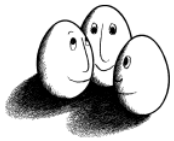
Suchstrategien

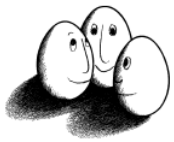
PG 402

Phillip Look
Christian Hüppe

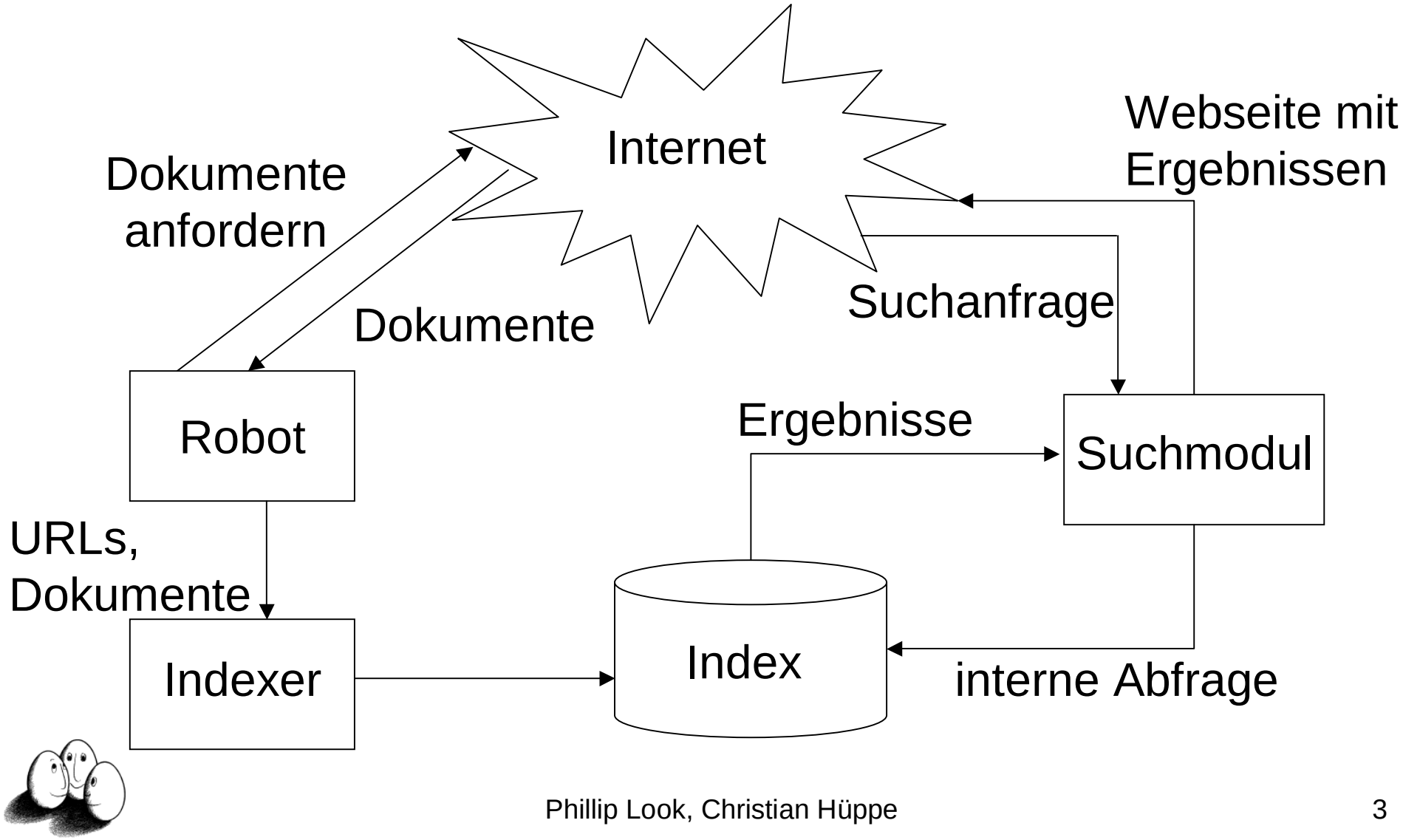


Überblick

- Einführung
- Untersuchung von 2 Suchmaschinen
- Verbesserung der Rankingfunktion mit Hilfe von Clickthrough-Daten
- Clustering von Query Logs
- Strukturorientierte Ansätze



Arbeitsweise von Suchmaschinen

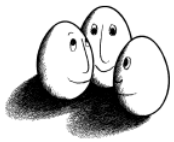


Ranking

manuell

- Handgepflegte Listen (Katalogsysteme)
- Häufigkeit der Schlagworte
- Formatierung der Schlagworte (Überschrift, fett, unterstrichen usw.)
- Link-Popularität
- Link-Qualität

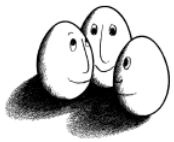
auto-
matisch



Kriterien einer guten Suchmaschine

- Geschwindigkeit der Antwort
- Anzahl u. Aktualität des Indexes
- Qualität des gelieferten Rankings
  - Absolute Qualität → Precision / Recall
  - Relative Qualität → Benutzerverhalten

Wie kann man das Ranking verbessern?



Verbesserung des Rankings

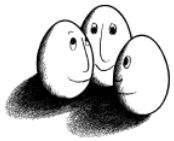
- Benutzerfeedback einbeziehen:

  Problem: schwer zugänglich, gering vorhanden[1]

  Lösung: System beobacht, welche Links benutzt werden

  →Clickthrough-Daten analysieren

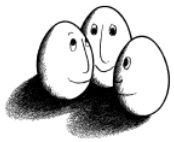
[1] WebWatcher: A Tour Guide for the World Wide Web,
T. Joachims, D. Freitag, T. Mitchell,
Proceedings of IJCAI97, August 1997



Clickthrough Daten

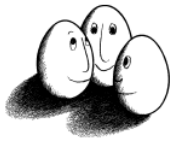
- Speicherung von
  - Name der anfragenden Maschine
  - IP-Adresse der anfragenden Maschine
  - Browser-Typ
  - Bildschirmauflösung
  - Benutzeranfragen
  - gewählte URLs

  } für uns interessant
  - usw.



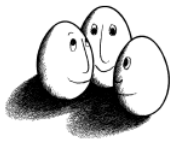
Beispiel für Clickthrough Daten

felony	jud13.flcourts.org/felony.html
missoula,+mt	missoula.bigsky.net/score/
feeding+infants+solid+foods	members.tripod.com/drlee90/solid.html
colorado+lotto+results	www.co-lotto.com/
northern+blot	www.invitrogen.com/expressions/1196-3.html
wildflowers	www.life.ca/nl/43/flowers.html
ocean+whales	playmaui.com/ocnraftn.html
ralph+lauren+polo	www.shopbiltmore.com/dir/stores/polo.htm
bulldog+hompage	www.adognet.com/breeds/2abulm01.html
lyrics	www.geocities.com/timessquare/cauldron/8071
churches+in+atlanta	acme-atlanta.com/religion/christn.html
retail+employment	www.crabtree-evelyn.com/employ/retail.html
illinois+mortgage+brokers	www.birdview.com/ypages2/c3.htm
stock+exchange+of+singapore	www.ses.com.sg
front+office+software	www.saleslogic.com/saleslogix.phtml
free+3d+home+architect	www.adfoto.com/ads1/homeplans.shtml
country+inns+sale	innmarketing.com/form.html
free+desktop+wallpaper	www.snap-shot.com/photos/fireworks/
automotive+marketing+research	www.barndoors.com/rcmresources.htm
router+basics	www.wheretobuy.com/prdct/706/55.html



Vorteile von Clickthrough-Daten

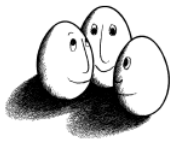
- man kommt sehr einfach an Daten
- die Datenmenge ist quasi unendlich
- unbeeinflusstes Benutzerverhalten wird aufgezeichnet
- man kann sinngemäße Queries / URLs finden
  - Benutzer wählen gleiche URLs bei unterschiedlichen Queries
  - Benutzer wählen unterschiedliche URLs bei gleicher Query



Experiment 1

Vergleichen von 2 Suchmaschinen A und B:

- Einheitliche Benutzeroberfläche
- Keine Einschränkung der Produktivität des Benutzers
- Benutzerbeurteilung soll klar demonstriert werden

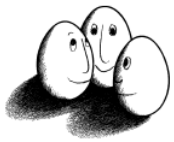


Experiment 1

- Query wird an Suchmaschine A und B gesandt
- Ranking A = (a_1, a_2, ...) und B = (b_1, b_2, ...) können in ein gemeinsames Ranking C = (c_1, c_2, ...) gemischt werden, so dass für jeden Rang n von C gilt

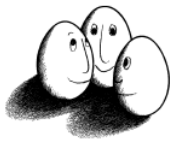
$$\{c_1, ..., c_n\} = \{a_1, ..., a_{n_a}\} \cup \{b_1, ..., b_{n_b}\}$$

mit $n_b \leq n_a \leq n_b + 1$.



Ranking generieren

```
public void combine( Vector a, Vector b, int na, int nb, Vector c ) {
  if ( na == nb ) {
      if ( !c.contains( a.get( na + 1 ) ) ) {
        c.add( a.get ( na + 1 ) );
      }
      this.combine( a, b, na+1, nb, c );
  }
  else {
      if ( !c.contains( b.get( nb + 1 ) ) ) {
        c.add( b.get( nb + 1 ) );
      }
      this.combine( a, b, na, nb +1, c );
  }
}
```



Vorteile dieses Aufbaus

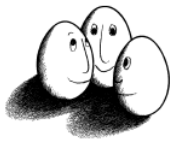
Annahme:

- Alle Links werden ohne Ausnahme betrachtet

  → Benutzer nimmt von Top-Ranking A und Top-Ranking B gleich viele Links wahr

- Benutzer wählt relevante Links
- Kurzzusammenfassung bietet genug Informationen → Wahl nicht zufällig

  → Klicks enthalten Informationen über die Qualität der Top n_a und n_b Links beider Suchmaschinen



Google Results:

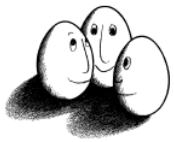
1. Kernel Machines
   http : //svm.first.gmd.de/
2. SVM-Light Support Vector Machine
   http : //ais.gmd.de/ ~ thorsten/svm_light/
3. Support Vector Machine and Kernel ... References
   http : //svm.....com/SVMrefs.html
4. Lucent Technologies: SVM demo applet
   http : //svm.....com/SVT/SVMsvt.html
5. Royal Holloway Support Vector Machine
   http : //svm.dcs.rhbnc.ac.uk/
6. Support Vector Machine - The Software
   http : //www.support−vector.net/software.html
7. Support Vector Machine - Tutorial
   http : //www.support−vector.net/tutorial.html
8. Support Vector Machine
   http : //jbolivar.freeservers.com/

MSNSearch Results:

1. Kernel Machines
   http : //svm.first.gmd.de/
2. Support Vector Machine
   http : //jbolivar.freeservers.com/
3. An Introduction to Support Vector Machines
   http : //www.support − vector.net/
4. Archives of SUPPORT-VECTOR-MACHINES ...
   http : //www.jiscmail.ac.uk/lists/SUPPORT...
5. SVM-Light Support Vector Machine
   http : //ais.gmd.de/ ~ thorsten/svm_light/
6. Support Vector Machine - The Software
   http : //www.support−vector.net/software.html
7. Lagrangian Support Vector Machine Home Page
   http : //www.cs.wisc.edu/dmi/lsvm
8. A Support ... - Bennett, Blue (ResearchIndex)
   http : //citeseer.../bennett97support.html

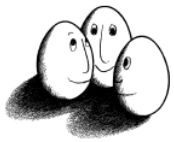
Combined Results:

1. **Kernel Machines**
   http : //svm.first.gmd.de/
2. Support Vector Machine
   http : //jbolivar.freeservers.com/
3. **SVM-Light Support Vector Machine**
   http : //ais.gmd.de/ ~ thorsten/svm_light/
4. An Introduction to Support Vector Machines
   http : //www.support − vector.net/
5. Support Vector Machine and Kernel Methods References
   http : //svm.research.bell − labs.com/SVMrefs.html
6. Archives of SUPPORT-VECTOR-MACHINES@JISCMAIL.AC.UK
   http : //www.jiscmail.ac.uk/lists/SUPPORT−VECTOR−MACHINES.html
7. **Lucent Technologies: SVM demo applet**
   http : //svm.research.bell − labs.com/SVT/SVMsvt.html
8. Royal Holloway Support Vector Machine
   http : //svm.dcs.rhbnc.ac.uk/
9. Support Vector Machine - The Software
   http : //www.support − vector.net/software.html
10. Lagrangian Support Vector Machine Home Page
    http : //www.cs.wisc.edu/dmi/lsvm



Google vs. MSNSearch

Ranging A	Ranging B	A besser	B besser	A = B	Total
Google	MSNSearch	34	20	69	123
Google	Standard	18	1	15	34
MSNSearch	Standard	16	2	5	23

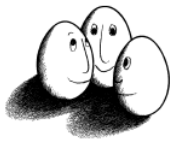


Benutzerfeedback aus Clickthrough-Daten Extrahieren

- Gegebenes Ranking: ($link_1$, $link_2$, $link_{3,...}$)
- C ist Ranking der angeklickten Links

$$link_i <_{r*} link_j$$

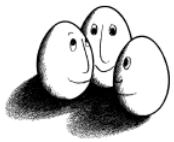
für alle Paare $1 \leq j < i$, mit $i \in C$ und $j \notin C$.



Lernen einer Rankingfunktion mit der SVM

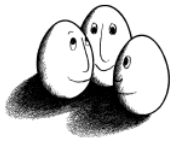
- Vereinfache das Problem in ein binäres Klassifikations-Problem
  - relevanter Link
  - unrelevanter Link
- Trainingmenge mit queries q und (optimales) Zielranking r*

$$(q_1,r^*_1), (q_2,r^*_2), \ldots, (q_n,r^*_n).$$



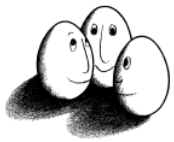
Attribute eines Links zum Lernen

- Link-Rang bei anderen Suchmaschinen
- Domain-Name in Query enthalten
- Länge der URL
- Länderkennung der URL
- Domain der URL
- Tilde in der URL
- Cosinusmaß zwischen URL-Worten und Query
- Cosinusmaß zwischen Titel und Query



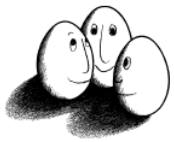
Vektorraummodell (1)

- Terme der Datenbasis spannen einen orthogonalen Vektorraum auf
- Dokumente und Queries sind Punkte in diesem Vektorraum
- Gesucht werden Dokumente, deren Vektoren ähnlich zum Queryvektor sind

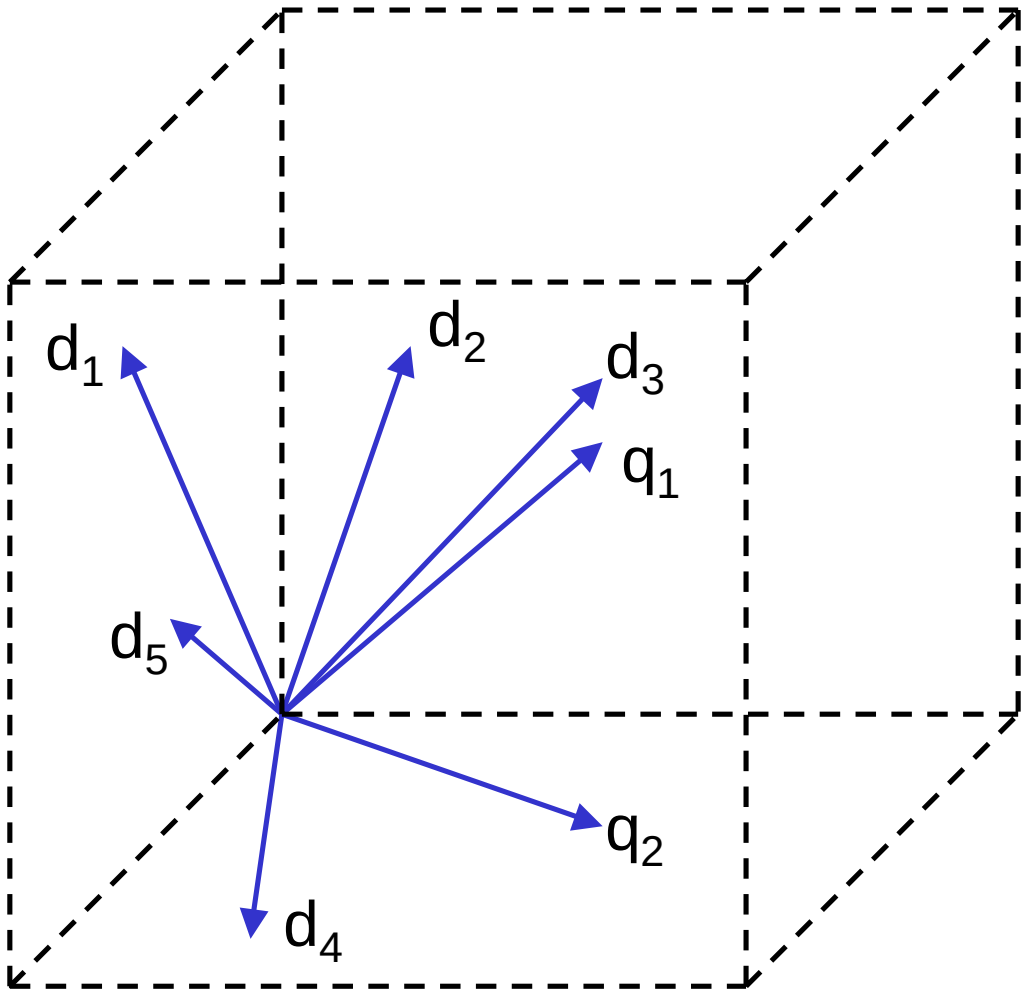


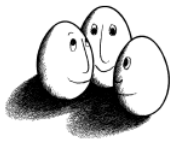
Vektorraummodell (2)

- Ähnlichkeit wird über Cosinusmaß bestimmt
- Cosinusmaß: Cosinus des Winkels zwischen zwei Dokumentvektoren bzw. zwischen Query- und Dokumentvektor
- Wenn Vektoren ähnlich ausgerichtet sind
  → Winkel entsprechend klein

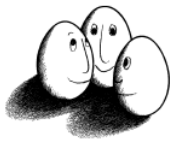


Vektorraummodell - Beispiel





Clustering von Query Logs



Clustering von Query Logs

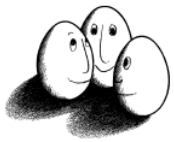
- Daten → Clickthroughdaten

  {Query, ausgewählte URLs}
- Betrachtung der Daten als Graph

  Konten → Queries und URLs

  Kanten → Zusammenhang: Query und URL
- Besonderheit

  Inhalt der URLs wird ignoriert

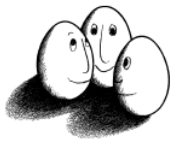


Clustering Algorithmus

- hierarchical agglomerative clustering (HAC)
- Inhalt wird entgegen normalen HAC-Algorithmen ignoriert

  Vorteil:

  - Clustering erfordert keine Speicherung von großen Datenmenge
  - Kann auch bei textfreien Seiten angewandt werden (z. B. Bilder)
  - Implementierung ist sehr viel einfacherer (Kein aufwendiges Preprocessing um Inhalte aufzuarbeiten)



Clustering Algorithmus

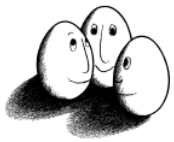
- Clustering von queries

  Beispiel:

  - Benutzer stellt Anfrage q
  - q $\in$ Cluster C

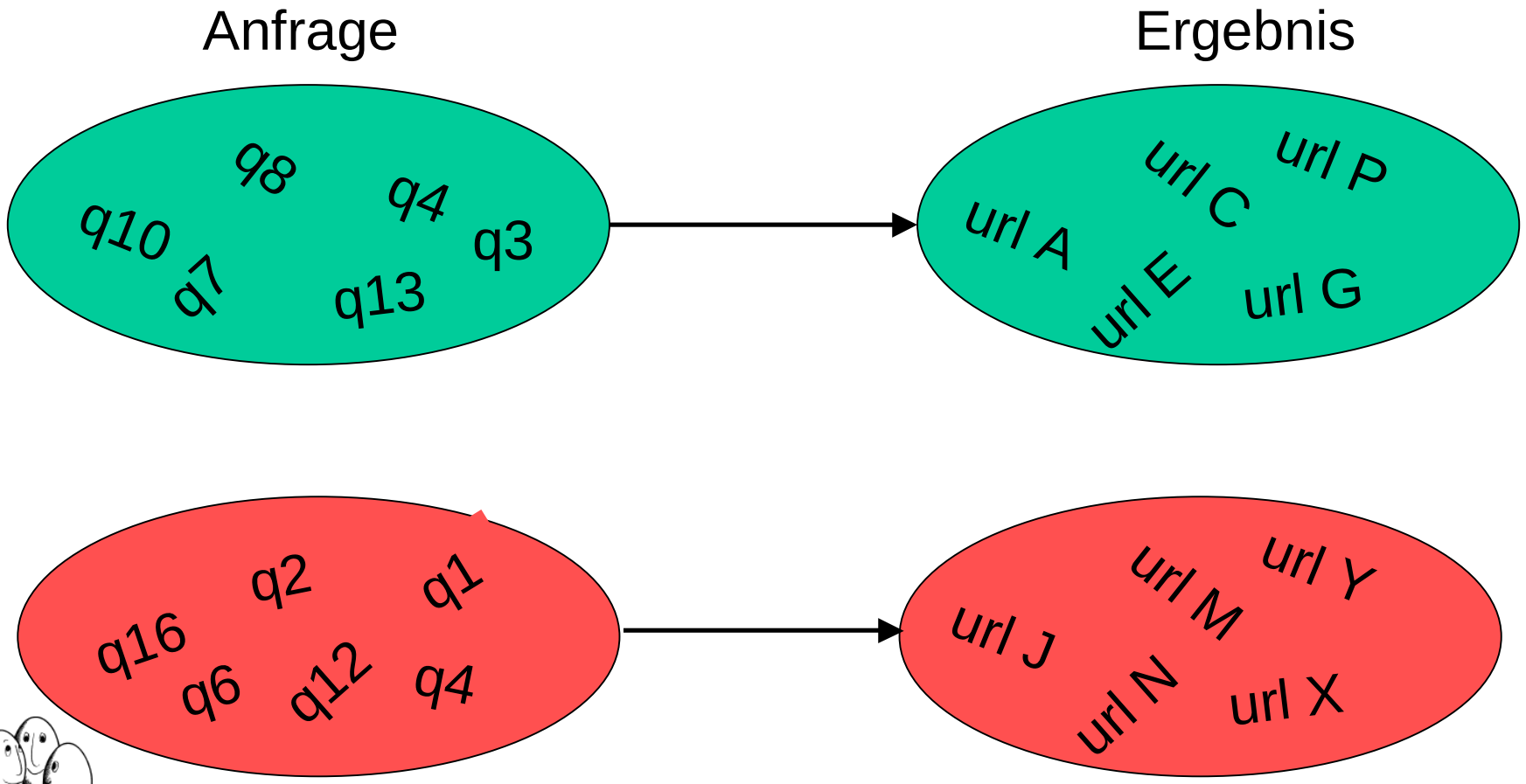
  →System schlägt andere queries aus C vor

  Gute Unterstützung, bei unbefriedigendem Ergebnis



Graph-Based Iterative Clustering

- Disjunkte Mengen von Anfragen
- Disjunkte Mengen von URLs

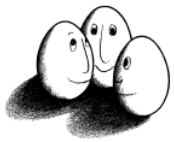


Algorithmus 1

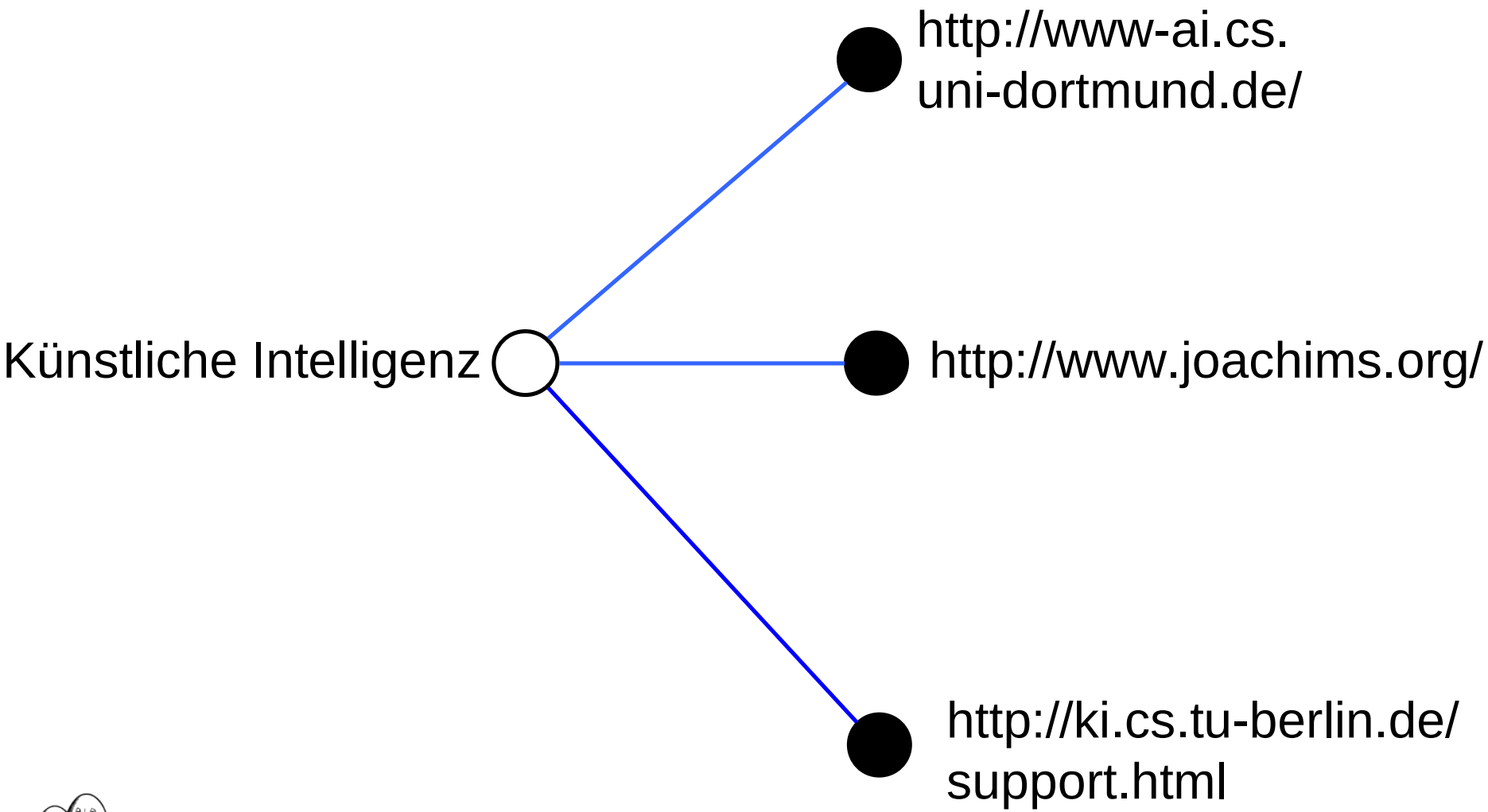
Input: Clickthrough-Daten C

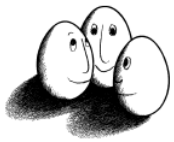
Output: Graph G

1. Sammle alle Anfragen aus C in der Menge Q
2. Sammle alle URLs aus C in der Menge U
3. Für jedes Element in Q erzeuge einen „weißen“ Knoten in G
4. Für jede URL in U erzeuge einen „schwarzen“ Knoten in G
5. Wenn Anfrage q mit einer URL u auftritt, dann lege eine Kante in G zwischen den schwarzen und den weißen Knoten



Beispiel zum Algorithmus 1





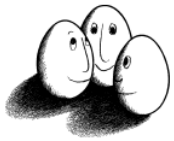
Ähnlichkeit zwischen Knoten

- N(x) = Menge benachbarter Knoten von x in einem Graphen
- Knoten y ähnelt x, wenn N(x) und N(y) eine große Überschneidung haben

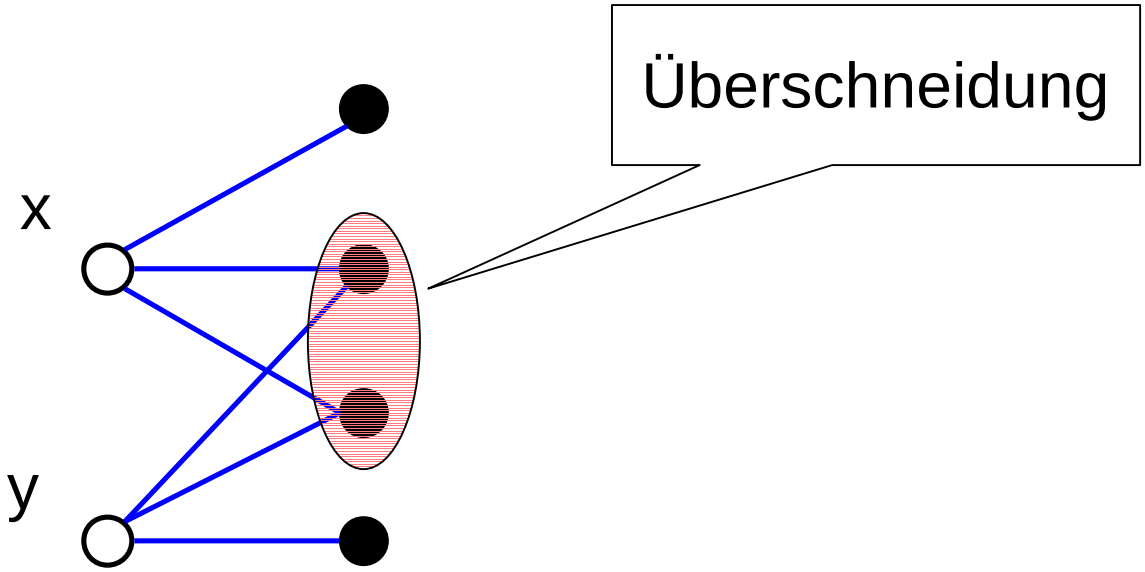
Formal:

$$\sigma(x,y) \stackrel{\text{def}}{=} \begin{cases} \dfrac{N(x) \cap N(y)}{N(x) \cup N(y)} \\ 0 \end{cases}$$

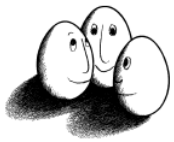
Mit $\left|N(x) \cup N(y)\right| > 0$



Beispiel zur Ähnlichkeit



Ähnlichkeit: $\sigma(x, y) = \frac{2}{4}$

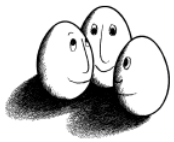


Algorithmus 2

Input: Graph G

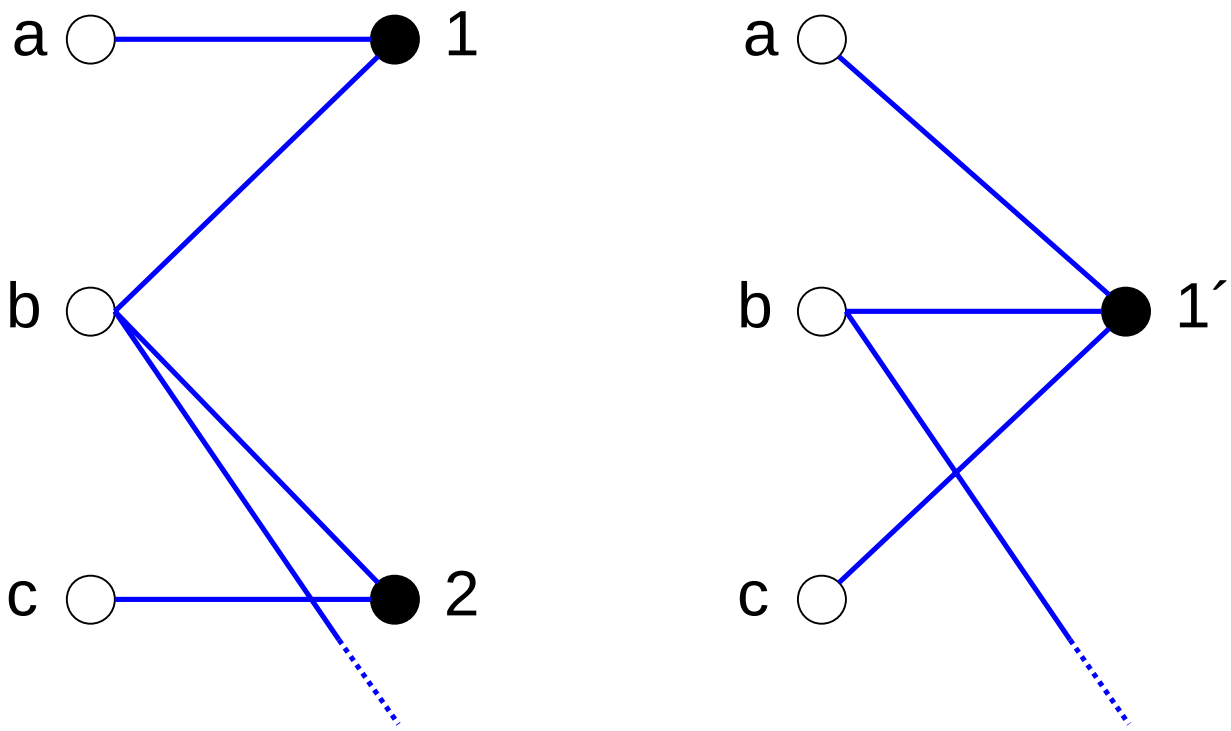
Output: Neuer Graph G´:

1. Bewerte alle weißen Knotenpaare in G
2. Füge die Knoten w_x, w_y *zusammen, für die* $\sigma(w_x, w_y)$ *am größten ist*
3. *Schritt 1 und 2 für schwarze Knoten durchführen*
4. *Bis zur Abbruchbedingung gehe zu 1*



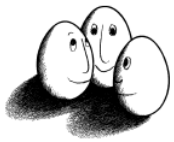
Iterativeransatz?

- Manche Ähnlichkeiten können nicht im Originalgraphen erkannt werden

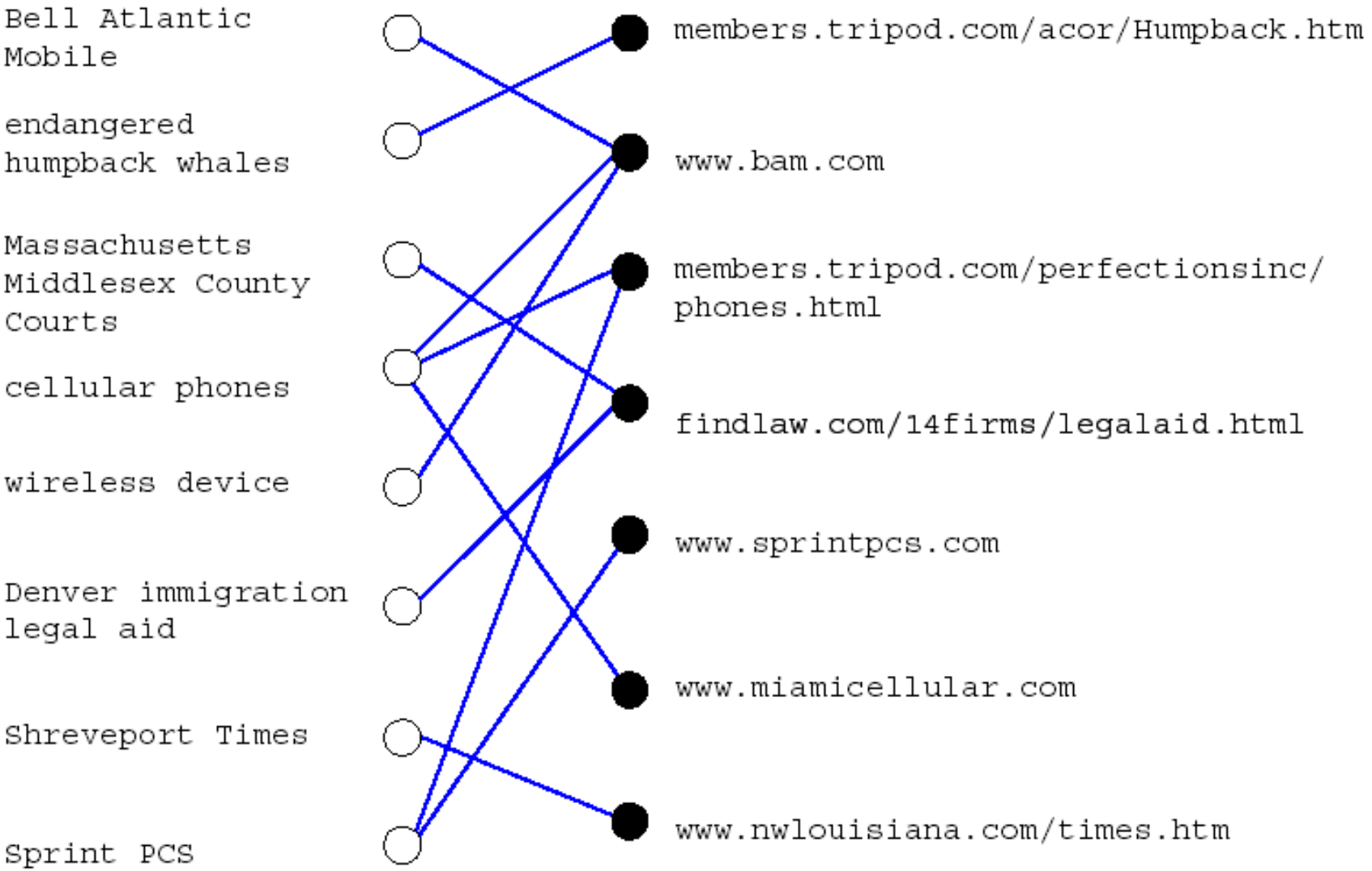


$\sigma(a, c)=0$

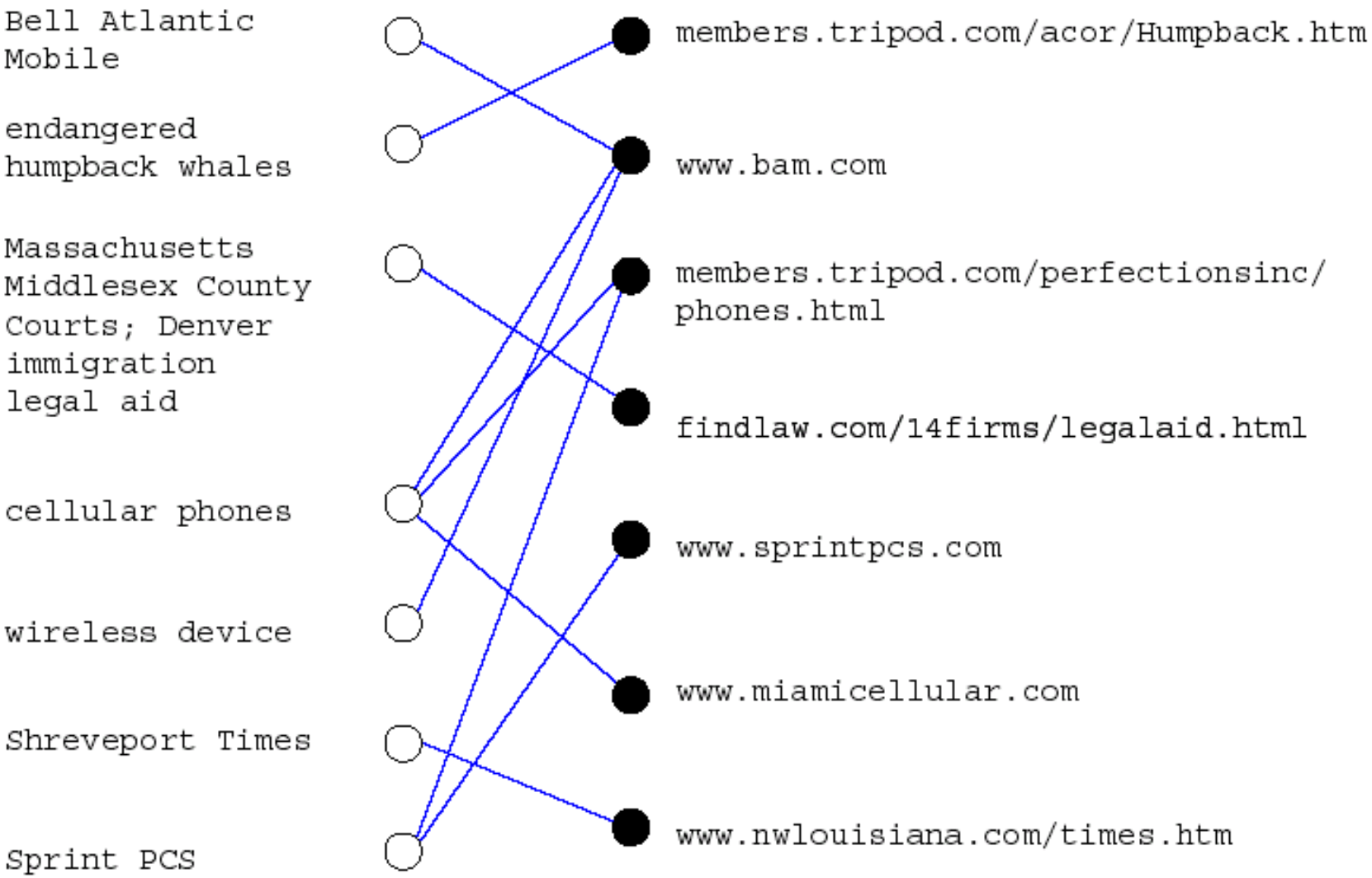
$\sigma(1, 2)=1/3 \longrightarrow \sigma(a, c)=1$

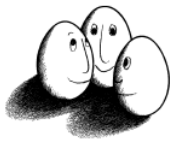


Originalgraph

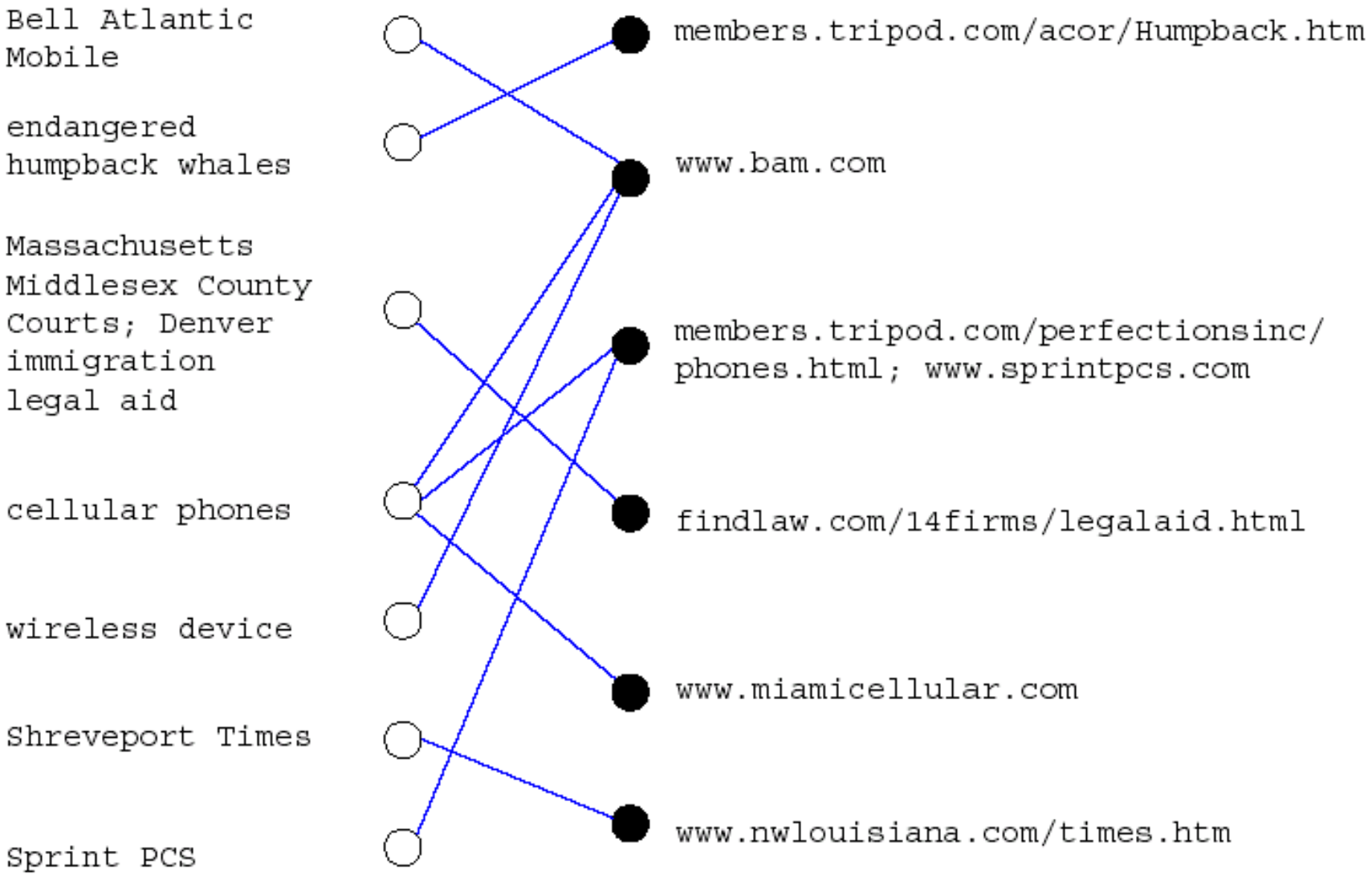


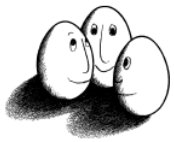
Query Clustering





URL Clustering

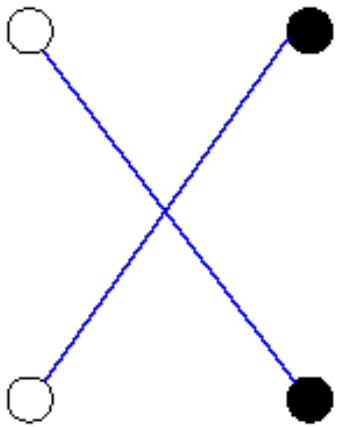




Fertige Cluster

Bell Atlantic
Mobile, wireless
device, cellular
phones, Sprint
PCS

endangered
humpback whales



members.tripod.com/acor/Humpback.htm

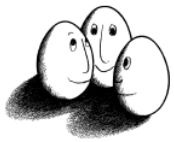
www.bam.com,
www.miamicellular.com,
members.tripod.com/perfectionsinc/
phones.html, www.sprintpcs.com

Massachusetts
Middlesex County
Courts, Denver
immigration
legal aid

findlaw.com/14firms/legalaid.html

Shreveport Times

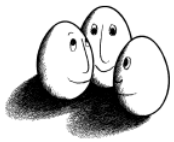
www.nwlouisiana.com/times.htm



Strukturorientierte Ansätze

Überblick:

- Fish-Search
- Enhanced Categorization
- Focused Crawling
- Related Pages
- Fourier Domain Scoring

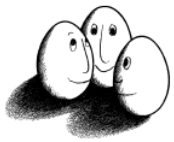


Fish-Search

zum Verfahren:

- Eingabe: URL (einzelner Fisch)
- folge den Links * (Fischschwarm)
- bewerte die Seite (Stärke des Fisches)
- gehe zu (*)

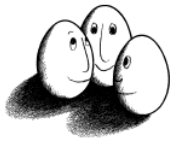
P.M.E. De Bra, R.G.J. Post **1994**



Fish-Search

Vor- und Nachteile:

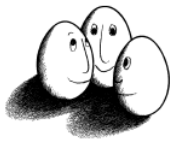
+ gefundenen Seiten existieren
+ keine Datenbank erforderlich
- hohe Netzlast
- nicht gelinkte Seiten werden nicht gefunden
- Sackgassen werden nicht verlassen



Fish-Search

für uns:

- ausgehende Links verfolgen
- URL als Ausgangspunkt

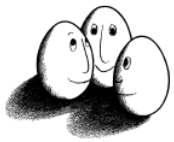


Enhanced Categorization

zum Verfahren:

- gegeben ist eine Unterteilung in Kategorien (Freizeit / Sport / Tennis / Turniere / ...)
- für jede Kategorie sind gute Beispiele vorhanden
- ermittle Diskriminanten und Rauschen

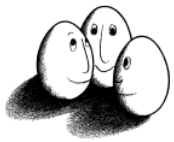
Soumen Chakrabarti, Byron Dom, Piotr Indyk **1998**



Enhanced Categorization

- erweitere Inhalt der Seite um:
  a) Text der benachbarten Seiten
  b) Pfad der benachbarten Seiten (besser)

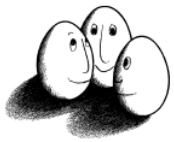
- Ordne die Seite in eine Kategorie ein.



Enhanced Categorization

Vor- und Nachteile:

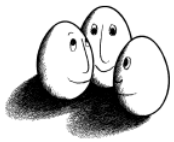
+ Nachbarschaft der Seite wird mit untersucht
- große Beispielmengen erforderlich (20.000 Seiten von Yahoo)
- Einteilung in Kategorien erforderlich
- benötigte Kategorien zu speziell



Enhanced Categorization

für uns:

- Diskriminanten finden
- aus- und eingehende Links untersuchen



Focused Crawling

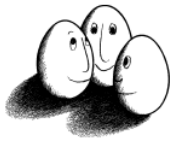
Verfahren:

- vereint die ersten beiden Verfahren
- Ausgangspunkt ist eine URL
- Nachkommen werden auch kategorisiert

Vor- und Nachteile:

+ Search-On-Demand spart Speicherplatz
- Interaktion

`Soumen Chakrabarti, Byron Dom, Martin van den Berg` **`1999`**



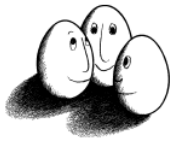
Related Pages

zum Verfahren:

- Eingabe: URL
- Ausgabe: ähnliche Seiten

- basiert nur auf der Linkstruktur
- zwei verschiedene Ansätze

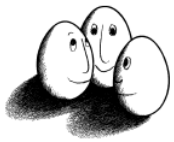
Jeffrey Dean, Monika Henzinger **1998**
basiert auf HITS Algo. von Kleinberg **1998**



Related Pages

zum Verfahren: (cocitation)

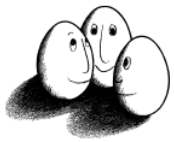
1. suche Eltern und Kinder der Start-URL
2. suche Geschwister
3. wähle die besten Geschwister



Related Pages

zum Verfahren: (companion)

1. baue vincinity-Graph
2. entferne Duplikate (Mirror Seiten)
3. bewerte Kanten
4. wähle die bestbewerteten Seiten



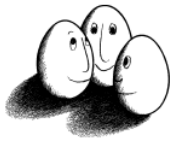
Related Pages

Vor- und Nachteile:

+ einfache zu implementieren
- es werden nicht alle Dokumente gefunden
- hohe Performance nur mit Cache

für uns:

- mehrere Generationen von Links verarbeiten



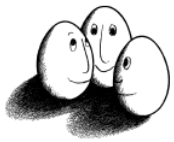
Fourier Domain Scoring

zum Verfahren:

- Idee: betrachte Wörter als Wellen
- Eingabe: Schlüsselwörter

1. teile Text in Abschnitte
2. berechne die Wellen
3. Fourier-Transformation -> Amplitude, Phase
4. vergleiche Dokument mit der Eingabe

Laurence Park, M. Palaniswami, R. Kotagiri, **2001**



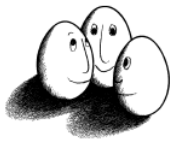
Fourier Domain Scoring

Vor- und Nachteile:

+ Häufigkeit und Position werden beachtet

für uns:

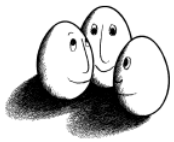
- als Post-Processing-Schritt möglich
- Unbekannte: Fourier-Transformation



Auswertung

Qualität der Links:

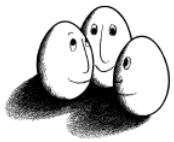
	gesamt	gleicher Host	guter Inhalt	gute Links
absolut (ein.):	15	12	3	5
relativ (ein.):	1,6	80%	20%	33%
absolut (aus.):	138	138	19	14
relativ (aus.):	15,5	100%	14%	10%



Auswertung

Links verfolgen:

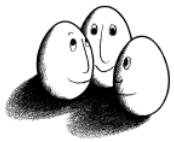
- es gibt kaum Links zu guten Treffern
- Sackgassen werden nicht verlassen
- Verfolgung der Links lohnt sich nicht



Auswertung

Diskriminanten finden:

- Schlüsselwörter: Alterspyramide, Deutschland
- Diskriminante: Bevölkerungsentwicklung
- Suchergebnis: viele gute Treffer
- Ansatz ist vielversprechend



Literatur

[**Fish-Search**] P.M.E. De Bra, R.G.J. Post, **1994**, „Information Retrieval in the World-Wide-Web: Making Client-based search feasible“

[**Enhanced Categorization**] Soumen Chakrabarti, Byron Dom, Piotr Indyk, **1998**, „Enhanced Hypertext Categorization using hyperlinks“

[**Focused Crawling**] Soumen Chakrabarti, Byron Dom, Martin van den Berg, **1999**, „Focused Crawling: a new approach to topic-specific Web resource discovery“

[**Related Pages**] Jeffrey Dean, Monika Henzinger, **1998**, „Finding Related Pages in the World Wide Web“

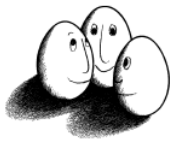
[**Fourier Domain Scoring**] Laurence Park, M. Palaniswami, R. Kotagiri, **2001**, „Internet Document Filtering using Fourier Domain Scoring“

[**Untersuchung von 2 Suchmaschinen**] Joachims Thorsten, **2002**, „Evaluating Search Engines using Clickthrough Data“

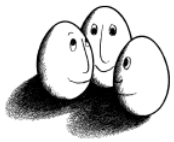
[**Verbesserung der Rankingfunktion mit Hilfe von Clickthrough-Daten**] Joachims Thorsten, **2002**, „Optimizing Search Engines using Clickthrough Data“

[**Clustering von Query Logs**] Beeferman Doug, Berger Adam, **200?**, „Agglomerative clustering of a search engine query log“

[**Einführung**] Babiak Ulrich, **1997**, „Effektive Suche im Internet“

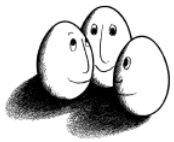


Ende



Google

- entwickelt von Lawrence Page u. Sergey Brin
- Prototyp 1997, Stanford-Universität
- Abdeckung: 1.2 Milliarden (ges. 2 Milliarden)*
- Strategie, Analyse:
  - Wie sind die Seiten vernetzt?
    Beispiel: Zitate in wissenschaftlichen Artikel
  - Welche Seite verweißt auf eine andere?
    Ranghöhe von Seite A bestimmt Ranghöhe von Seite B



* Stand: Feb. 2001

- Menge von Dokumenten $D = \{d_1, ..., d_m\}$
- optimales Ranking r*