

Detecting Near Duplicates for Web Crawling

Gurmeet Singh Manku et al., WWW
2007*

Detecting Near Duplicates for Web Crawling

- Finde „near-duplicates“ in großen Repositories
 - Mehrere Milliarden Web Dokumente
 - Identischer Inhalt
 - Kleine, irrelevante Unterschiede, z.B. Werbung
 - Charikar's simhash: Ähnliche Dokumente haben ähnliche Hash Werte (fingerprints)
- Effizientes Finden aller fingerprints mit Hamming Abstand $\leq k$ (online & batch)

Brute Force and Indexed Approaches to Pairwise Document Similarity Comparisons with MapReduce

Jimmy Lin, SIGIR 2009*

Efficient Search Ranking in Social Networks

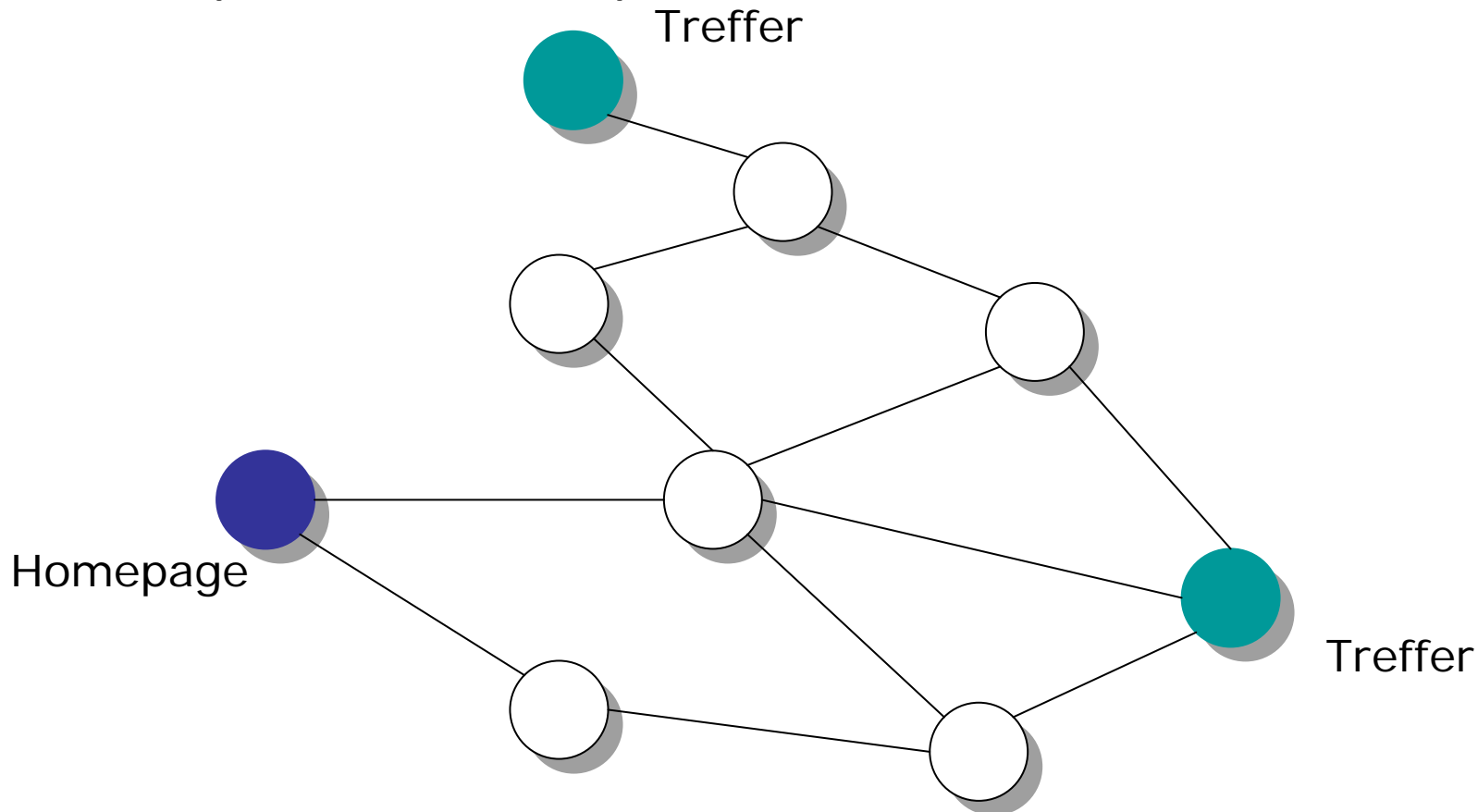
Monique V. Vieira et al.,
CIKM 2007*

Efficient Search Ranking in Social Networks

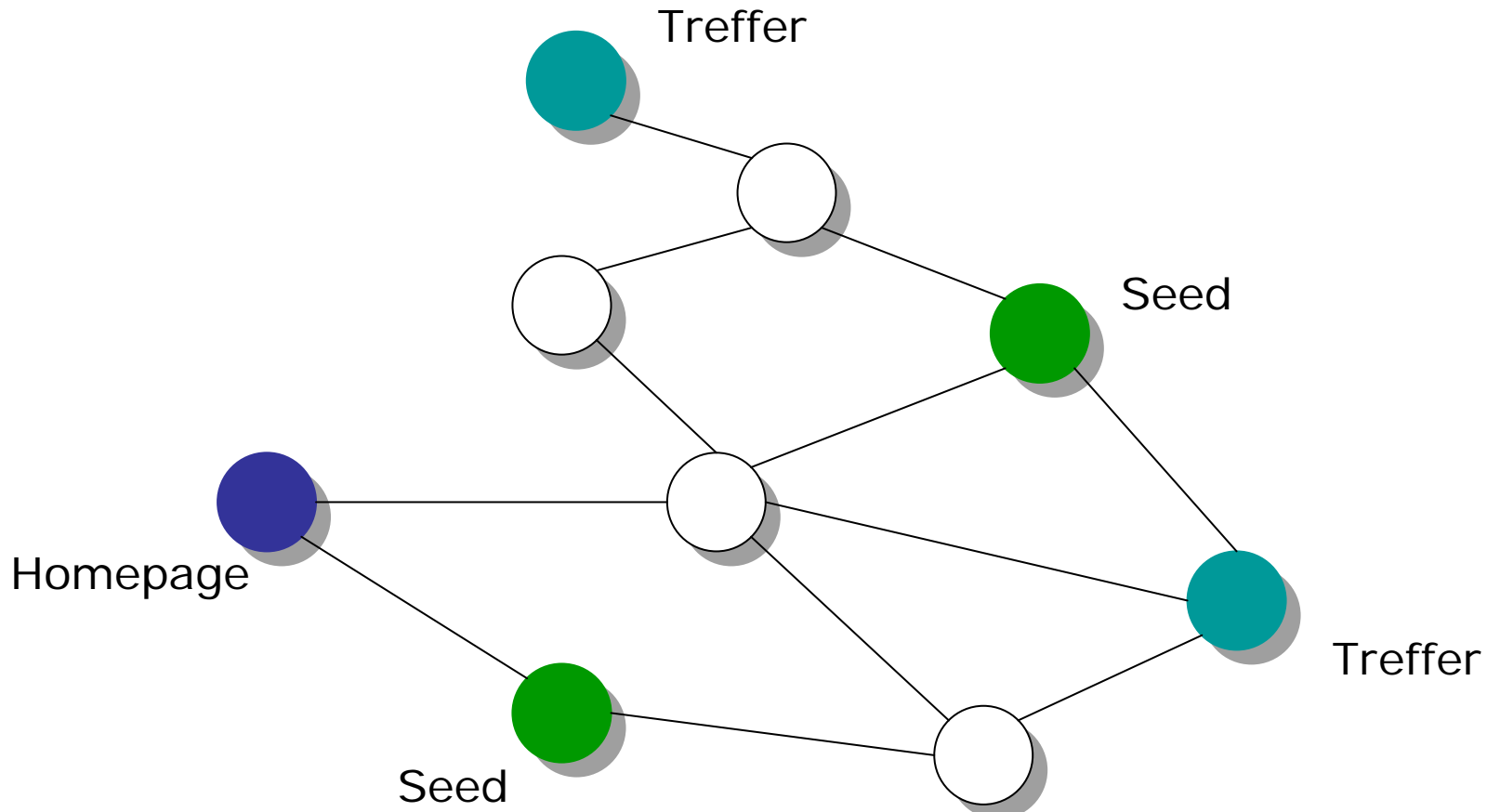
- Berücksichtigung der Wikipedia-Linkstruktur für Ranking
- Beispiel
 - Startseite ist „Kraftfahrzeug“
 - Suche nach „Golf“: „VW Golf“ höher gerankt als „Golfball“
- Vorgehen
 - Mindestabstand zu vorgegebener Startseite beeinflusst Ranking
 - Annahme: Nahe Seiten sind relevanter für den Nutzer

Efficient Search Ranking in Social Networks

- DBpedia Link Graph



Efficient Search Ranking in Social Networks



Graph Twiddling in a MapReduce World

Jonathan Cohen, Computing in
Science and Engineering, 2009

Graph Twiddling in a MapReduce World

- Bekannte Graph-Algorithmen mit MapReduce
 - Breitensuche, Tiefensuche, Zusammenhangskomponenten
- Graph Traversal in MapReduce?
- Graph Repräsentation in MapReduce?
- Siehe auch Google Lecture: [Cluster Computing and MapReduce Lecture 5](#)

The PageRank Citation Ranking: Bringing Order to the Web

Page, Lawrence and Brin, Sergey and
Motwani, Rajeev and Winograd,
Terry, Technical Report, Stanford
InfoLab, 1999

Scalable Distributed Reasoning using Map Reduce

Jacopo Urbani et al., ISWC 2009*

Scalable Distributed Reasoning using Map Reduce

- RDF als Datenmodell für die Repräsentation der extrahierten Daten
- RDF Schema mit Klassen, Properties, Constraints

The screenshot shows a Mozilla Firefox browser window displaying the Wikipedia page for "Linked Data". The browser's address bar shows the URL `http://dbpedia.org/page/Linked_Data`. The page content includes the title "About: [Linked Data](#)" and the subtitle "An Entity in Data Space: dbpedia.org". A red circle highlights the introductory paragraph: "Linked Data is a sub-topic of the Semantic Web. The term Linked Data is used to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web." Another red circle highlights the "dbpprop:abstract" property value, which is a list of links. A third red circle highlights the "Categories" section, which includes "Semantic Web" and "World Wide Web". A fourth red circle highlights the "rdf:type" property value, which is "vago-class:Buzzwords".

Property	Value
dbpprop:abstract	<ul style="list-style-type: none">Linked Data is a sub-topic of the Semantic Web. The term Linked Data is used to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web.鏈連資料是正在快速發展的語義網的一系列的活動，它描述了一套在全球資訊網上發佈、分享和連接數據的方法。
dbpprop:hasPhotoCollection	<ul style="list-style-type: none">http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Linked_Data
dbpprop:reference	<ul style="list-style-type: none">http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkingOpenData.pdfhttp://www.w3.org/2005/ajar/tabhttp://linkeddata.orghttp://community.linkeddata.org/http://www.linkeddataplanet.com/http://dataviewer.zitgist.com/http://www.ted.com/talks/tim_berners_lee_on_the_next_web.htmlhttp://www.w3.org/DesignIssues/LinkedData.htmlhttp://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/http://virtuoso.openlinksw.com/Whitepapers/html/VirtLinkedDataDeployment.htmlhttp://virtuoso.openlinksw.com/presentations/Creating_Deploying_Exploiting_Linked_Data2http://isweb.uni-koblenz.de/Research/lena/http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/http://linkeddata.uriburner.com/odehttp://linkeddata.org/slides/2008-02-amsterdam-catch.pdfhttp://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdfhttp://virtuoso.openlinksw.com/presentations/Virtuoso_Deploying_Linked_Data/Virtuoso_Dehttp://esw.w3.org/topic/TaskForces/CommunityProjects/LinkingOpenData/Datasetshttp://www2008.org/papers/pdf/p1265-bizer.pdf
rdf:type	<ul style="list-style-type: none">vago-class:Buzzwords

Scalable Distributed Reasoning using Map Reduce

- dbpedia.org/page/Linked_Data **rdf:type**
dbpedia.org/class/yago/Buzzwords **rdfs:subClassOf**
dbpedia.org/class/yago/Buzzword106608277 **rdfs:subClassOf**
dbpedia.org/class/yago/Nonsense106607339
- Regelbasiertes RDFS-Reasoning mit MapReduce
 - **s rdf:type X & X rdfs:subClassOf Y =>**
s rdf:type Y
 - ...
- Example: DBpedia
 - Input 150.1 M, Output 172.0 M
 - Time 5'20''

DisCo: Distributed Co-clustering with Map-Reduce

Spiros Papadimitriou and
Jimeng Sun, ICDM 2008*

DisCo: Distributing Mastering with

Miss University of Florida



Formation	1956
Type	Beauty Pageant
Headquarters	Gainesville, Florida
Location	United States
Key people	Florida Blue Key
Website	Official website

Elysian Airlines



IATA	ICAO	Callsign
-	-	-
Founded		
Hubs	Nsimalen International Airport, Conakry International Airport	
Destinations	9	
Company slogan	The African Dream	
Headquarters	Yaoundé, Cameroon	
Key people	Dr Yomi Ojetunde, Prof Julius Oben	
Website	http://www.elysianairlines.com/	

Hughes Airwest



IATA	ICAO	Callsign
RW	RW	Airwest
Founded		
	1941 (as West Coast Airlines)	
	1941 (as Southwest Airways)	
	1945 (as Bonanza Air Lines)	
Commenced operations	1968	
Ceased operations	1980 (merged with Republic Airlines).	
Destinations		
Headquarters		
Key people	Howard Hughes - Owner	

Aloha Airlines



IATA	ICAO	Callsign
AQ	AAH	ALOHA
Founded		
	July 26, 1946 (as Trans-Pacific Airlines) ^[1]	
Commenced operations	July 26, 1946	
Ceased operations	March 31, 2008 (All operations transferred to Aloha Air Cargo for cargo flights)	
Hubs	Honolulu International Airport	
Frequent flyer program	AlohaPass	
Member lounge	Ali'i Club	
Fleet size	22	
Destinations	12	
Parent company	Aloha Air Group	
Headquarters	Honolulu CDP, Hawaii	
Key people	David Banmiller (President & CEO)	
Website	alohaairlines.com	

DisCo: Distributed Co-clustering with Map-Reduce

- Infoboxen von "Elysian Airlines" und "Hughes Airwest" beschreiben das gleiche Thema, haben aber nur ein Attribut gemeinsam
- Co-occurrence: Attribute kommen häufig in anderen Infoboxen vor, die das gleiche Thema beschreiben

Company Slogan Hubs Destinations Headquarters Website Key people ...

	Company Slogan	Hubs	Destinations	Headquarters	Website	Key people ...
Elysian Airlines	Gray	Gray	Gray	Gray	White	White
Aloha Airlines	White	Green	Green	Green	Gray	Gray
Hughes Airwest	White	White	White	Gray	Gray	White
Miss University of Florida	White	White	Green	Green	White	Gray

Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

Petr Krajca and Vilem Vychodil,
IDA 2009*

Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

- Begriffe (Konzepte) formen in der Kontexttabelle maximal gefüllte Rechtecke
 - Vertauschung von Zeilen und Spalten erlaubt
 - Airline: Hubs, Destinations, Headquarters, Website, Key people

Elysian Airlines																						
Aloha Airlines																						
Hughes Airwest																						
Miss University of Florida																						

The table illustrates a context table with four rows and 22 columns. The first column contains the names of the entities: Elysian Airlines, Aloha Airlines, Hughes Airwest, and Miss University of Florida. The second column is shaded gray. Two overlapping red rectangles are shown, representing maximal filled rectangles (concepts) in the table. The first red rectangle covers the first four columns of the first three rows. The second red rectangle covers the next four columns of the first three rows. The remaining columns in the first three rows are shaded gray, and the bottom row is mostly shaded gray with a few white cells.

Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

	0	1	2	3	4	5	6
0		×	×		×	×	×
1	×		×	×			×
2	×	×					
3	×		×	×		×	
4	×		×		×		×

$$C_1 = \langle \{0, 1, 2, 3, 4\}, \{\} \rangle,$$

$$C_2 = \langle \{1, 2, 3, 4\}, \{0\} \rangle,$$

$$C_3 = \langle \{2\}, \{0, 1\} \rangle,$$

$$C_4 = \langle \{\}, \{0, 1, 2, 3, 4, 5, 6\} \rangle,$$

$$C_5 = \langle \{1, 3, 4\}, \{0, 2\} \rangle,$$

$$C_6 = \langle \{1, 3\}, \{0, 2, 3\} \rangle,$$

$$C_7 = \langle \{3\}, \{0, 2, 3, 5\} \rangle,$$

$$C_8 = \langle \{1\}, \{0, 2, 3, 6\} \rangle,$$

$$C_9 = \langle \{4\}, \{0, 2, 4, 6\} \rangle,$$

$$C_{10} = \langle \{1, 4\}, \{0, 2, 6\} \rangle,$$

$$C_{11} = \langle \{0, 2\}, \{1\} \rangle,$$

$$C_{12} = \langle \{0\}, \{1, 2, 4, 5, 6\} \rangle,$$

$$C_{13} = \langle \{0, 1, 3, 4\}, \{2\} \rangle,$$

$$C_{14} = \langle \{0, 4\}, \{2, 4, 6\} \rangle,$$

$$C_{15} = \langle \{0, 3\}, \{2, 5\} \rangle,$$

$$C_{16} = \langle \{0, 1, 4\}, \{2, 6\} \rangle.$$