

# Linkbasierte Strukturierung von Web-Seiten

Konzepte und Methoden

Dr. Michael Brinkmeier

Institut für Theoretische Informatik  
Technische Universität Ilmenau

28. September 2006

# Welche Informationen sind im WWW enthalten?

## Binsenweisheiten

- Das WWW ist eine der größten und wichtigsten Informationsquellen.
- Das WWW entwickelt sich unkontrolliert und damit sind die Informationen unstrukturiert.

## Kernfrage

Welche Informationen sind im WWW enthalten und wie können sie extrahiert/handhabbar gemacht werden?

# Welche Informationen sind im WWW enthalten?

- **Inhalte (textuell und multimedial)**

- ▶ „klassischer“ Ansatz  $\Rightarrow$  Texte
- ▶ Textbasierte Suchmaschine  $\hat{=}$  [Reverse Word Index](#),

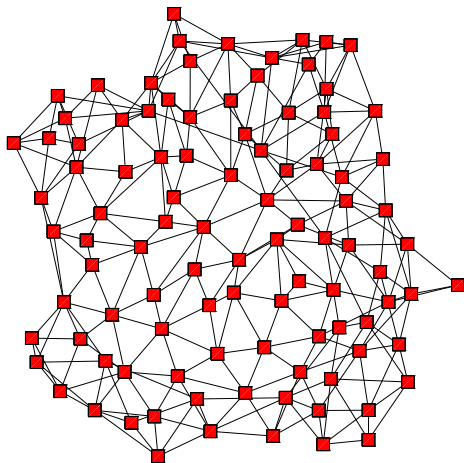
- In Zukunft: **Semantische Informationen (Semantic Web)**

- **Links**

- ▶ Bisher im Wesentlichen für das Ranking benutzt (z.B. PageRank, HITS)
- ▶ Links werden von den Autoren **bewußt** gesetzt.
- ▶ Ein Link stellt eine Beziehung zwischen zwei Seiten her
  - ★ **inhaltlicher Zusammenhang**
  - ★ soziale Beziehung (Freunde, Bekannte, Kollegen ...)
  - ★ wirtschaftliche Beziehung (Partnerfirmen, Kunden, Lieferanten ...)
  - ★ ...

# Mathematische Formalisierung

**Netzwerk/Graph** aus **Knoten** (Dokumente) und **Kanten** (Links)



# Kohäsive Gruppen

## Grundidee

Finde Gruppen von Knoten, die in irgendeiner Form dichter miteinander verbunden sind.

## Kohäsive Gruppen oder Communities

- Der Begriff **kohäsive (Unter-)Gruppe** stammt aus der Soziometrie, genauer aus der **Analyse sozialer Netzwerke**
  - ▶ Knoten = **Akteure / Individuen**
  - ▶ Kanten = **direkte Beziehungen**
- Eine Gruppe von **Individuen** heißt **kohäsiv**, wenn die Beziehungen zwischen ihnen im Vergleich zum gesamten Netzwerk in irgendeinem Sinn relativ **stark, direkt, intensiv, häufig** oder **positiv** sind.

# Kohäsive Gruppen

- Verschiedenste (mathematische) Definitionen, jede mit spezifischen Vor- und Nachteilen  
(z.B. Cliques, Kerne, LS-Mengen,  $\lambda$ -Mengen, bipartite Kerne, FLG-Communities, ...)
- Problem 1: (Effiziente) Berechnung dieser Gruppen
  - ▶ Algorithmen
  - ▶ verteilte Berechnung
  - ▶ Datenstrukturen
- Problem 2: sehr große Netzwerke
  - ▶ Speicherplatzintensiv
  - ▶ Entschärfung durch verteilte Berechnung möglich?

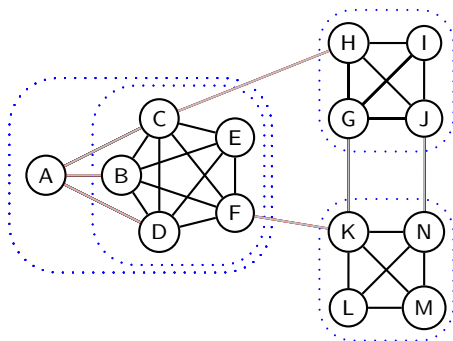
# Nutzung von kohäsiven Gruppen im WWW

- Nicht nur reine Auflistung der gefundenen Seiten nach Relevanz
- Gruppierung der Ergebnisse nach **Communities**
- Zusätzliche Anzeige von besonders relevanten Seiten, die sich in den gleichen Communities befinden (enthalten nicht unbedingt genau den gesuchten Begriff)
- Durchsuchen des WWW durch „Zusammenfalten“ von uninteressanten Gruppen oder Untergruppen

# Der von uns verwendete Begriff der Kohäsion

Kohäsion = Kantenzusammenhang

Kohäsion = minimale Anzahl von Kanten, die entfernt werden müssen, damit die Gruppe in (mindestens) zwei Teile zerfällt.



Diese Art der Kohäsion führt zu einer **Hierarchie** von Communities.



# Die Web-Domain tu-ilmenau.de

72.907 Knoten, 434.625 Kanten, 0:37:28 Stunden

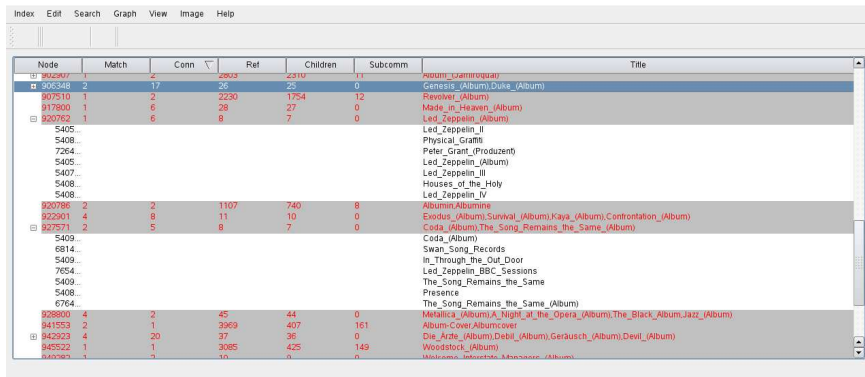
Beispiel: Suche nach **mbrinkme** in der URL

Node	Match	Conn	Ref	Children	Subcomm	Title
root						
17169	7	2	58180	9972	40	http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/documents/ma.pdf
29552	2	3	42276	9631	44	http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/documents/communitie...
61138	15	10	20	19	0	http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/publications.php
63732	1	2	296	37	1	http://eiche.theoinf.tu-ilmenau.de/%7Embrinkme/de/studienarbeiten.html
62232	1	2	25	4	1	http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/documents/communities.html
63769	1	3	259	250	1	http://eiche.theoinf.tu-ilmenau.de/~mbrinkme
61136	1	9	22	2	1	http://eiche.theoinf.tu-ilmenau.de/%7Embrinkme/de/index.php
61138		10	20	19	0	
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/
7..						http://eiche.theoinf.tu-ilmenau.de/lehre/quiz
7..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/publications.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/privat/games.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/clearindex.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/research.php
2..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/diplom.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/cv.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/links.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/en/lectures.php
7..						http://wcms1.rz.tu-ilmenau.de/fakia/Institut_fuer_Theore634.0.html
8..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/publications.php
6..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/research.php
5..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/lectures.php
4..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/links.php
4..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/cv.php
1..						http://eiche.theoinf.tu-ilmenau.de/~mbrinkme/de/diplom.php
3..						http://kt3.theoinf.tu-ilmenau.de/quiz/login.php
7..						http://kt3.theoinf.tu-ilmenau.de/quiz/
65938						http://eiche.theoinf.tu-ilmenau.de/%7Embrinkme/de/index.php

# Die deutsche Wikipedia

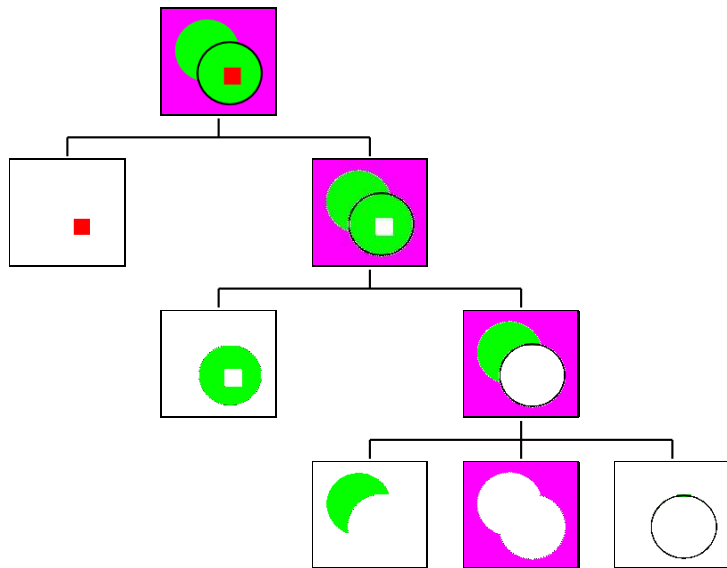
978.888 Knoten, 12.850.009 Kanten, ca. 4 Stunden

Beispiel: Suche nach **Album** im Titel des Eintrags



Node	Match	Conn	Ref	Children	Subcomm	Title
902907	1	2	2803	2310	11	Album_Damiroquai
906348	2	17	26	25	0	Genesis_(Album),Duke_(Album)
907510	1	2	2230	1754	12	Revolver_(Album)
917800	1	6	28	27	0	Made_in_Heaven_(Album)
920762	1	6	8	7	0	Led_Zeppelin_(Album)
5405...						Led_Zeppelin_II
5408...						Physical_Graffiti
7264...						Peter_Grant_(Produzent)
5405...						Led_Zeppelin_(Album)
5407...						Led_Zeppelin_III
5408...						Houses_of_the_Holy
5408...						Led_Zeppelin_IV
920786	2	2	1107	740	8	Albumin,Albumine
922901	4	8	11	10	0	Exodus_(Album),Survival_(Album),Kaya_(Album),Confrontation_(Album)
927571	2	5	8	7	0	Coda_(Album),The_Song_Remains_the_Same_(Album)
5409...						Coda_(Album)
6814...						Swan_Song_Records
5409...						In_Through_the_Out_Door
7854...						Led_Zeppelin_BBC_Sessions
5409...						The_Song_Remains_the_Same
5408...						Presence
6764...						The_Song_Remains_the_Same_(Album)
928900	4	2	45	44	0	Metallica_(Album),A_Night_at_the_Opera_(Album),The_Black_Album,Jazz_(Album)
941553	2	1	3969	407	161	Album-Cover,Albumcover
942923	4	20	37	38	0	Die_Arzte_(Album),Debil_(Album),Geräusch_(Album),Devil_(Album)
945522	1	1	3085	425	149	Woodstock_(Album)
946382	1	2	10	0	0	Welcome_Innocente_Misconceps_(Album)

# Ein Bild sagt mehr als tausend Worte



# Aktuelle und zukünftige Arbeiten

- Verbesserung der Algorithmen/Implementierung
- Verbesserung der Vor- und Nachverarbeitung
- Abfrageschnittstelle
- Visualisierung
  - ▶ Layout der Netzwerke auf der Basis der Communities
- Erweiterung des Konzeptes auf sog. Hypergraphen
  - ▶ Wort-Dokument-Beziehungen
- Experimente mit verschiedenen Netzwerken

# Danke für Ihre Aufmerksamkeit!

Dr. Michael Brinkmeier

Institut für Theoretische Informatik  
Technische Universität Ilmenau

`mbrinkme@tu-ilmenau.de`