Detecting Near Duplicates for Web Crawling

Gurmeet Singh Manku et al., WWW 2007*

* 16th international conference on World Wide Web

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)

- Effizentes Finden aller fingerprints mit Hamming Abstand <= k (online & batch)

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

Jimmy Lin, SIGIR 2009*

* 32nd Conference on Research and Development in Information Retrieval

Efficient Search Ranking in Social Networks

Monique V. Vieira et al.,
CIKM 2007*

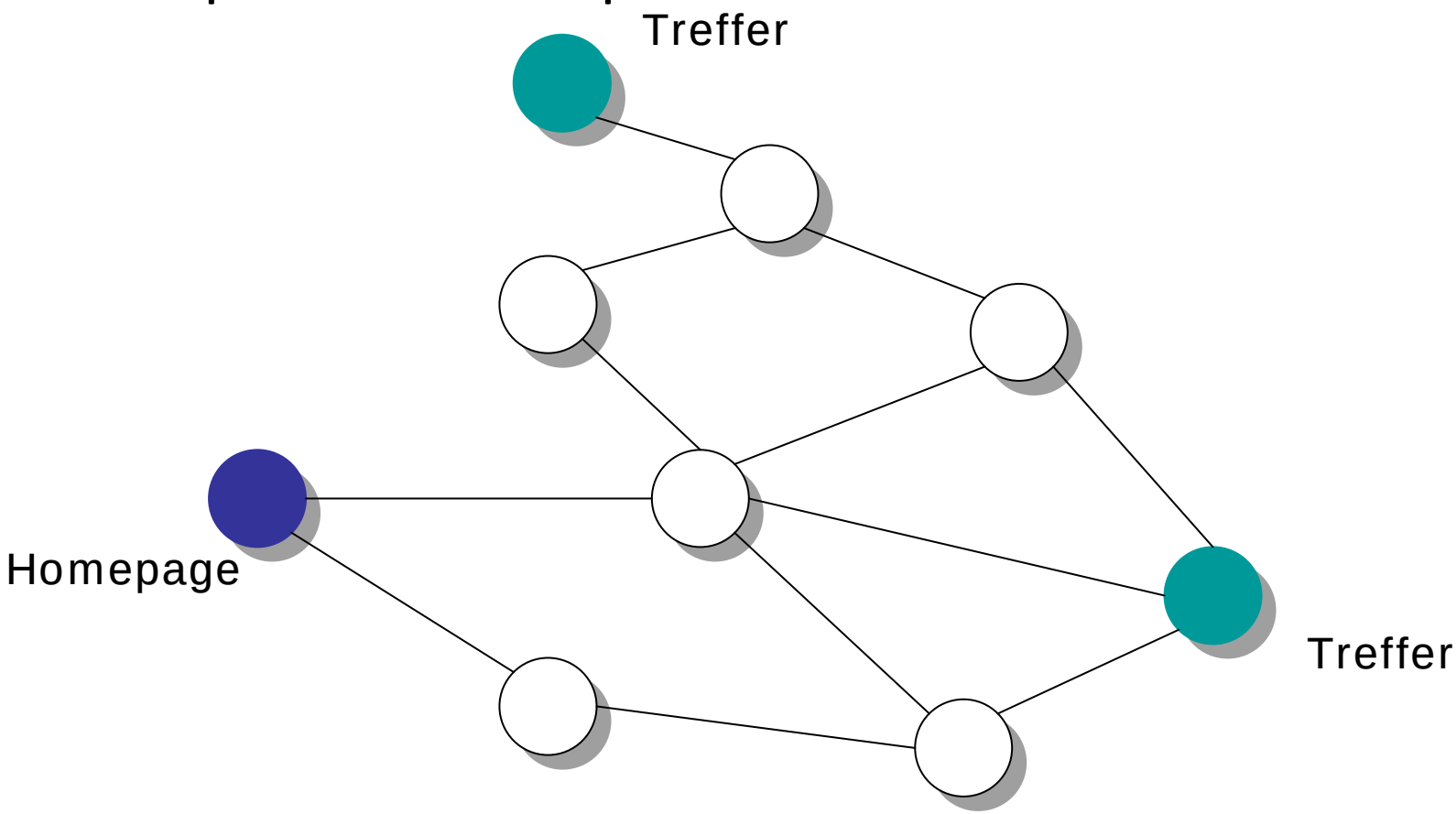
* 16th International Conference on Information and Knowledge Management

Efficient Search Ranking in Social Networks

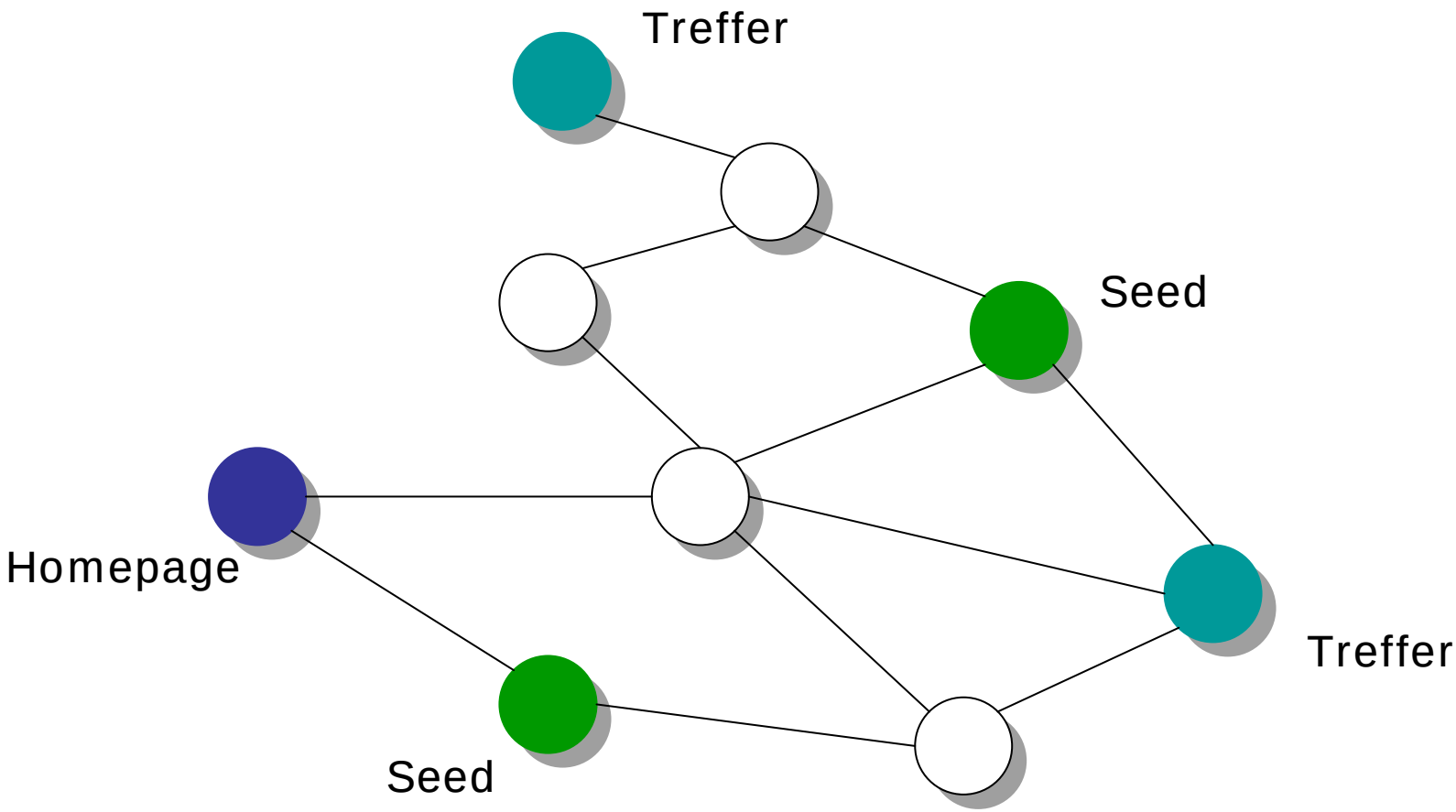
- Berücksichtigung der Wikipedia-Linkstruktur für Ranking
- Beispiel
  - Startseite ist „Kraftfahrzeug“
  - Suche nach „Golf“: „VW Golf“ höher gerankt als „Golfball“
- Vorgehen
  - Mindestabstand zu vorgegebener Startseite beeinflusst Ranking
  - Annahme: Nahe Seiten sind relevanter für den Nutzer

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, Zusammenhangs-komponenten
- 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

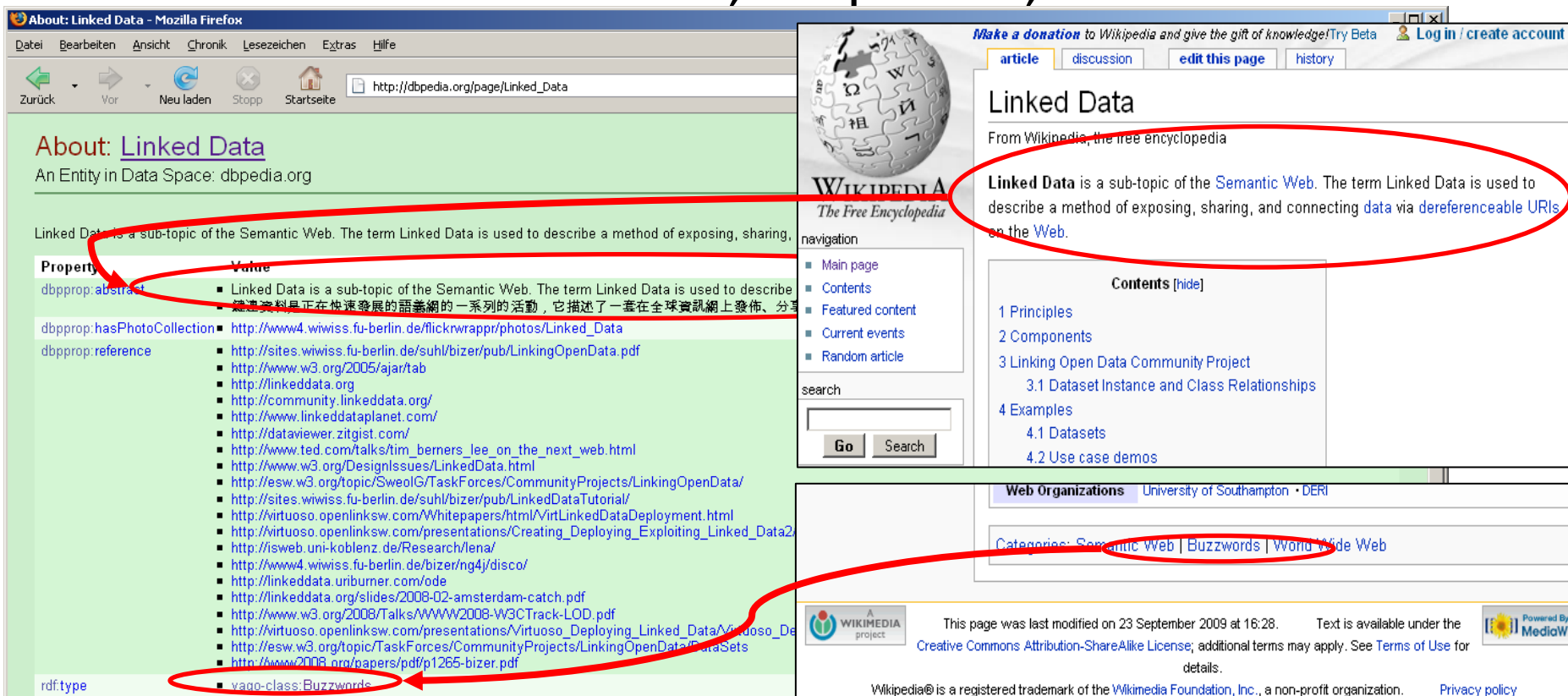
Scalable Distributed Reasoning using Map Reduce

Jacopo Urbani et al., ISWC 2009*

* 8th International Semantic Web Conference

Scalable Distributed Reasoning using Map Reduce

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



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*

* 8th IEEE International Conference on Data Mining

DisCo: Distri[illegible]ustering with [illegible]e

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

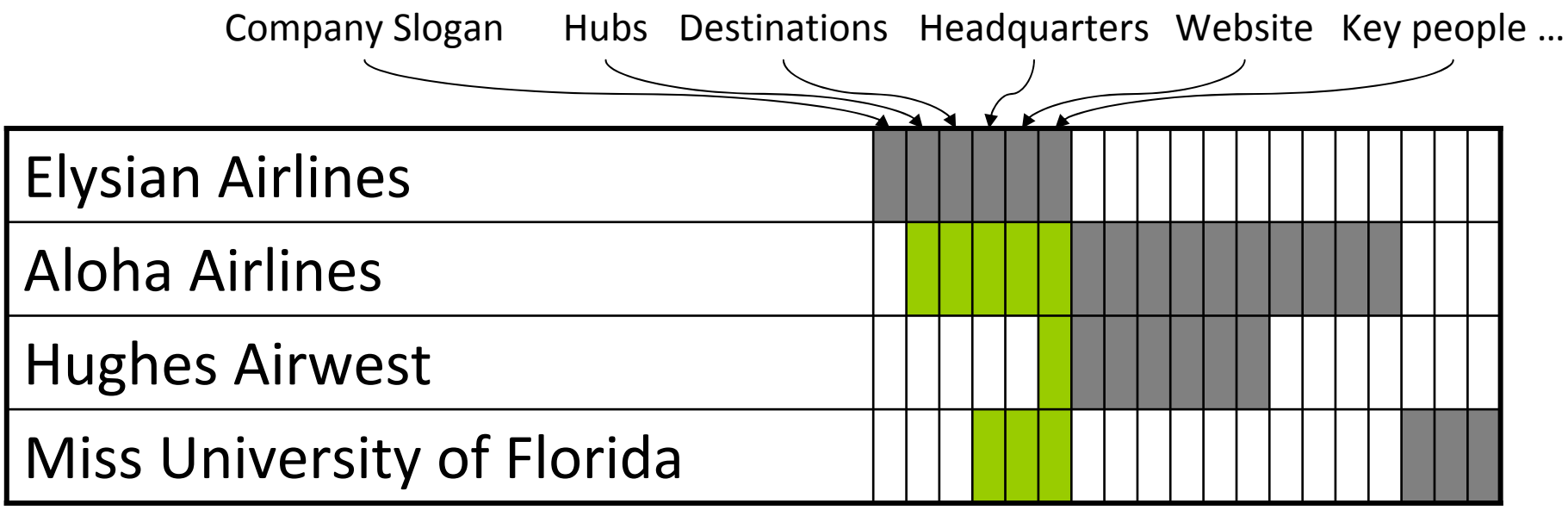
Aloha Airlines

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-occurence: Attribute kommen häufig in anderen Infoboxen vor, die das gleiche Thema beschreiben



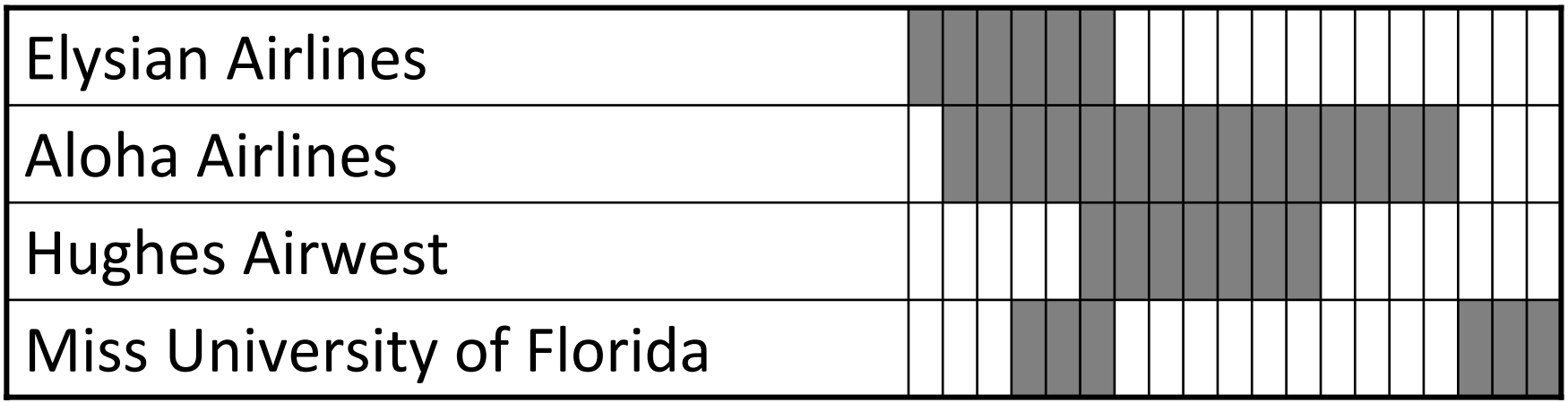
Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

Petr Krajca and Vilem Vychodil,
IDA 2009*

* 8th International Symposium on Intelligent Data Analysis

Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

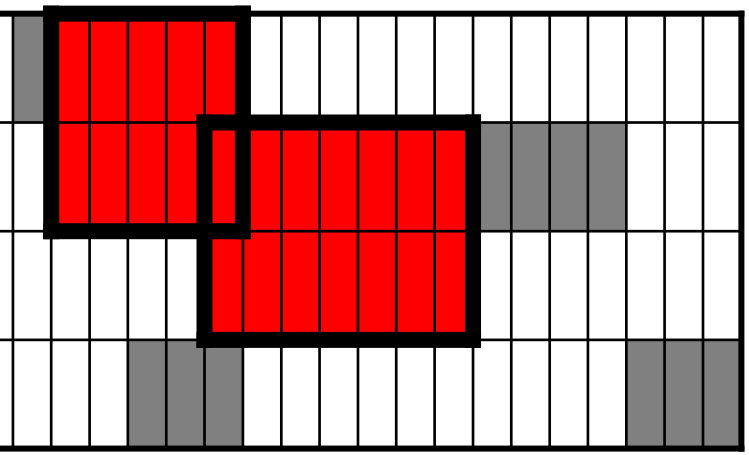
- Formale Begriffsanalyse
- Kontexttabelle
  - Binäre Relation
  - Objekte als Zeilen und Merkmale als Spalten



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	



Distributed Algorithm for Computing Formal Concepts Using Map-Reduce Framework

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

$C_1 = \langle\{0,1,2,3,4\},\{\}\rangle$,
$C_2 = \langle\{1,2,3,4\},\{0\}\rangle$,
$C_3 = \langle\{2\},\{0,1\}\rangle$,
$C_4 = \langle\{\},\{0,1,2,3,4,5,6\}\rangle$,
$C_5 = \langle\{1,3,4\},\{0,2\}\rangle$,
$C_6 = \langle\{1,3\},\{0,2,3\}\rangle$,
$C_7 = \langle\{3\},\{0,2,3,5\}\rangle$,
$C_8 = \langle\{1\},\{0,2,3,6\}\rangle$,
$C_9 = \langle\{4\},\{0,2,4,6\}\rangle$,
$C_{10} = \langle\{1,4\},\{0,2,6\}\rangle$,
$C_{11} = \langle\{0,2\},\{1\}\rangle$,
$C_{12} = \langle\{0\},\{1,2,4,5,6\}\rangle$,
$C_{13} = \langle\{0,1,3,4\},\{2\}\rangle$,
$C_{14} = \langle\{0,4\},\{2,4,6\}\rangle$,
$C_{15} = \langle\{0,3\},\{2,5\}\rangle$,
$C_{16} = \langle\{0,1,4\},\{2,6\}\rangle$.