



Fokussierte Informationssuche in sozialen Netzen

Ralf Schenkel

*Gemeinsame Arbeit mit Tom Crecelius, Mouna Kacimi, Sebastian Michel,
Thomas Neumann, Josiane Parreira, Gerhard Weikum*

Social Tagging Networks

Definition: Social Tagging Network

Webseite, wo Menschen

- Information **veröffentlichen** und **annotieren**
- Information **bewerten** und begutachten
- ihre **Interessen** veröffentlichen
- ein Netzwerk von **Freunden** unterhalten
- mit Freunden **interagieren**

Bekannte Beispiele:

- Flickr (Fotos) 
- YouTube (Videos) 
- del.icio.us (Bookmarks)
- Librarything (Bücher) 
- Discogs (CDs) 
- CiteULike (Publikationen) 
- Facebook 
- Myspace (Medien) 

Showcase: librarything.com

Bücher

	Author	Date	Tags	Ratings	Google Books	Andere
						Shared
	Larry Wasserman	2004	statistics, math, statistical inference	☆☆☆☆☆	Partial view	23
	Stuart J. Russell	2002	artificial intelligence	☆☆☆☆☆	Book info	476/2
	Ian H. Witten	2005	data mining, machine learning	★★★★☆	Partial view	90
	Hector Garcia-Molina	2001	database systems	☆☆☆☆☆	Book info	18
	Dennis Shasha	2002	database tuning	★★★★★	Partial view	8
	Phil Collins	2007		☆☆☆☆☆	Book info	2
	David A. Grossman	2004	information retrieval, ir, algorithms	★★★☆☆	Partial view	17/2
	Kate Mosse	2007		★★★★★	Book info	1661/63

librarything.com: Soziale Interaktion

Member: ralf.schenkel

Library 20 books — see library

Reviewed None so far

Clouds tag cloud, author cloud

Tags statistics (3), math (3), travel (3), algorithms (2), china (2), database systems (2), artificial intelligence (1), probabilistic methods (1), queueing theory (1), tree locking (1) — see all tags

Groups None

Favorite None specified (how to add)

Account type public, free (upgrade to paid account)

URLs <http://www.librarything.com/profile/ralf.schenkel> (profile)
<http://www.librarything.com/catalog/ralf.schenkel> (library)

Kommentare 8

Comments from other LibraryThing-ers

[disable comments](#) | [show archived comments](#) | [archive all comments](#)

No comments posted.

Leave your comment

You can use some HTML tags, such as ``, `<i>` and `<a>`.

 [Edit profile/account settings](#)

 [Put your library on your blog](#)

Ähnliche Benutzer

Members with your books

Showing weighted | raw

scg (8/633), flint63 (7/879), tvassilakis (2/5), intel4004 (3/63), alunufal (3/50), rmeindl (5/635), the5eeker (3/100), Almohandes (2/4), stibbits (2/83), c0dekhan (2/114), kingpiko (3/151), da_zhuang (3/185), paulbatt (3/183), martinkenny (3/272), iena (2/83), macrakis (2/208), ceperez (2/88), ohenzo (2/23), ExpatJane (2/112), fantamic (2/74), Sappetta (2/121), jensgram (2/87), harro (2/62), crono (2/9), (show more)

Explizite Freunde

Member connections

Interesting libraries: kingpiko, martinkenny, rmeindl, scg

Friends (pending): ranwar
(Edit/see other members' connections)

Random books from ralf.schenkel's library

Database Tuning: Principles

librarything.com: Suche

LibraryThing^{BETA}

ralf.schenkel [sign

What's on your book

Your library Add books Your profile Tags Search Tools

Talk Groups Local Zeitgeist About

Tag info: travel

Includes: travel, reizen (what?)

Tag and its aliases used 87,027 times by 9,747 users. You use this tag 3 times ([see tagged books](#))

Most often tagged travel

- Notes from a small island by Bill Bryson (631)
- In a sunburned country by Bill Bryson (569)
- 1,000 places to see before you die by Patricia Schultz (399)
- The lost continent : travels in small-town America by Bill Bryson (439)
- A walk in the woods : rediscovering America on the... by Bill Bryson (603)
- Neither here nor there : travels in Europe by Bill Bryson (420)
- I'm a stranger here myself : notes on returning to America... by Bill Bryson (348)
- Blue highways : a journey into America by William Least Heat-Moon (225)
- The great railway bazaar : by train through Asia by Paul Theroux (187)
- The art of travel by Alain de Botton (202)
- In Patagonia by Bruce Chatwin (184)
- London by Michael Leapman (132)
- A year in Provence by Peter Mayle (235)
- Riding the iron rooster : by train through China by Paul Theroux (156)
- The kingdom by the sea : a journey around Great Britain by Paul Theroux (142)
- A...
- T...
- T...
- G...
- Under the Tuscan Sun by Frances Mayes (214)

Suchergebnisse unabhängig vom anfragenden Benutzer (und seinem sozialen Kontext)

([see more](#)|[see raw count](#))

[your tags](#) | [LT tag cloud](#) | [LT a](#)

Related tags (show number)

- [adventure](#) [Africa](#) [America](#)
- [architecture](#) [Art](#) [Asia](#) [australia](#)
- [autobiography](#) [Biography](#)
- [Britain](#) [California](#) [China](#) [Culture](#)
- [England](#) [essays](#) [Europe](#)
- [fiction](#) [food](#) [France](#)
- [geography](#) [Germany](#) [Greece](#)
- [guide](#) [guidebook](#) [Hiking](#)
- [History](#) [Humor](#) [humour](#)
- [India](#) [Ireland](#) [Italy](#) [Japan](#)
- [literature](#) [London](#) [maps](#) [memoi](#)
- [nature](#) [non-fiction](#) [own](#)
- [Paris](#) [Photography](#) [read](#)
- [reference](#) [Spain](#) [travel guide](#)
- [unread](#) [USA](#)

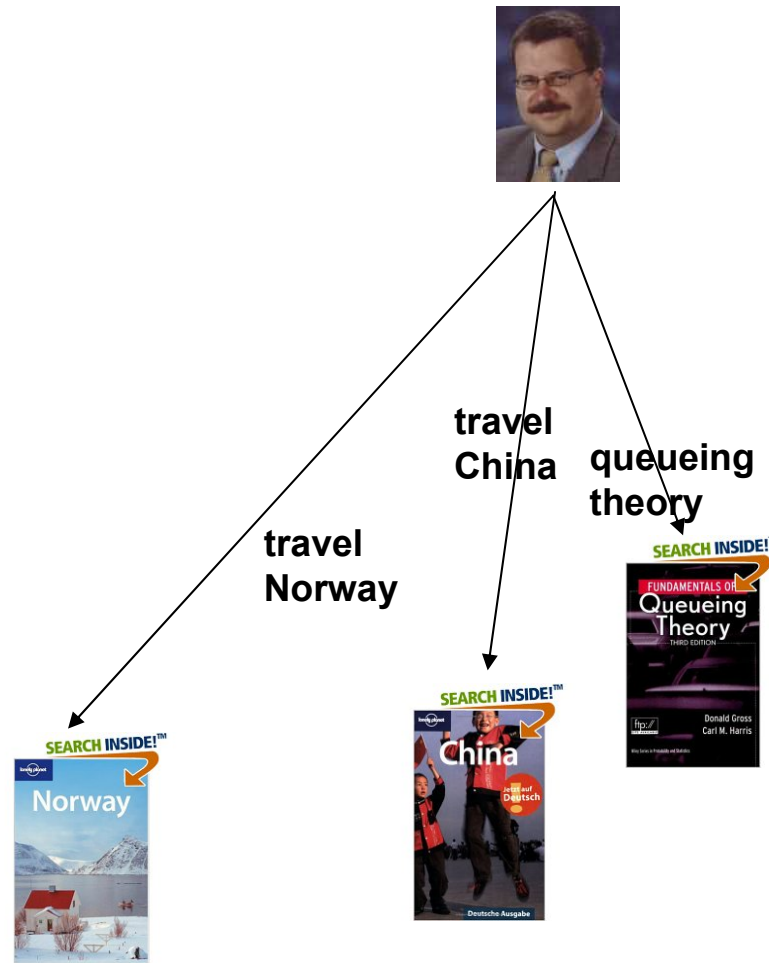
Related subjects

- [Voyages and travels](#) (2,141)
- [United States > Description and travel](#) (1,745)
- [Large type books](#) (1,541)

Outline

- **Modellierung von Social Tagging Networks**
 - Graphmodell
 - Verschiedene Informationsbedürfnisse
- Effektives Scoring für Anfragen
- Effiziente Ausführung von Anfragen
- Zusammenfassung und offene Fragen

Modellierung von Social Tagging Networks

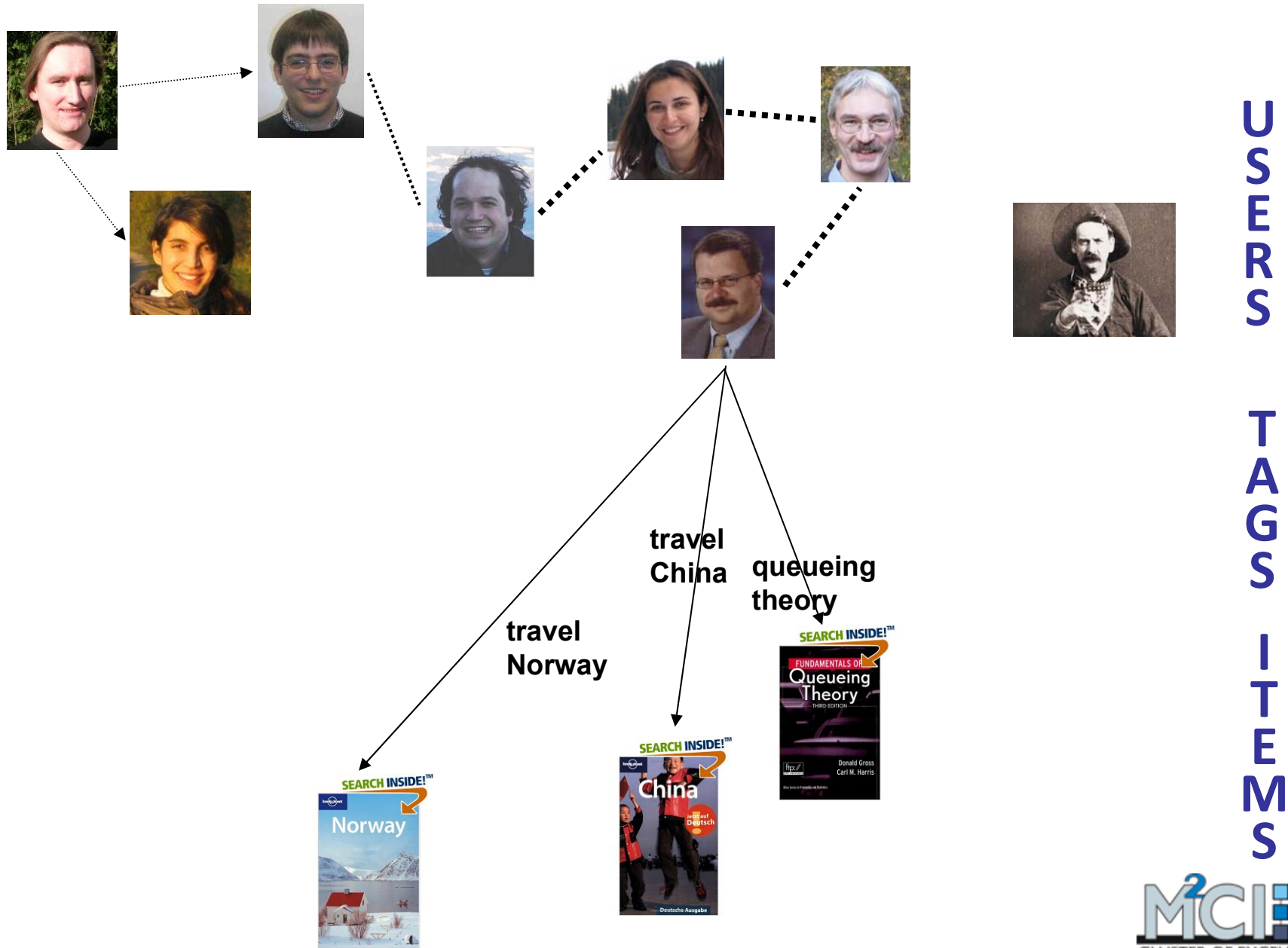


U
S
E
R
S

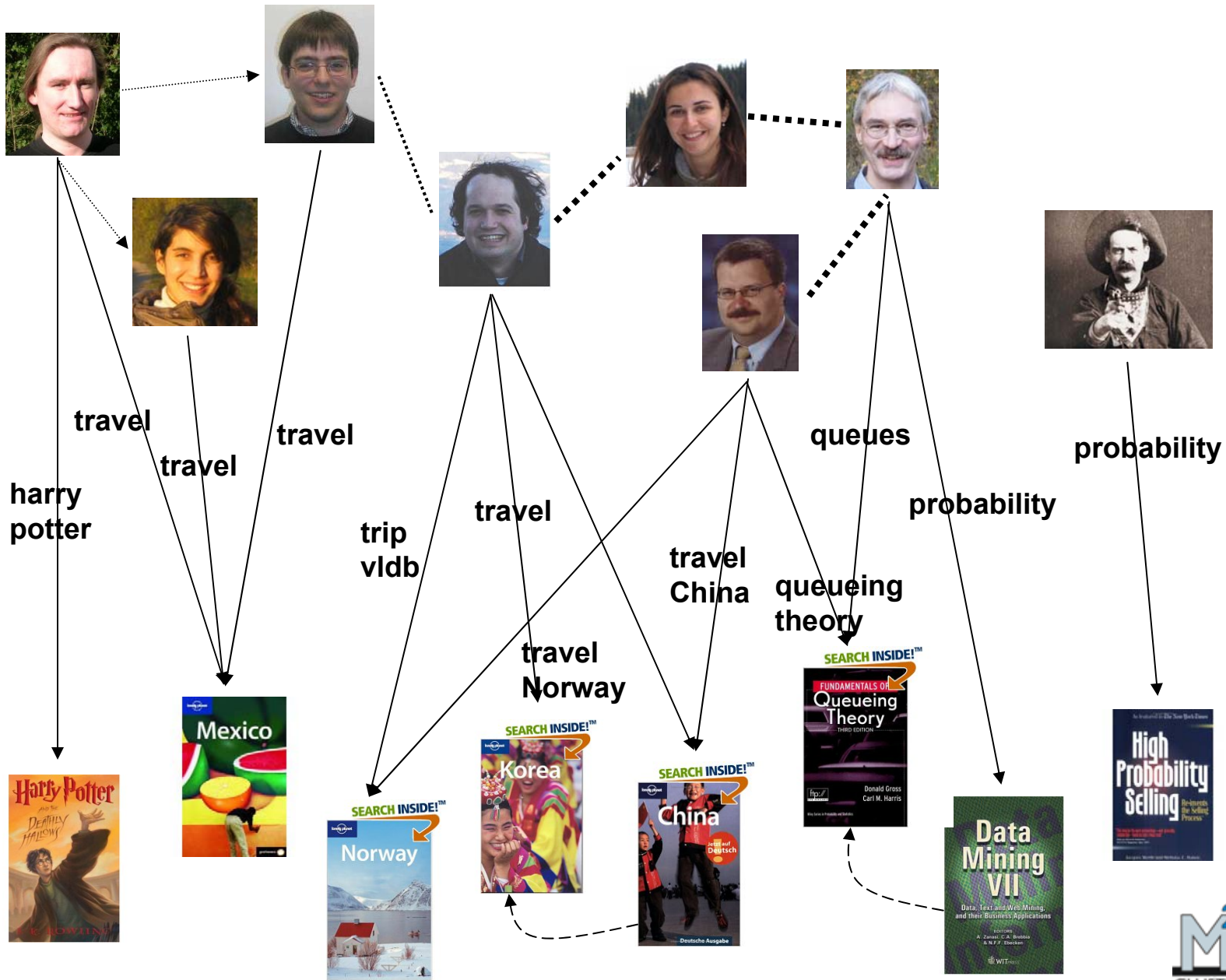
T
A
G
S

I
T
E
M
S

Modellierung von Social Tagging Networks



Modellierung von Social Tagging Networks



USERS

TAGS

ITEMS

Bausteine eines Social Tagging Network

Graph $G=(U \cup I, E_U \cup E_I \cup E_{UI})$ mit

- 2 Arten von Knoten:
 - Users U (optional gewichtet)
 - Items I (optional gewichtet)
- 3 Arten von Kanten:
 - E_U : User-User (optional gewichtet)
 - E_I : Item-Item (optional gewichtet)
 - E_{UI} : User-Item (beschriftet mit Tags T , opt. gewichtet)

Informationsbedarf 1: Global



USERS

TAGS

ITEMS

Informationsbedarf 2: Ähnliche Benutzer



Wunschliste: Social-Aware Social Search

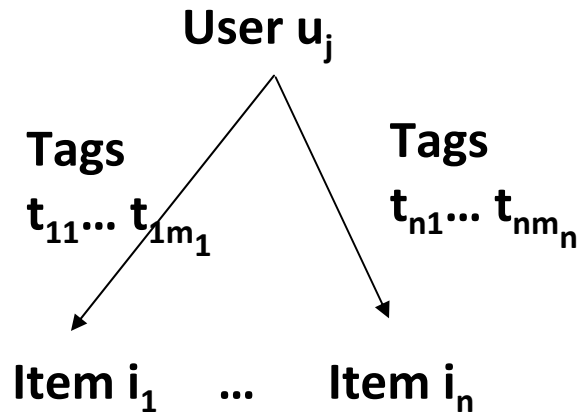
- Ergebnisse einer Suche hängen ab von
 - **Globaler** Popularität von Items
 - **Inhaltlichem Kontext** des anfragenden Benutzers (Items, Tags)
 - **Sozialem Kontext** des anfragenden Benutzers (Freunde)
- Automatische **Anfrageexpansion** (nicht nur Synonyme)
- **Skalierbare** Anfrageauswertung
- **Erklärung** der Ergebnisse

(Ähnliche Wunschliste für soziale Empfehlungssysteme)

Outline

- Modellierung von Social Tagging Networks
- **Effektives Scoring von Anfragen**
 - **Quantizierung von Freundschaften**
 - **Benutzerspezifische Scores**
 - **Experimentelle Evaluierung**
- Effiziente Ausführung von Anfragen
- Zusammenfassung und offene Fragen

Notation



tags(u): Tags, die von User u vergeben wurden

items(u): Items, die von User u getaggt wurden

items(t): Items, die mit Tag t getaggt wurden

df(t): Anzahl von Items, die mit Tag t getaggt wurden

tf_u(i,t): Häufigkeit, mit der Benutzer u Item i mit Tag t getaggt hat

tf(i,t): Häufigkeit, mit der Item i mit Tag t getaggt wurde

Quantifizierung von Freundschaftsgraden

- Globaler „Freundschafts“grad:

$$P_{global}(u, u') = \frac{1}{|U|}$$

- Inhaltsbasierter Freundschaftsgrad
- Graphbasierter Freundschaftsgrad
- Integrierter Freundschaftsgrad

Inhaltsbasierter Freundschaftsgrad

Verschiedene Möglichkeiten:

- **Gemeinsamkeiten der Tagverwendung:**

$$P_{content}(u, u') = \frac{2 | tags(u) \cap tags(u') |}{| tags(u) | + | tags(u') |}$$

- **Gemeinsamkeiten der annotierten Items:**

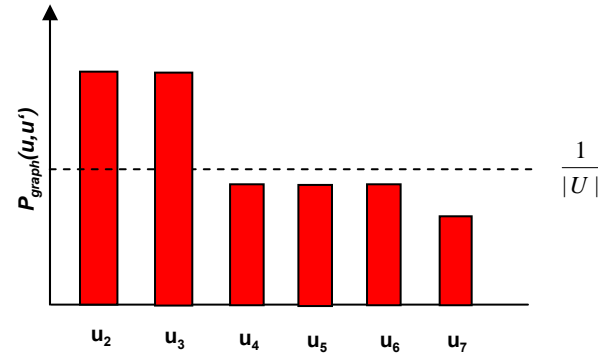
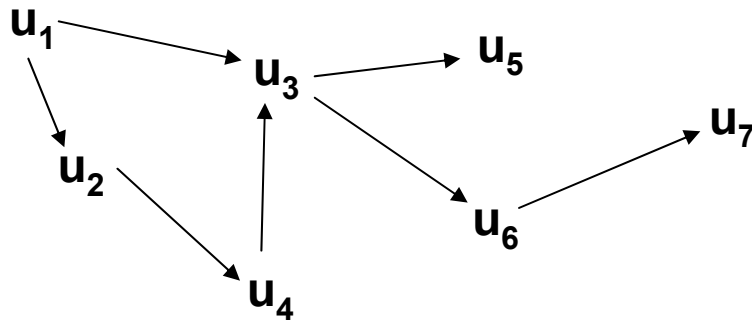
$$P_{content}(u, u') = \frac{2 | items(u) \cap items(u') |}{| items(u) | + | items(u') |}$$

Für beide:

- $P_{content}(u, u) := 0$

- **Normalisierung so dass** $\sum_{u' \neq u} P_{content}(u, u') = 1$

Graphbasierter Freundschaftsgrad



Ungewichtete Kanten:

$$w(u_i, u_{i+1}) = 1$$

$$w(u_{i_1}, \dots, u_{i_j}) = \sum_{k=1}^{j-1} w(u_{i_k}, u_{i_{k+1}}) = j - 1$$

$$P_{graph}(u, u') = \frac{1}{\min_{\text{path } p=u \dots u'} w(p)}$$

Für beide:

- $P_{graph}(u, u) := 0$

- **Normalisierung so dass** $\sum_{u' \neq u} P_{graph}(u, u') = 1$

Kanten gewichtet mit $P_{content}$:

$$w(u_i, u_{i+1}) = P_{content}(u_i, u_{i+1})$$

$$w(u_{i_1}, \dots, u_{i_j}) = \prod_{k=1}^{j-1} w(u_{i_k}, u_{i_{k+1}})$$

$$P_{graph}(u, u') = \max_{\text{path } p=u \dots u'} w(p)$$

Integrierter Freundschaftsgrad

Kombination von

- Inhaltsbasiertem Freundschaftsgrad
- Graphbasiertem Freundschaftsgrad
- Hintergrundmodell (globalem Freundschaftsgrad)

$$F(u, u') = \alpha \cdot P_{graph}(u, u') + \beta \cdot P_{content}(u, u') + (1 - \alpha - \beta) \frac{1}{|U|}$$

$$(0 \leq \alpha, \beta \leq 1; \alpha + \beta \leq 1)$$

Exkurs: Scores im Text Retrieval

Allgemeines Framework:

$$score(i, t) \propto tf(i, t) \cdot idf(t)$$

Wichtigkeit von t für Item i
(je häufiger, desto besser)

Wichtigkeit von t im Corpus
(je seltener, desto besser)

Handoptimierte Instanz: Okapi BM25

$$score(i, t) = \frac{(k_1 + 1)tf(i, t)}{k_1 + tf(i, t)} \cdot \log \frac{|I| - df(t) + 0.5}{df(t) + 0.5}$$

Linearkombination zum Scoring von Anfragen

$$score(i, t_1 \dots t_n) = \sum_{j=1}^n score(i, t_j)$$

Benutzerspezifisches Scoring

$$\begin{aligned}tf(i, t) &= \sum_{u \in U} tf_u(i, t) \\ &= |U| \cdot \underbrace{\sum_{u \in U} \frac{1}{|U|} tf_u(i, t)}\end{aligned}$$

Globaler Freundschaftsgrad

Definition einer benutzerspezifischen *social frequency*:

$$sf_u(i, t) = |U| \cdot \sum_{u' \in U} F(u, u') tf_u(i, t)$$

Definition eines benutzerspezifischen *social score*:

$$score_u(i, t) = \frac{(k_1 + 1) sf_u(i, t)}{k_1 + sf_u(i, t)} \cdot \log \frac{|I| - df(t) + 0.5}{df(t) + 0.5}$$

Automatische Tagexpansion

Problem: Benutzer verwenden verschiedene Tags für ähnliche Dinge

⇒ schlechte Ausbeute (fehlende relevante Ergebnisse)

Beispiel:

MPI, MPIO, MPI-INF, MPI-CS, Max-Planck-Institut, D5, AG5, DB&IS, UdS, ...




Lösung:

1. Definition von **ähnlichen Tags**
2. **Expansion** von Anfragen mit ähnlichen Tags
3. **Modifikation** der Scoringfunktion für expandierte Anfragen

Experimentelle Evaluierung: Güte

Systematische Evaluierung der Ergebnisgüte schwierig

Drei mögliche Ansätze:

- **Benutzerstudien** (Anfragen und Ergebnisbewertungen durch Menschen) 
- Anfragen und Bewertungen aus **externer Information** abgeleitet (Beispiel: DMOZ Kategorien) 
- **Automatisierte Bewertungen** aus dem Kontext des Benutzers 
 - Items, die von Freunden getaggt wurden
 - Items, die in der Zukunft getaggt werden

Prototyp [SIGIR 2008 Demo]

Results for query: mystery / user: PandorasRequiem - Opera

File Edit View Bookmarks Widgets Tools Help

New tab Results for query: mys...

http://info5701.ag5.mpi-sb.mpg.de:8081/experiments/servlet/LTServlet Google

global [advanced settings](#)
 social spiritual
 hybrid, $\alpha =$ $\beta =$

mystery

Select User for Query: 'mystery'

#Docs: 106	#Friends: 89	PandorasRequiem
#Docs: 1	#Friends: 134	virgingloves
#Docs: 29	#Friends: 90	cnnrenner
#Docs: 38	#Friends: 90	LolaWalser
#Docs: 29	#Friends: 89	AsYouKnow_Bob
#Docs: 106	#Friends: 89	PandorasRequiem
#Docs: 122	#Friends: 65	ggchickapee

user: 'PandorasRequiem' - Docs

	Fat Tuesday (1997) by Sandra Brown	tags by 'PandorasRequiem': new, pop, brown, unread, thriller, mystery , orleans, romance, suspense, fiction,
	The Stand (2004) by Stephen King	tags by 'PandorasRequiem': post-apocalyptic, sf, end, sff, tbr, the, vs., dark, epic, fantasy, fiction, mystery , science, shelved, ...

Resulting Books for Query: 'mystery':

	Contributing Tags: mystery , ..more 1. Strong Poison (1995) tags by 'SeriousGrace': women, sayers, booklust, fiction, mystery , dorothy,	by Dorothy L. Sayers
	Contributing Tags: mystery , ..more 2. Busman's Honeymoon (1995) tags by 'SeriousGrace': women, sayers, booklust, fiction, mystery , dorothy,	by Dorothy L. Sayers
	Contributing Tags: mystery , ..more 4. The daughter of time , (1953) tags by 'Eurydice': age, iii, alan, 1950s, crime, grant, hospital-bound, solves, mystery , richard, inspector, historical, golden	by Josephine Tey

Benutzerstudie: LibraryThing

- Registrieren der Autoren bei Librarything
- Hinzufügen (eines Teils der) eigenen Bücher
- Auswählen von Freunden mit ähnlichem Profil
- Definieren von 33 Anfragen
- Crawlen (eines Teils) von LibraryThing
- Poolen und Bewerten der Ergebnisse
- Bestimmen von NDCG[10] für die unbekanntenen Ergebnisse (~gewichtete Präzision nach 10 Ergebnissen)

Benutzerstudie: LibraryThing

β (Inhaltsähnlichkeit)

	0,0	0,1	0,2	0,5	0,8	1,0
0,0	0,58	0,61	0,62	0,61	0,61	0,60
0,1	0,63	0,63	0,64	0,63	0,63	
0,2	0,62	0,62	0,63	0,64	0,64	
0,5	0,58	0,59	0,60	0,60		

α (Freundschaftsgraph)

- Güte insgesamt sehr hoch (1.0 ist optimal)
- Kombination von sozialer und inhaltlicher Ähnlichkeit gewinnt

Outline

- Modellierung von Social Tagging Networks
- Effektives Scoring von Anfragen
- **Effiziente Auswertung von Anfragen**
 - **Threshold-Algorithmen**
 - **ContextMerge**
 - **Experimentelle Evaluierung**
- Zusammenfassung & offene Fragen

Algorithmischer Überblick

- Input: Anfrage $q=\{t_1\dots t_n\}$ von Benutzer u
- Output: k Items mit höchsten Scores
- Ziele:
 - Unnötige Berechnungen vermeiden
 - Minimierung von Disk-I/O und CPU-Last
 - Verwenden von vorberechneten Daten auf Platte

Exkurs: Threshold-Algorithmen für Text IR

Input:

- Anfrage $q = \{t_1 \dots t_n\}$
- Listen $L(t_p)$ mit Paaren $\langle i, \text{score}(i, t_p) \rangle$, sortiert nach $\text{score}(i, t_p) \downarrow$

Output: k Dokumente mit höchsten aggregiertem Score

Algorithmus:

**Vorberechnung von $\text{score}_u(i, t)$ unmöglich
(hohe Dynamik, zu viele und zu lange Listen)
 \Rightarrow Threshold-Algorithmen nicht anwendbar**

Zerlegung der Social Frequency ($\beta=0$)

$$\begin{aligned} sf_u(i, t) &= |U| \cdot \sum_{u' \in U} F(u, u') tf_u(i, t) \\ &= |U| \cdot \sum_{u' \in U} \left[\alpha P_{graph}(u, u') + (1 - \alpha) \frac{1}{|U|} \right] tf_u(i, t) \\ &= |U| \left[\sum_{u' \in U} \alpha P_{graph}(u, u') tf_u(i, t) + (1 - \alpha) \sum_{u \in U} \frac{tf_u(i, t)}{|U|} \right] \\ &= \underbrace{\alpha |U| \sum_{u' \in U} P_{graph}(u, u') tf_u(i, t)}_{\text{Abhängig von User u}} + \underbrace{(1 - \alpha) tf(i, t)}_{\text{Unabhängig von User u}} \end{aligned}$$

Berechne $sf_u(i, t)$ „on the fly“ aus $tf(i, t)$, Freuden von u und ihren getaggtten Items

ContextMerge

Vorberechnete Listen:

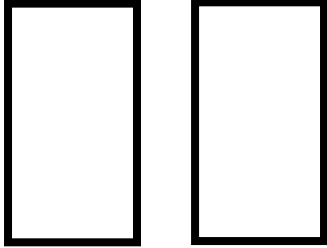
- **ITEMS(t)**: Paare $\langle i, tf(i, t) \rangle$, sortiert nach $tf(i, t) \downarrow$
- **FRIENDS(u)**: Paare $\langle u', P_{graph}(u, u') \rangle$, sort. nach $P_{graph}(u, u') \downarrow$
- **USERITEMS(u', t)**: Paare $\langle i, tf_{u'}(i, t) \rangle$, unsortiert

Anpassung von Threshold-Algorithmen für Query $u, t_1 \dots t_n$:

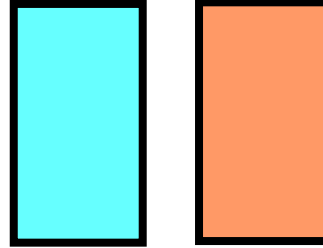
- Scanne **ITEMS(t_p)** und n Kopien von **FRIENDS(u)**, wähle „beste“ Liste
 - Falls **ITEMS(t_p)**: lies nächsten Eintrag
 - Falls **FRIENDS(u, p)**: Lies **USERITEMS(u', t_p)** für nächsten Freund u'
 - Aktualisiere Kandidaten und top-k
 - Prüfe Abbruchbedingung

ContextMerge: Schematische Ausführung

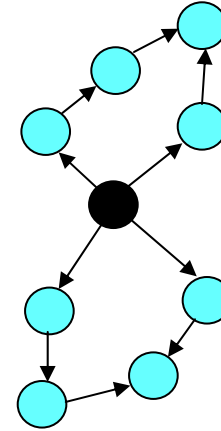
Items(t_1) Items(t_2)



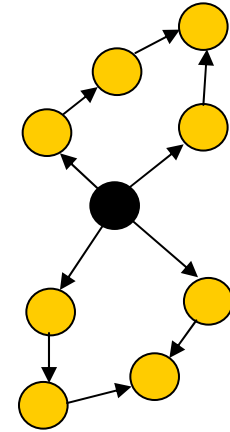
Friends(u, t_1) Friends(u, t_2)



considered
USERITEMS(u', t_1)

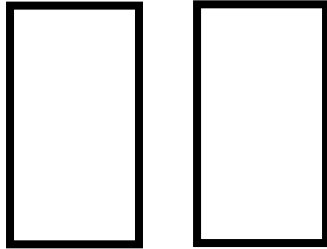


considered
USERITEMS(u', t_2)

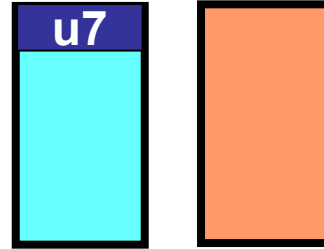


ContextMerge: Schematische Ausführung

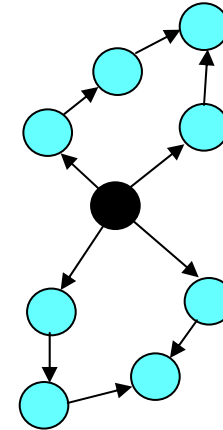
Items(t_1) Items(t_2)



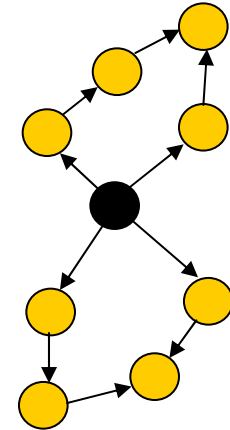
Friends(u, t_1) Friends(u, t_2)



considered
USERITEMS(u', t_1)

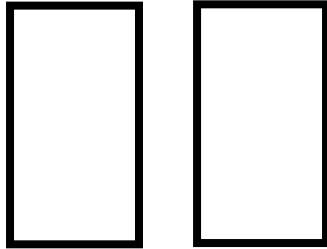


considered
USERITEMS(u', t_2)

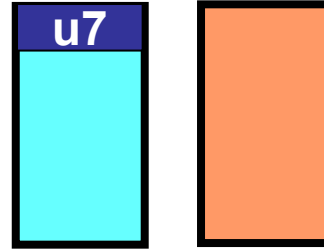


ContextMerge: Schematische Ausführung

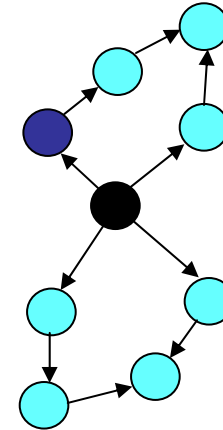
Items(t_1) Items(t_2)



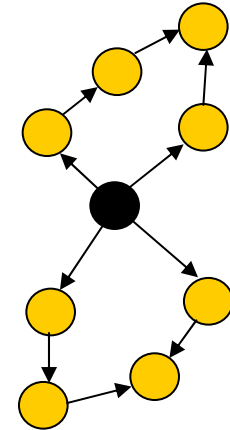
Friends(u, t_1) Friends(u, t_2)



considered
USERITEMS(u', t_1)

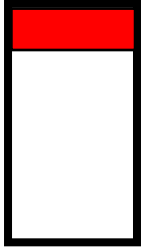


considered
USERITEMS(u', t_2)

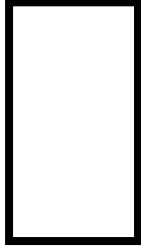


ContextMerge: Schematische Ausführung

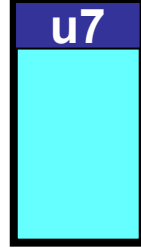
Items(t_1)



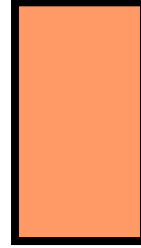
Items(t_2)



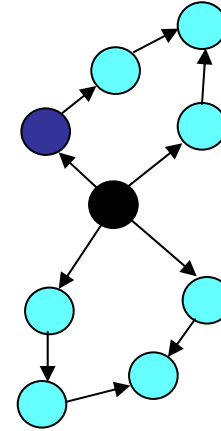
Friends(u, t_1)



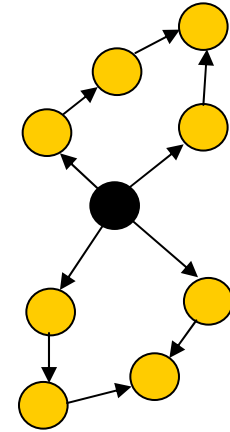
Friends(u, t_2)



considered
USERITEMS(u', t_1)

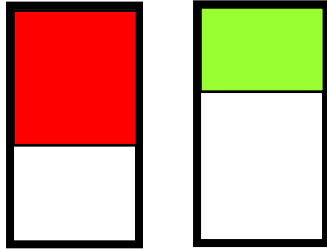


considered
USERITEMS(u', t_2)

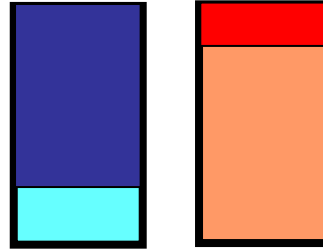


ContextMerge: Schematische Ausführung

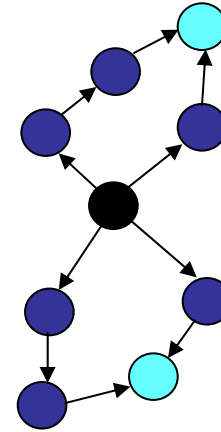
Items(t_1) Items(t_2)



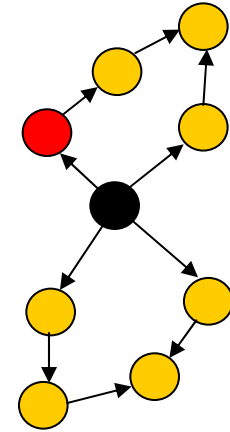
Friends(u, t_1) Friends(u, t_2)



considered
USERITEMS(u', t_1)



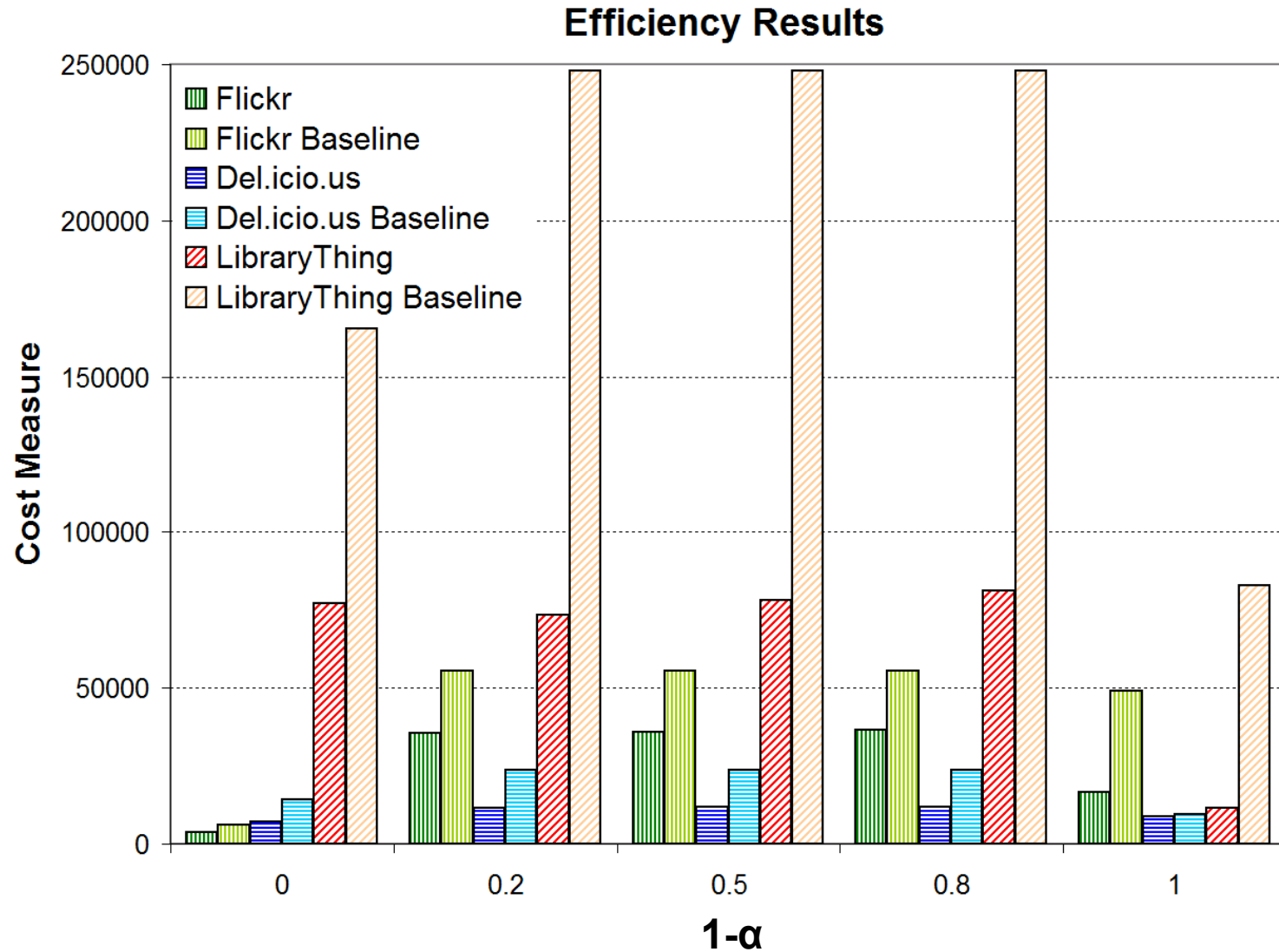
considered
USERITEMS(u', t_2)



Experimentelle Evaluierung: Effizienz

- Testbed: 3 große Crawls realer sozialer Netze
 - Flickr: 10 mio Fotos, ~50,000 Benutzer
 - Del.icio.us: ~175,000 Bookmarks, ~12,000 Benutzer
 - Librarything: ~6.5 mio Bücher, ~10,000 Benutzer
- Anfragen:
 - ~150 häufige Tagpaare in jedem Crawl
 - Für jede Anfrage: wähle Benutzer mit „genügend“ Ergebnissen & Freunden
- Gütemaß: $\#seq.$ Zugriffe + $100\#wahlfr.$ Zugriffe
- Baseline: full join + sort

Experimentelle Evaluierung: Effizienz



Outline

- Modellierung von Social Tagging Networks
- Effektives Scoring von Anfragen
- Effiziente Ausführung von Anfragen
- **Zusammenfassung und offene Fragen**

Zusammenfassung

- **Social-Aware Social Search** ist notwendig
- **Kontextabhängiger Score**
 - integriert globale Popularität mit inhaltlichem und sozialem Kontext des Benutzers
 - Unterstützt dynamische Tagexpansion
- **ContextMerge**: skalierbare Implementierung

Offene Fragen

- Aussagekräftiger und verfügbarer **Benchmark**
- **Queryklassifikation** und **Autoparametrisierung**
- Unterstützung inkrementeller **Änderungen**
- Erweiterung auf **Ratings**, Gewichte, ...
- Erweiterung auf **Nicht-Tags** (Bildfeatures, ...)
- Nachvollziehbare **Erklärung** der Ergebnisse
- Ausnutzen der **Dynamik** (aktuelle Themen, wachsende Gruppen, ...)

Social-Aware Search & Recommendations at planet scale