

Ranking by Paths through the Web

Matthias Kleine, Ingmar Rötzer

23.07.2009

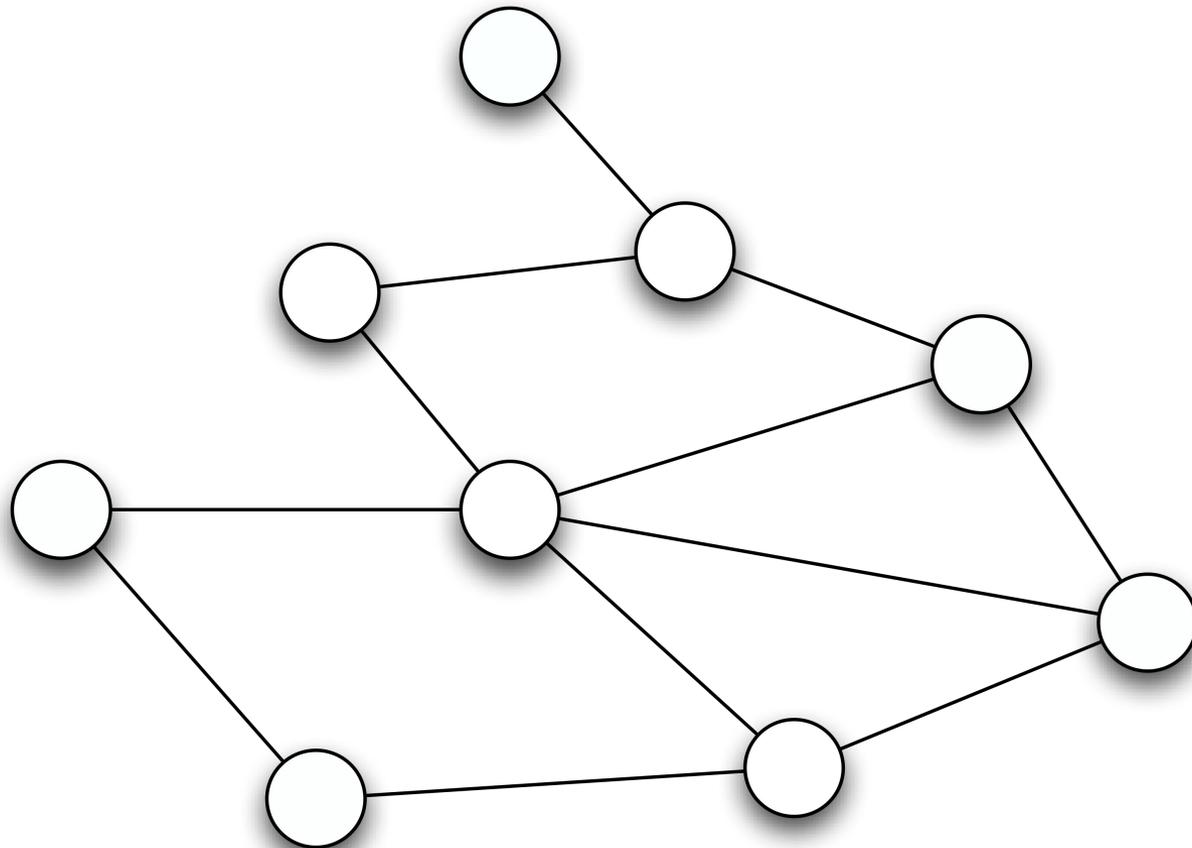
Ranking by Paths

2

- Wikipedia Suche
 - Berücksichtigung der Linkstruktur für Ranking
- Implementierung
 - Mindestabstand zu vorgegebener Startseite beeinflusst Ranking
 - Annahme: Nahe Seiten sind relevanter für den Nutzer
- Beispiel
 - Startseite ist „Kraftfahrzeug“
 - Suche nach „Golf“
 - „VW Golf“ höher gerankt als „Golfball“

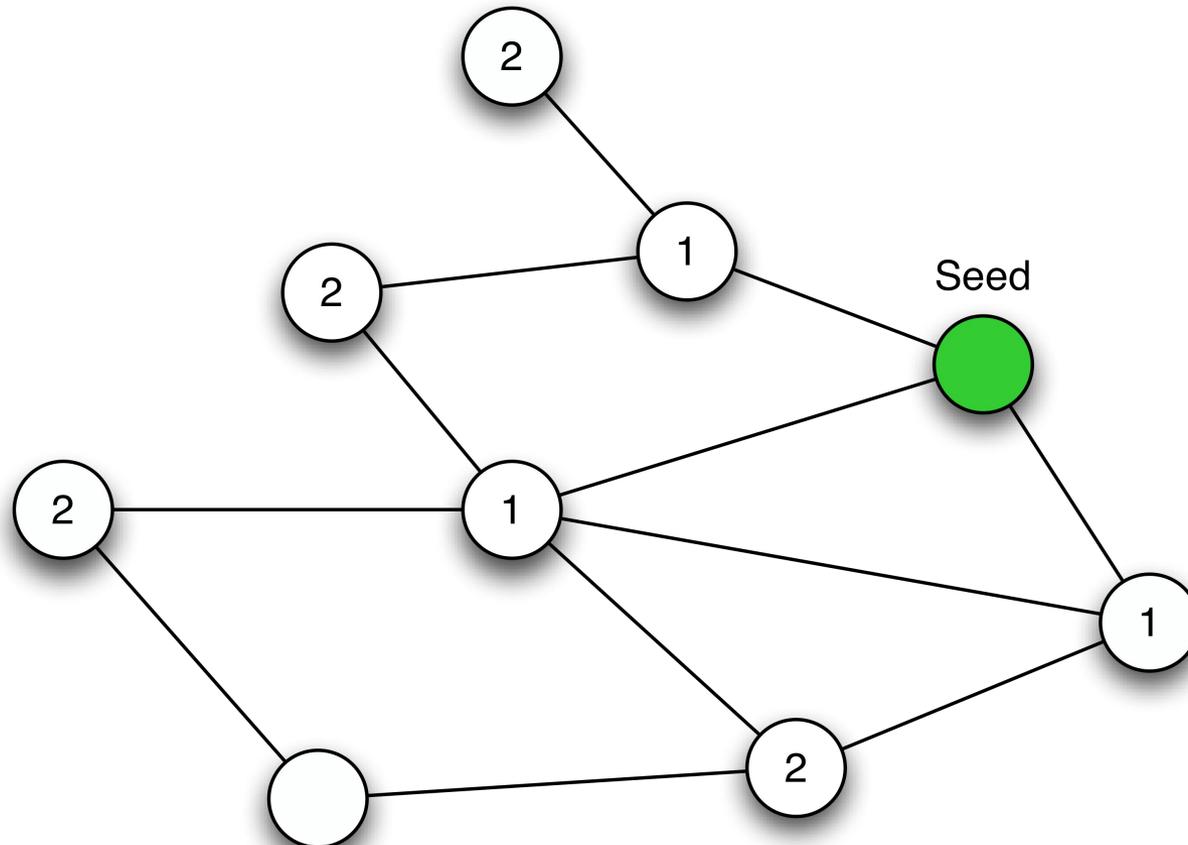
Wikipedia Link Graph

3



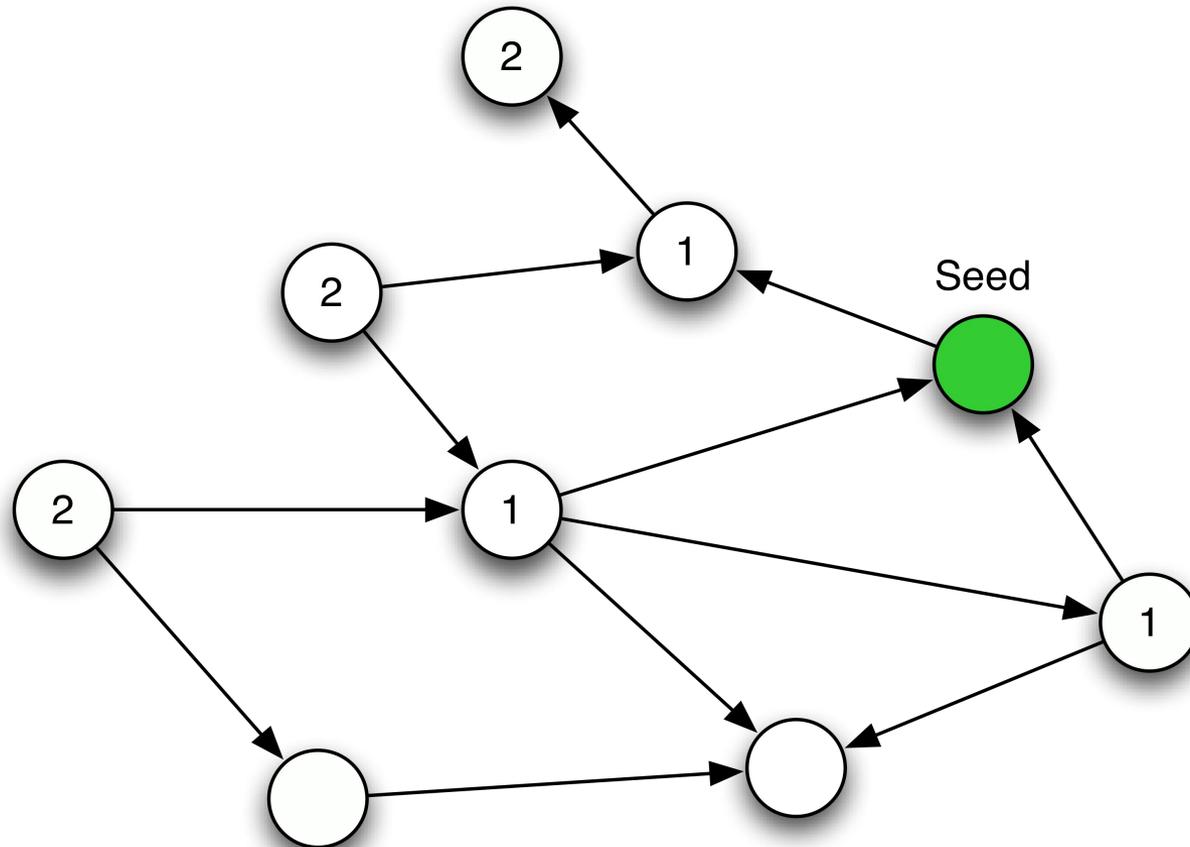
Abstände: Ungerichtet

4



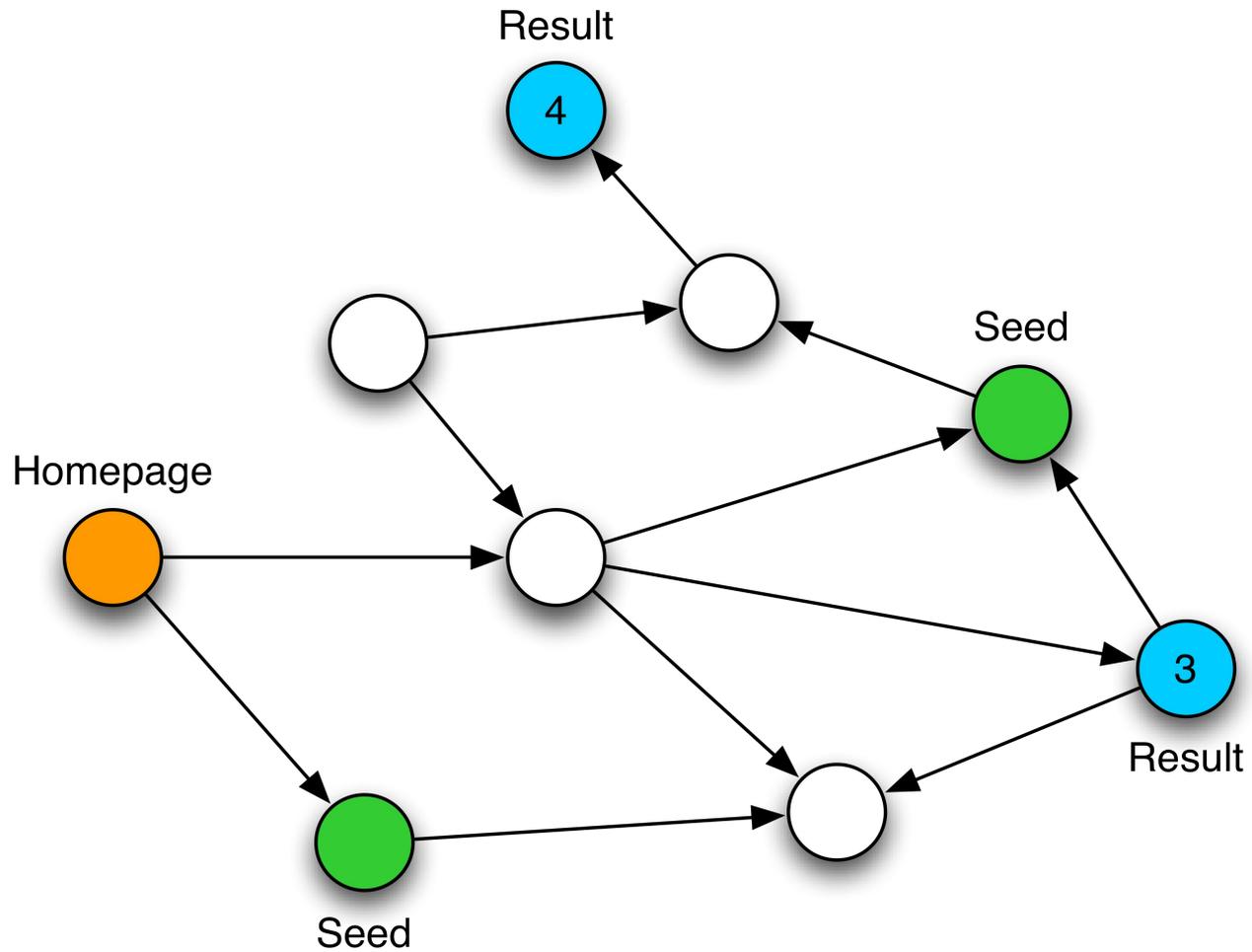
Abstände: Gerichtet

5



Ranking

6



Unser Ergebnis

7

- Funktionierende Implementierung
- Größtenteils mittels SQL (MySQL, DB2)
 - Kompakt
 - Optimierung durch DBMS
- Vorberechnung
 - Läuft in akzeptabler Zeit
 - 10.000 Seeds in einer Stunde
- Query
 - Antwortzeit zu hoch für interaktive Nutzung
 - Bei 10.000 Ergebnissen ca. 10 Sekunden

Demo

8

Erkenntnisse

9

- **Verwendeter Algorithmus ursprünglich für Freundschaftsbeziehungen in Social Networks**
 - Andere Linkstruktur
 - 100 Freunde vs. 270.000 Links zur Seite „USA“
 - Viele Seiten mit extrem wenig Links (z.B. Redirects)
- **Viele allgemeine Seiten oft verlinkt**
 - Tauchen häufig im Ergebnis auf
 - z.B. Jahreszahlen, Ländernamen, Universitäten, Fachrichtungen

Future Work

10

- Performance nicht akzeptabel
 - Vereinigung von Anfrage und Ranking im DBMS
 - Massive Parallelisierung
 - Einschränkung der Ergebnismenge durch zusätzliches Ranking
- Weitere Untersuchungen möglich
 - Gerichteter vs. Ungerichteter Graph
 - Seedauswahl
 - Gewichtung von Seiten nach Linkanzahl