Learning with Social Media

Tom Chao Zhou @Thesis Defense

Thesis Committee:
Prof. Yu Xu Jeffrey (Chair)
Prof. Zhang Sheng Yu (Committee Member)
Prof. Yang Qiang (External Examiner)

Supervisors:
Prof. Irwin King
Prof. Michael R. Lyu

- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

- Introduction
  - Background
  - Item Recommendation with Tagging Ensemble
  - User Recommendation via Interest Modeling
  - Item Suggestion with Semantic Analysis
  - Item Modeling via Data-Driven Approach
  - Conclusion and Future Work

Social Media

- What is Social Media?
  - Create, share, exchange; virtual communities
- Some Data
  - 45 million reviews in a travel forum TripAdvisor [Source]
  - 218 million questions solved in Baidu Knows [Source]
  - Twitter processed one billion tweets in Dec 2009, averages almost 40 million tweets per day [Source]
  - Time spent on social media in US: 88 billion minutes in July 2011, 121 billion minutes in July 2012 [Source]

Examples of Social Media

- Rating System

amazon.com and you're done.™ America's largest online retailer

淘宝网 Taobao.com The largest C2C website in China, over 2 billion products

IMDb The Internet Movie Database The biggest movie site on the planet, over 1,424,139 movies and TV episodes

Examples of Social Media

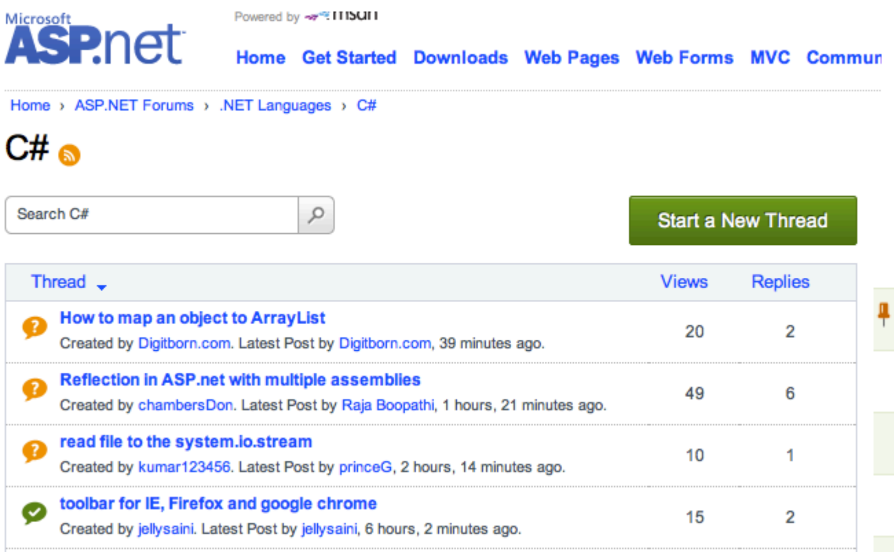
- Social Tagging System

del.icio.us social bookmarking — The largest social bookmarking website

flickr — The best online photo management and sharing application in the world

Examples of Social Media

- Online Forum



Examples of Social Media

- Community-based Question Answering

YAHOO! ANSWERS

10 questions and answers are posted per second

Baidu知道

218 million questions have been solved

Quora

A popular website with many experts and high quality answers

Challenges in Social Media

- Astronomical growth of data in Social Media
- Huge, diverse and dynamic
- Drowning in information, information overload

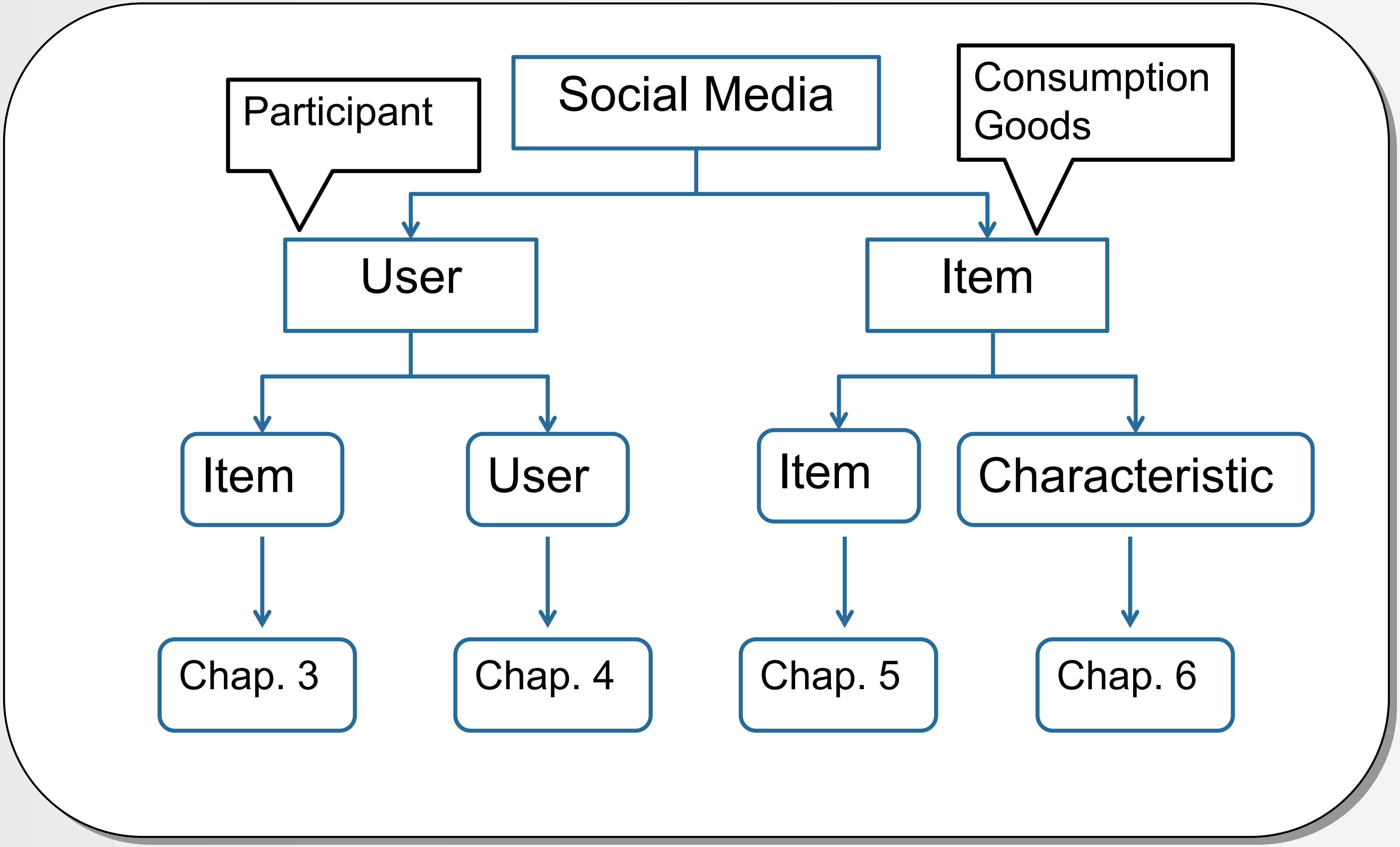
Objective of Thesis

- Establish automatic and scalable models to help social media users find their information needs more effectively

Objective of Thesis

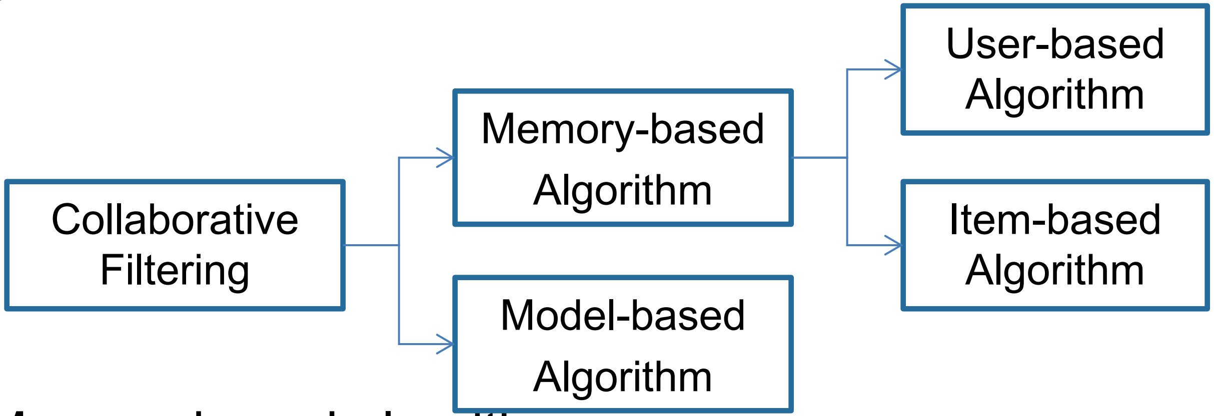
- Modeling users' interests with respect to their behavior, and recommending items or users they may be interested in
  - Chapter 3, 4
- Understanding items' characteristics, and grouping items that are semantically related for better addressing users' information needs
  - Chapter 5, 6

Structure of Thesis



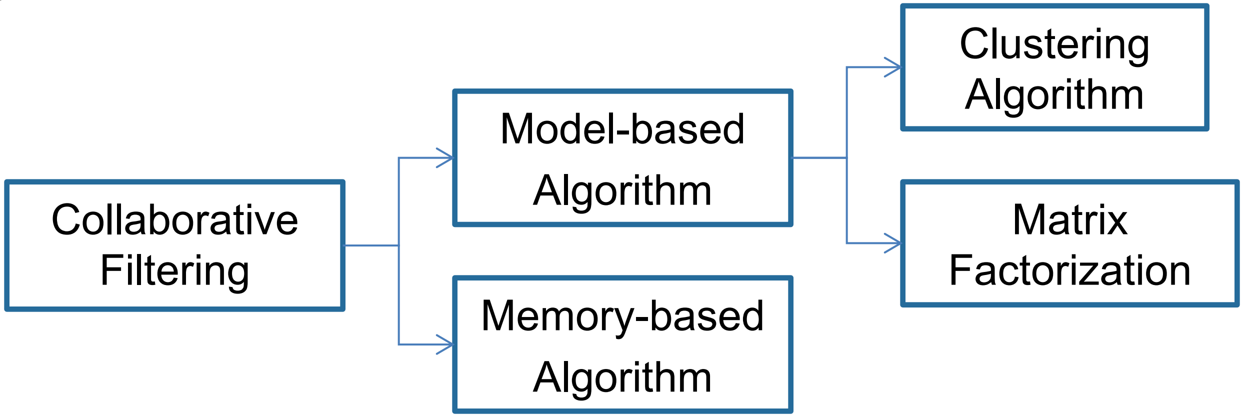
- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Recommender Systems



- Memory-based algorithms
  - User-based
  - Item-based
- Similarity methods
  - Pearson correlation coefficient (PCC)
  - Vector space similarity (VSS)
- Disadvantage of memory-based approaches
  - Recommendation performances deteriorate when the rating data is sparse

Recommender Systems



- Model-based algorithms
  - Clustering methods
  - Matrix factorization methods
- Disadvantage of traditional model-based approaches
  - Only use the user-item rating matrix, ignore other user behavior
  - Suffer the problem of data sparsity

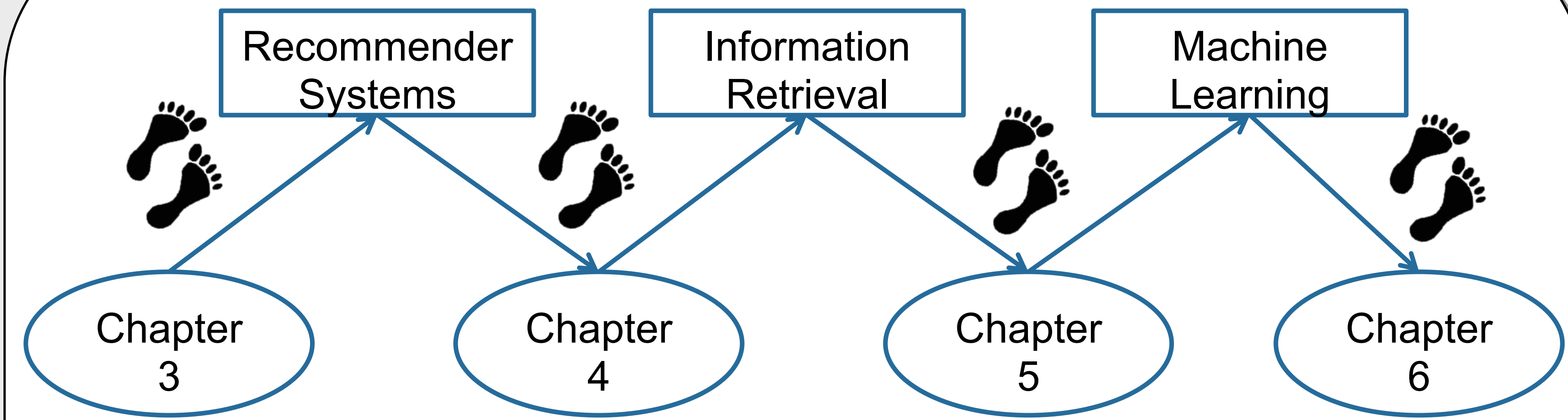
Machine Learning

- Whether the training data is available
- Yes? Supervised learning
  - Naive Bayes, support vector machines
- Some? Semi-supervised learning
  - Co-training, graph-based approach
- No? Unsupervised learning
  - Clustering, Latent Dirichlet Allocation

Information Retrieval

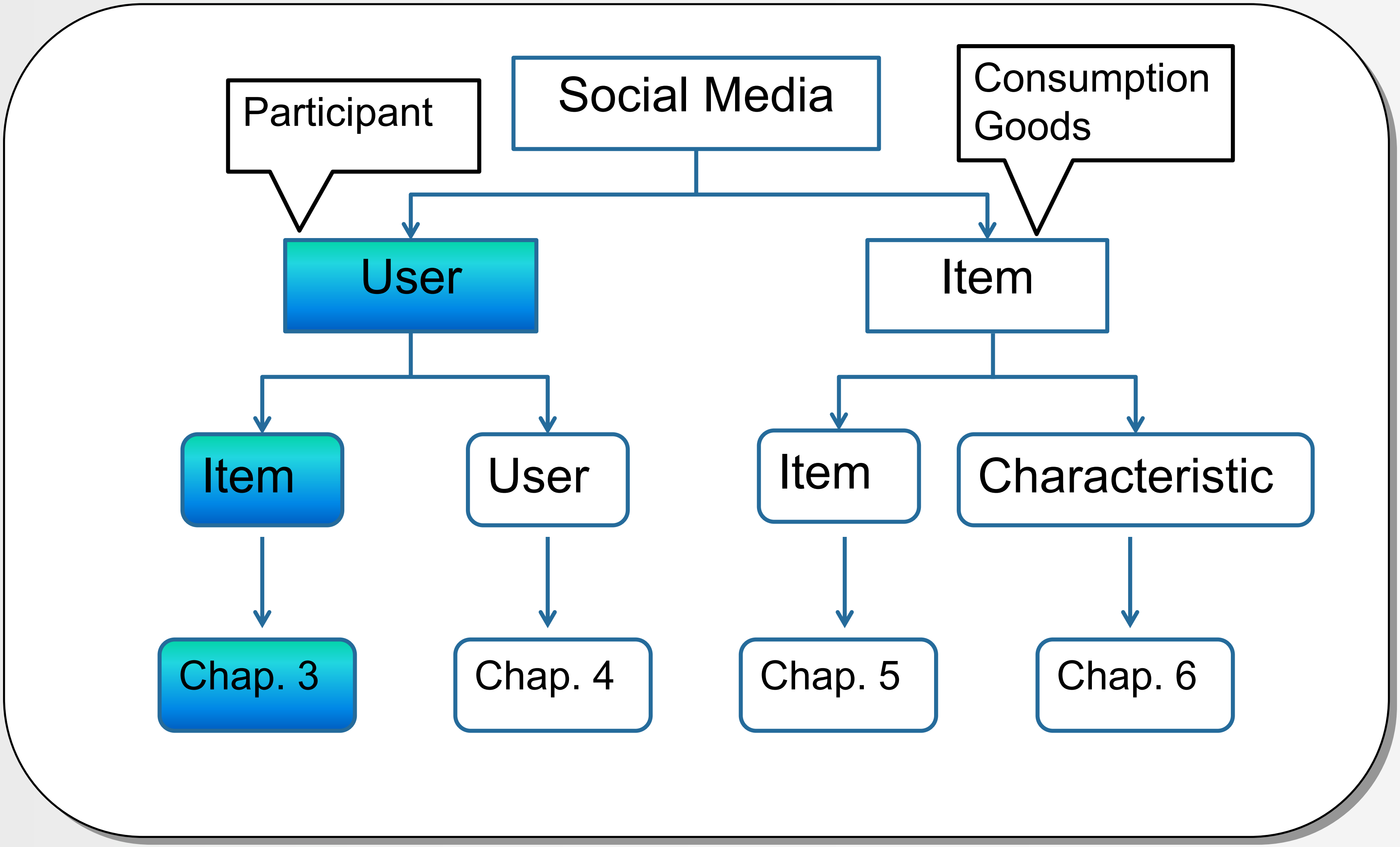
- Information Retrieval Models
  - Seek an optimal ranking function
- Vector Space Model
  - Weighting (TF-IDF)
- Probabilistic Model and Language Model
  - Binary independence model, query likelihood model
- Translation Model
  - Originated from machine translation
  - Solve the lexical gap problem

Techniques Employed



- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Structure of Thesis



A Toy Example

	The Godfather	Inception	Forrest Gump
Alex	4	?	5
Bob	4	2	?
Tom	?	2	4

1: Strong dislike, 2: Dislike, 3: It’s OK, 4: Like, 5: Strong like

Challenge

- Rating matrix is very sparse, density of ratings in commercial recommender system is less than 1%
- Performance deteriorates when rating matrix becomes sparse

Problem

Task: Predicting the missing values

User-item rating matrix

	i_1	i_2	i_3	i_4	i_5
u_1	3	5	2	?	?
u_2	?	4	?	4	?
u_3	3	4	1	?	?
u_4	?	?	?	3	5
u_5	?	5	?	4	?

Fact:

Ratings reflect users' preferences

Challenge:

Rating matrix is very sparse, only use rating information not enough

Thought:

Whether there exists contextual information that can also reflect users' judgments?

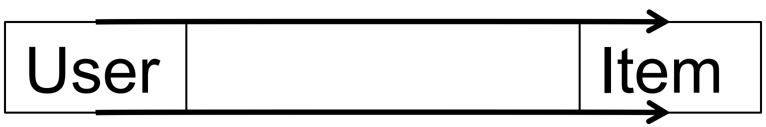
How can we utilize that kind of contextual information to improve the prediction quality?

Motivation

- Social tagging is to collaboratively creating and managing tags to annotate and categorize content
- Tags can represent users' judgments and interests about Web contents quite accurately

Motivation

Rating:
preference

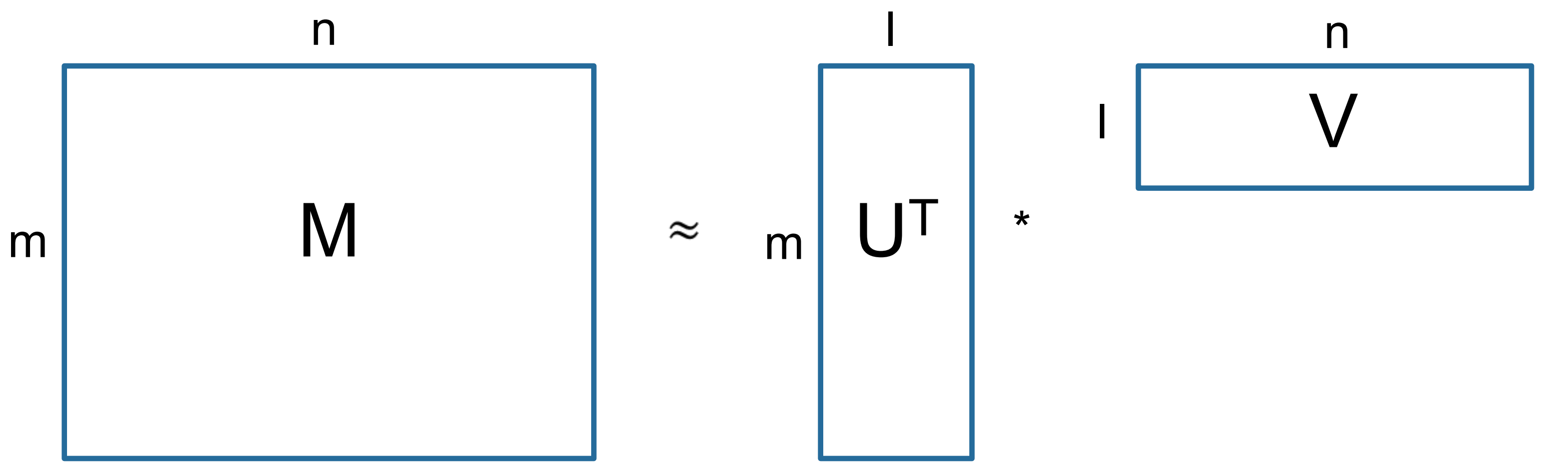


Tagging:
interest

To improve the recommendation quality and tackle the data sparsity problem, fuse tagging and rating information together

Intuition of Matrix Factorization

- $M = U^T * V, M \in R^{m*n}, U \in R^{l*m}, V \in R^{l*n}, l \ll (m, n)$



- Physical meaning of each row in U and V is a latent semantic dimension
- E.g., action, comedy, if M is a user-movie rating matrix

User-Item Rating Matrix Factorization

Conditional distributions over the observed

$$p(R|U,V,\sigma_R^2) = \prod_{i=1}^{m}\prod_{j=1}^{n}[\mathcal{N}(r_{ij}|g(U_i^TV_j),\sigma_R^2)]^{I_{ij}^R}$$

- U: user latent feature matrix.
- V: item latent feature matrix.
- $U_i^TV_j$: predicted rating (user i to item j).

	i_1	i_2	i_3	i_4	i_5
u_1	3	5	2		
u_2		4		4	
u_3	3	4	1		
u_4				3	5
u_5		5		4	

User-Item Rating Matrix R

Zero-mean spherical Gaussian priors are placed on the user latent feature matrix and the item latent feature matrix

$$p(U|\sigma_U^2) = \prod_{i=1}^{m}\mathcal{N}(U_i|0,\sigma_U^2\mathbf{I})$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{n}\mathcal{N}(V_j|0,\sigma_V^2\mathbf{I})$$

Posterior distributions of U and V based only on observed ratings

$$p(U,V|R,\sigma_V^2,\sigma_U^2,\sigma_R^2)$$

User-Tag Tagging Matrix Factorization

Conditional over the observed tagging data

$$p(C|U,T,\sigma_C^2) = \prod_{i=1}^{m}\prod_{k=1}^{o}[\mathcal{N}(c_{ik}|g(U_i^T T_k),\sigma_C^2)]^{I_{ik}^C}$$

- ➢U: user latent feature matrix,
- ➢T: tag latent feature matrix.
- ➢$U_i^T T_{k:}$ predicted value of the model.

	t_1	t_2	t_3	t_4	t_5
u_1	4	32	5		
u_2		4		4	
u_3	3	33	12		
u_4				3	5
u_5		5		4	

User-Tag Tagging Matrix C

Posterior distributions of U and T

$$p(U,T|C,\sigma_U^2,\sigma_T^2,\sigma_C^2)$$

Jack:

action (20), animation (20), romantic (1)

Item-Tag Tagging Matrix Factorization

	t_1	t_2	t_3	t_4	t_5
i_1	14	20	15		
i_2		4		4	
i_3	13	23	12		
i_4				13	5
i_5		15		14	

Item-Tag Tagging Matrix D

Titanic:

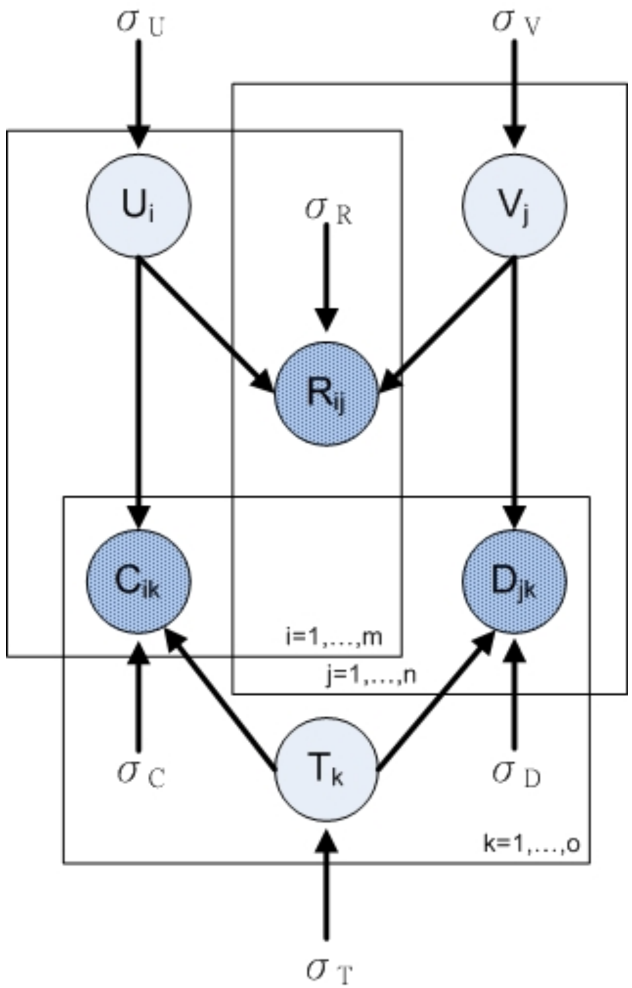
romance (20), bittersweet (20), action (1)

Posterior distributions of V and T

$$p(V, T | D, \sigma_D^2, \sigma_T^2, \sigma_V^2)$$

TagRec Framework

U	User latent feature matrix
V	Item latent feature matrix
T	Tag latent feature matrix
R	User-item rating matrix
C	User-tag tagging matrix
D	Item-tag tagging matrix



Experimental Analysis

- MovieLens 10M/100K data set:
  - Provided by GroupLens research
  - Online movie recommender service MovieLens (http://movielens.umn.edu)
- Statistics:
  - Ratings: 10,000,054
  - Tags: 95,580
  - Movies: 10,681
  - Users: 71,567

Experimental Analysis

- MAE comparison with other approaches (a smaller MAE means better performance)

Training Data	Baseline Methods		Dimensionality = 10			Dimensionality = 20		
	UMEAN	IMEAN	SVD	PMF	TagRec	SVD	PMF	TagRec
80%	0.7686	0.7379	0.6169	0.6162	**0.6159**	0.6167	0.6156	**0.6145**
50%	0.7710	0.7389	0.6376	0.6354	**0.6352**	0.6349	0.6337	**0.6307**
30%	0.7742	0.7399	0.6617	0.6599	**0.6528**	0.6570	0.6569	**0.6494**
20%	0.7803	0.7416	0.6813	0.6811	**0.6664**	0.6776	0.6766	**0.6650**
10%	0.8234	0.7484	0.7315	0.7127	**0.6964**	0.7264	0.7089	**0.6962**

UMEAN: mean of the user's ratings

IMEAN: mean of the item's ratings

SVD: A well-know method in Netflix competition

PMF: Salakhutdinov and Mnih in NIPS'08

Experimental Analysis

- RMSE comparison with other approaches (a smaller RMSE value means a better performance)

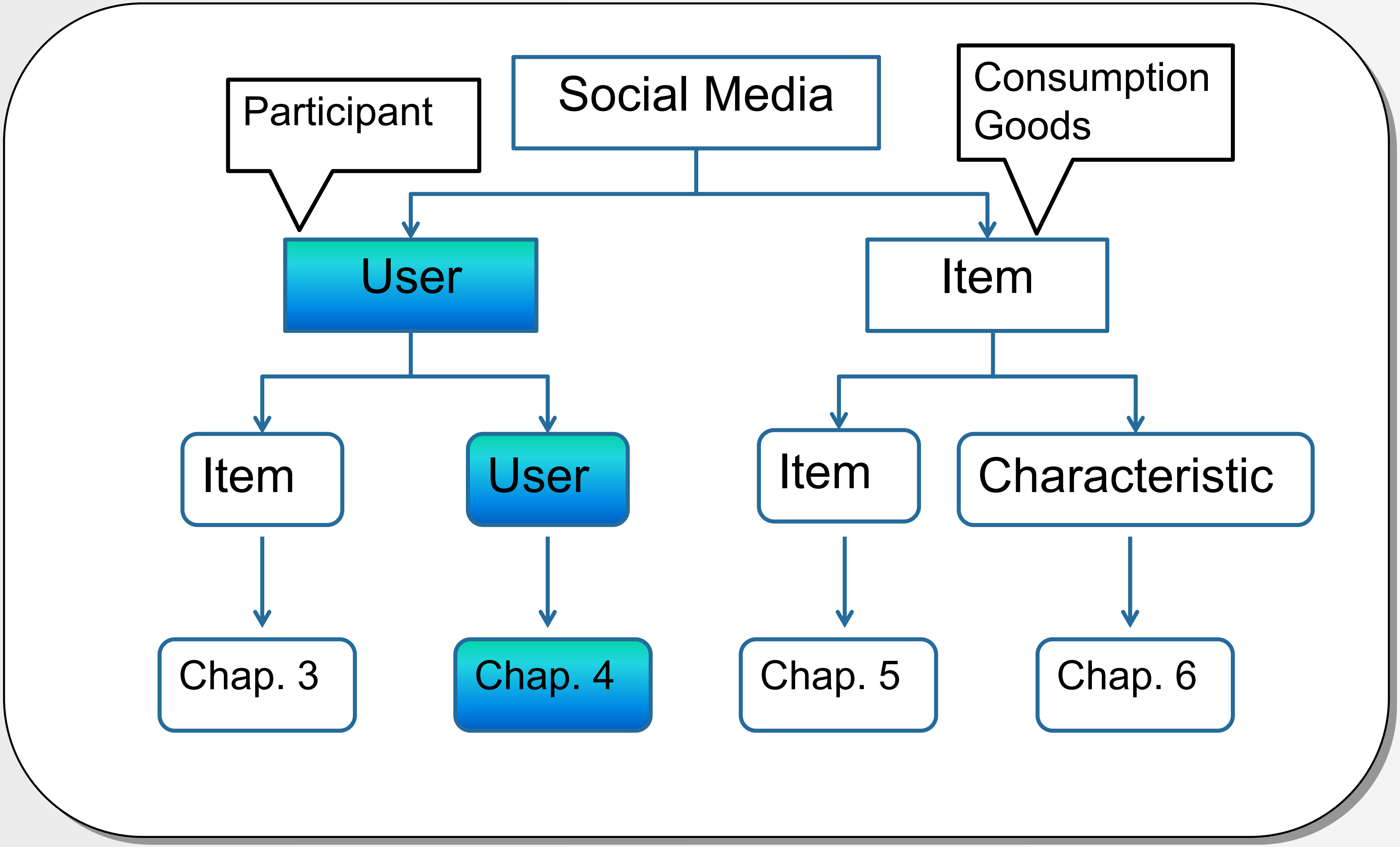
Training Data	Baseline Methods		Dimensionality = 10			Dimensionality = 20		
	UMEAN	IMEAN	SVD	PMF	TagRec	SVD	PMF	TagRec
80%	0.9779	0.9440	0.8087	0.8078	**0.8077**	0.8054	0.8025	**0.8022**
50%	0.9816	0.9463	0.8330	0.8326	**0.8321**	0.8289	0.8252	**0.8217**
30%	0.9869	0.9505	0.8636	0.8587	**0.8492**	0.8575	0.8553	**0.8450**
20%	1.0008	0.9569	0.8900	0.8824	**0.8659**	0.8857	0.8791	**0.8639**
10%	1.1587	0.9851	0.9703	0.9236	**0.9038**	0.9638	0.9183	**0.9031**

Contribution of Chapter 3

- Propose a factor analysis approach, referred to as TagRec, by utilizing both users' rating information and tagging information based on probabilistic matrix factorization
- Overcome the data sparsity problem and non-flexibility problem confronted by traditional collaborative filtering algorithms

- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Structure of Thesis



Problem and Motivation

- Social Tagging System

delicious
social bookmarking

Calendar **World Cup 2010** by MARCA.com
www.marca.com/deporte/futbol/mundial/sudafrica-2010/calendario-english.html
worldcup · calendar · visualization · football · 2010 — 3610

FIFA.com - The matches of **2010** FIFA **World Cup** South Africa
www.fifa.com/worldcup/matches/index.html
worldcup · fifa · soccer · football · 2010 — 606

World Cup 2010 Twitter replay | **Football** | guardian.co.uk
www.guardian.co.uk/football/world-cup-match-replay
twitter · visualization · worldcup · infographics · guardian — 1130

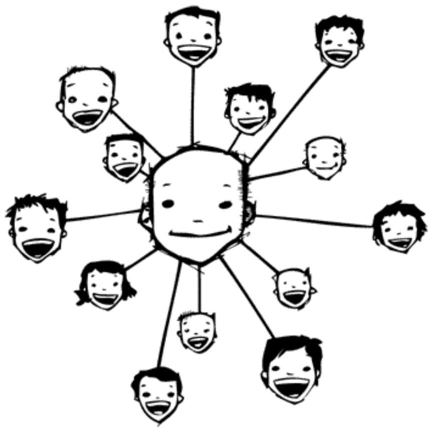
FIFA.com - Fédération Internationale de **Football** Association (FIFA)
www.fifa.com/
soccer · sports · fifa · sport · football — 1234

Problem and Motivation

- Tagging:
  - Judgments on resources
  - Users' personal interests

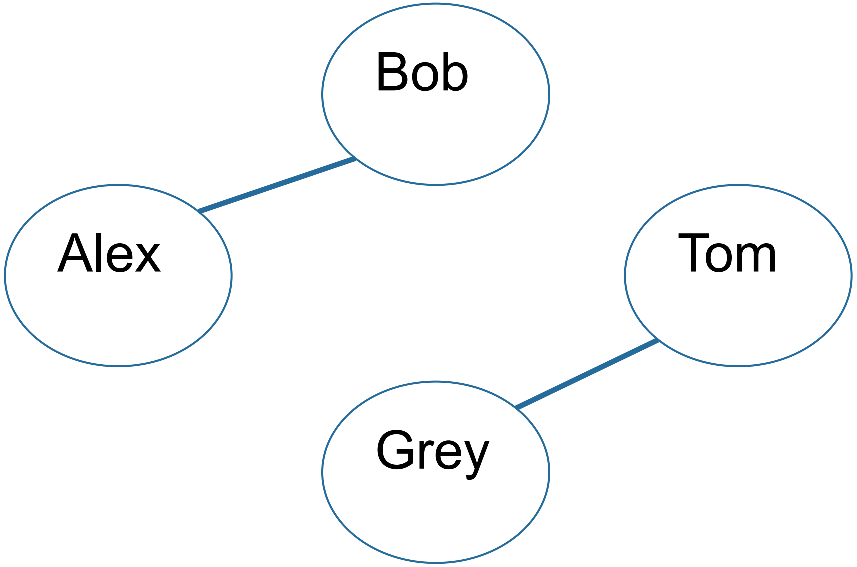
Problem and Motivation

- Providing an automatic interest-based user recommendation service



Challenge

- How to model users' interests?
- How to perform interest-based user recommendation?



UserRec: User Interest Modeling

- Triplet: user, tag, resource

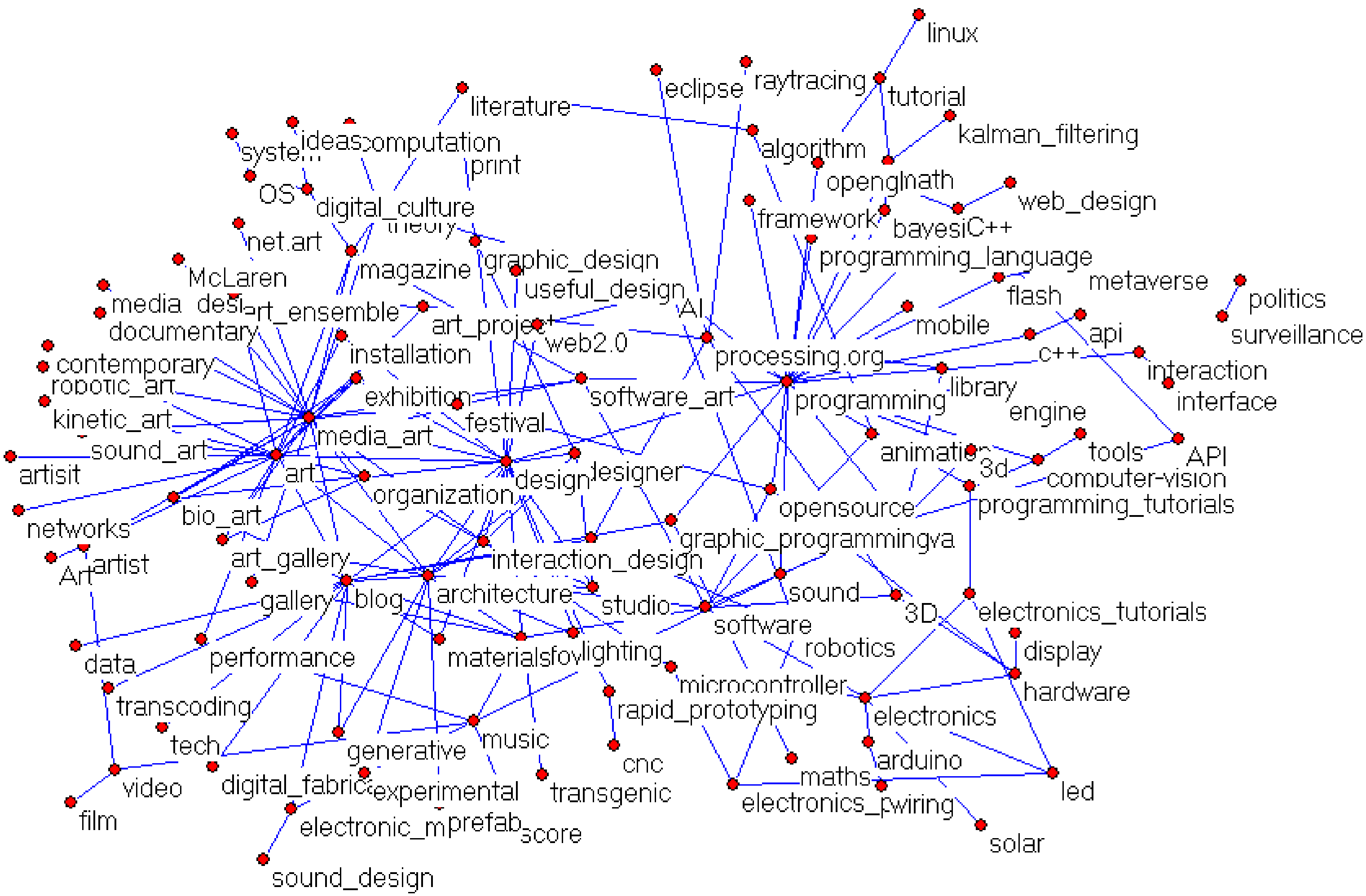
URL	http://www.nba.com
Tags of user 1	Basketball, nba
Tags of user 2	Sports, basketball, nba

- Observations of tagging activities:
  - Frequently used user tags can be utilized to characterize and capture users' interests
  - If two tags are used by one user to annotate one URL at the same time, it is very likely that these two tags are semantically related

UserRec: User Interest Modeling

- User Interest Modeling:
  - Generate a weighted tag-graph for each user
  - Employ a community discovery algorithm in each tag-graph

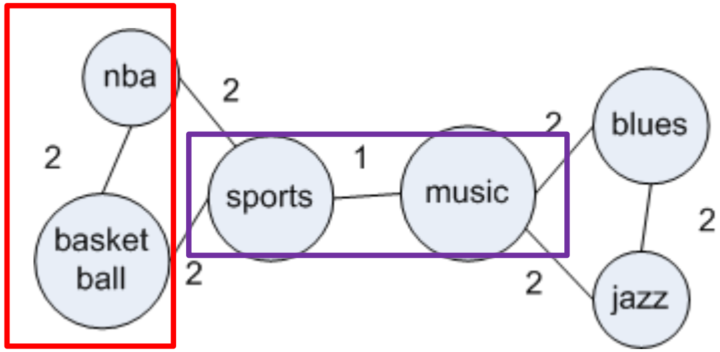
UserRec: User Interest Modeling



UserRec: User Interest Modeling

- Generate a weighted tag-graph for each user:

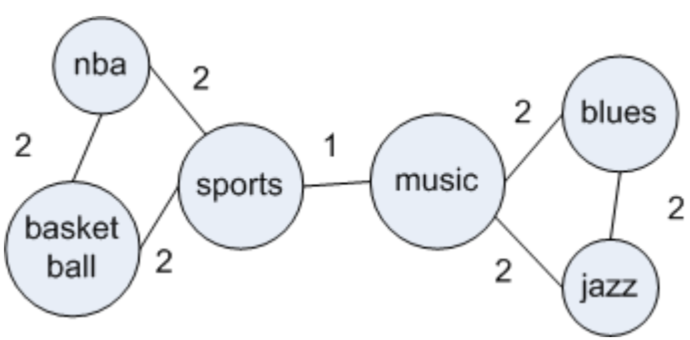
http://espn.go.com	basketball, nba, sports
http://msn.foxsports.com	basketball, nba, sports
http://www.ticketmaster.com	sports, music
http://freemusicarchive.org	music, jazz, blues
http://www.wwoz.org	music, jazz, blues



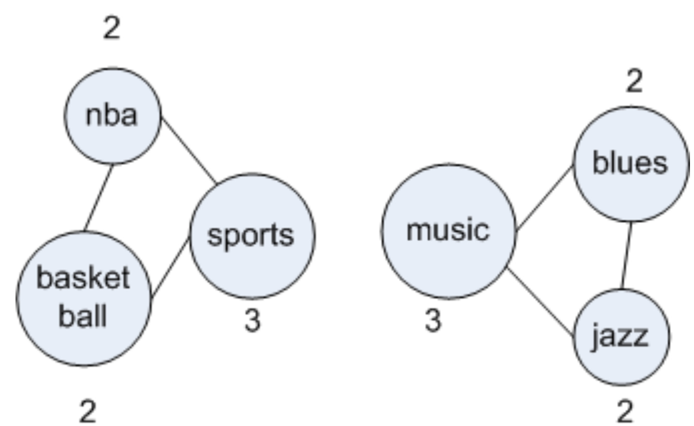
tag-graph

UserRec: User Interest Modeling

- Employ community discovery in tag-graph
  - Optimize modularity
  - If the fraction of within-community edges is no different from what we would expect for the randomized network, then modularity will be zero
  - Nonzero values represent deviations from randomness



tag-graph



two communities

Interest-based User Recommendation

- Representing topics of user with a random variable
  - Each community discovered is considered as a topic
  - A topic consists of several tags
  - Importance of a topic is measured by the sum of number of times each tag is used in this topic
  - Employ maximum likelihood estimation to calculate the probability value of each topic of a user
- A Kullback-Leibler divergence (KL-divergence) based method to calculate the similarity between two users based on their topics' probability distributions

Experimental Analysis

- Data Set:
  - Delicious
- Statistics:

Users	Bookmarks	Network*	Fans**
366,827	49,692,497	425,069	395,415

* This is the total number of users in all users' personal networks.

** This is the total number of fans of all users.

Experimental Analysis

- Memory-based collaborative filtering methods:
  - Person correlation coefficient (PCC)
  - PCC-based similarity calculation method with significance weighting
- Model-based collaborative filtering methods:
  - Probabilistic matrix factorization
  - Singular value decomposition
  - After deriving the latent feature matrices, we still need to use memory-based approaches on derived latent feature matrices: SVD-PCC, SVD-PCCW, PMF-PCC, PMF-PCCW

Experimental Analysis

Comparison with approaches those are based on URLs (a larger value means a better performance for each metric)

Metrics	Memory-Based Approaches		Model-Based Approaches				UserRec
	PCC	PCCW	SVD-PCC	SVD-PCCW	PMF-PCC	PMF-PCCW	
Precision@R	0.0717	0.1490	0.0886	0.0907	0.1136	0.1322	**0.3272**
MAP	0.1049	0.1874	0.1218	0.1245	0.1491	0.1745	**0.3752**
Bpref	0.0465	0.1148	0.0568	0.0582	0.0765	0.1029	**0.2913**
MMVRR	0.0626	0.1154	0.0710	0.0736	0.0858	0.1088	**0.2345**

Comparison with approaches those are based on Tags (a larger value means a better performance for each metric)

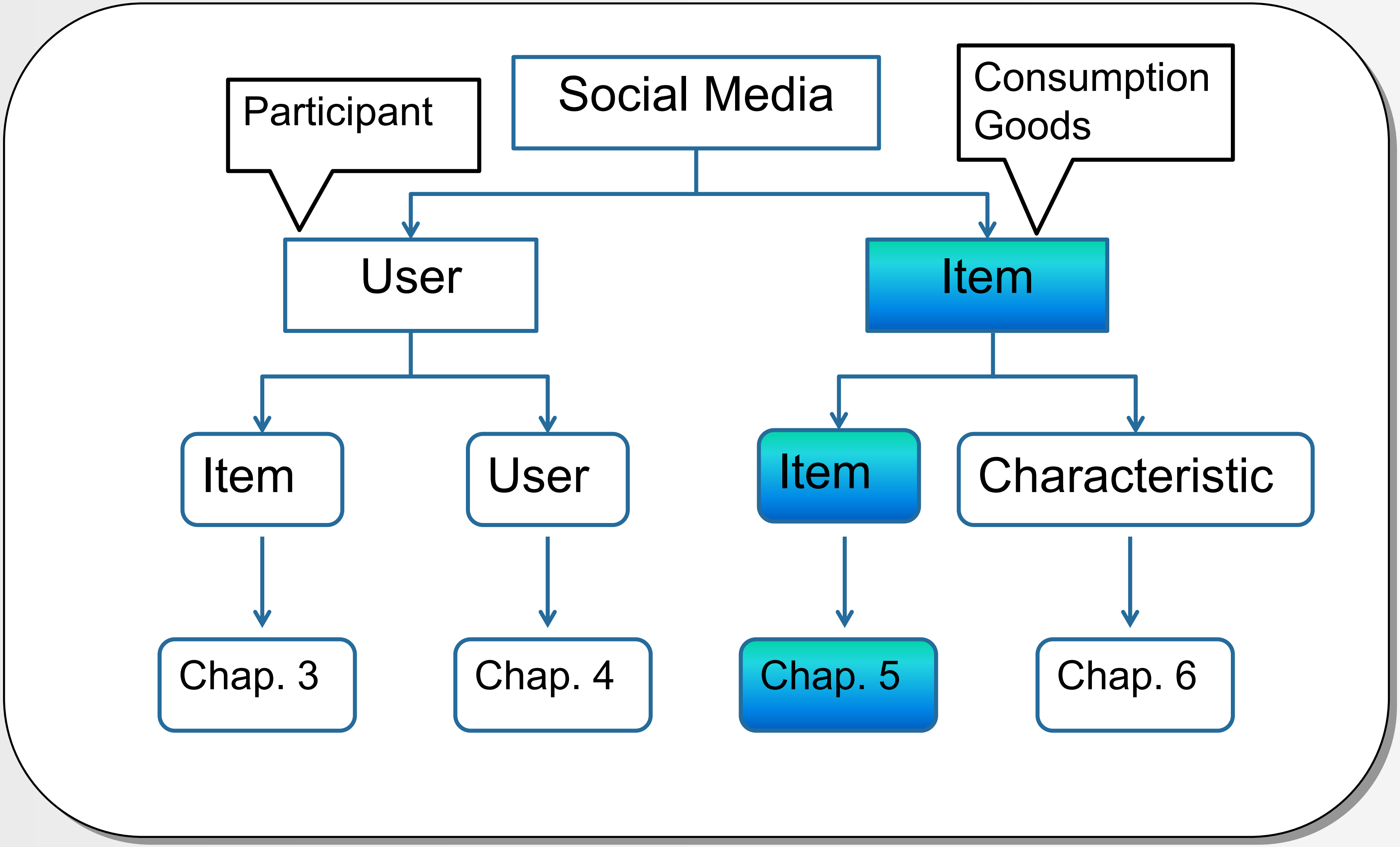
Metrics	Memory-Based Approaches		Model-Based Approaches				UserRec
	PCC	PCCW	SVD-PCC	SVD-PCCW	PMF-PCC	PMF-PCCW	
Precision@R	0.1495	0.3168	0.1540	0.2042	0.1875	0.2084	**0.3272**
MAP	0.1816	0.3444	0.1898	0.2469	0.2084	0.2440	**0.3752**
Bpref	0.1132	0.2395	0.1170	0.1479	0.1376	0.1707	**0.2913**
MMVRR	0.1129	0.1943	0.1151	0.1397	0.1300	0.1550	**0.2345**

Contribution of Chapter 4

- Propose the User Recommendation (UserRec) framework for user interest modeling and interest-based user recommendation
- Provide users with an automatic and effective way to discover other users with common interests in social tagging systems

- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Structure of Thesis



Problem and Motivation

- Social media systems with Q&A functionalities have accumulated large archives of questions and answers
  - Online Forums
  - Community-based Q&A services

Problem and Motivation

Query:
Q1: How is Orange Beach in Alabama?

Question Search:
Q2: Any ideas about Orange Beach in Alabama?

Question Suggestion:
Q3: Is the water pretty clear this time of year on Orange Beach?
Q4: Do they have chair and umbrella rentals on Orange Beach?

Topic: travel in orange beach

Results of Our Model

- Why can people only use the air phones when flying on commercial airlines, i.e. no cell phones etc.?
- Results of our model:

1. Why are you supposed to keep cell phone off during flight in commercial airlines? (Semantically equivalent)
2. Why don't cell phones from the ground at or near airports cause interference in the communications of aircraft? (Semantically related)
3. Cell phones and pagers really dangerous to avionics? (Semantically related)

Interference of aircraft

Problem and Motivation

- Benefits
  - Explore information needs from different aspects
    - “Travel”: beach, water, chair, umbrella
  - Increase page views
    - Enticing users’ clicks on suggested questions
  - Relevance feedback mechanism
    - Mining users’ click through logs on suggested questions

Challenge

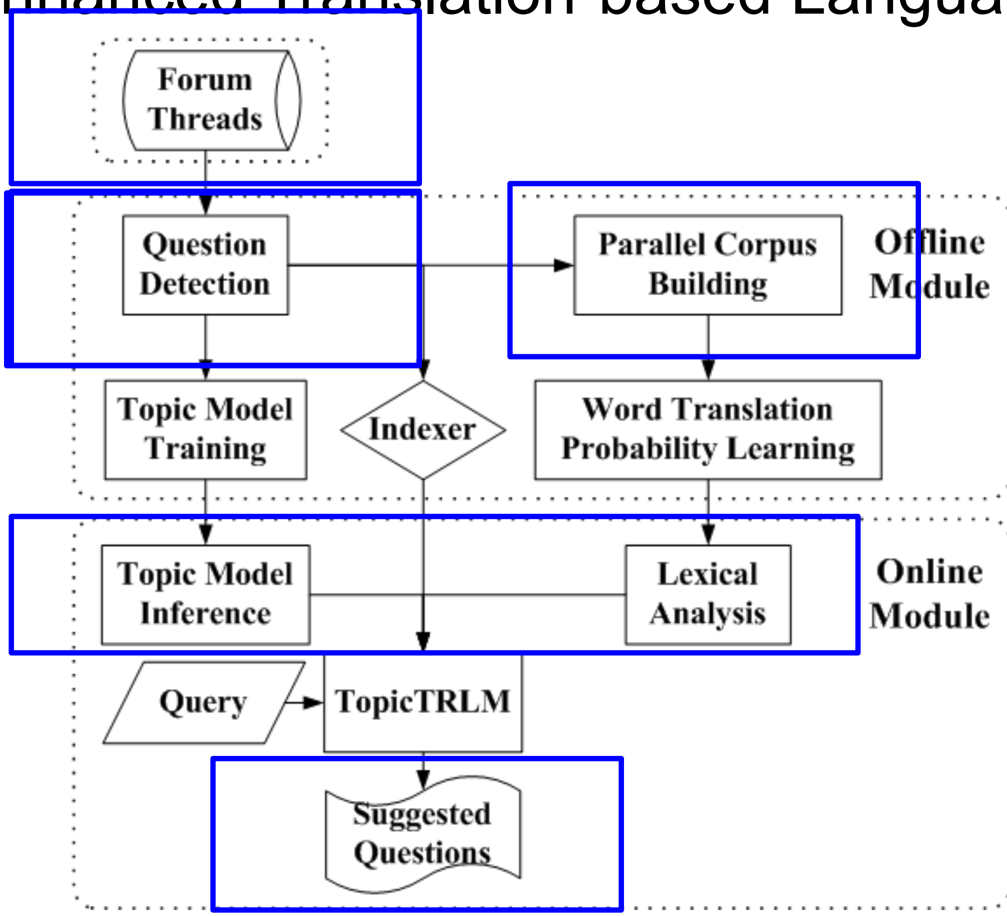
- Traditional bag-of-words approaches suffer from the shortcoming that they could not bridge the lexical chasm between semantically related questions

Document Representation

- Document representation
  - Bag-of-words
    - Independent
    - Fine-grained representation
    - Lexically similar
  - Topic model
    - Assign a set of latent topic distributions to each word
    - Capturing important relationships between words
    - Coarse-grained representation
    - Semantically related

TopicTRLM in Online Forum

- TopicTRLM
  - Topic-enhanced Translation-based Language Model



TopicTRLM in Online Forum

TRLM score: BoW

LDA score: topic model

$$P(q \mid D) = \prod_{w \in q} P(w \mid D)$$

$$P(w \mid D) = \boxed{\gamma P_{trlm}(w \mid D)} + \boxed{(1-\gamma) P_{lda}(w \mid D)}$$

- q: a query, D: a candidate question
- w: a word in query
- γ : parameter balance weights of BoW and topic model
- Jelinek-Mercer smoothing

TopicTRLM in Online Forum

- TRLM

$$P_{trlm}(w \mid D) = \frac{\mid D \mid}{\mid D \mid + \lambda} P_{mx}(w \mid D) + \frac{\lambda}{\mid D \mid + \lambda} P_{mle}(w \mid C)$$

$$P_{mx}(w \mid D) = \beta P_{mle}(w \mid D) + (1-\beta) \sum_{t \in D} \boxed{T(w \mid t)} P_{mle}(t \mid D)$$

  - C: question corpus, λ: Dirichlet smoothing parameter
  - T(w|t): word to word translation probabilities
- Use of LDA

$$P_{lda}(w \mid D) = \sum_{z=1}^{K} P(w \mid z) P(z \mid D)$$

- K: number of topics, z: a topic

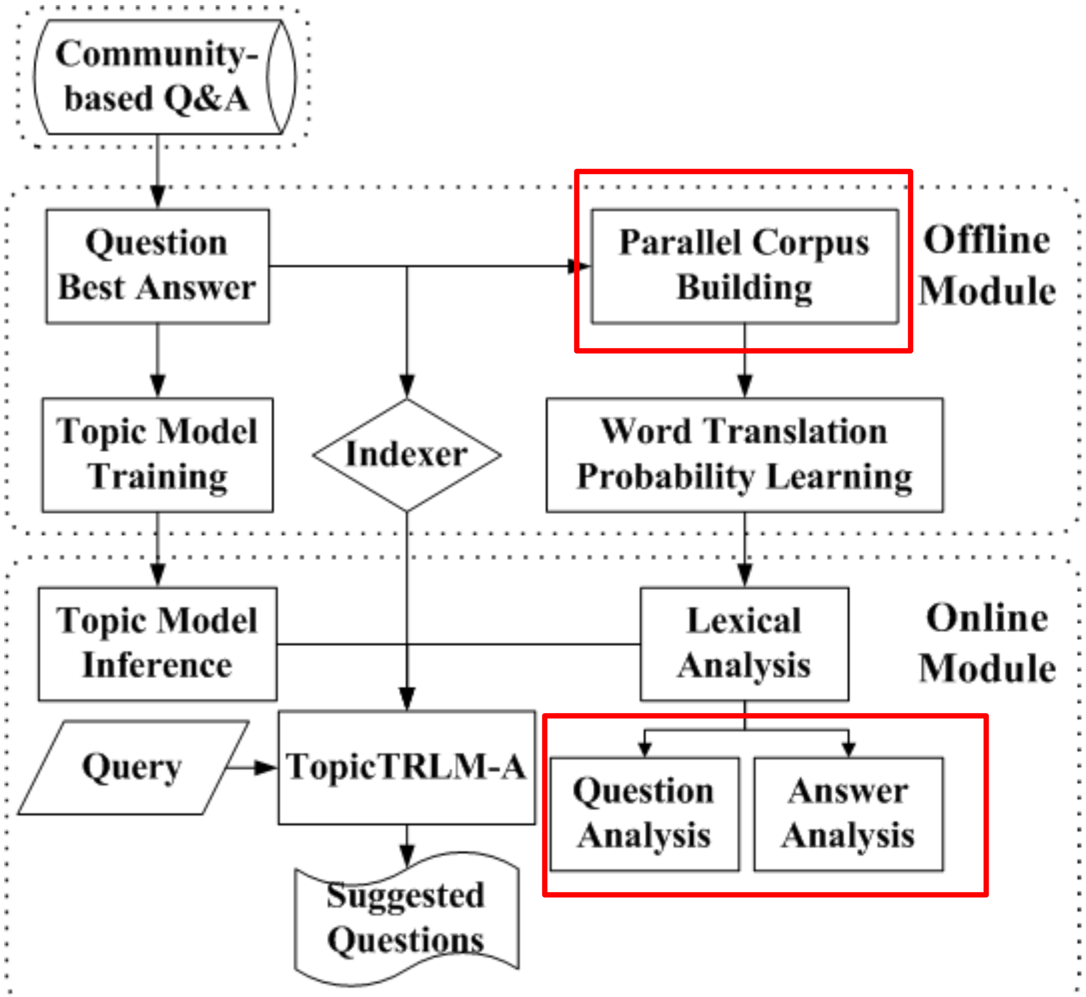
TopicTRLM in Online Forum

- Estimate T(w|t)
  - IBM model 1, monolingual parallel corpus
  - Questions are focus of forum discussions, questions posted by a thread starter (TS) during the discussion are very likely to explore different aspects of a topic
- Build parallel corpus
  - Extract questions posted by TS, question pool Q
  - Question-question pairs, enumerating combinations in Q
  - Aggregating all q-q pairs from each forum thread

TopicTRLM-A in Community-based Q&A

- Best answer for each resolved question in community-based Q&A services is always readily available
- Best answer of a question could also explain the semantic meaning of the question
- Propose TopicTRLM-A to incorporate answer information

TopicTRLM-A in Community-based Q&A



Experiments in Online Forum

- Data set
  - Crawled from TripAdvisor
  - TST_LABEL: labeled data for 268 questions
  - TST_UNLABEL: 10,000 threads at least 2 questions posted by thread starters
  - TRAIN_SET: 1,976,522 questions, 971,859 threads
    - Parallel corpus to learn T(w|t)
    - LDA training data
    - Question repository

Experiments in Online Forum

- Performance comparison (a larger value in metric means better performance)

Metrics	LDA	QL	TR	TRLM	TopicTRLM
$P@R$	0.2411	0.3370	0.4135	0.4555	**0.5140**
MAP	0.3684	0.4089	0.4629	0.5029	**0.5885**
MRR	0.5103	0.5277	0.5311	0.5317	**0.5710**

- LDA performs the worst, coarse-grained
- TRLM > TR > QL
- TopicTRLM outperforms other approaches

Experiments in Community-based Q&A

- Date Set
  - Yahoo! Answers
  - “travel” category
  - “computers & internet” category

Experiments in Community-based Q&A

Performance of different models on category “computers & internet” (a larger metric value means a better performance)

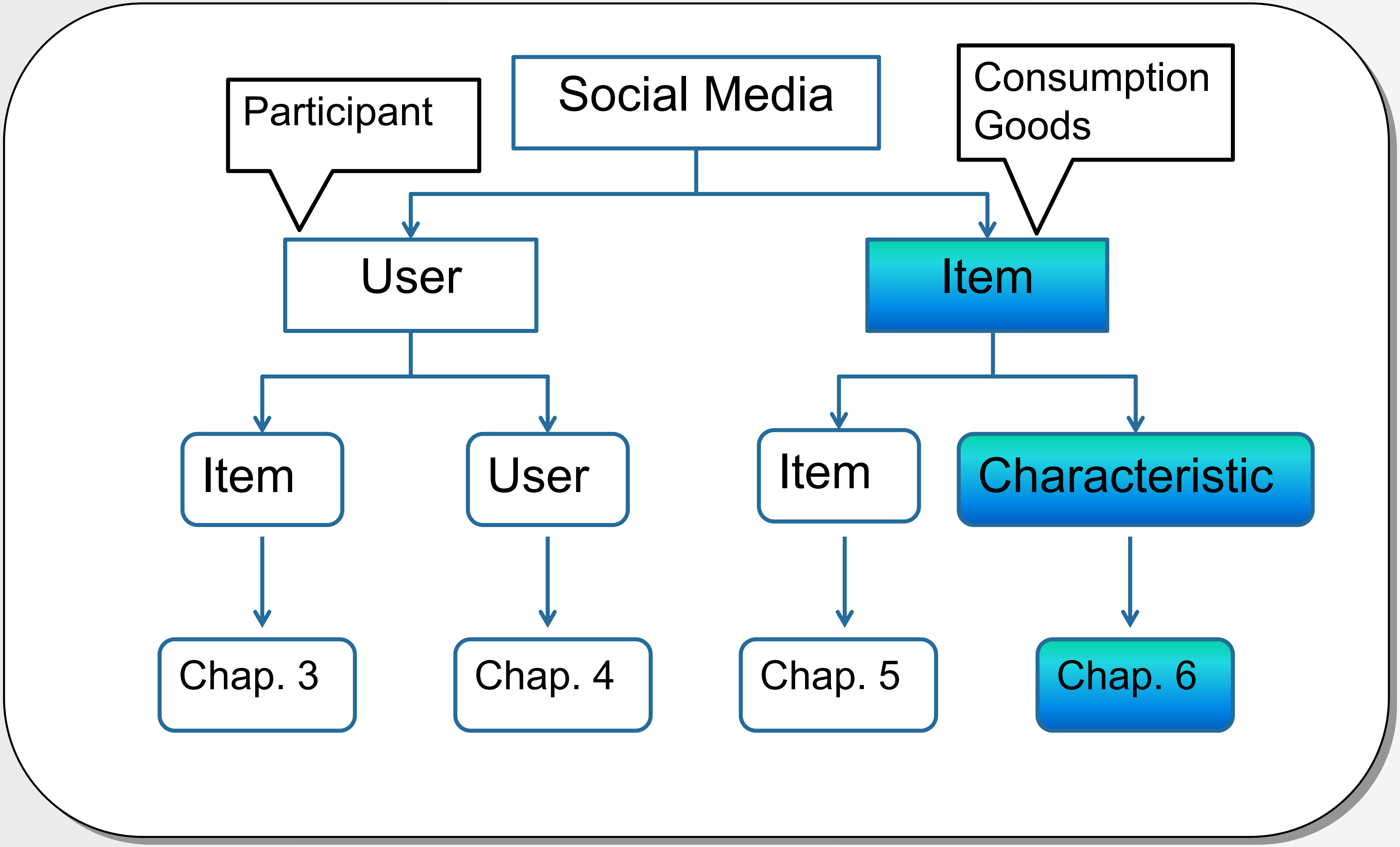
Methods	MAP	Bpref	MRR	P@R
LDA	0.2397	0.136	0.2767	0.1594
QL	0.346	0.2261	0.416	0.2594
TRLM	0.3532	0.2368	0.4271	0.2777
TopicTRLM	0.4235	0.2755	0.5559	0.3197
TopicTRLM-A	**0.6228**	**0.4673**	**0.7745**	**0.5467**

Contribution of Chapter 5

- Propose question suggestion, which targets at suggesting questions that are semantically related to a queried question
- Propose the TopicTRLM which fuses both the lexical and latent semantic knowledge in online forums
- Propose the TopicTRLM-A to incorporate answer information in community-based Q&A

- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Structure of Thesis



Challenge of Question Analysis

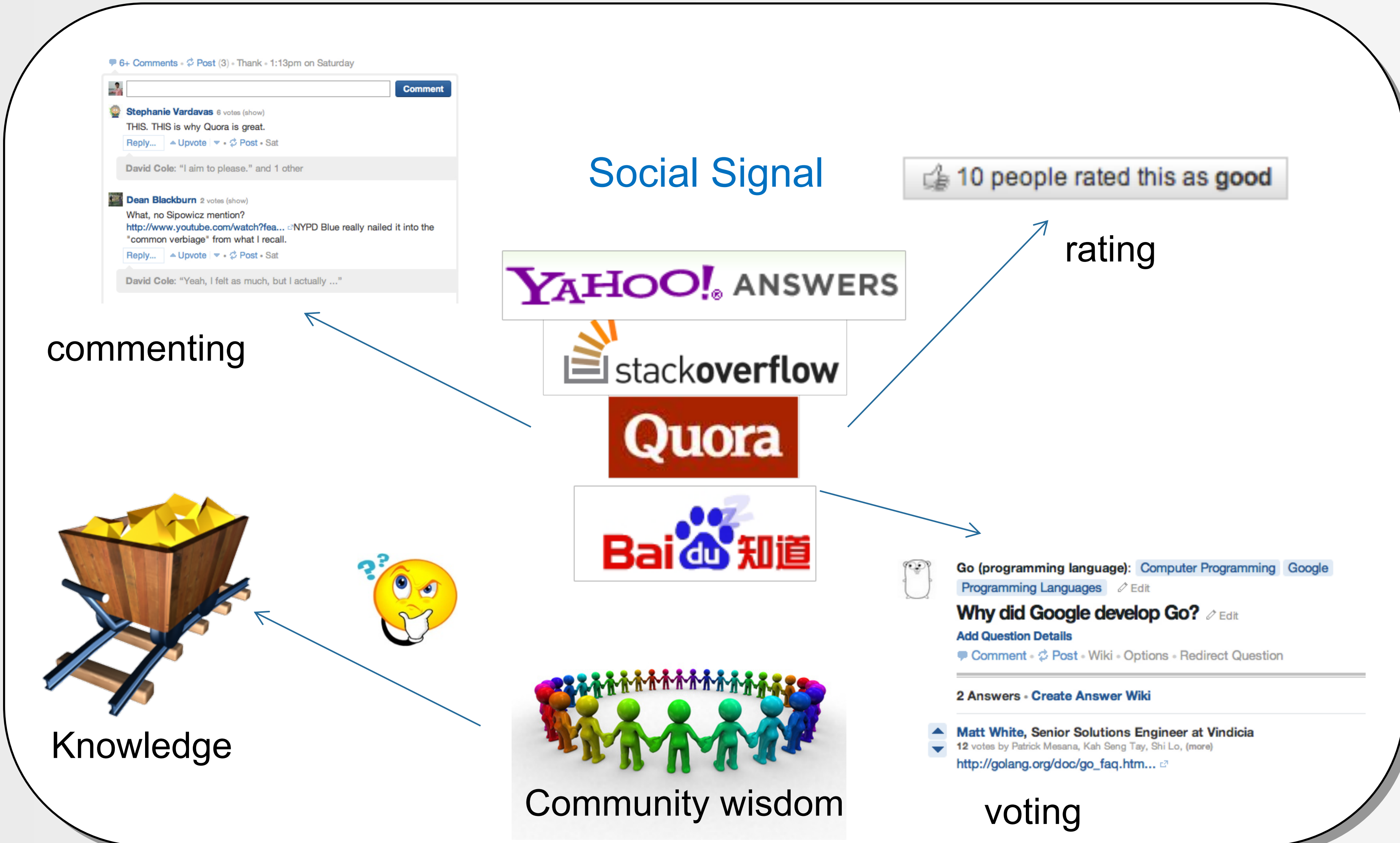
- Questions are ill-phrased, vague and complex
  - Light-weight features are needed
- Lack of labeled data

Problem and Motivation

- "Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn't available."

-- Alon Halevy,
Peter Norvig and
Fernando Pereira

Problem and Motivation



Problem and Motivation

- Whether we can utilize social signals to collect training data for question analysis with NO manual labeling
- Question Subjectivity Identification (QSI)
- Subjective Question
  - One or more subjective answers
  - What was your favorite novel that you read?
- Objective Question
  - Authoritative answer, common knowledge or universal truth
  - What makes the color blue?

Social Signal

- Like: like an answer if they find the answer useful
- Subjective
  - Answers are opinions, different tastes
  - Best answer receives similar number of likes with other answers
- Objective
  - Like an answer which explains universal truth in the most detail
  - Best answer receives higher likes than other answers

Social Signal

- Vote: users could vote for best answer
- Subjective
  - Vote for different answers, support different opinions
  - Low percentage of votes on best answer
- Objective
  - Easy to identify answer contains the most fact
  - Percentage of votes of best answer is high

Social Signal

- Source: references to authoritative resources
  - Only available for objective question that has fact answer
- Poll and Survey
  - User intent is to seek opinions
  - Very likely to be subjective

Social Signal

- Answer Number: the number of posted answers to each question varies
- Subjective
  - Post opinions even they notice there are other answers
- Objective
  - May not post answers to questions that have received other answers since an expected answer is usually fixed
- A large answer number indicates subjectivity
- HOWEVER, a small answer number may be due to many reasons, such as objectivity, small page views

Feature

- Word
- Word n-gram
- Question Length
- Request Word
- Subjectivity Clue
- Punctuation Density
- Grammatical Modifier
- Entity

Experiments

- Dataset
  - Yahoo! Answers, 4,375,429 questions with associated social signals
  - Ground truth: adapted from Li, Liu and Agichtein 2008

Experiments

Method	Precision
Supervised	0.6596
CoCQA	0.6861 (+4.20%)
L + V + PS + AN + S	0.6626 (+0.45%)
L	0.5714 (−13.37%)
V + PS + AN + S	0.6981 (+5.84%)
PS + AN + S	0.6915 (+4.84%)
V + PS + AN	**0.7214 (+9.37%)**
V + AN	0.7201 (+9.17%)
AN + S	0.7038 (+6.70%)

CoCQA utilizes some amount of unlabeled data, but it could only utilize a small amount (3, 000 questions)

Effectiveness of collecting training data using well-designed social signals

These social signals could be found in almost all CQA

Experiments

Method/Feature	Word	Word n-gram
Supervised	0.6380	0.6596 (+3.39%)
CoCQA	0.6432	0.6861 (+6.66%)
V + PS + AN	0.6707	0.7214 (+7.56%)
V + AN	0.6265	0.7201 (+14.94%)
AN + S	0.6157	0.7038 (+14.31%)

Better performance using word n-gram compared with word

Social signals achieve on average 12.27% relative gain

Experiments

Precision	ngram	ngram + qlength	ngram + rword	ngram + sclue
	0.6596	0.6896	0.6834	0.6799
ngram + pdensity	**ngram + gmodifier**	**ngram + entity**	**heuristic features**	**ngram + heuristic**
0.7000	0.6950	0.6801	0.6995	**0.7337**(+**11.23**%)

Adding any heuristic feature to word n-gram improve precision

Combining heuristic feature and word n-gram achieves 11.23% relative gain over n-gram

Contribution of Chapter 6

- Propose an approach to collect training data automatically by utilizing social signals in community-based Q&A sites without involving any manual labeling
- Propose several light-weight features for question subjectivity identification

- Introduction
- Background
- Item Recommendation with Tagging Ensemble
- User Recommendation via Interest Modeling
- Item Suggestion with Semantic Analysis
- Item Modeling via Data-Driven Approach
- Conclusion and Future Work

Conclusion

- Modeling users’ interests with respect to their behavior, and recommending items or users they may be interested in
  - TagRec
  - UserRec
- Understanding items’ characteristics, and grouping items that are semantically related for better addressing users’ information needs
  - Question Suggestion
  - Question Subjectivity Identification

Future Work

- TagRec
  - Mine explicit relations to infer some implicit relations
- UserRec
  - Develop a framework to handle the tag ambiguity problem
- Question Suggestion
  - Diversity the suggested questions
- Question Subjectivity Identification
  - Sophisticated features: semantic analysis

Publications: Conferences (7)

1. **Tom Chao Zhou**, Xiance Si, Edward Y. Chang, Irwin King and Michael R. Lyu. A Data-Driven Approach to Question Subjectivity Identification in Community Question Answering. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12), pp 164-170, Toronto, Ontario, Canada, July 22 - 26, 2012.
2. **Tom Chao Zhou**, Michael R. Lyu and Irwin King. A Classification-based Approach to Question Routing in Community Question Answering. In Proceedings of the 21st International Conference Companion on World Wide Web, pp 783-790, Lyon, France, April 16 - 20, 2012.
3. **Tom Chao Zhou**, Chin-Yew Lin, Irwin King, Michael R. Lyu, Young-In Song and Yunbo Cao. Learning to Suggest Questions in Online Forums. In Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-11), pp 1298-1303, San Francisco, California, USA, August 7 - 11, 2011.
4. Zibin Zheng, **Tom Chao Zhou**, Michael R. Lyu, and Irwin King. FTCloud: A Ranking-based Framework for Fault Tolerant Cloud Applications. In Proceedings of the 21st IEEE International Symposium on Software Reliability Engineering (ISSRE 2010), pp 398-407, San Jose CA, USA, November 1- 4, 2010.

Publications: Conferences (7)

5. **Tom Chao Zhou**, Hao Ma, Michael R. Lyu, Irwin King. UserRec: A User Recommendation Framework in Social Tagging Systems. In Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI-10), pp 1486-1491, Atlanta, Georgia, USA, July 11 - 15, 2010.
6. **Tom Chao Zhou**, Irwin King. Automobile, Car and BMW: Horizontal and Hierarchical Approach in Social Tagging Systems. In Proceedings of the 2ndWorkshop on SocialWeb Search and Mining (SWSM 2009), in conjunction with CIKM 2009, pp 25-32, Hong Kong, November 2 - 6, 2009.
7. **Tom Chao Zhou**, Hao Ma, Irwin King, Michael R. Lyu. TagRec: Leveraging Tagging Wisdom for Recommendation. In Proceedings of the 15th IEEE International Conference on Computational Science and Engineering (CSE-09), pp 194199, Vancouver, Canada, 29-31 August, 2009.

Publications: Journals (2), Under Review (1)

- Journals
  1. Zibin Zheng, **Tom Chao Zhou**, Michael R. Lyu, and Irwin King. Component Ranking for Fault-Tolerant Cloud Applications, IEEE Transactions on Service Computing (TSC), 2011.
  2. Hao Ma, **Tom Chao Zhou**, Michael R. Lyu and Irwin King. Improving Recommender Systems by Incorporating Social Contextual Information, ACM Transactions on Information Systems (TOIS), Volume 29, Issue 2, 2011.
- Under Review
  1. **Tom Chao Zhou**, Michael R. Lyu and Irwin King. Learning to Suggest Questions in Social Media. Submitted to Journal of the American Society for Information Science and Technology (JASIST).

- Thanks!
- Q & A

FAQ

- [FAQ: Chapter 3]
- [FAQ: Chapter 4]
- [FAQ: Chapter 5]
- [FAQ: Chapter 6]

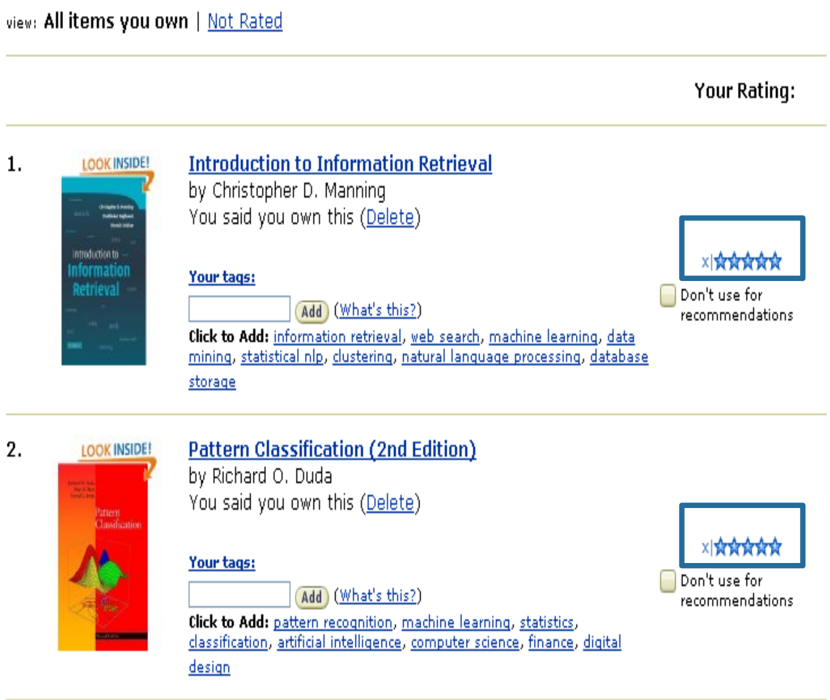
FAQ: Chapter 3

- An example of a recommender system
- MAE and RMSE equations
- Parameter sensitivity
- Tag or social network
- Intuition of maximize the log function of the posterior distribution in Eq. 3.10 of thesis

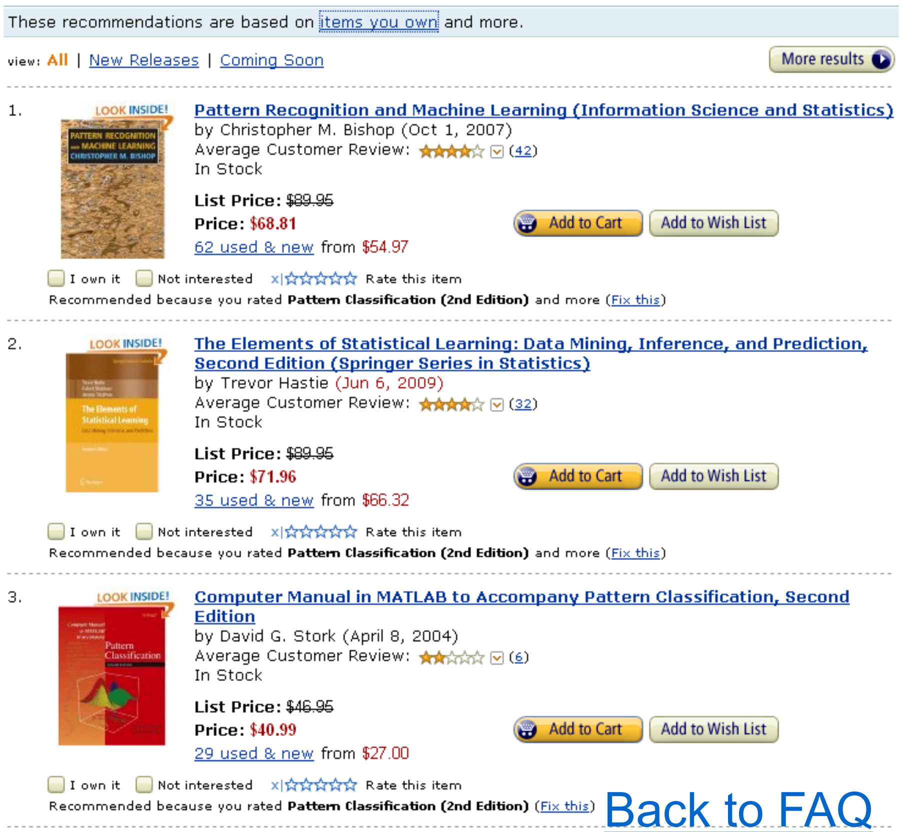
Back to FAQ

An Example of A Recommender System

Have some personal preferences.



Get some recommendations.



Back to FAQ

MAE and RMSE

- Mean absolute error (MAE)

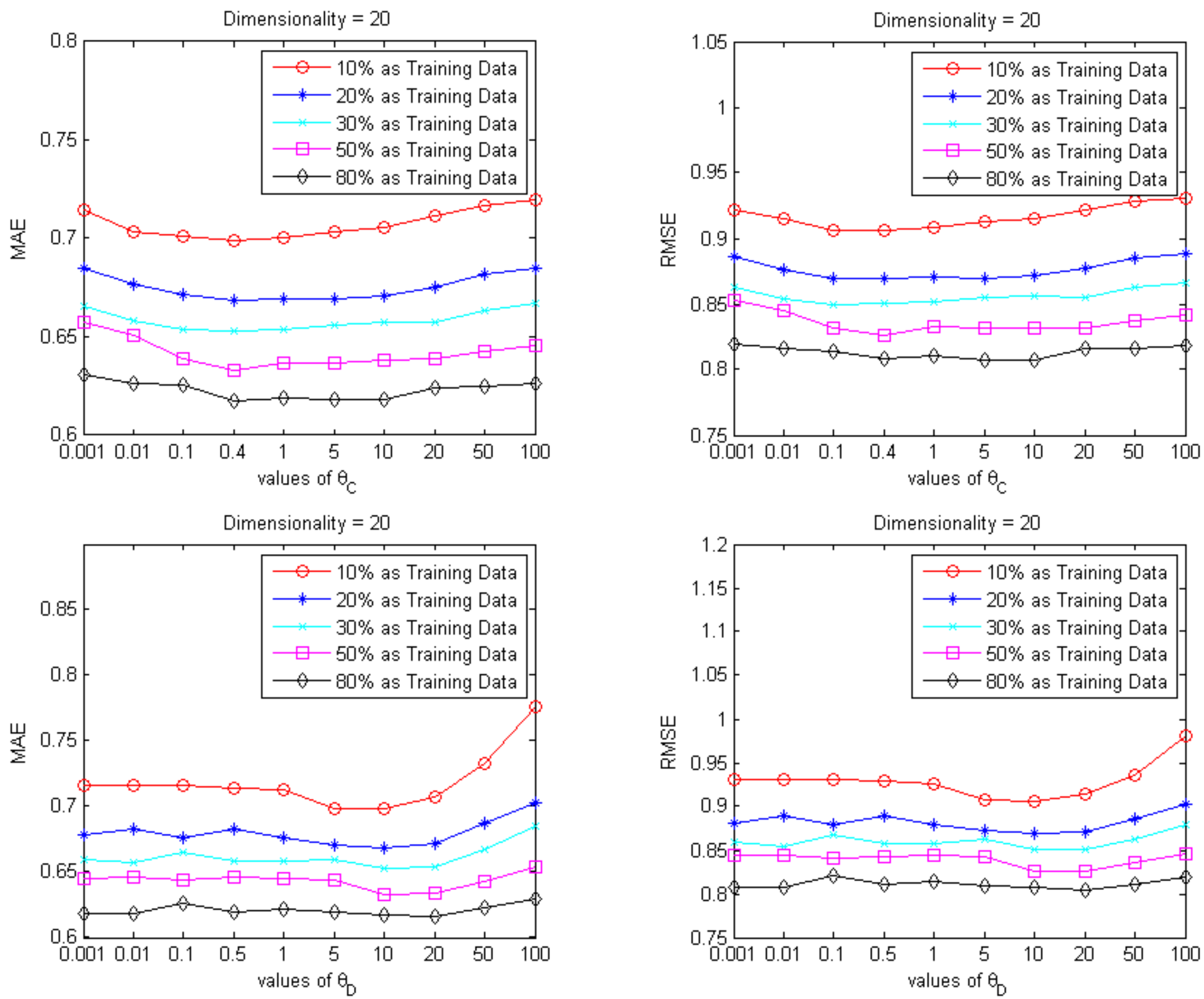
$$MAE = \frac{\sum_{ij} |r_{ij} - \hat{r}_{ij}|}{N}$$

- Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i,j} (r_{i,j} - \hat{r}_{i,j})^2}{N}}$$

Back to FAQ

Parameter Sensitivity



Back to FAQ

Tag or Social Network?

- What is the difference of incorporating tag information and social network information?
- Answer: both tagging and social networking could be considered as user behavior besides rating. They explain users' preferences from different angles. The proposed TagRec framework could not only incorporate tag information, but also could utilize social network information in a similar framework.

Back to FAQ

Intuition of maximize the log function of the posterior distribution in Eq. 3.10 of thesis

- The idea of maximize the log function of the posterior distributions is equivalent to maximize the posterior distributions directly, because the logarithm is a continuous strictly increasing function over the range of the likelihood. The reason why I would like to maximize the posterior distributions is that after Bayesian inference, I need to calculate the conditional distributions to get the posterior distributions, e.g.: p(R|U,V), R is the observed ratings, and U, V are parameters. To estimate the U, V, I use the maximum likelihood estimation to estimate the parameter space, thus I need to maximize the conditional distributions P(R|U,V). So this is the reason why I have to maximize the log function in my approach

Back to FAQ

FAQ: Chapter 4

- What is modularity?
- Comparison on Precision@N
- Comparison on Top-K accuracy
- Comparison on Top-K recall
- Distribution of number of users in network
- Distribution of number of fans of a user
- Relationship between # fans and # bookmarks
- Why we use the graph mining algorithm instead of some simple algorithms, e.g. frequent mining

Back to FAQ

What is Modularity?

- The concept of modularity of a network is widely recognized as a good measure for the strength of the community structure

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

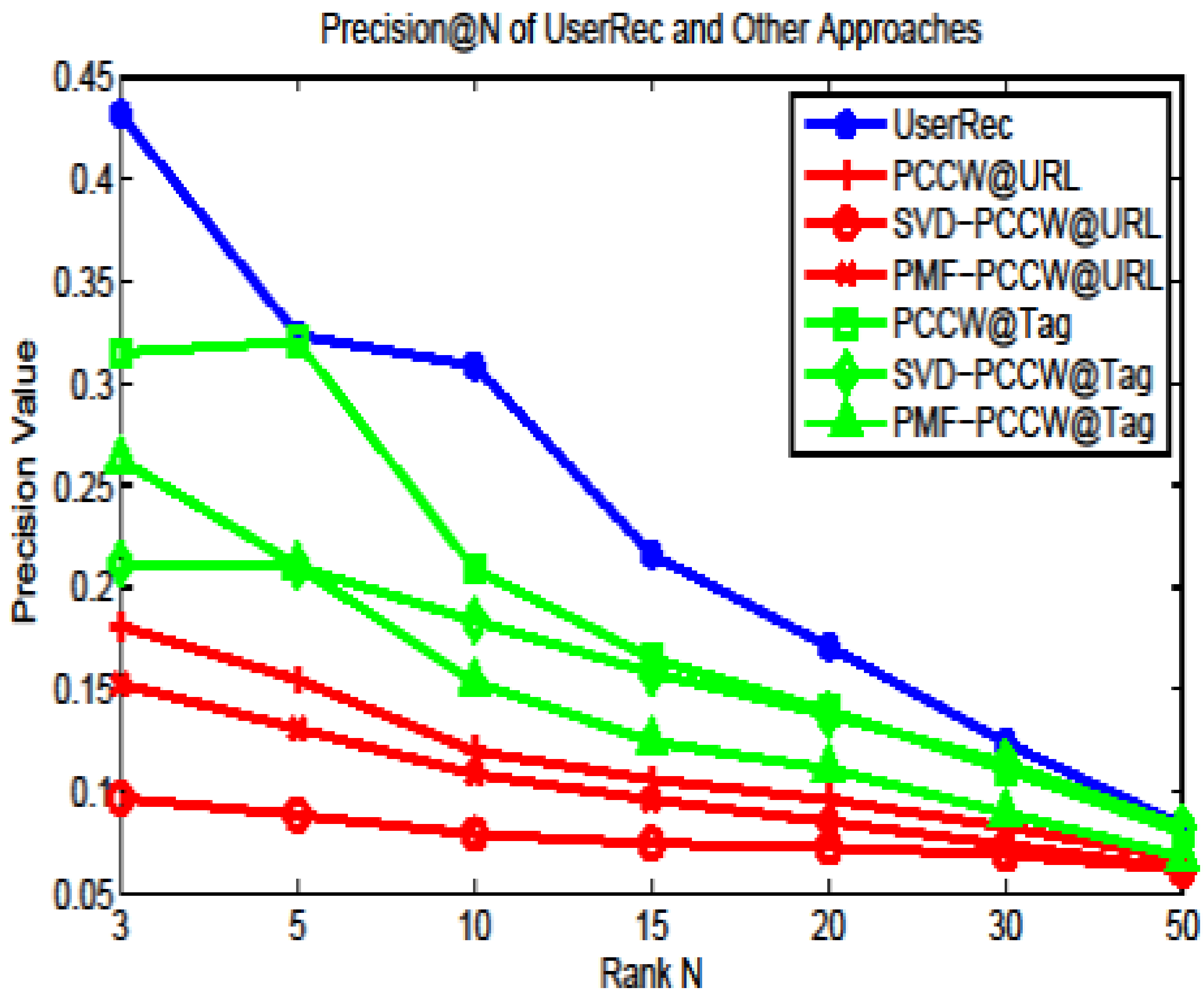
$$k_i = \sum_k A_{ik} \quad m = \tfrac{1}{2} \textstyle\sum_{ij} A_{ij}$$

A_{ij} is the weight between node i and node j

$\delta(c_i, c_j)$ is 1 if node i and node j belong to the same community

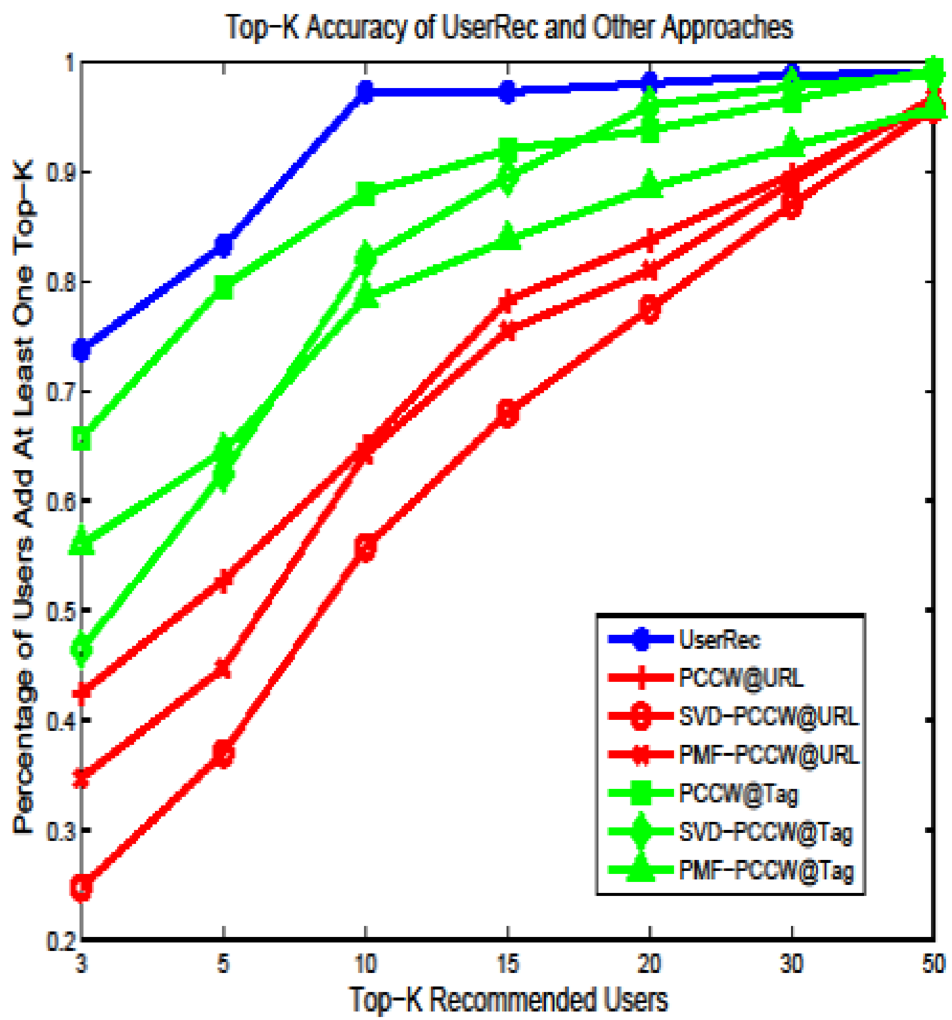
Back to FAQ

Comparison on Precision@N



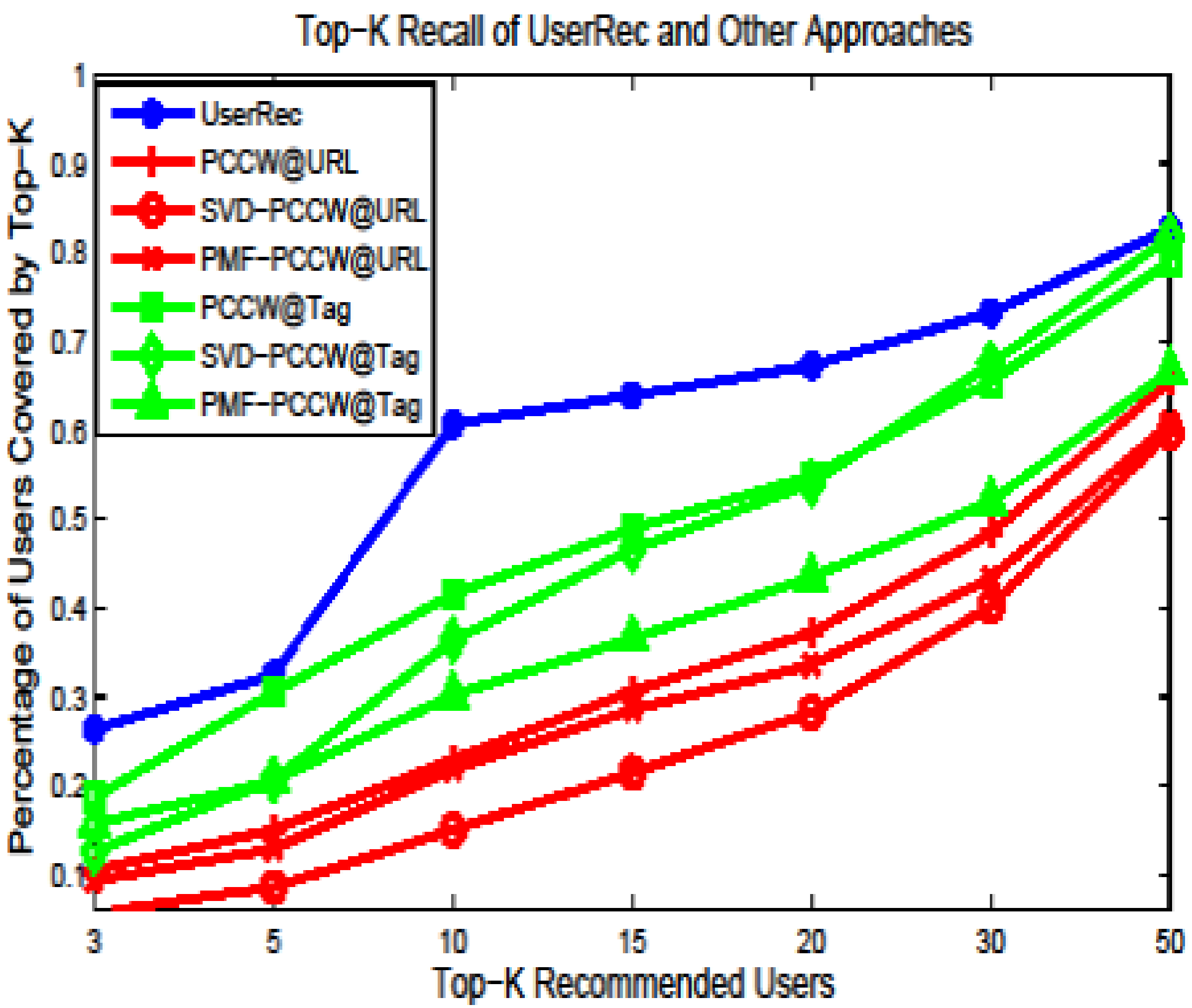
[Back to FAQ]

Comparison on Top-K Accuracy



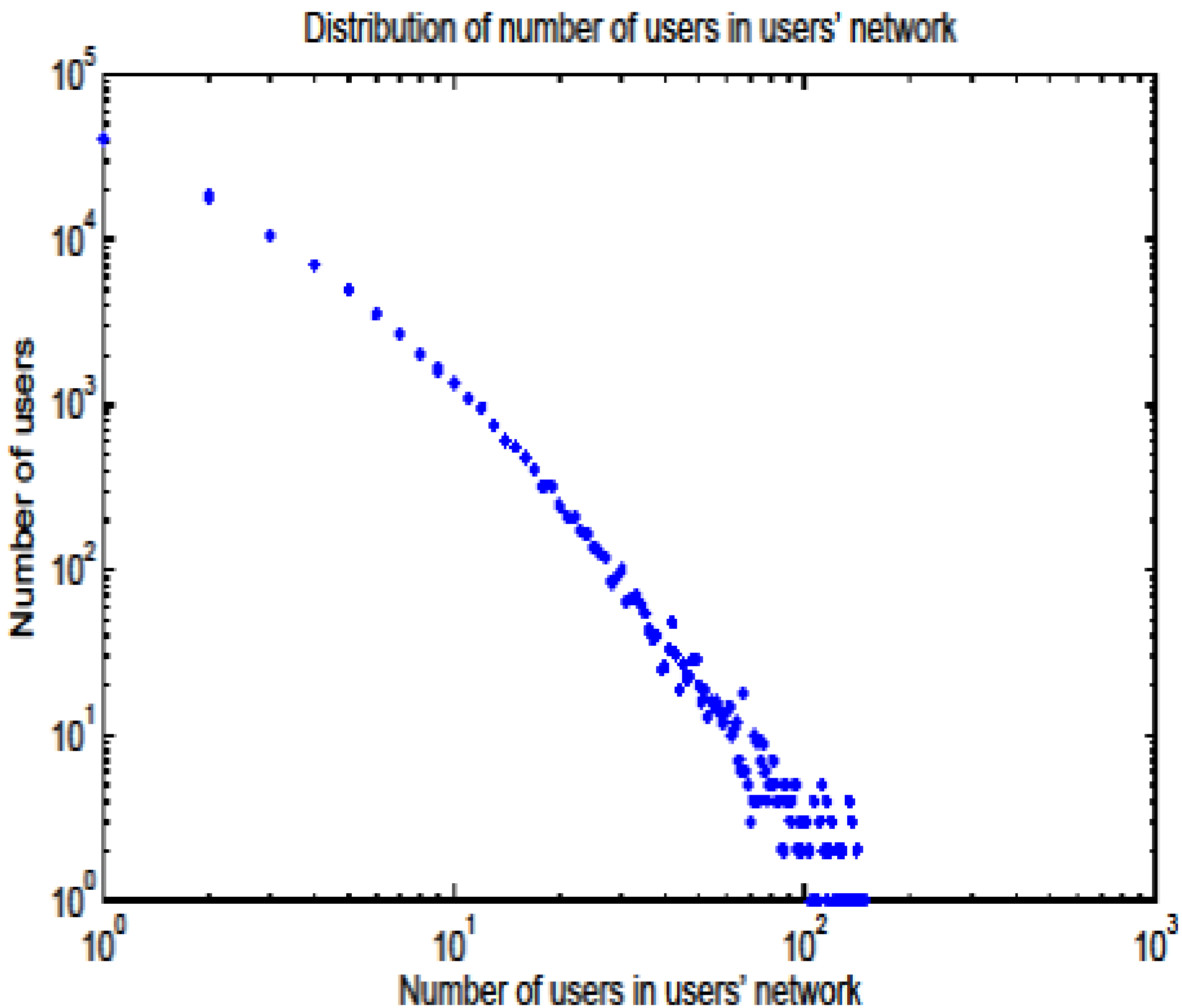
[Back to FAQ]

Comparison on Top-K Recall



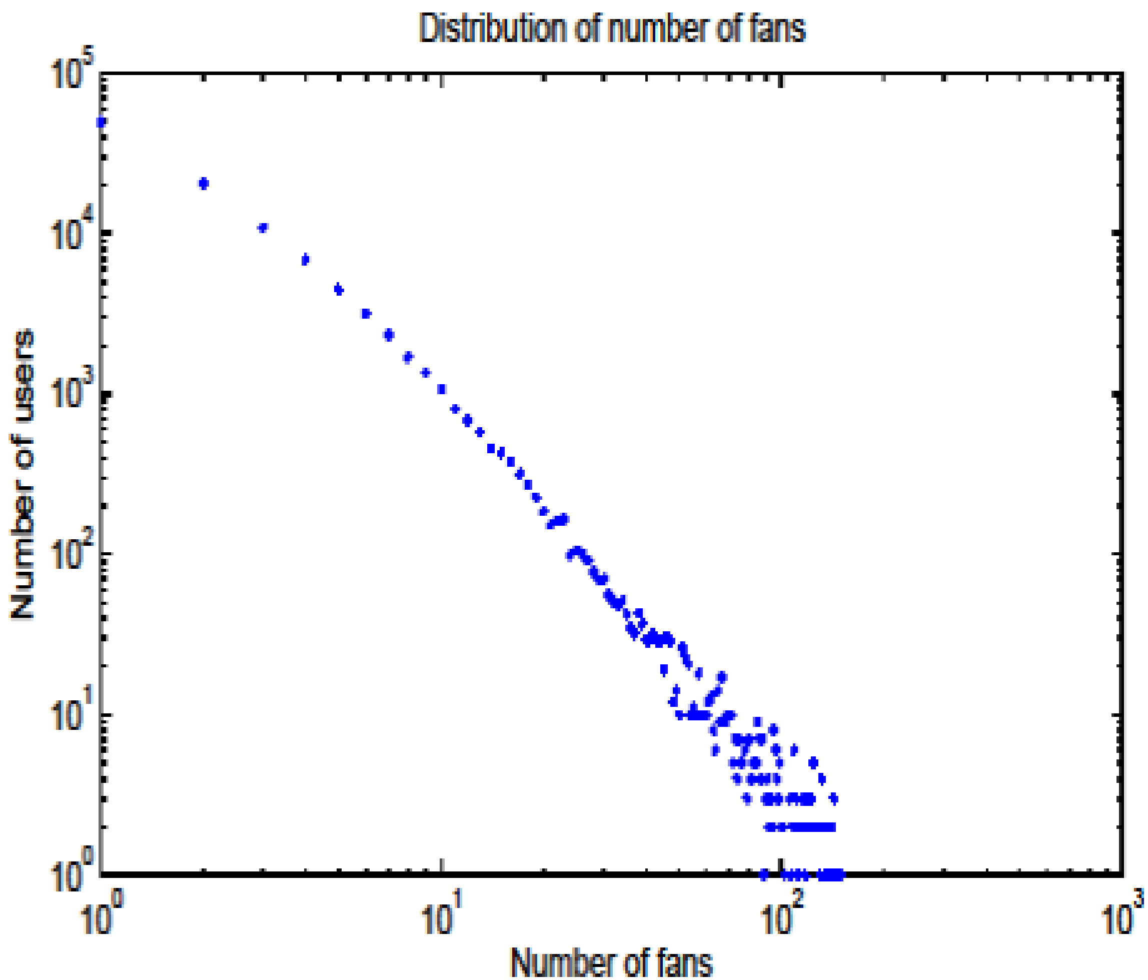
Back to FAQ

Distribution of Number of Users in Network



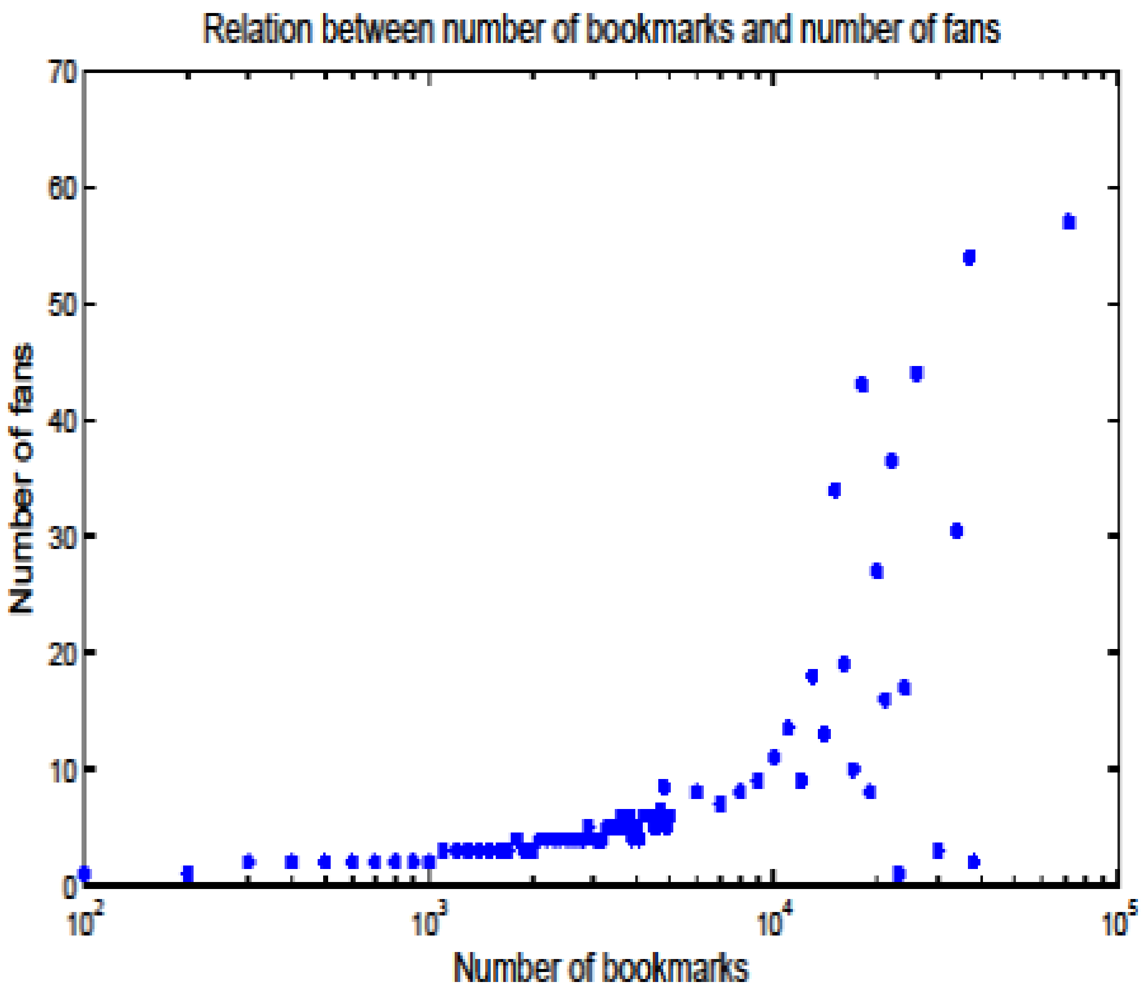
Back to FAQ

Distribution of Number of Fans of A User



Back to FAQ

Relationship Between # Fans, # bookmarks



[Back to FAQ]

Why we use the graph mining algorithm instead of some simple algorithms, e.g. frequent itemset mining

- We use community discovery algorithm on each tag-graph, and could accurately capture users' interests on different topics. The algorithm is efficient, and the complexity is $O(n\log^2 n)$. While frequent itemset mining is suitable for mining small itemset, e.g., 1, 2, 3 items in each set. However, each topic could contain many tags.

Back to FAQ

FAQ: Chapter 5

- Experiments on word translation
- Dirichlet smoothing
- Build monolingual parallel corpus in community-based Q&A
- An example from Yahoo! Answers
- Formulations of TopicTRLM-A
- Data Analysis in online forums
- Performance on Yahoo! Answers “travel”

Back to FAQ

Experiments on Word Translation

- Word translation

Words	shore		park		condo		beach	
Rank	IBM 1	LDA	IBM 1	LDA	IBM 1	LDA	IBM 1	LDA
1	shore	shore	park	park	condo	condo	beach	beach
2	beach	groceri	drive	hotel	beach	south	resort	slope
3	snorkel	thrift	car	stai	area	north	what	jet
4	island	supermarket	how	time	unit	shore	hotel	snowboard
5	kauai	store	area	area	island	pacif	water	beaver
6	condo	nappi	where	recommend	maui	windward	walk	huski
7	area	tesco	walk	beach	rent	seaport	area	steamboat
8	water	soriana	time	nation	owner	alabama	room	jetski
9	boat	drugstor	ride	tour	shore	opposit	snorkel	powder
10	ocean	mega	hotel	central	rental	manor	restaur	hotel

- IBM 1: semantic relationships of words from semantically related questions
- LDA: co-occurrence relations in a question

Back to FAQ

Dirichlet Smoothing

- Bayesian smoothing using Dirichlet priors
  - A language model is a multinomial distribution, for which the conjugate prior for Bayesian analysis is the Dirichlet distribution
  - Choose the parameters of the Dirichlet to be

$$(\mu p(w_1 \mid C),\ \mu p(w_2 \mid C), \ldots, \mu p(w_n \mid C))$$

  - Then the model is given by

$$p_\mu(w \mid d) = \frac{c(w; d) + \mu p(w \mid C)}{\sum_{w' \in V} c(w'; d) + \mu}$$

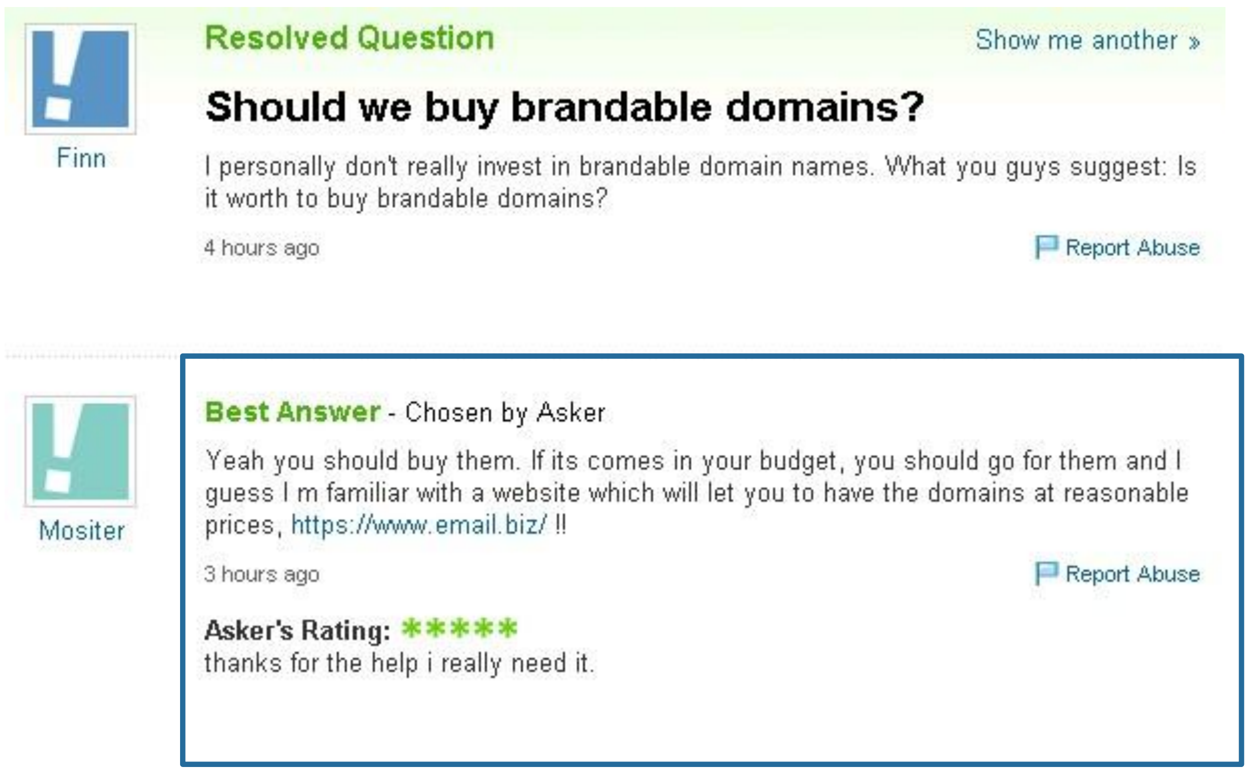
Back to FAQ

Build Monolingual Parallel Corpus in Community-based Q&A

- Aggregate question title and question detail as a monolingual parallel corpus

Back to FAQ

An Example from Yahoo! Answers



Best answer available

Back to FAQ

TopicTRLM-A in Community-based Q&A

$$P(q|(Q,A)) = \prod_{w \in q} P(w|(Q,A)),$$

$$P(w|(Q,A)) = \boxed{\epsilon P_{trlm}(w|(Q,A))} + \boxed{(1-\epsilon)P_{lda}(w|Q)}$$

Lexical score   Latent semantic score

[Back to FAQ]

TopicTRLM-A in Community-based Q&A

$$P_{trlm}(w|(Q,A)) = \frac{|(Q,A)|}{|(Q,A)|+\lambda} P_{mx}(w|(Q,A)) + \frac{\lambda}{|(Q,A)|+\lambda} P_{mle}(w|C),$$

Dirichlet smoothing

$$P_{mx}(w|(Q,A)) = \boxed{\eta P_{mle}(w|Q)} + \boxed{\theta \sum_{t \in Q} T(w|t) P_{mle}(t|Q)} + \boxed{\mu P_{mle}(w|A)}$$

Question LM score

Question translation model score

Answer ensemble

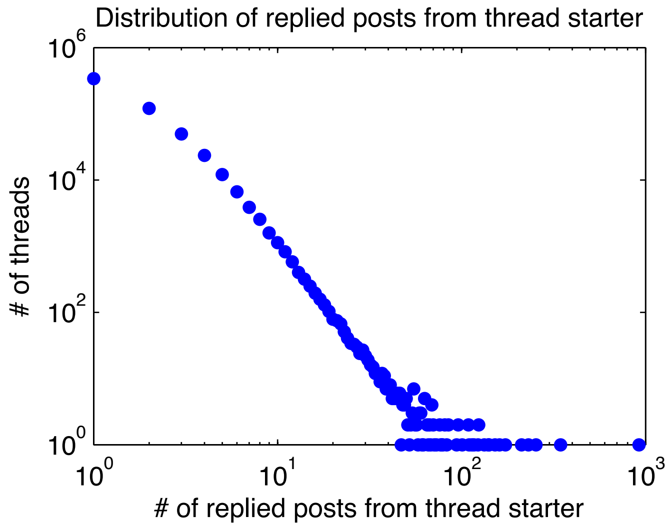
Back to FAQ

Data Analysis in Online Forums

- Data Analysis
- Post level

# Threads	# Threads that have replied posts from TS	Average # replied posts from TS
1,412,141	566,256	1.9

- Forum discussions are quite interactive
- Power law



Back to FAQ

Performance on Yahoo! Answers “travel”

Performance of different models on category “travel”
(a larger metric value means a better performance)

Methods	MAP	Bpref	MRR	P@R
LDA	0.1345	0.0612	0.1616	0.0675
QL	0.316	0.1902	0.388	0.2048
TRLM	0.3222	0.2034	0.3923	0.2234
TopicTRLM	0.3615	0.244	0.4406	0.2644
TopicTRLM-A	**0.467**	**0.3167**	**0.5963**	**0.387**

Back to FAQ

FAQ: Chapter 6

- Examples of subjective, objective questions
- Benefits of performing question subjectivity identification
- How to define subjective and object questions

Back to FAQ

Examples of Subjective,Objective Questions

- Question subjectivity identification
- Subjective
  - What was your favorite novel that you read?
  - What are the ways to calm myself when flying?
- Objective
  - When and how did Tom Thompson die? He is one of the group of Seven.
  - What makes the color blue?

Back to FAQ

Benefits of Performing QSI

- More accurately identify similar questions
- Better rank or filter the answers
- Crucial component of inferring user intent
- Subjective question --> Route to users
- Objective question --> Trigger AFQA

Back to FAQ

How to define subjective and object questions

- Ground truth data was created using Amazon’s Mechanical Turk service. Each question was judged by 5 qualified Mechanical Turk workers. Subjectivity was decided using majority voting
- Linguistic people are good at manual labeling
- Compute science people should focus on how to use existing data to identify subjective/objective questions, such as social signals, answers, etc. Not focus on manual labeling

Back to FAQ