Predictive Modeling for Navigating Social Media

Dissertation Defense by
HU Meiqun
School of Information Systems
Singapore Management University

November 30, 2011

Dissertation Committee

LIM Ee-Peng

Professor of Information Systems, Chair

JIANG Jing

Assistant Professor of Information Systems

David LO

Assistant Professor of Information Systems

Christopher KHOO Soo Guan

Associate Professor, Division of Information Studies,
Wee Kim Wee School of Communication and Information,
Nanyang Technological University

delicious
social bookmarking

It's Free! Join Now Sign In

The tastiest bookmarks on the web.
Save your own or see what's fresh now!

Learn More

HIDE INTRO

Search the biggest collection of bookmarks in the universe... Search Delicious Search

Popular Bookmarks | Explore Tags

The most popular bookmarks on Delicious right now

See more Popular bookmarks

New bookmarks saved in the last minute 1 0 0

Tags

Technology Review: 3-D Printing for the Masses SAVE — 62
prototyping printing 3dprinting fabrication 3d

Back Up a Web Server - Webmonkey SAVE — 61
backup linux sysadmin web server

mixi for iPhoneから発掘されたmixi日記投稿用API « ku SAVE — 81
mixi api iPhone atompub wsse

Cut-Your-Spending-by-$500-Per-Month: Personal Finance News from Yahoo! Finance SAVE — 61
money finance personal tips savings

Movie box office charts SAVE — 122
visualization movies lisp film design

AppleInsider | Ten step guide to sharing your iPhone's connection with — 68

Popular Tags

design
blog
video
software
tools
music
programming
webdesign
reference
tutorial
art
web
howto
javascript
free
linux
web2.0

All
acoustic
ambient
blues
classical
country
electronic
emo
folk
hardcore
hip hop
indie
jazz
latin
metal
pop
pop punk
punk
reggae

Popular Music on Last.fm

Popular | Hyped | Popular in Singapore

Coldplay
213,068,934 plays (3,799,877 listeners)

Similar to: Snow Patrol, Keane, Travis, Kings of Leon, The Killers

rock

Radiohead
310,596,782 plays (5,546,453 listeners)

Similar to: Thom Yorke, Jonny Greenwood, Sigur Rós, Portishead, Blur

alternative

Adele
49,757,829 plays (1,395,878 listeners)

Similar to: Amy Winehouse, Jessie J, Florence + The Machine, Beyoncé, Duffy

soul

Tags

citeulike

Browse | FAQ | Log in

Search citeulike

citeulike is a free service for managing and discovering scholarly references

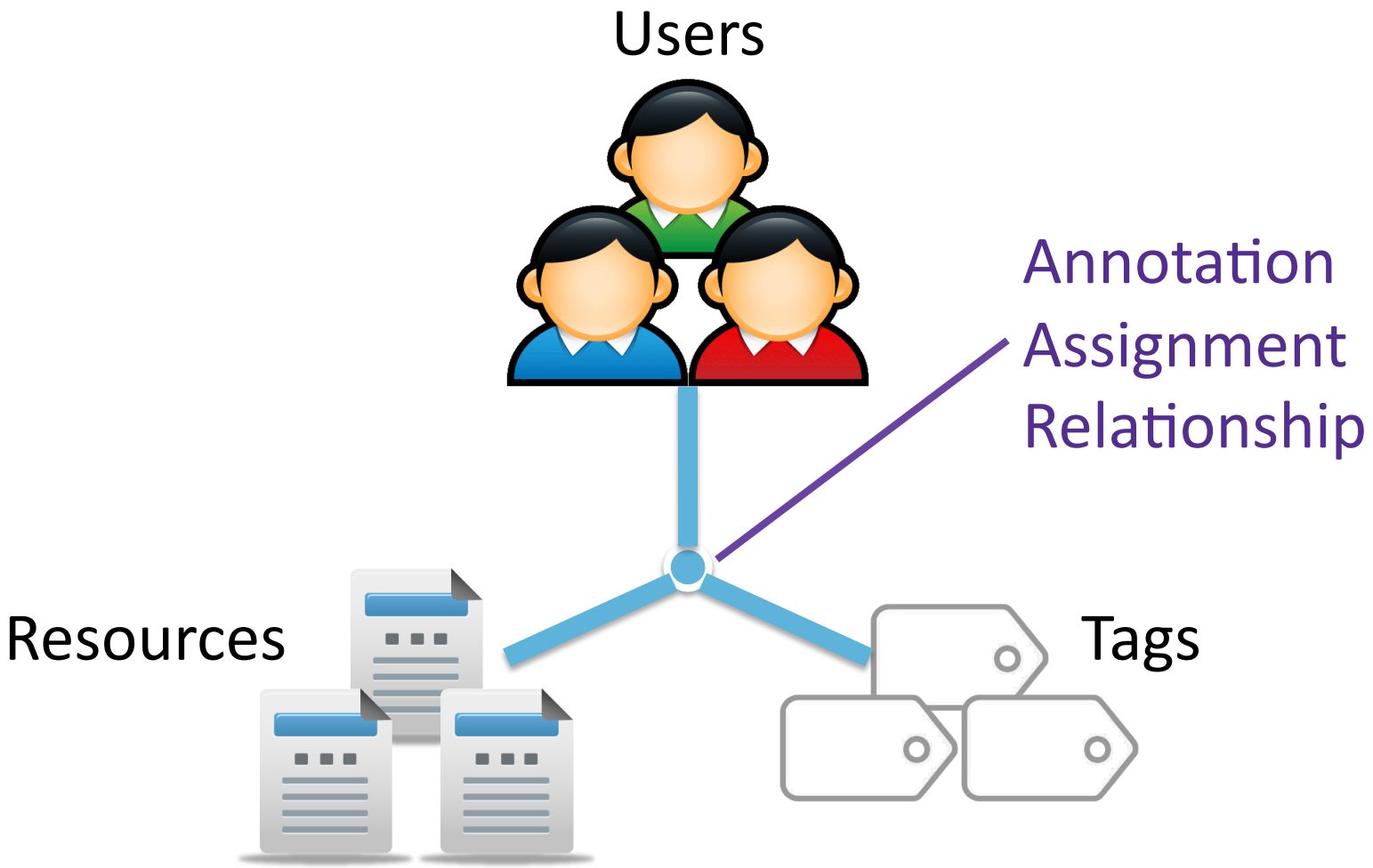
5,686,199 articles - 2,716 added today.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendations NEW
- Share references with your peers
- Find out who's reading what you're reading
- Store and search your PDFs

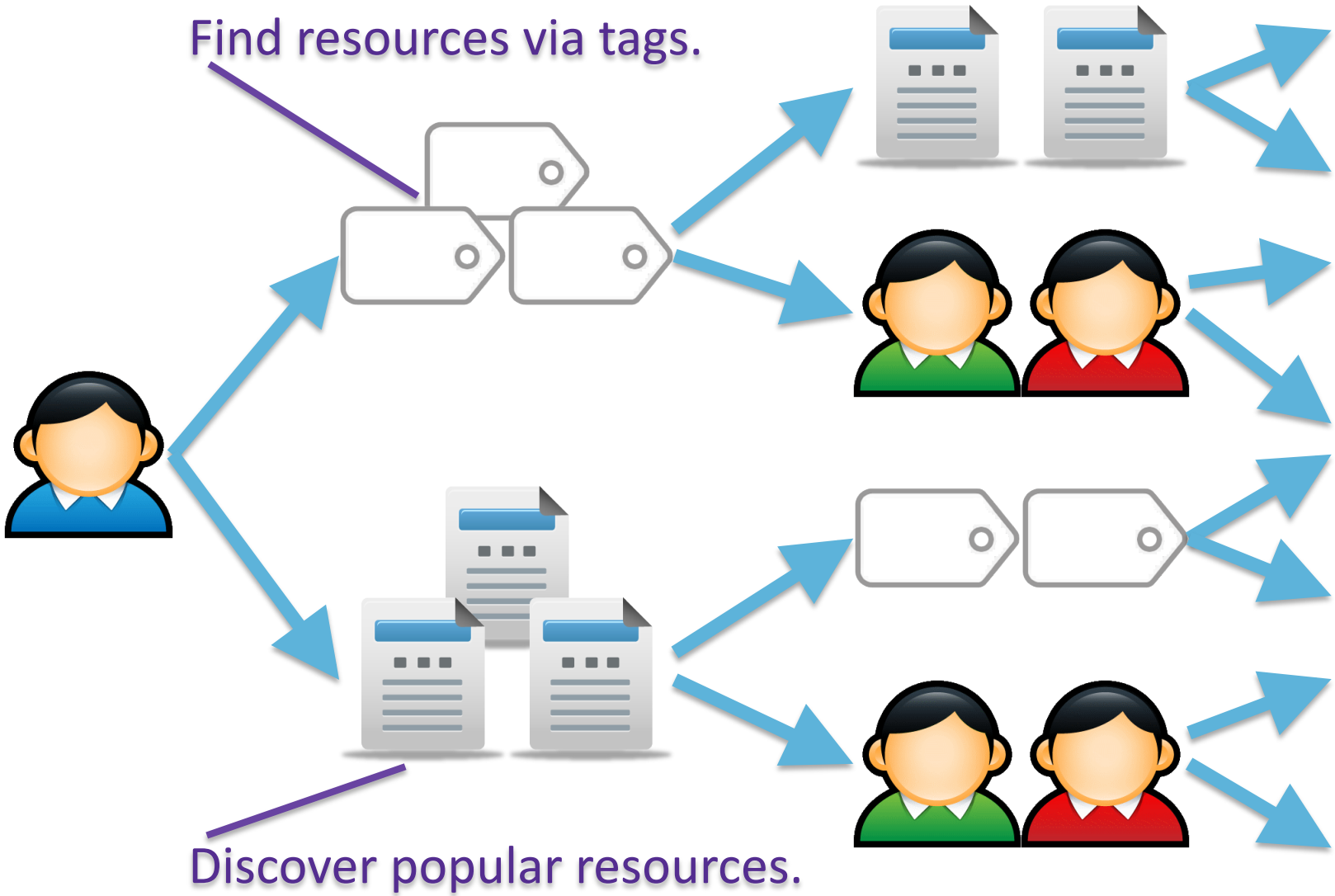
Tagging Widely Adopted in Social Media

Images from the Web

Social Tagging Space



Navigation in Social Tagging Space



Challenges for Effective Navigation

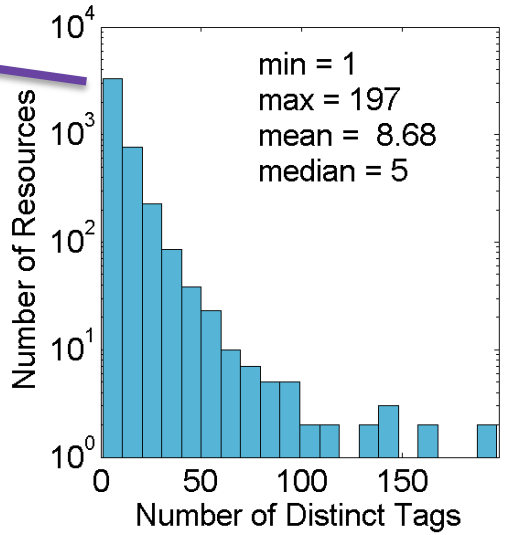
- Challenge 1: Tag Sparseness
- Challenge 2: Personalization
- Challenge 3: Resource Ranking

Challenge 1: Tag Sparseness

- Many resources are untagged.

Study	Resource Type	Amount Untagged
[Abel et al. 2008]	URLs, Photos, Publications, etc.	≥50% (GroupMe!)
[Hu et al. 2010]	News Articles	≥65% (Delicious)

- Many resources have few tags.
- Tag sparseness affects finding relevant resources.
- Task 1: Tag Prediction



How Social Media & Game Mechanics Can Motivate Students

May 27, 2011 by Patrick Supanc

Save a Bookmark

delicious Save a Bookmark

Signed in as

URL http://mashable.com/2011/05/26/social-media-games-education/ Mark as Private

TITLE How Social Media & Game Mechanics Can Motivate Students

NOTES 1000 chars

TAGS

SEND Send your bookmarks Clear

Tags Send

Popular Tags: click to add from popular tags on Delicious

education socialmedia social gaming gamification media

Recommended Tags

Save Cancel

Ads by

Be Mc

Hypnosi

Patrick

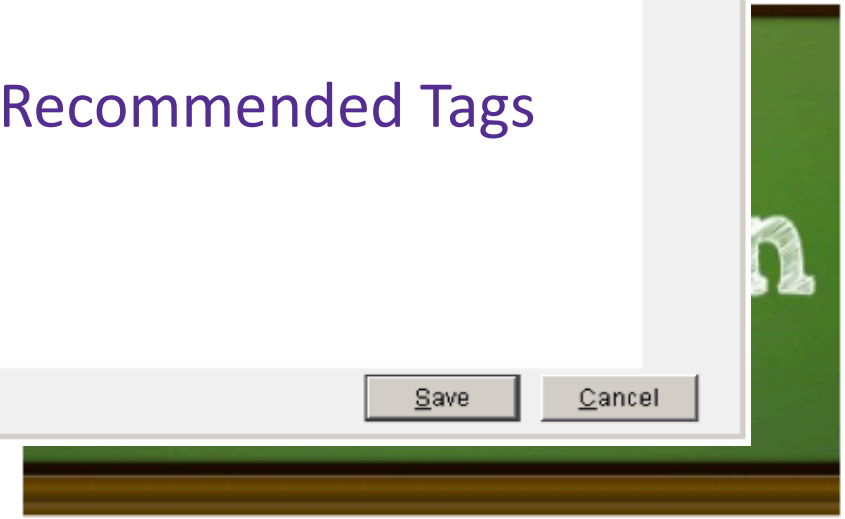
Readir

develo

been a

the Wc

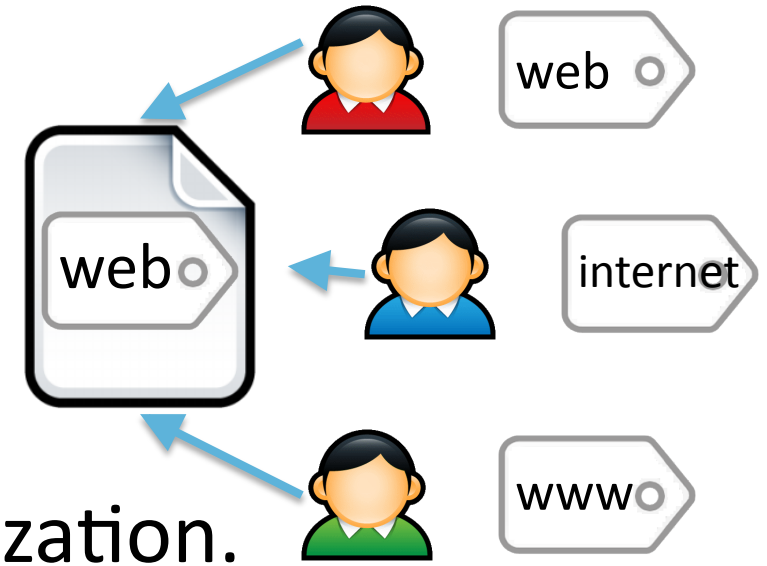
Blackb



Social media and online games have the potential to convey 21st century skills that aren't necessarily part of school

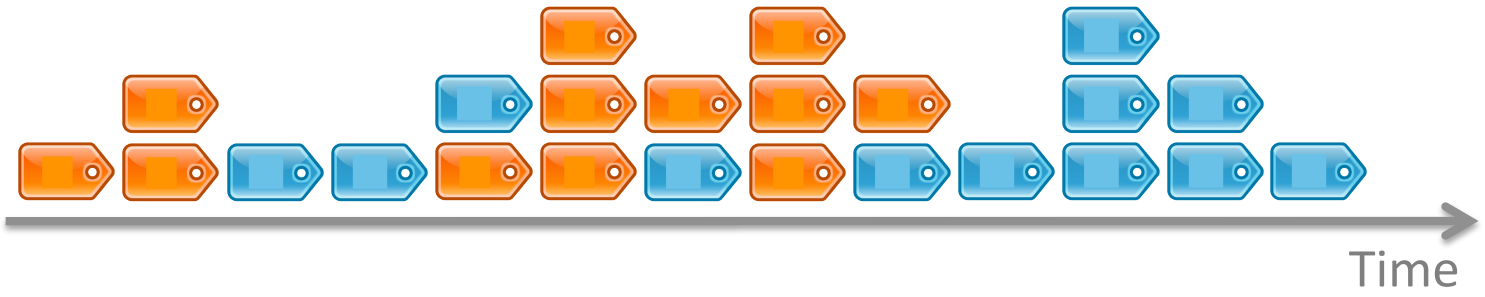
Challenge 2: Personalization

- Tag Recommendation
  - Target User
  - Target Resource
- Ease tagging; But
- Low utility without personalization.
  - Individual Tagging Preference
- Task 2: Personalized Tag Recommendation



Challenge 3: Resource Ranking

- Tag assignments are useful for resource ranking:
  - Popularity = number of annotations
  - Recency = time of the latest annotation
- Tags are assigned at different time.
- Temporal profiles are not analyzed.
- Task 3: Trend Discovery



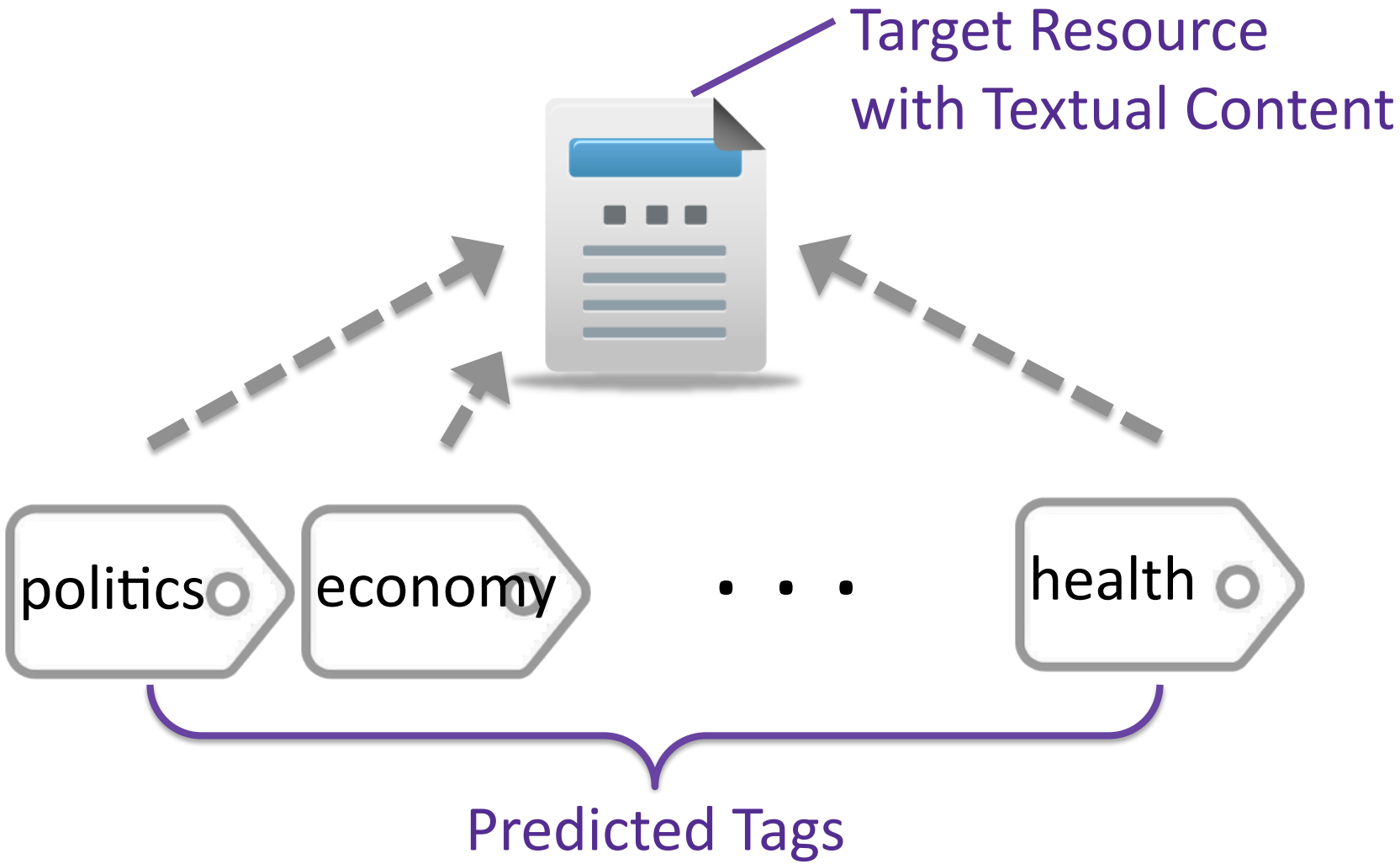
Main Contributions

- Study three prediction tasks that address the current challenges for navigating the social tagging space.
  - Tag Prediction
  - Personalized Tag Recommendation
  - Trend Discovery
- Propose and develop holistic methods for solving these prediction tasks.

Outline

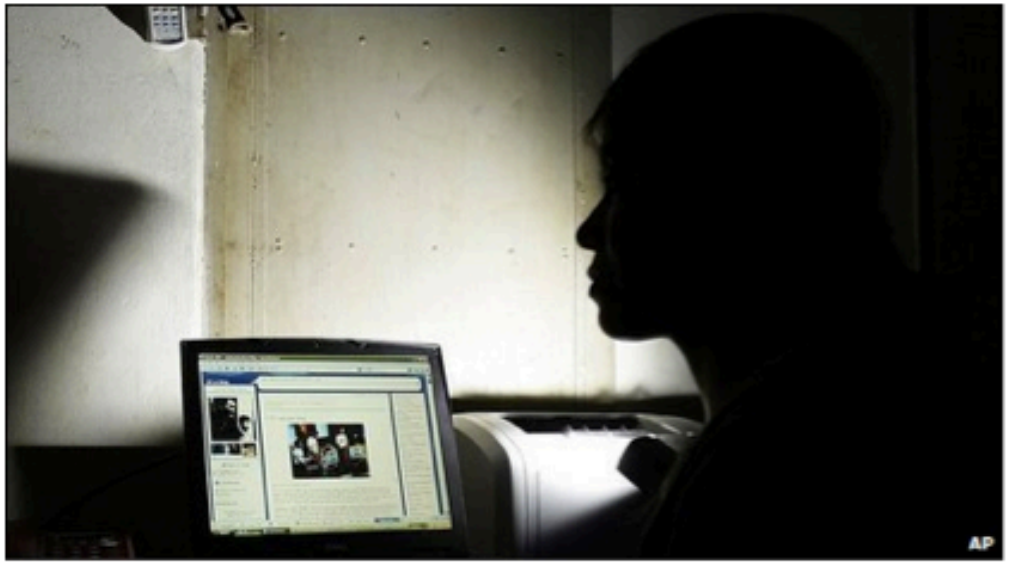
- Motivation
- Studies and Results:
  - Tag Prediction
  - Personalized Tag Recommendation
  - Trend Discovery
- Conclusion

Tag Prediction Task



An Example

Nato's cyber defence warriors



By Frank Gardner
Security correspondent, BBC News

Nato officials have told the BBC their computers are under constant attack from organisations and individuals bent on trying to hack into their secrets.

The attacks keep coming despite the establishment of a co-ordinated cyber defence policy with a quick-reaction cyber team on permanent standby.

The cyber defence policy was set up after a wave of cyber attacks on Nato member Estonia in 2007, and more recent attacks on Georgia - so what are they defending against and how do they do it?

Tags on Delicious

Topic: security

- security
- cyber
- cybersecurity
- cyberattack
- cyberwarfare

…

Topic: nato

- nato
- otan
- sota
- u.s.
- europe
- dossier_otan

…

- bbc

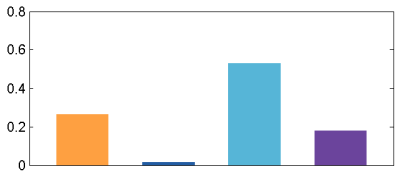
Intuitions and Research Questions

- Tags ≠ Content Words
- Tags form topics.

- Research Questions:
  - How to solve tag prediction using topics?
  - How to model the topics of tags?
  - How to relate the topics of tags to the target resource?

security
cyber
cybersecurity
cyberattack
cyberwarfare
…
nato
otan
sota
u.s.
europe
dossier_otan
…
bbc

Topic Mixture of Words



Correspondence

Nato's cyber defence warriors

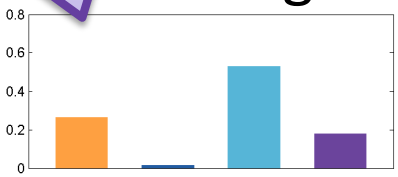
By Frank Gardner
Security correspondent, BBC News

Nato officials have told the BBC their computers are under constant attack from organisations and individuals bent on trying to hack into their secrets.

The attacks keep coming despite the establishment of a co-ordinated cyber defence policy with a quick-reaction cyber team on permanent standby.

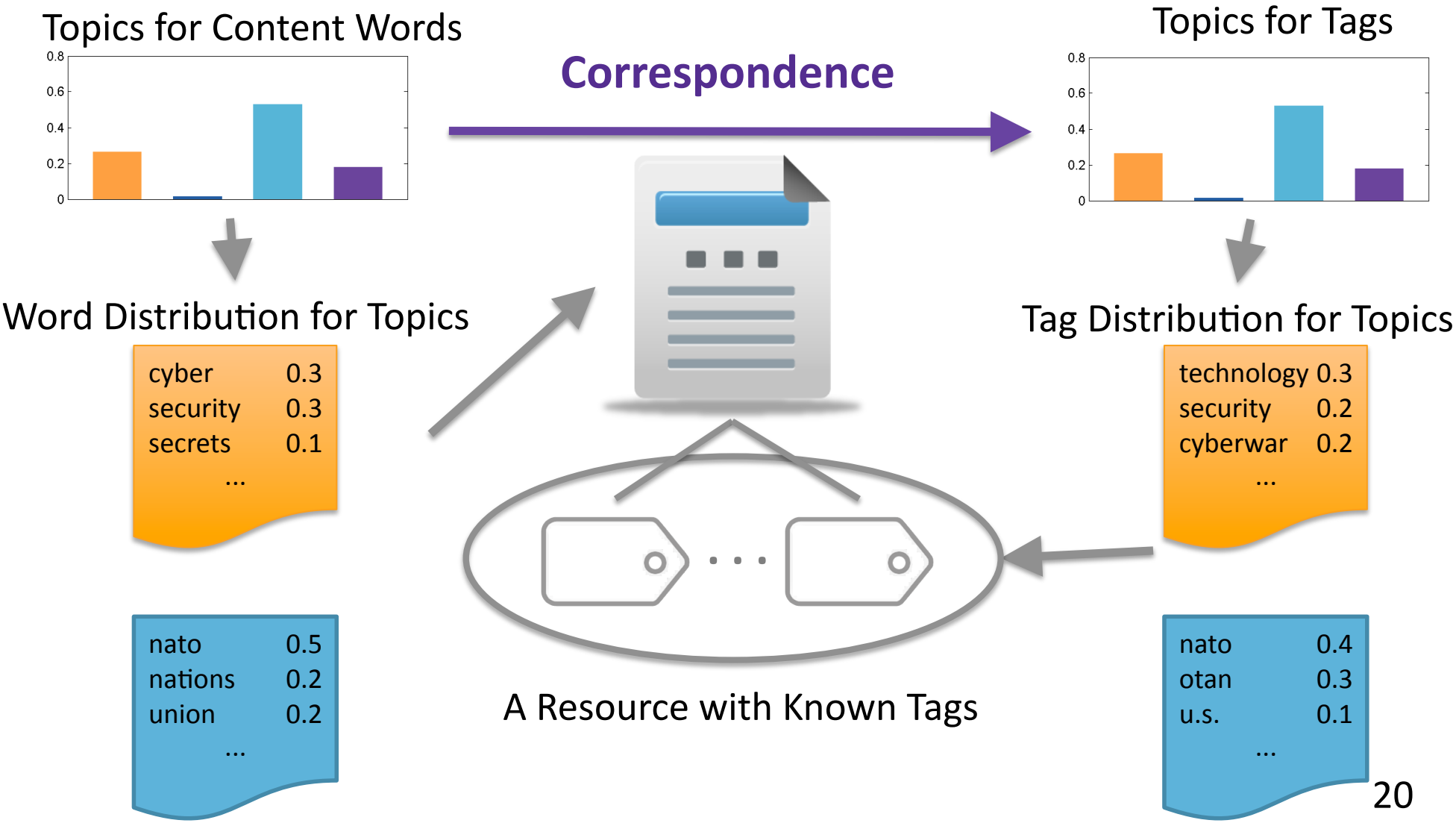
The cyber defence policy was set up after a wave of cyber attacks on Nato member Estonia in 2007, and more recent attacks on Georgia - so what are they defending against and how do they do it?

Topic Mixture of Tags



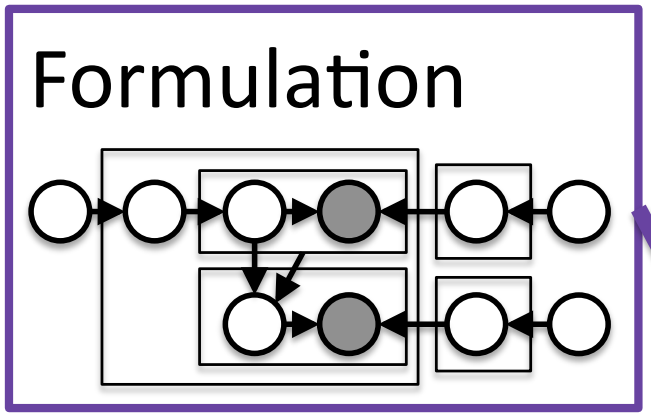
security
cyber
cybersecurity
cyberattack
cyberwarfare
...
nato
otan
sota
u.s.
europe
union
dossier_otan
...
bbc

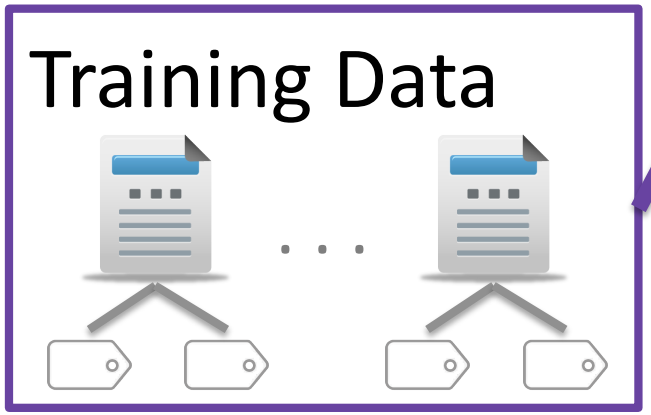
LDAtgg Model Assumptions

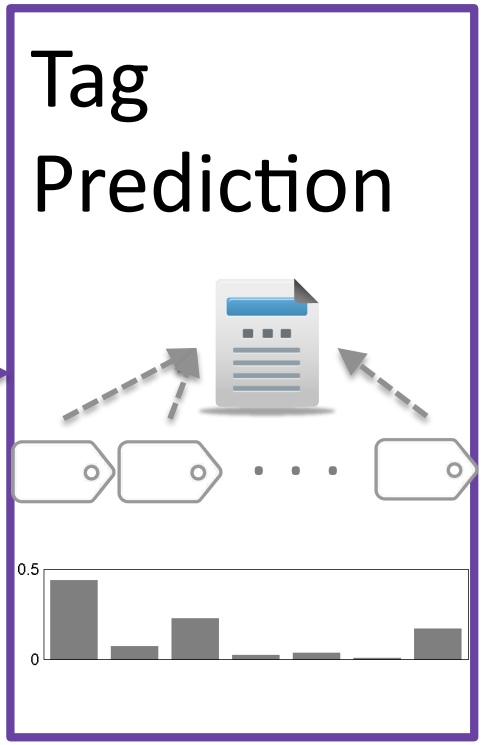


LDAtgg Approach Overview

Encodes the latent variables and relationship that generate the training data.







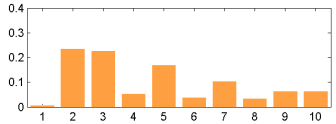
Tag Distribution for Topics

Word Distribution for Topics

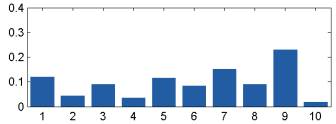
Tag Prediction using LDAtgg

Word Distribution for Topics

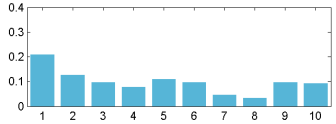
Topic 1



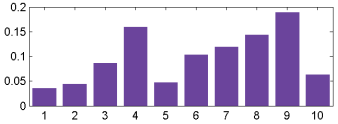
Topic 2



Topic 3

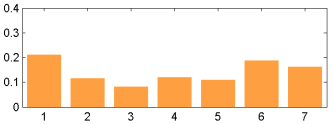


Topic 4

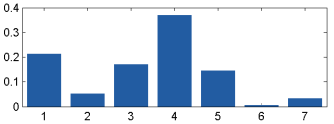


Tag Distribution for Topics

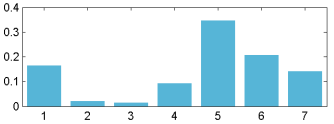
Topic 1



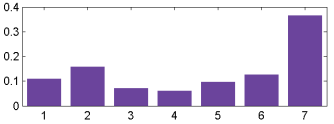
Topic 2



Topic 3

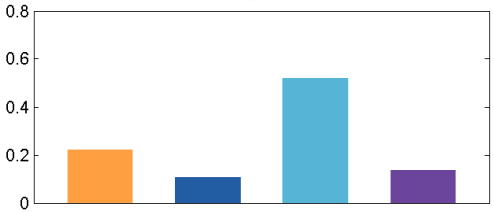


Topic 4



Target Resource

1



Topic Mixture for Target Resource

2

Probability for Predicted Tags

Methods Compared

Method Group	Method	Source of Candidate Tags	Assumptions on the Coupling between Topics of Tags and Topics of Words
Content-based	tf	Words	None
	tf-idf	Words	
Topic-model-based	tagLDA	Tags	Conditional Independence
	LDAtgg	Tags	Correspondence

Our Proposed Method

Dataset from Delicious

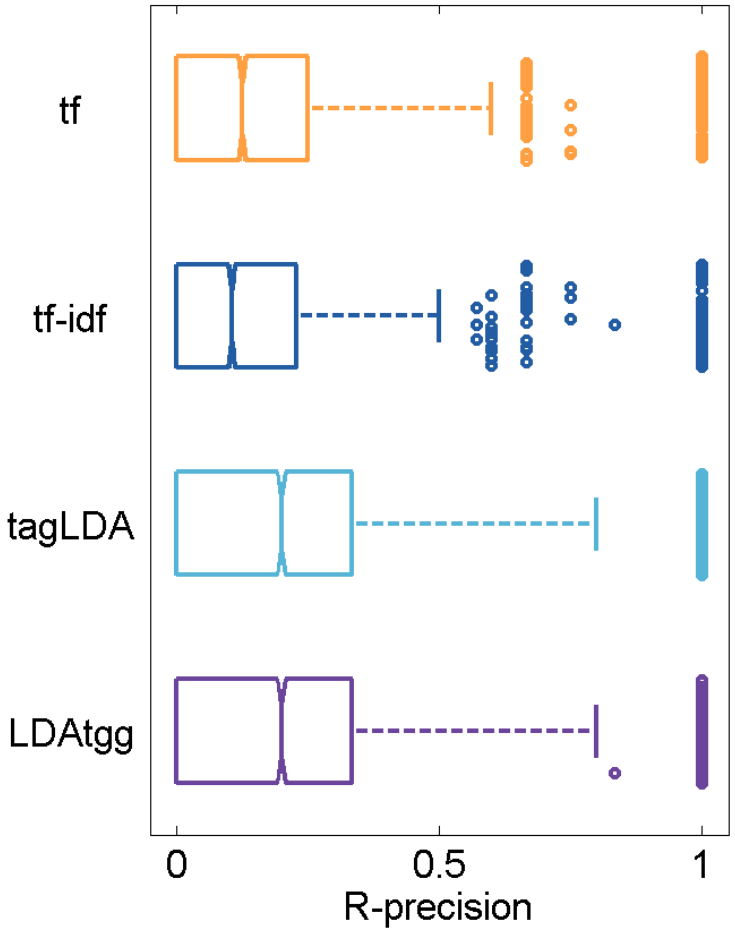
Dataset Statistics	
News Articles	BBC, CNN, USAToday
Number of Resources	4,493
Average word tokens per resource	344
Average tag tokens per resource	17
Size of word vocabulary	24,322
Size of tag vocabulary	12,468

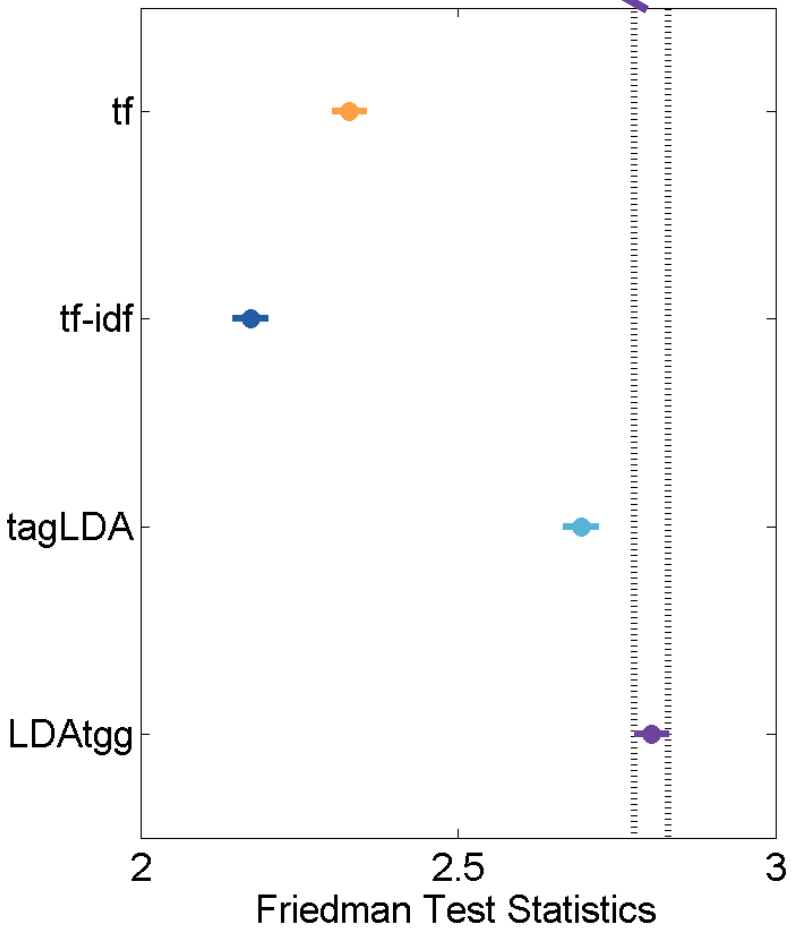
Data collected during April 2009.

We split the dataset into 5 folds for cross-validation.

R-precision and Significance Test for Tag Prediction Performance

LDAtgg outperforms other methods significantly.





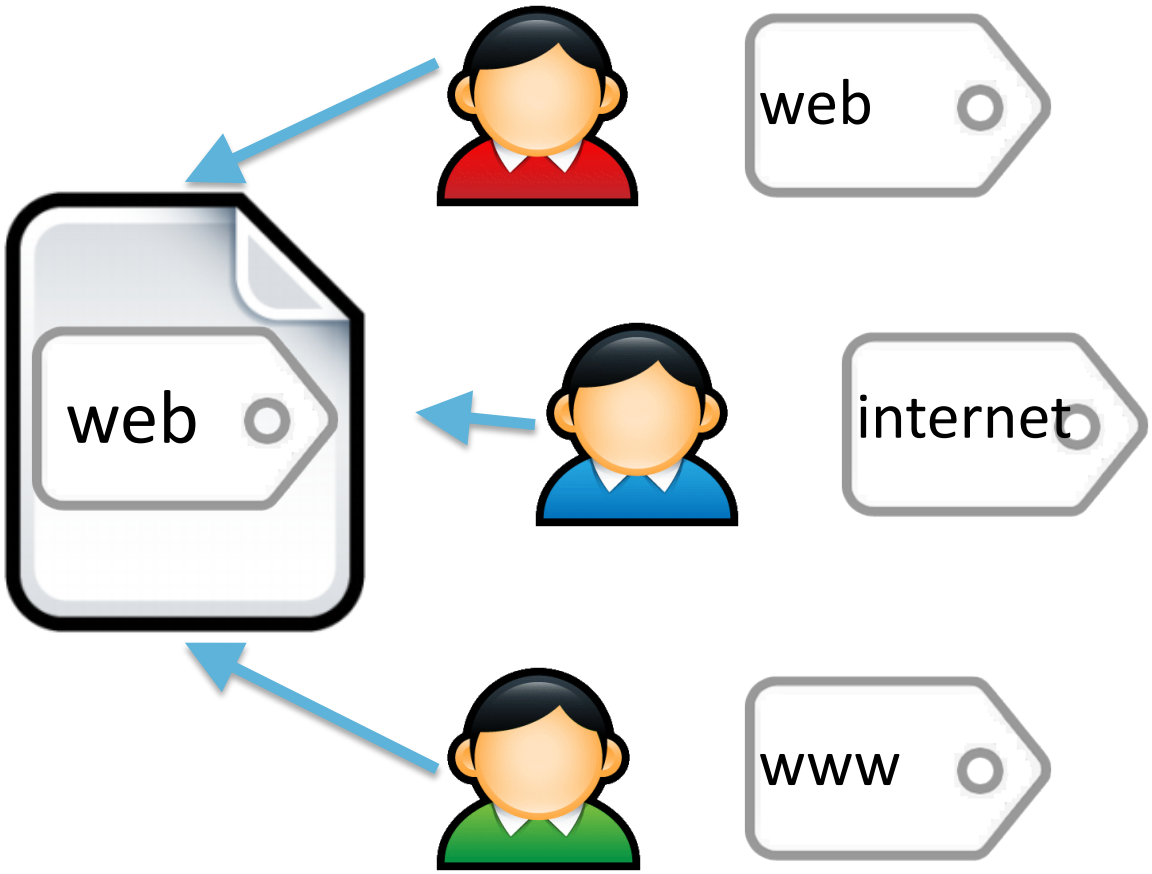
Contributions for Study 1

- Proposed LDAtgg Model
  - Model topics, tags and words.
  - Assume correspondence between the topics of tags and topics of words.
- Evaluated Tag Prediction
  - Online news articles.
  - Tags from Delicious.
  - LDAtgg outperforms baselines.

Outline

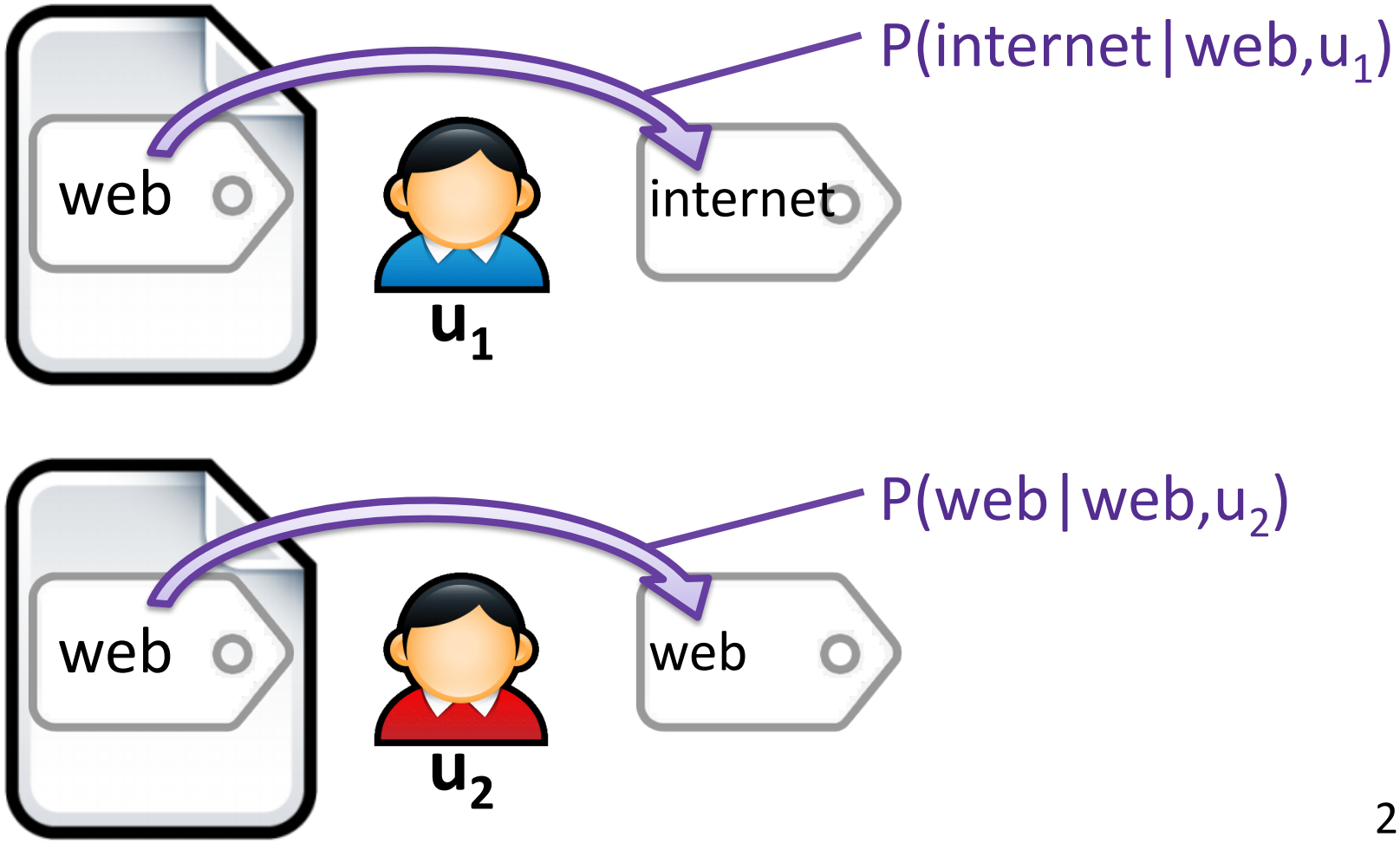
- Motivation
- Studies and Results:
  - Tag Prediction
  - Personalized Tag Recommendation
  - Trend Discovery
- Conclusion

Personal Tagging Preferences



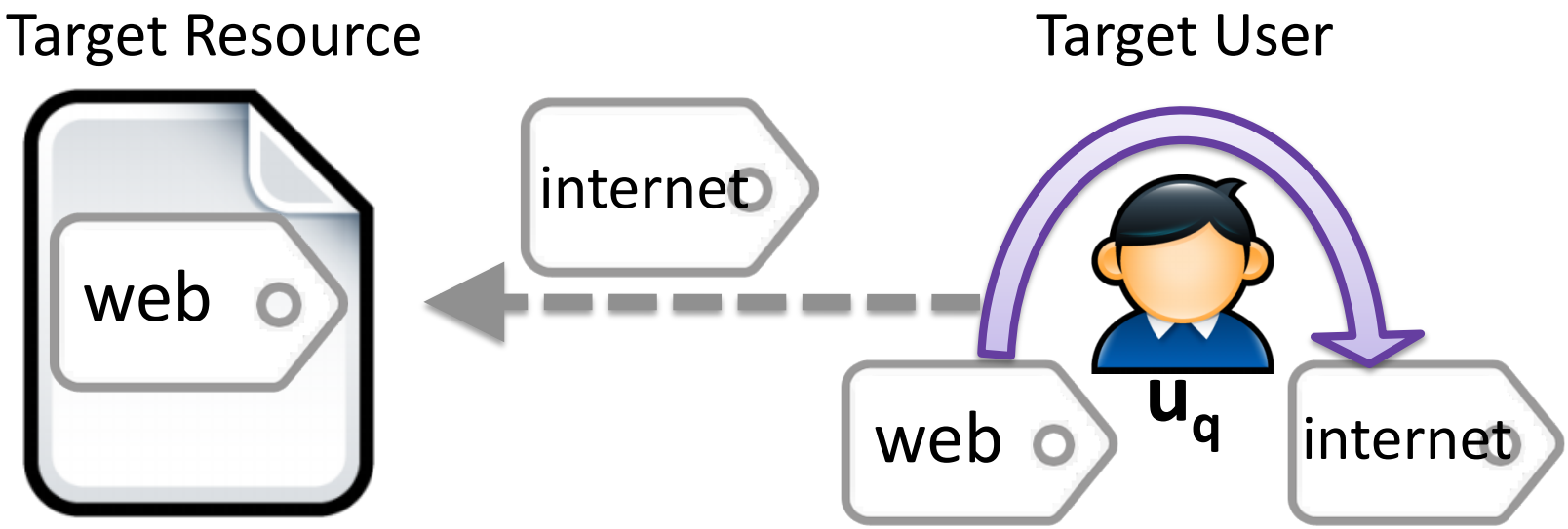
Deriving Preference Patterns

P(personal tag | resource tag , user)



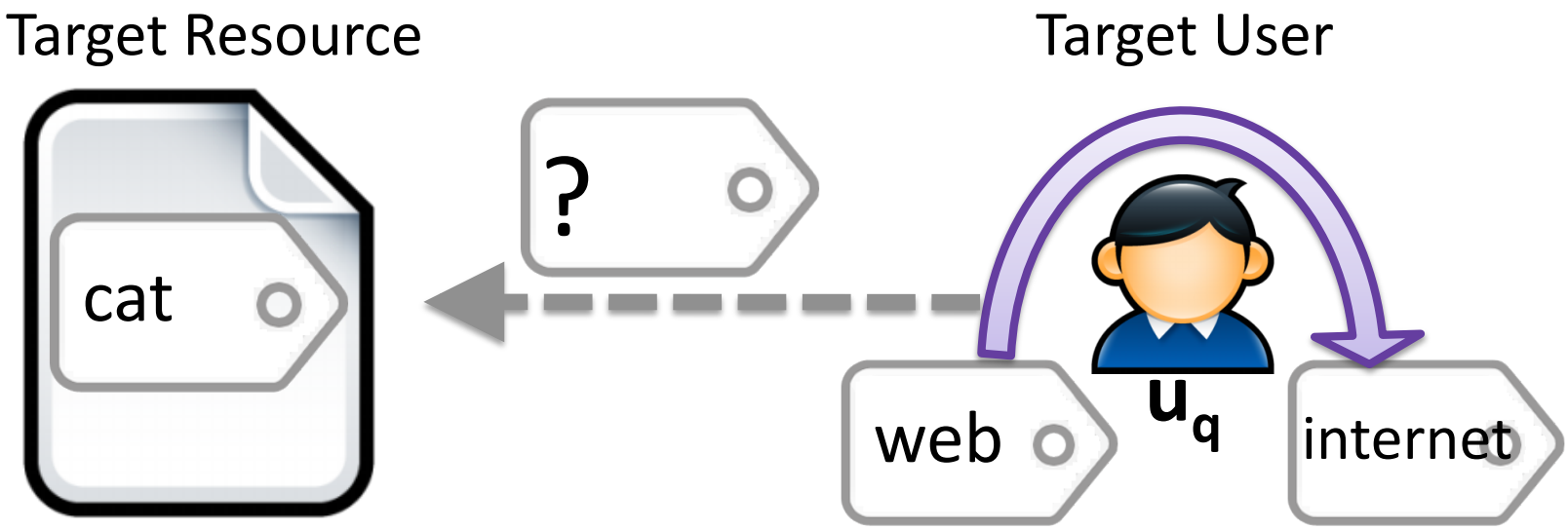
Applying Preference Patterns

$$P(\text{internet} \mid \text{web} , u_q)$$



When Preference Patterns are Missing...

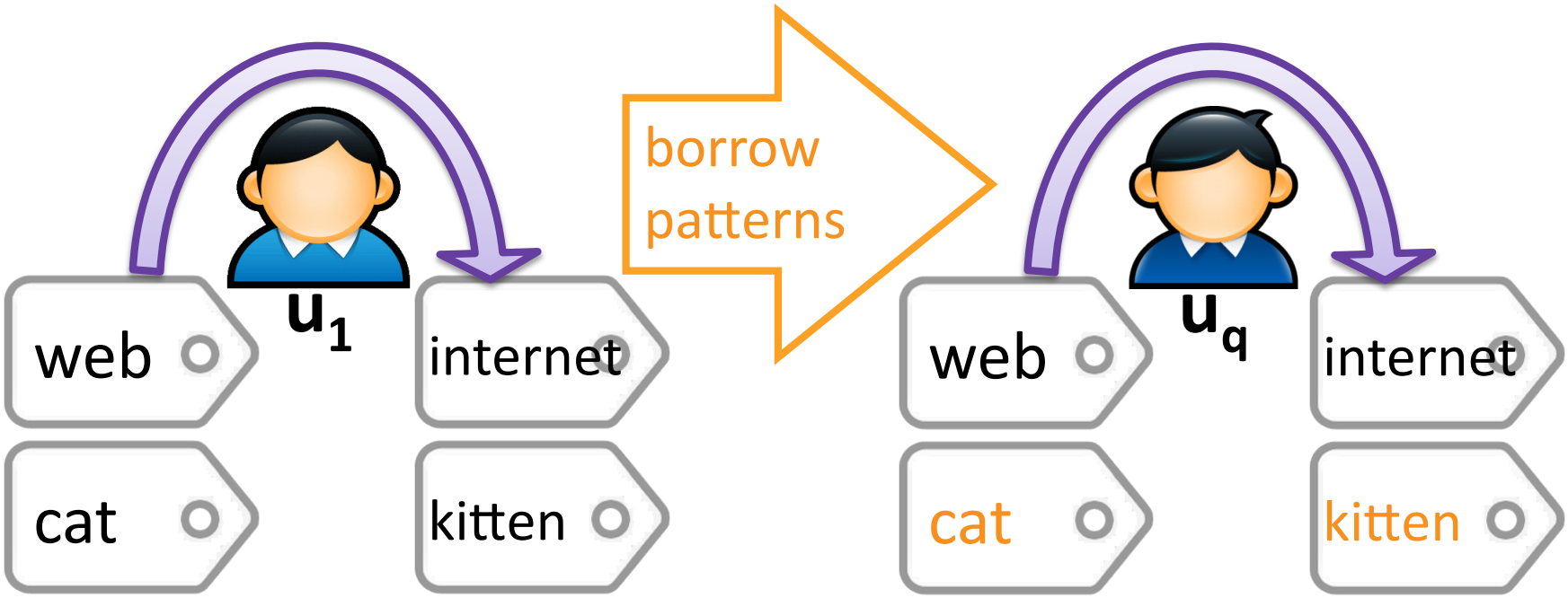
$P(? \mid cat , u_q)$ is not seen for the target user.



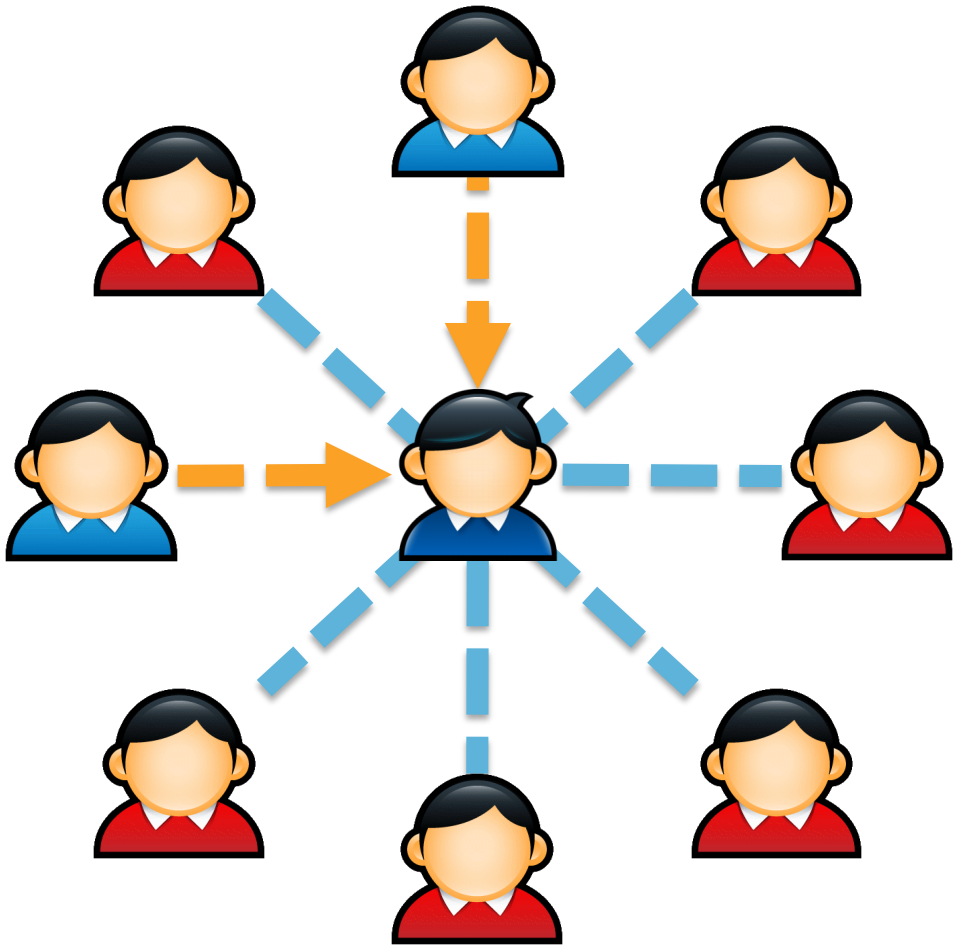
Borrowing Preference Patterns

Intuition:

Borrowed tagging preference patterns can help recommend more tags to the target user.



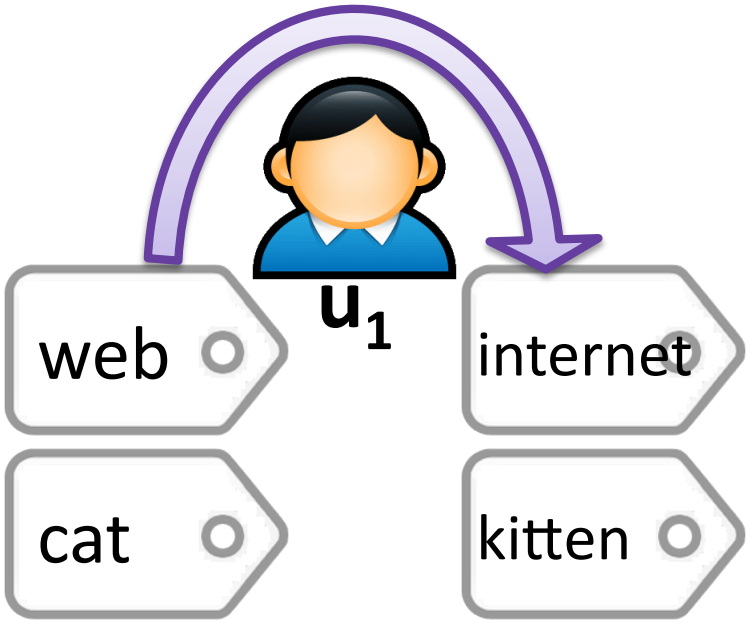
Finding Like-minded Users

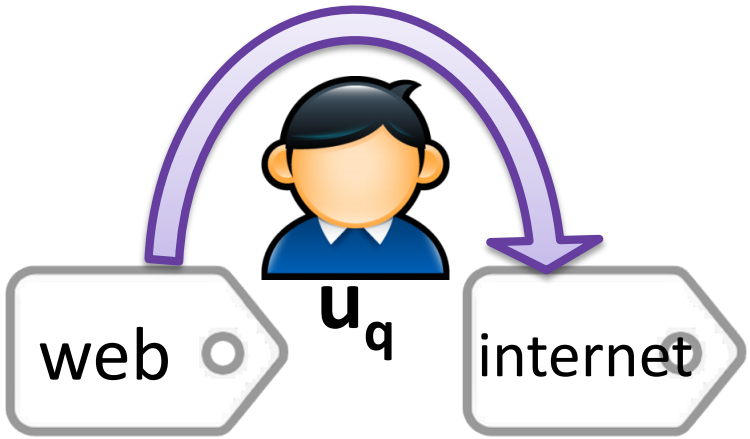


Like-minded Users

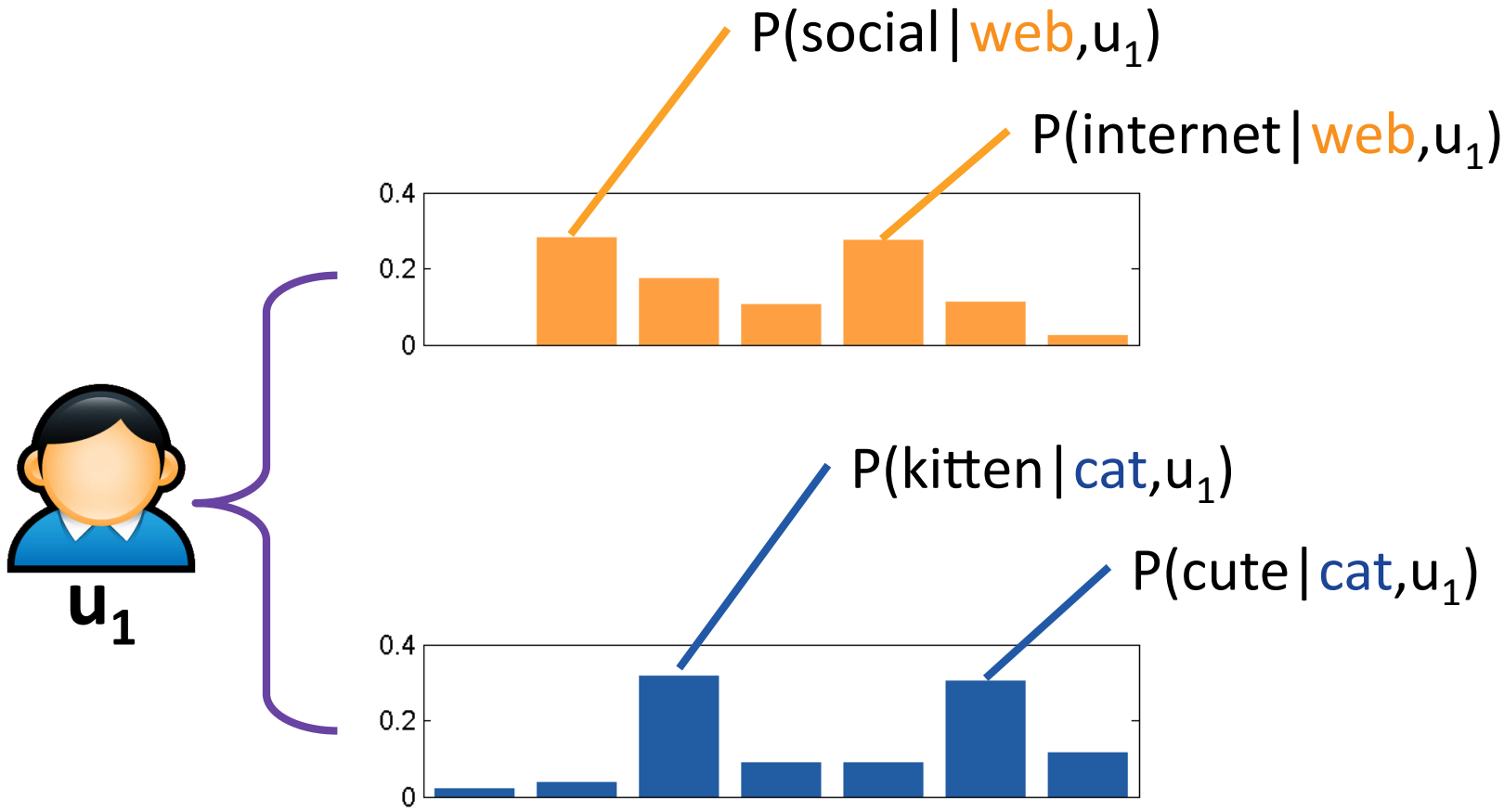
Assumption:

Users are like-minded if they show similar tagging preference patterns.

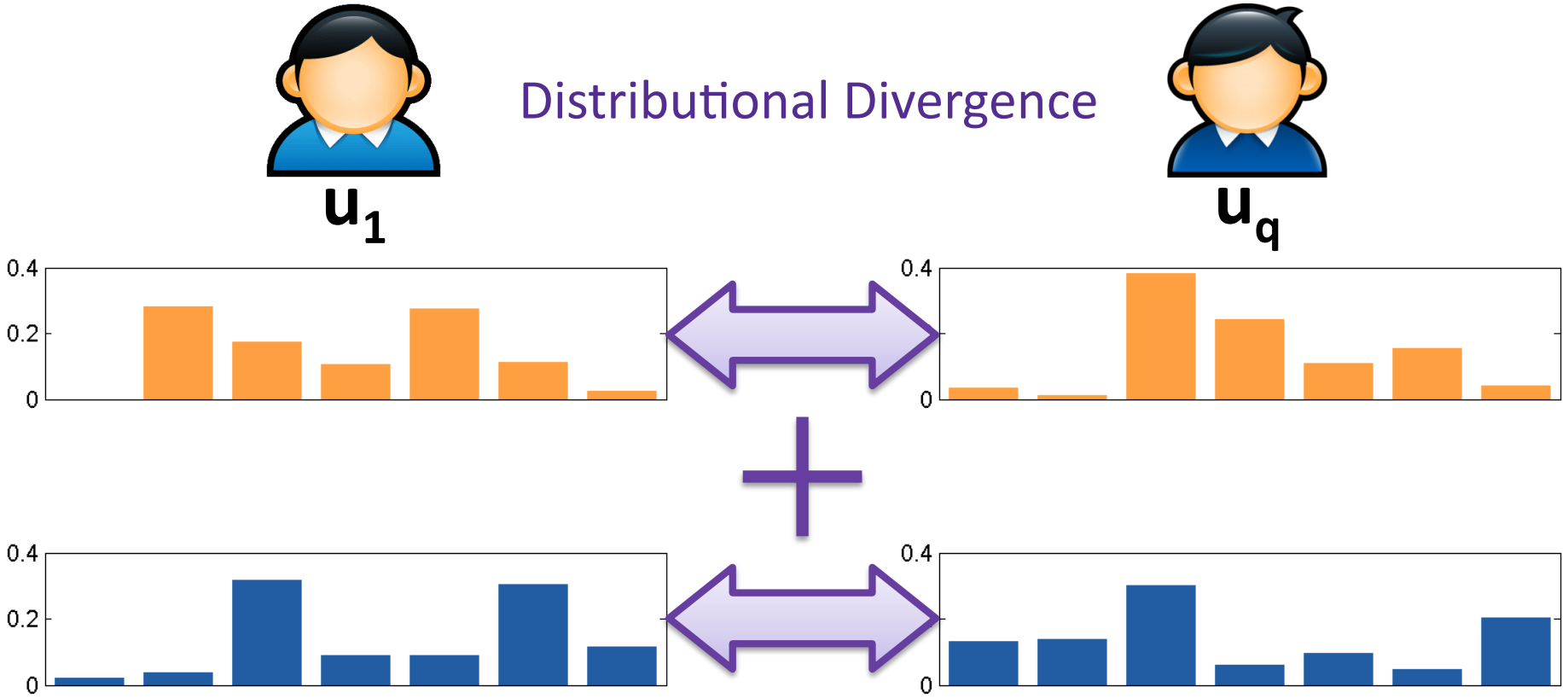




Profiling Users by Preference Patterns



Measuring Similarity Between Users



Methods Compared

Methods	Intuitions	
	Extract Tagging Preference Patterns	Leverage Like-minded Users
knn		✔
trans-U	✔	
trans-N	✔	✔

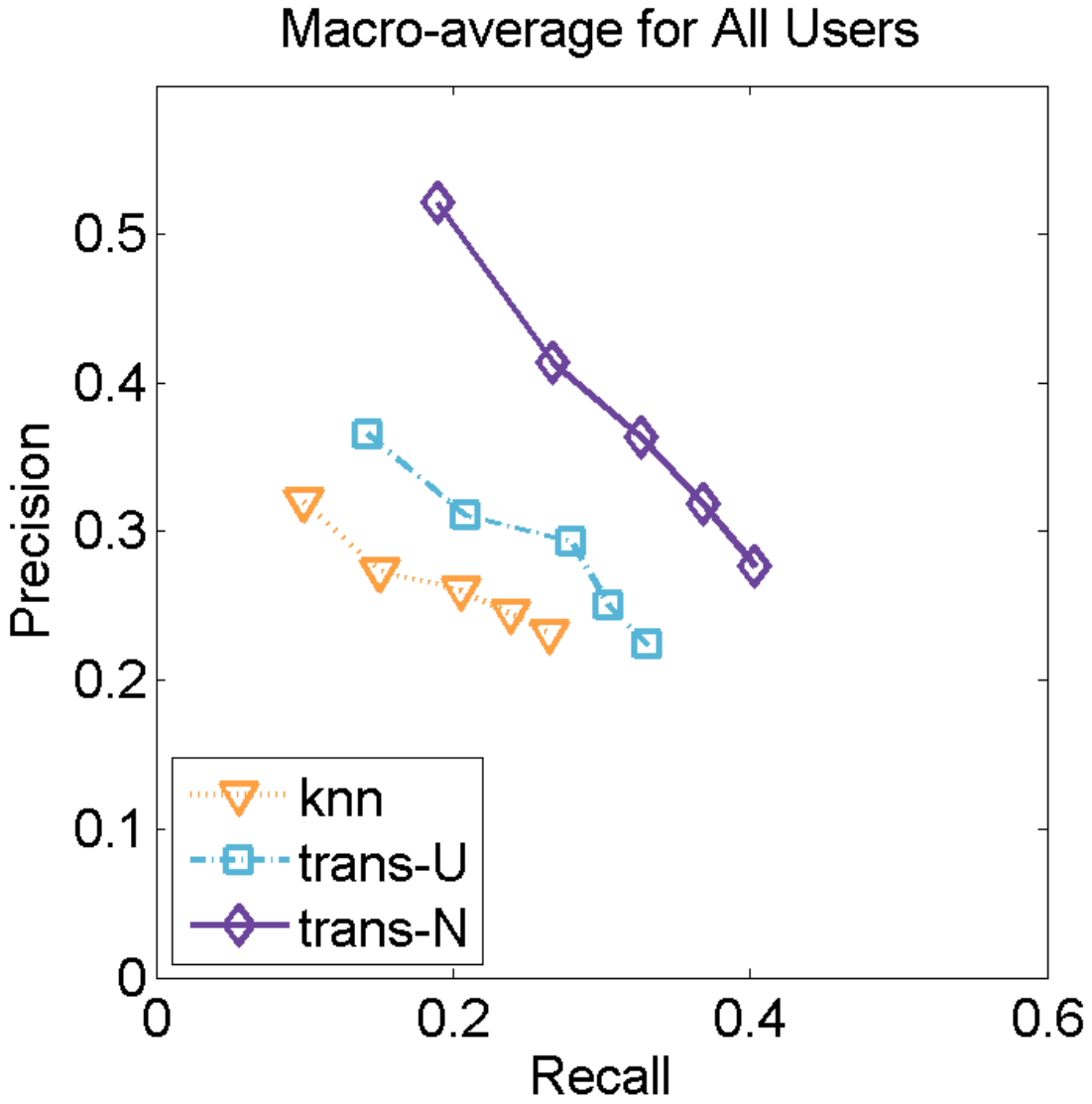
Our Proposed Method

Dataset from BibSonomy

	Train	Validation	Test
Time Frame	Feb 06 ~ Dec 08	Jan 09 ~ Jun 09	Jul 09 ~ Dec 09
Number of Users	1,185	136	57
Number of Bookmarks	64,120	775	279

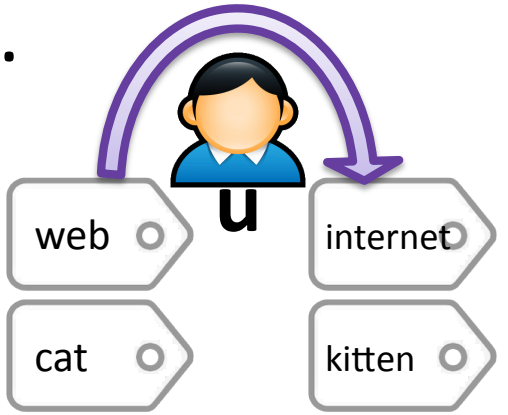
- Users in the test set are those
  - who appear in both training and validation sets.
  - whose tags used in the test are seen in the training set, but may not be used by themselves.

Precision-Recall Curve for Recommendation Performance

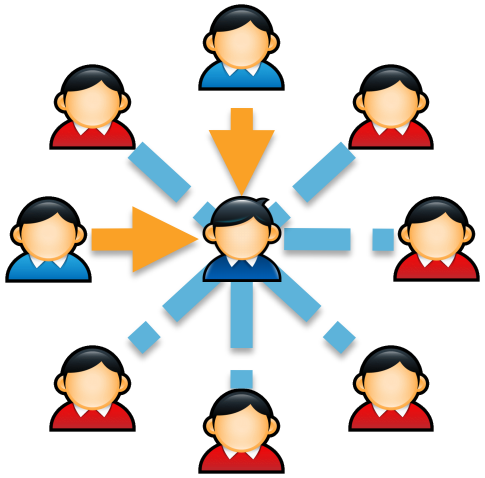


Contributions for Study 2

- Proposed a probabilistic framework.
  - Extract tagging preference patterns.
  - Find like-minded users.
  - Borrow tagging preference patterns.



- Evaluated Tag Recommendation
  - Dataset from BibSonomy.
  - trans-N outperforms baselines.



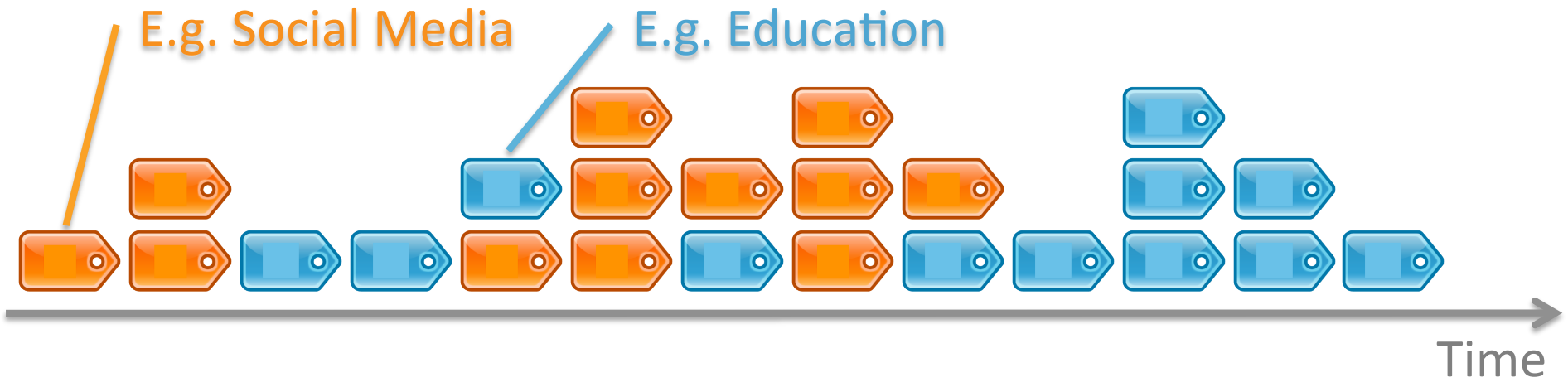
Outline

- Motivation
- Studies and Results:
  - Tag Prediction
  - Personalized Tag Recommendation
  - Trend Discovery
- Conclusion

Intuitions

- Tags arrive at different times…
  - Tagging captures interest in the annotated resources.
  - Tags carry interpretations of the annotators.

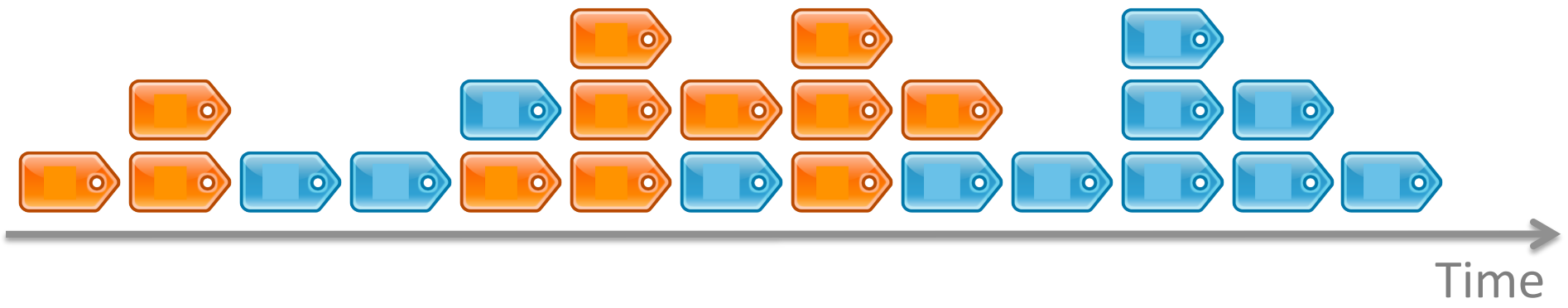
    → From different aspects of interest.



Motivating Questions

Given an annotated resource,

- How trending (amount of interest) are the resources?
- When did the trends emerge?
- How fast was the emergence?
- Which aspect of trends (e.g. topics)?



Research Objectives

Task: Trend Discovery using Social Annotations

Research Questions:

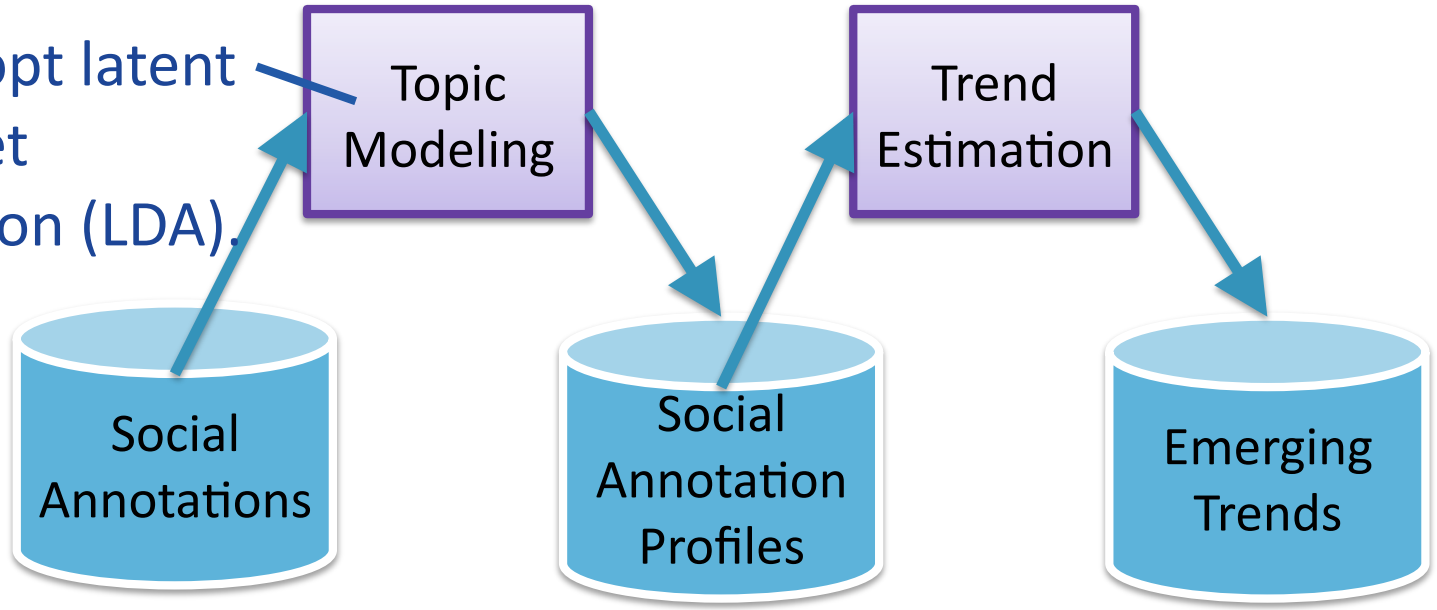
- How to discover emerging trends of the resource using social annotations?
- How to use the discovered trends to perform resource ranking tasks?

Proposed Trend Discovery Process

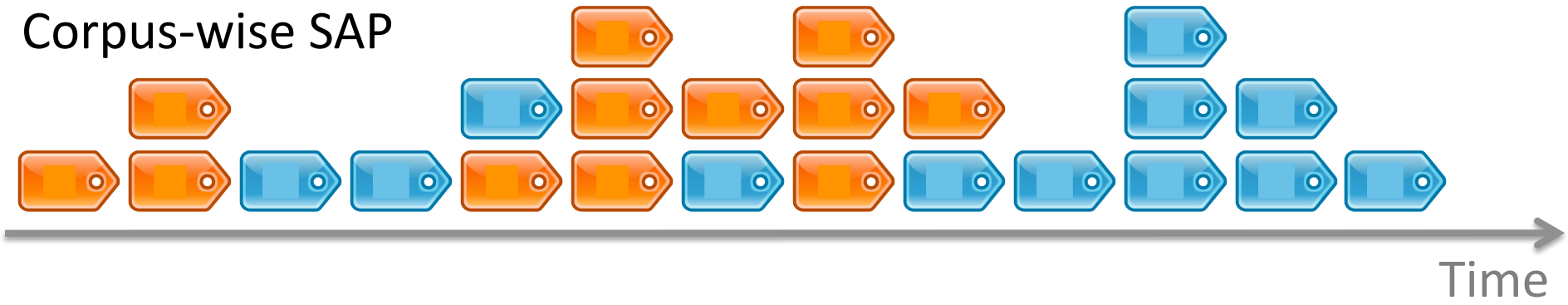
1. Topic Modeling:
To analyze the multiple aspects of interest in the annotation content.

2. Trend Estimation:
To parameterize the characteristics of emerging trends.

We adopt latent Dirichlet allocation (LDA).



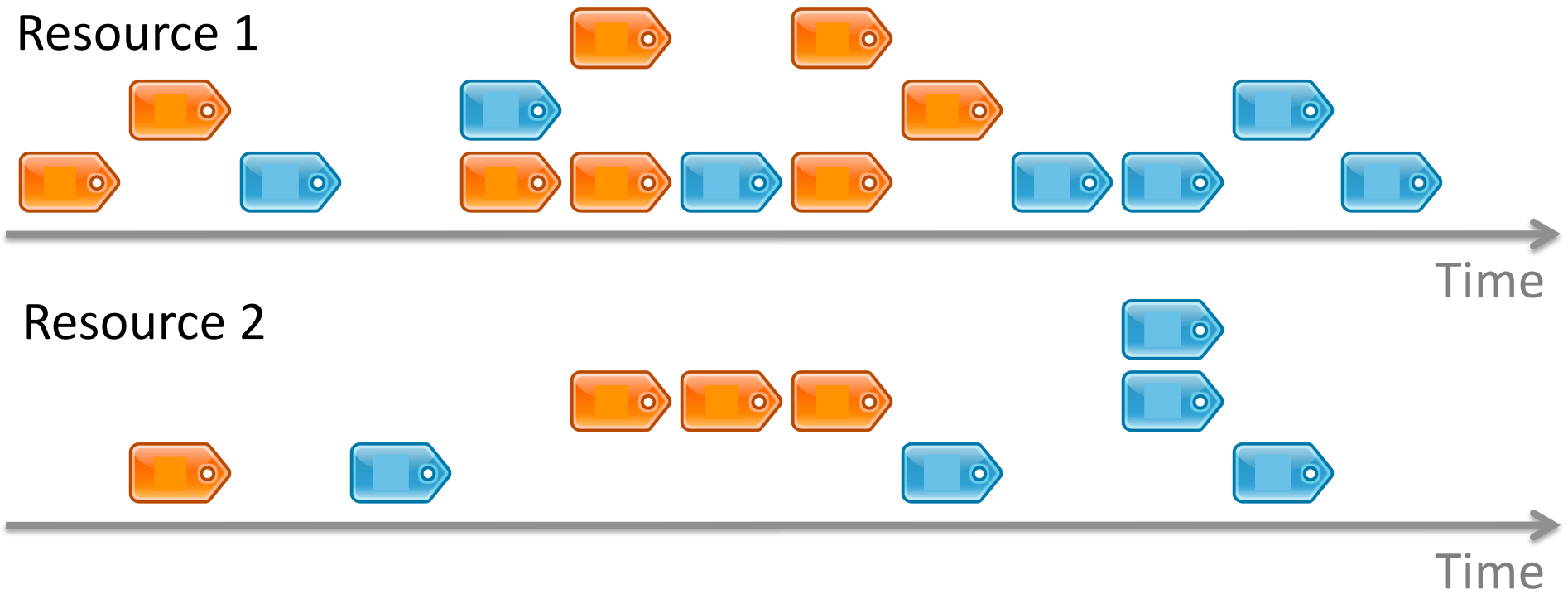
Social Annotation Profiles (SAP)



To slice the corpus-wise SAP by two dimensions.

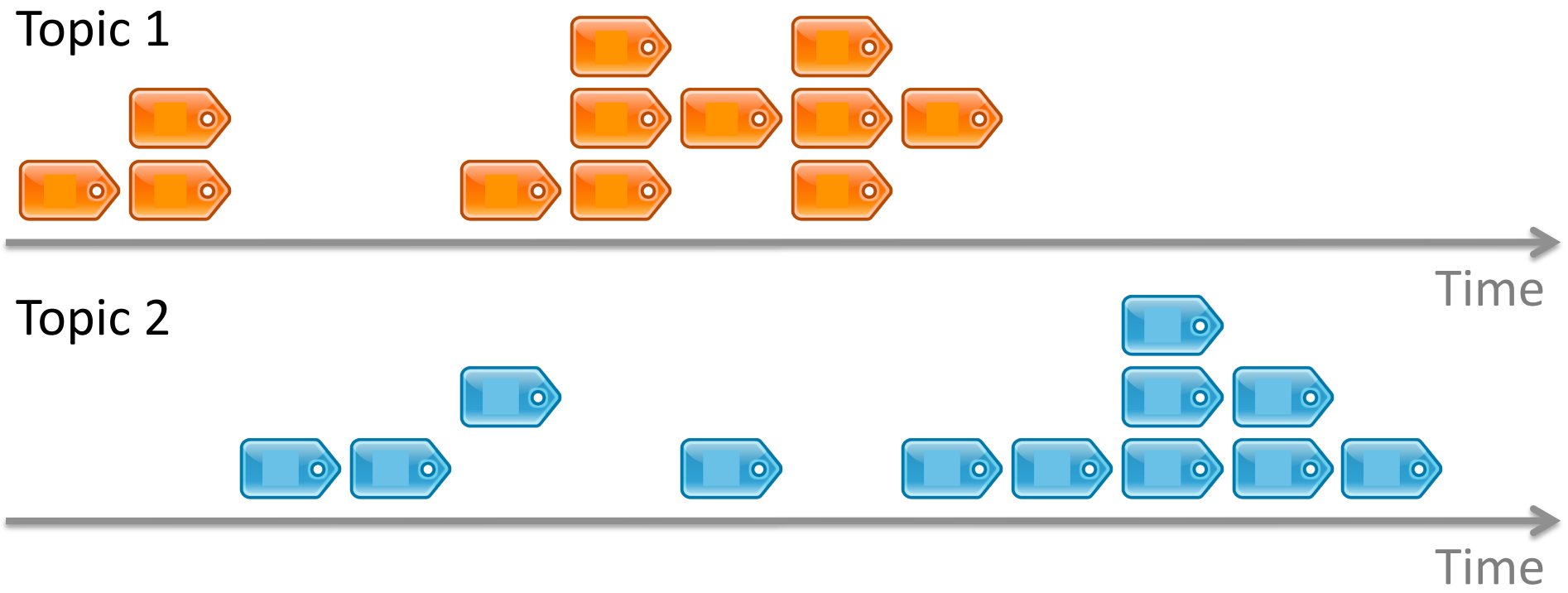
Corpus-wise SAP		**By Target Resource**	
		Regardless	Specific
By Target Topic	Regardless	-	**Resource-specific SAP**
	Specific	**Topic-specific SAP**	**Resource-Topic-specific SAP**

Resource-specific SAP



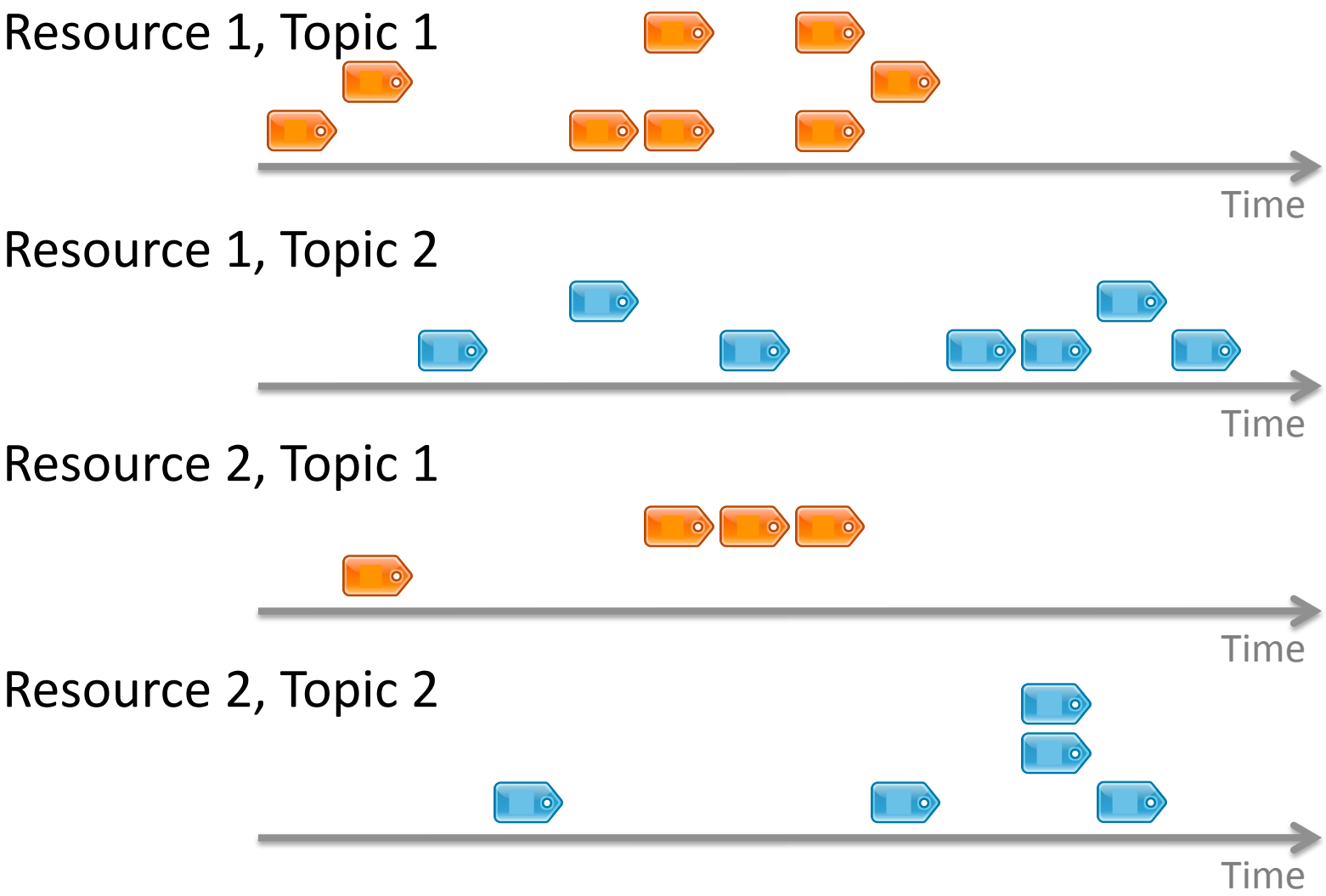
- All tags are assigned to the same resource.
  - Regardless of topic.

Topic-specific SAP

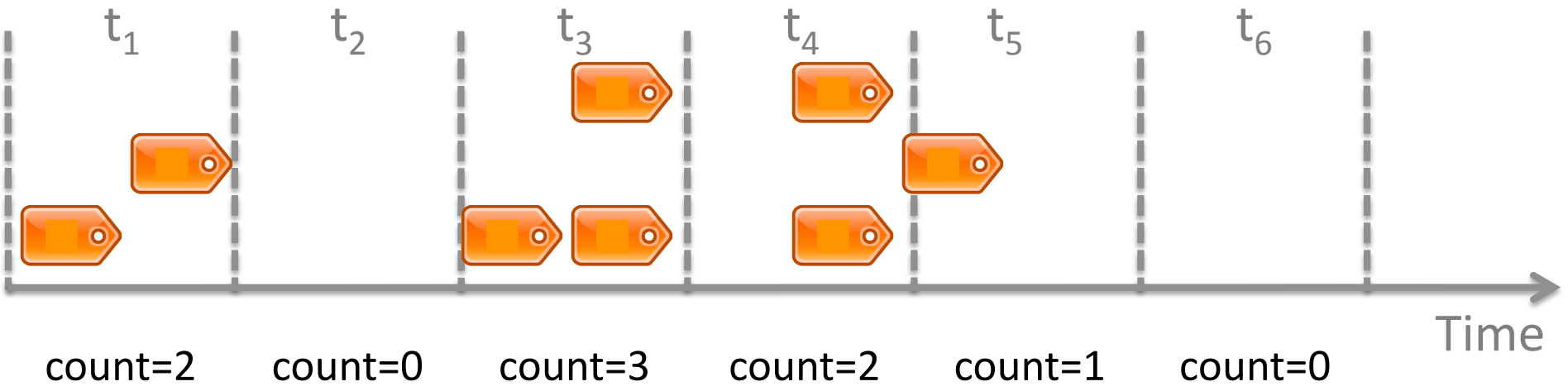


- All tags are for the same topic.
  - Regardless of target resource.

Resource-Topic-specific SAP



Social Annotation Time Series (Q)



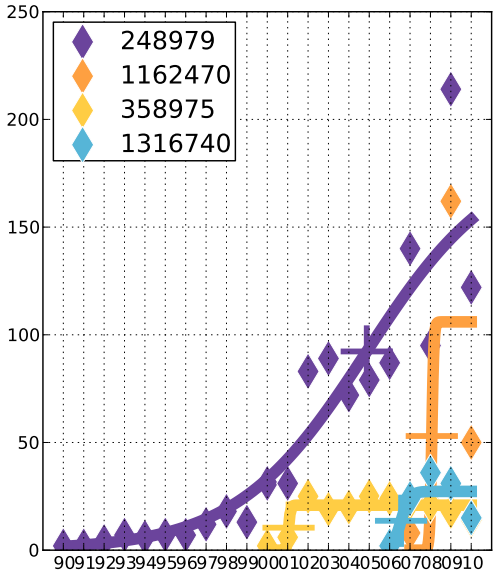
- Q = { (t_1,2),(t_2,0),(t_3,3),(t_4,2),(t_5,1),(t_6,0) }
- Trend Estimation on Q

Proposed Trend Estimator

- The Sigmoid Estimator

$$\hat{Q}(t) = \frac{\lambda}{1 + e^{-\sigma(t-\tau)}}$$

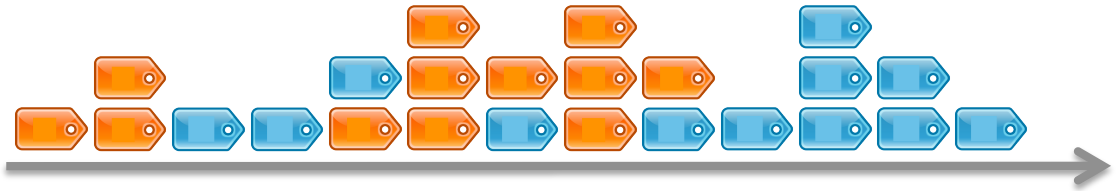
  - amplitude : λ → how trending
  - offset time : τ → when
  - gradient at t=τ : ¼$\lambda\sigma$ → how fast



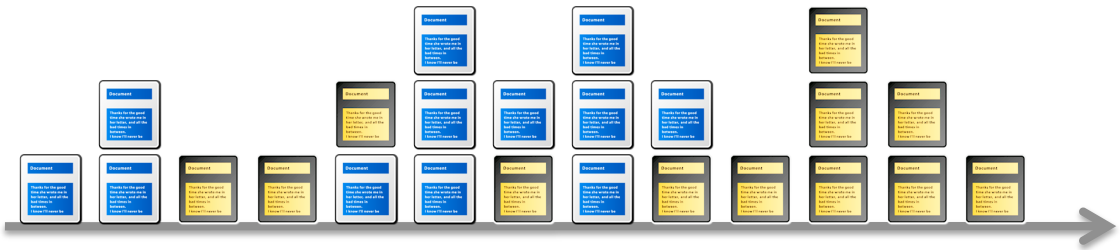
Fit to Data

Experiments

- Target Resources: Scientific Publications
- Social Annotations
  - Social Tags from CiteULike



  - Citing Documents from ACM DL



Common Characteristics:

1. From the annotators;
2. Assigned to the target publications;
3. Arrive at different time.

Trend Analysis Tasks

Trend Analysis Task	Social Annotation Profile Used
Corpus-wise Trending Topics	Topic-specific SAP
Resource-specific Topic Trends	Resource-Topic-specific SAP
Topic-specific Resource Ranking	

Corpus-wise Trending Topics

Topic ID	Topic Tags	How trending	When	How fast
122	acm vldb sigmod	124.1	SEP 07	35.1
157	social community	62.7	NOV 05	18.9
089	recommender personalization collaborativefiltering	55.2	MAY 07	18.6
027	ir retrieval relevancefeedback	50.2	NOV 06	29.6
103	hci interaction interface ui user	44.5	APR 06	11.8

Results using CiteULike data.

Resource-specific Topic Trends

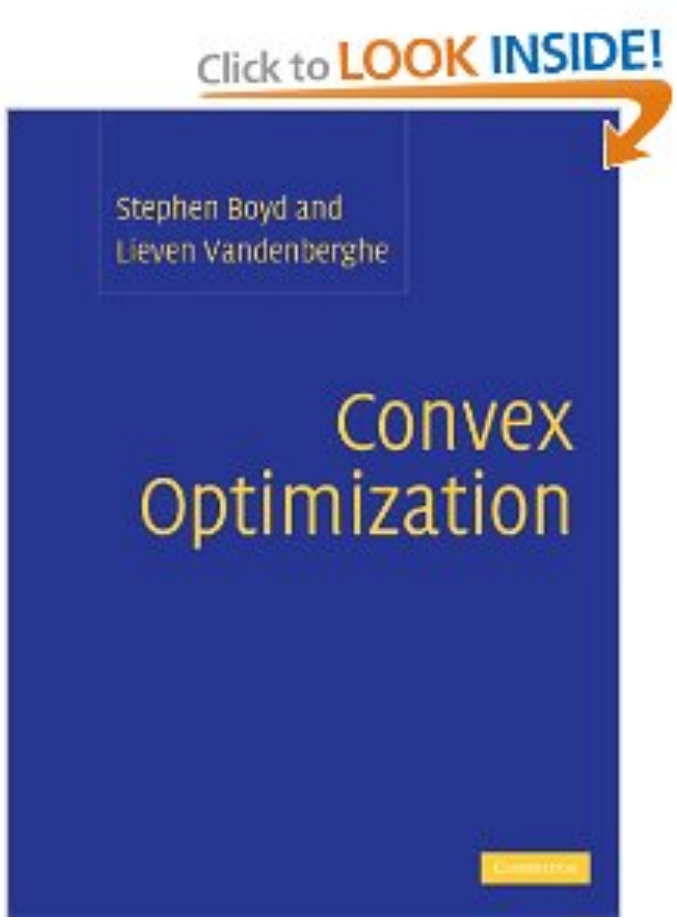
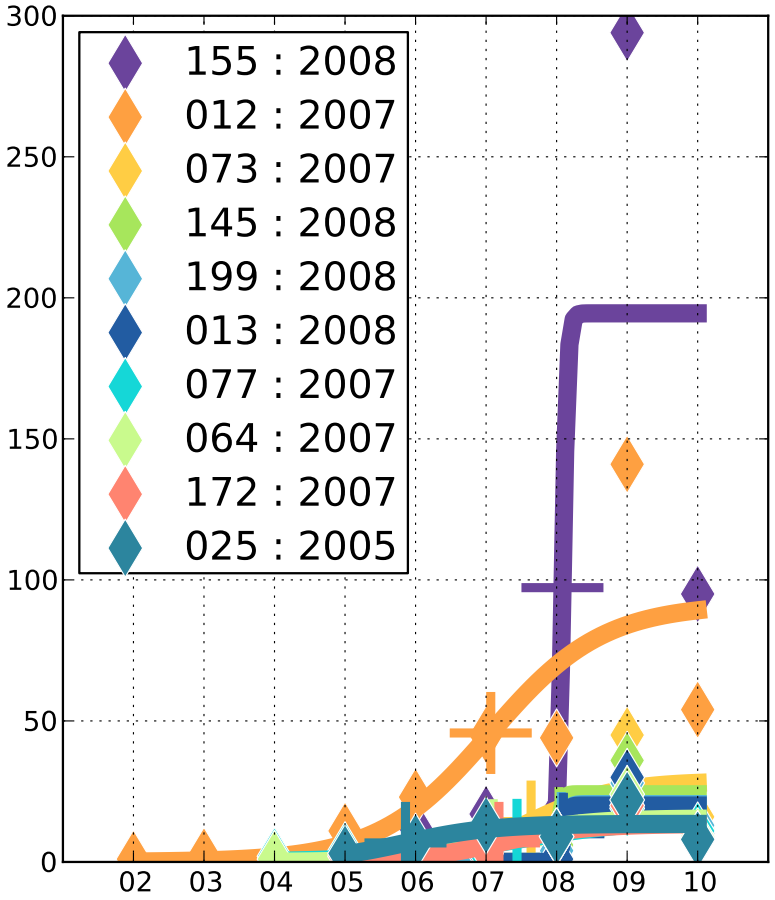


Image from Amazon™



Top Emerging Topics for the Book
Convex Optimization

Resource-specific Topic Trends (2)

For Resource: Convex Optimization				
Topic ID	**Topic Keywords**	**How trending**	**When**	**How fast**
155	channel capacity	194.5	2008	1378.9
012	optimization problem	91.4	2007	28.7
073	wireless networks	28.6	2007	11.6
145	sensor networks	24.0	2008	190.9
199	noise signal filters	20.5	2008	110.0

Results using ACM DL data.

Topic-specific Resource Ranking

For Topic 155: channel capacity				
Title	**How trending**	**When**	**How fast**	**Citations**
Elements of Information Theory	200.0	2008	1018.6	2410
Convex Optimization	194.5	2008	1378.9	1239
On limits of wireless communications in a fading environment when using multiple antennas	146.5	2008	782.5	487
NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey	93.0	2008	562.5	242
Matrix Computations (3rd ed.)	43.5	2008	224.2	1121

Results using ACM DL data.

Contributions for Study 3

- Trend Discovery Process
  - Analyze social annotation profiles.
  - Characterize emerging trends.
  - Perform trend analysis tasks.
- Experiments for Scientific Publications
  - Dataset from CiteULike and ACM DL
    - Corpus-wise Trending Topics
    - Resource-specific Topic Trends
    - Topic-specific Resource Ranking

Outline

- Motivation
- Studies and Results:
  - Tag Prediction
  - Personalized Tag Recommendation
  - Trend Discovery
- Conclusion

Conclusion

- Studies in this dissertation address current challenges for navigating the social tagging space:
  - Tag prediction addresses tag sparseness for resources.
  - Personalized tag recommendation addresses the tagging preferences of individual users.
  - Trend discovery using social annotations is a novel task for resource ranking.

Summary of Contributions

- Tag Prediction
  - Model topics, tags and words.
  - Assume correspondence.
- Personalized Tag Recommendation
  - Extract tagging preference patterns.
  - Find like-minded users.
- Trend Discovery
  - Analyze temporal profiles.
  - Characterize emerging trends.

Publications

Journal Submission:

- Meiqun Hu, Ee-Peng Lim and Jing Jiang, **Social Tag Prediction using Latent Topics**. Submitted to the Journal of the American Society of Information Science and Technology. Under review.

Conference Publications:

- Meiqun Hu, Ee-Peng Lim and Jing Jiang, **Using Social Annotations for Trend Discovery in Scientific Publications**. In HCIR 2011.
- Meiqun Hu, Ee-Peng Lim and Jing Jiang, **A Probabilistic Approach to Personalized Tag Recommendation**. In SocialCom 2010.

Included in Course References, COMP621U, Spring 2011, HKUST.

Doctoral Consortium:

- Meiqun Hu, Ee-Peng Lim and Jing Jiang, **A Topic Modeling Approach to Social Tag Prediction**. In Bulletin of IEEE Technical Committee on Digital Libraries, 6, Fall 2010.

Other Publications

Conference Publications:

- Meiqun Hu, Ee-Peng Lim and Ramayya Krishnan, **Predicting Outcome for Collaborative Featured Article Nomination in Wikipedia**. In ICWSM 2009.
- Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw and Ba-Quy Vuong, **Measuring Article Quality in Wikipedia: Models and Evaluation.** In CIKM 2007.
- Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady W. Lauw and Ba-Quy Vuong, **On Improving Wikipedia Search using Article Quality**. In WIDM 2007.

Included in Course References, CS 598, Spring 2008, UIUC.

Thank you!