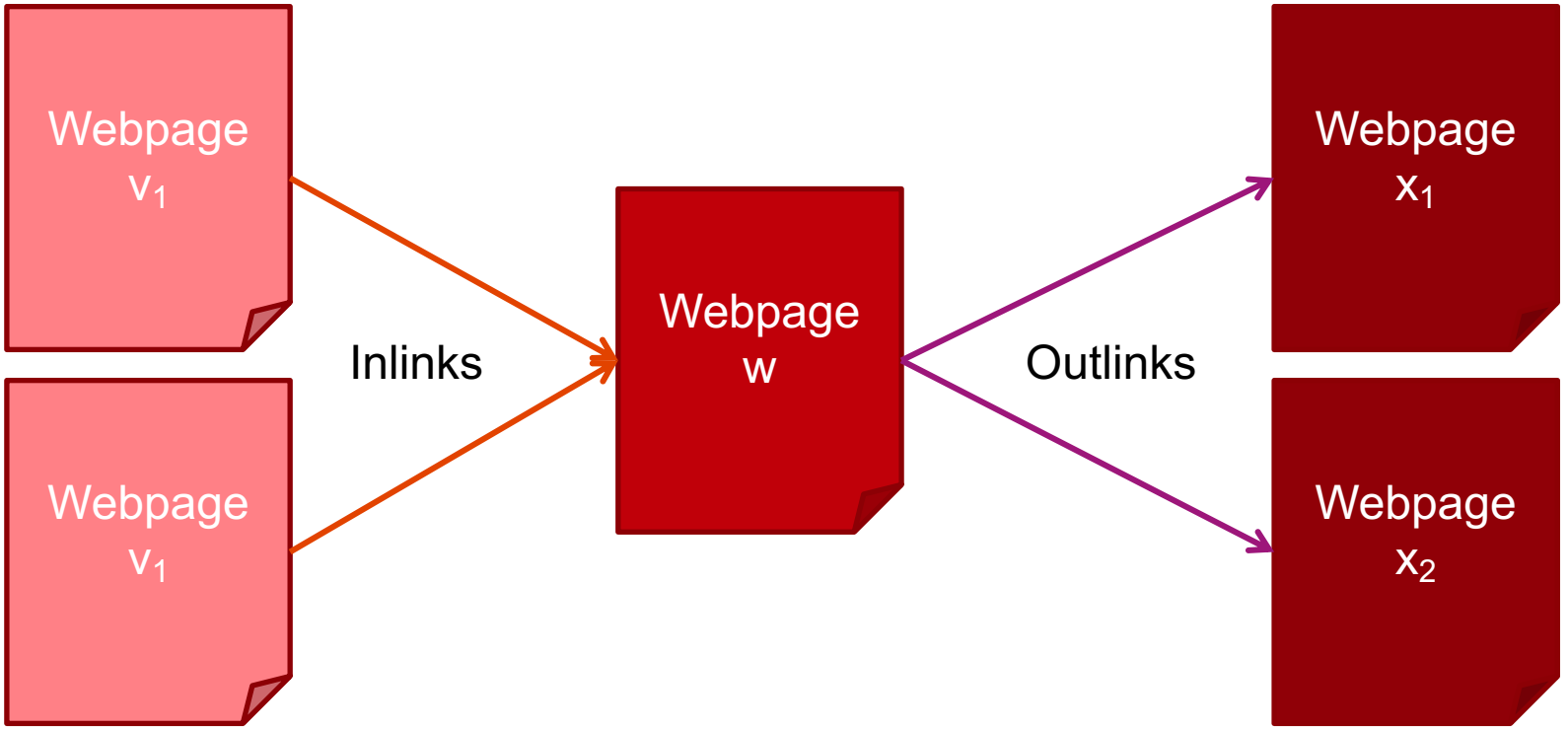
Gute Knoten zeigen nicht auf böse Knoten

- Wenn du auf einen bösen Knoten zeigst, bist du böse
- Wenn ein guter Knoten auf dich zeigt, bist du gut



PageRank: Die zentrale Idee (1)

- Jede Webpage hat eine Anzahl an **ausgehenden Links** (*Forward Links, Outlinks*) und eine Anzahl an **eingehenden Links** (*Backlinks, Inlinks*).



PageRank: Die zentrale Idee (2)

Webpages unterscheiden sich stark in der **Anzahl ihrer Inlinks**.

- So hat bspw. die Webseite www.spiegel.de/index.html mehr als 14 Millionen Inlinks*
- Viele andere Webseiten besitzen nur wenige Inlinks.

Die Annahme ist, dass diese **Seiten mit vielen Inlinks wichtiger sind**, als diese mit wenigen Inlinks…

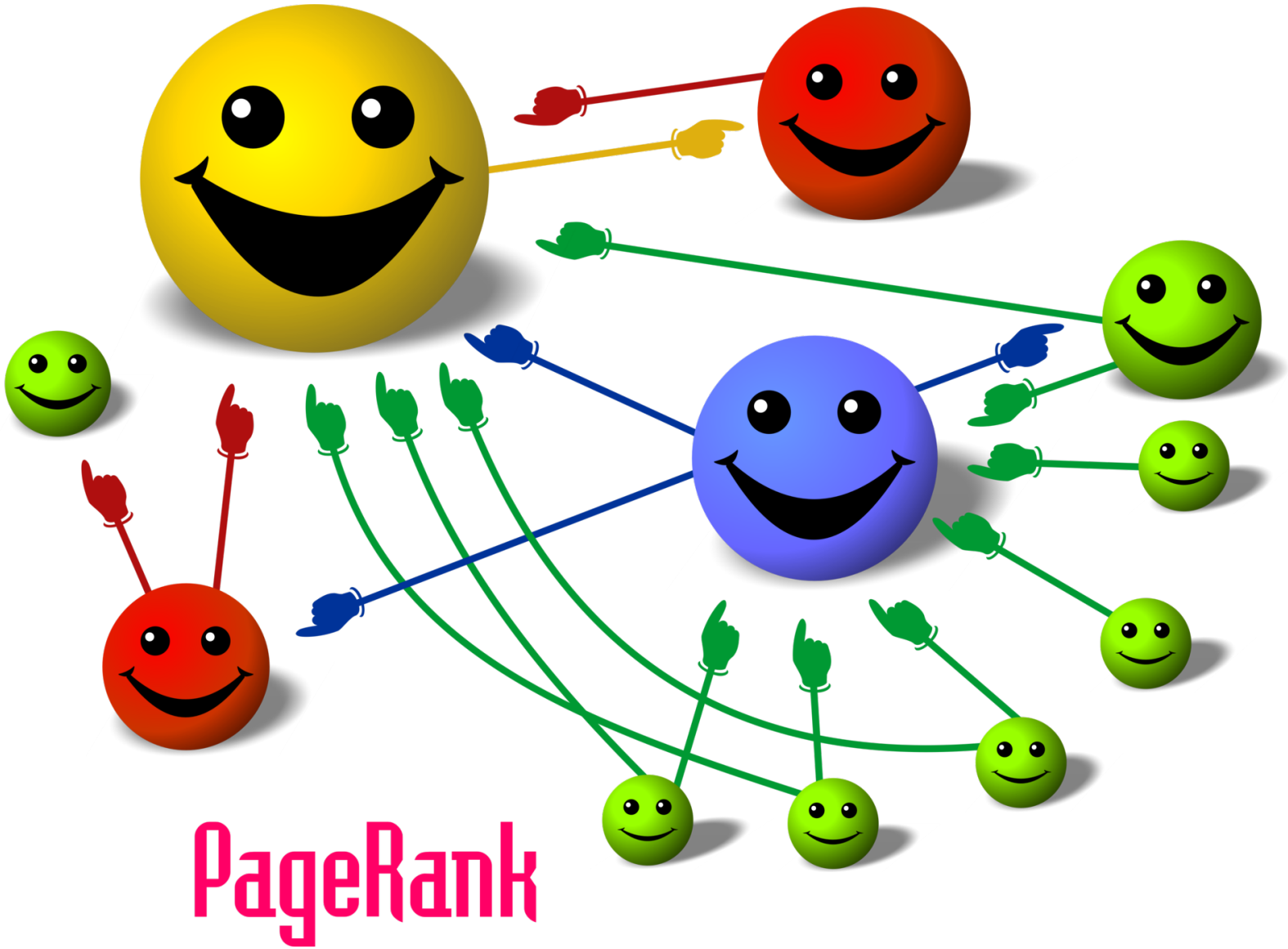
* Quelle: http://www.seokicks.de/backlinks/www.spiegel.de

PageRank: Die zentrale Idee (3)

Weiterhin könnte man davon ausgehen, dass Inlinks, die von einer „wichtigen“ Seiten wiederrum mehr „Wichtigkeit“ ausdrücken, als Inlinks von „unwichtigen“ Seiten.

- Verlinkt bspw. die die Webseite www.spiegel.de auf eine andere Webseite, hat diese vielleicht nur einen einzigen Inlinks, aber dieser ist ein sehr aussagekräftiger und wird ggf. auch sehr oft geklickt.

Zusammengefasst: Eine Webpage hat einen hohen PageRank, wenn die Summe der PageRanks der Inlinks ebenfalls hoch ist. Dies umfasst die Fälle, dass eine Webpage viele Inlinks hat, als auch dass sie wenige, aber dafür „wichtige“ Inlinks besitzt.



http://en.wikipedia.org/wiki/PageRank#mediaviewer/File:PageRank-hi-res.png

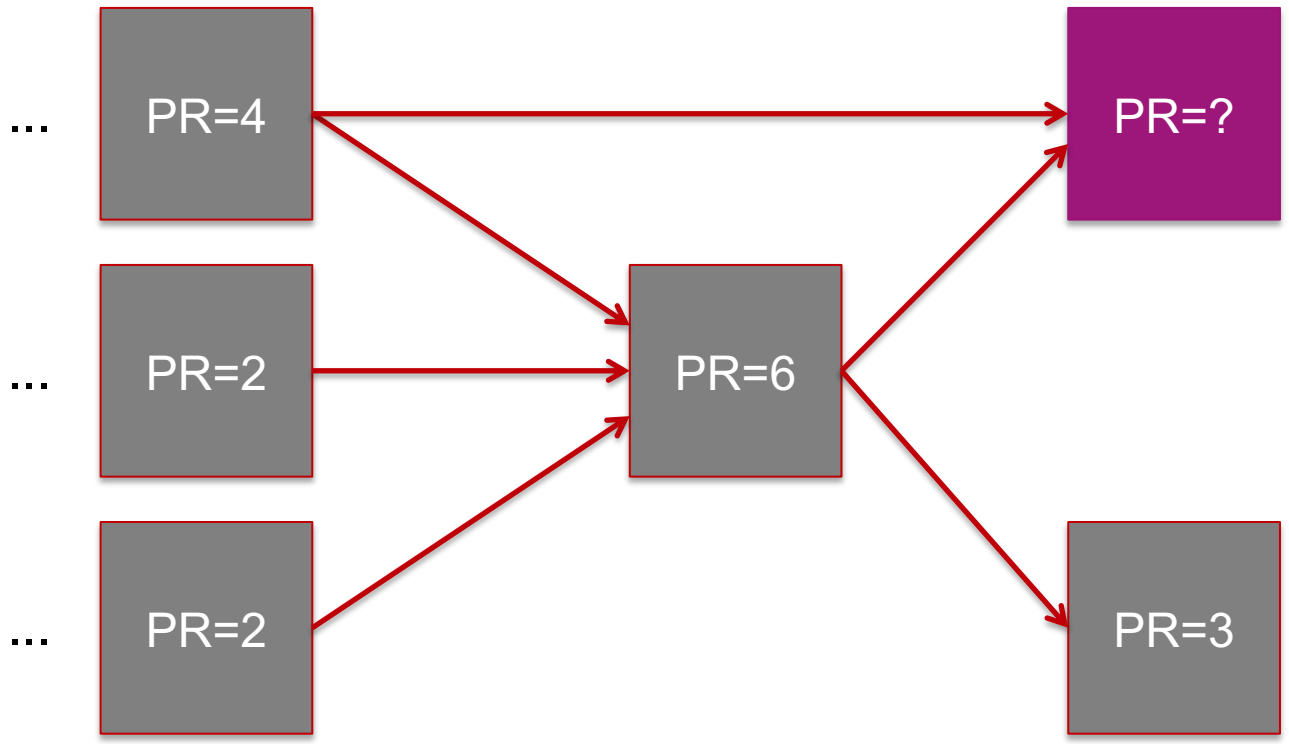
$PageRank_{simple}$: Einfache Definition

$PR_s(w)$	PageRank einer Webseite w
IN_w	Die Menge von Webseiten, die auf w zeigen (inlinks)
c_i	Die Anzahl der Links, die von i ausgehen (outlinks)

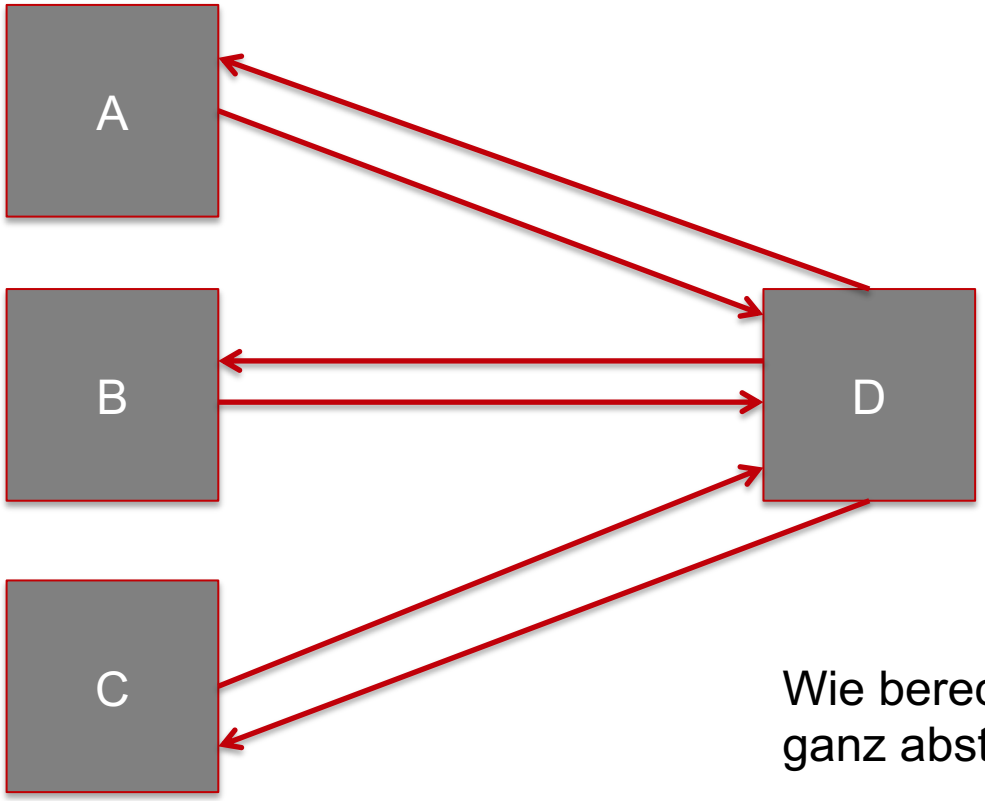
$$PR_s(w) = \sum_{i \in IN_w} \frac{PR_s(i)}{c_i}$$

- Die Gleichung ist **rekursiv**, kann aber mit jeder Menge von Rankings berechnet und iterativ gelöst werden

Ein Beispiel



Ein abstrakteres Beispiel / Übung



Wie berechnen Sie den PageRank,
ganz abstrakt, ohne Zahlen?

PageRank: Problem des Rank Sink

- Die vorherige Definiton hat ein Problem: *rank sink*
  - Wenn Webpages aufeinander zeigen, allerdings zu keinen anderen, wird eine **Schleife** in Gang gesetzt.
  - Ein hoher Rank wird **akkumuliert**, allerdings nie auf andere Seiten übertragen.

Quelle: http://europa.nvc.cs.vt.edu/~cegyhazy/cs5604/pagerank.ppt

PageRank: Definition

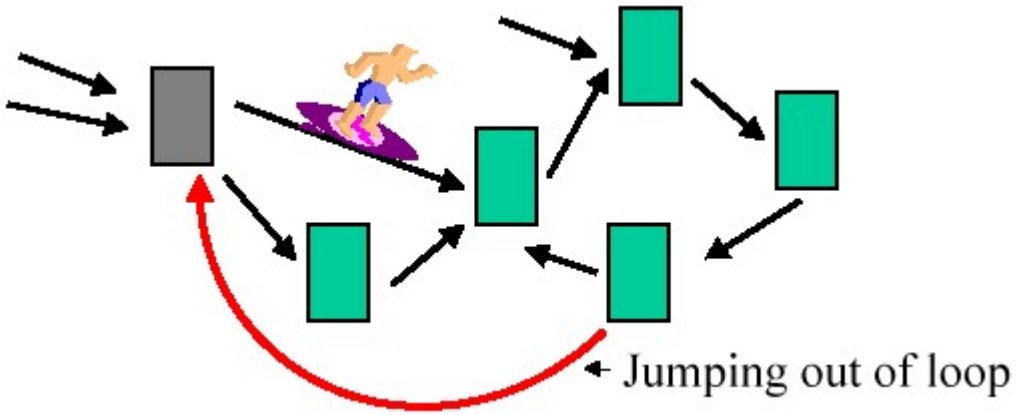
PR(w)	PageRank einer Webseite w
IN_w	Die Menge von Webseiten, die auf w zeigen (inlinks)
c_i	Die Anzahl der Links, die von i ausgehen (outlinks)
n	Anzahl aller Webseiten
d	Dämpfungsfaktor zwischen [0..1], meist 0,85

$$PR(w) = \frac{1-d}{n} + d \sum_{i \in IN_w} \frac{PR_s(i)}{c_i}$$

- Die Gleichung ist **rekursiv**, kann aber mit jeder Menge von Rankings berechnet und iterativ gelöst werden, siehe z.B. https://www.youtube.com/watch?v=4c3DAxQXzLI
- Die Summe aller Pageranks nach dieser Definition ist 1!

Das Random Surfer-Modell

Diese Definition entspricht der **Wahrscheinlichkeitsverteilung** eines zufälligen Weges durch den Web-Graphen.



- Man kann sich den **Dämpfungsfaktor 1-d** als die Wahrscheinlichkeit vorstellen, mit der der Surfer irgendwann gelangweilt ist und zu beliebigen anderen Seiten springt.
- So kann er niemals in einer Schleife feststecken.

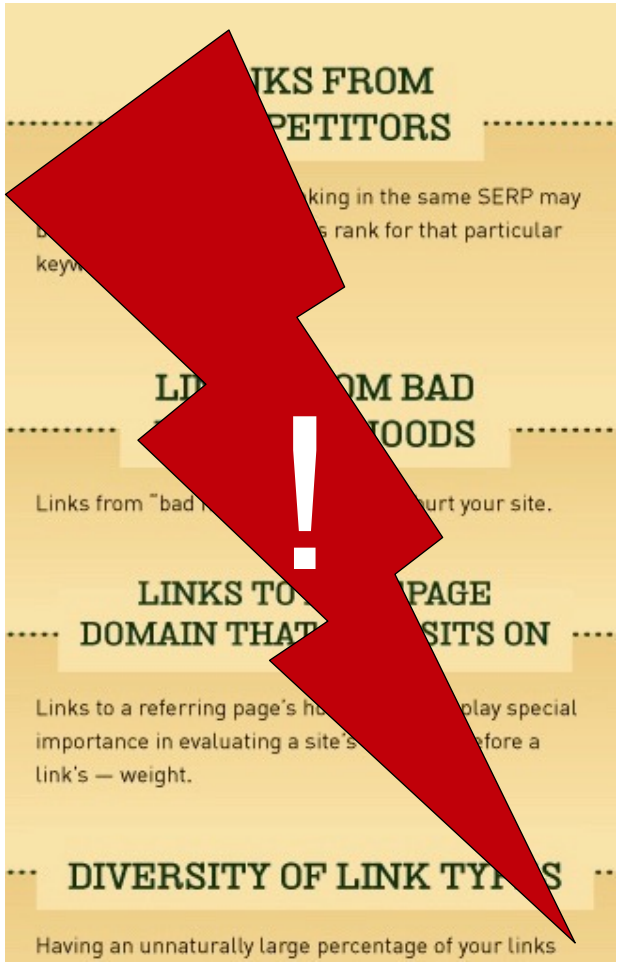
Praktische Anwendung des PageRank

1. PageRank für alle Webseiten wird berechnet – Bevor jemals eine Suche abschickt wurde.
2. Nutzer stellt **Anfrage** an Websuchmaschine.
3. Auf Grundlage der Suchterme wird zunächst eine **ungeordnete Menge** von potentiell relevanten Webseiten zusammengestellt (boolesche Anfrage)
4. Anschließend wird die Menge an Webseiten **in eine geordnete Liste überführt**, indem absteigend nach einem Score sortiert wird, der u.a. auf dem PageRank basiert.

- Natürlich fließen in die tatsächliche Berechnung des Score noch **viele andere Faktoren** mit ein...

Ranking Signals

- Insgesamt fließen in das Ranking einer wirklichen Web-Suchmaschine wie Google, viele 100 „Signals“ ein.
- Dies können z.B. sein
  - **Page-Level Signale** wie z.B. Title-Tags oder Texte in Link.
  - **Site-Level Signale** wie z.B. Trust, Verschlüsselung etc.
  - Signale aus **sozialen Netzwerken**
  - **uvm. – siehe Link zur Infografik!**
  - **PageRank nur noch von kleiner Relevanz...**



http://cdn.alleywatch.com/wp-content/uploads/2014/08/ranking_factors_infographic_2.jpg

Ausnutzung der Schwächen

Link-Farmen

- Als Linkfarm wird eine Ansammlung von Webseiten im Web bezeichnet, die primär dem Zweck dient**, möglichst viele Hyperlinks** auf eine andere Webpräsenz zu legen.
- Die Erstellung solcher Linkfarmen dient der **Suchmaschinenoptimierung** (SEO) bzw. der Manipulalon von Suchmaschinen, d. h., die verlinkte Website soll für Suchanfragen auf einen der ersten Plätze der Trefferliste gebracht werden. Dabei sind die einzelnen Seiten einer solchen Linkfarm vielfach einander sehr ähnlich oder identisch.

- Google selbst beschreibt seine Gegenmaßnahmen z.B. hier: https://support.google.com/webmasters/answer/93713

Quelle: http://de.wikipedia.org/wiki/Linkfarm

Weitere Schwächen

- Entscheidend ist **nicht das Interesse der Leser**, sondern lediglich das anderer Webseitenbetreiber.
- Finanzkräftige Seitenbetreiber können sich **Inlinks erkaufen**. Dies führt dazu, dass statt qualitativ hochwertigem Inhalt oft die finanziellen Möglichkeiten über die Reihenfolge der Suchergebnisse entscheiden.
- Webmaster sehen oft im **PageRank das einzige Bewertungskriterium** für den Linktausch. Der Inhalt der verlinkten Seiten gerät in den Hintergrund.
- Der PageRank liefert **keinen Beitrag zur qualitativen Einordnung von Websites**.

http://de.wikipedia.org/wiki/PageRank

Zusammenfassung Web Retrieval

- Das Web ist nicht mit „klassischem“ IR zu vergleichen und muss anders angegangen werden...
- Robots durchsuchen das Web und bauen den Index von Web-Suchmaschinen auf.
- PageRank ist ein **globales Ranking**, dass auf der Struktur des World Wide Web basiert.
- PageRank verwendet Informationen über **Backlinks/Inlinks** um das Web zu ordnen.
- PageRank verwendet ein sogenanntes **Random Surfer-Modell**.

- **Aber**: Heute spielt der PageRank laut Google selbst nur noch eine untergeordnete Rolle für das Ranking von Webseiten.