CQARank:Jointly Model Topics and Expertise in Community Question Answering

Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang,
Huiping Sun, Zhong Chen

Peking University

Singapore Management University

Community Question Answering

- Open platforms for sharing expertise
- Large repositories of valuable knowledge

北京大学 PEKING UNIVERSITY

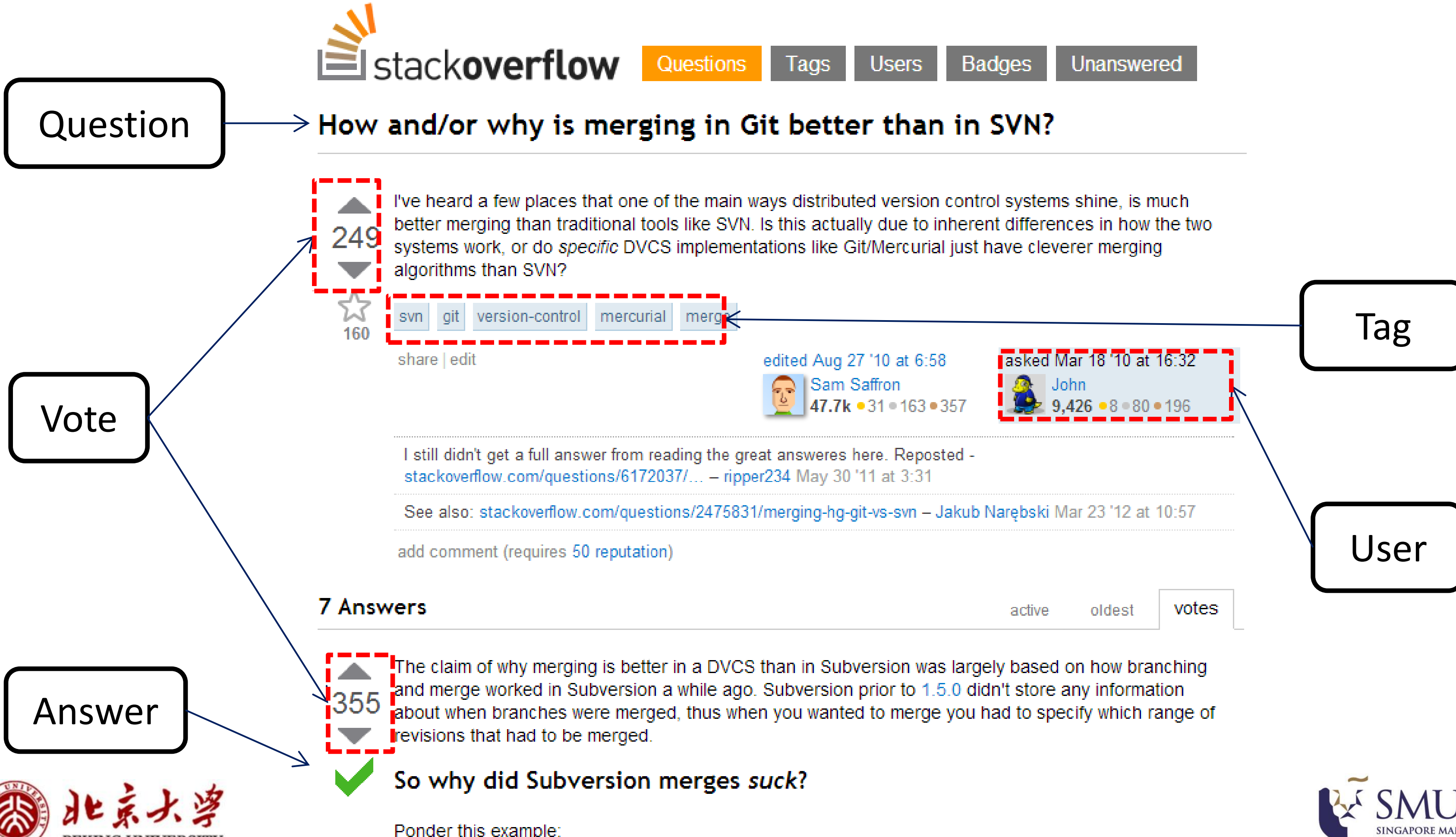
SMU SINGAPORE MANAGEMENT UNIVERSITY

Existing CQA Mechanism Challenges

- Poor expertise matching
- Low-quality answers
- Under-utilized archived questions
- Fundamental question: how to model *topics* and *expertise* in CQA sites

Motivation

- A case study of Stack Overflow



北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Motivation

- Propose a principle approach to jointly model topics and expertise in CQA
  - No one is expert in all topical interests
  - Each new question should be routed to answerers interested in related topics with the right level of expertise
- Achieve better understanding of both user topical interest and expertise by leveraging tagging and voting information
  - Tags are important user-generated category information of Q&A posts
  - Votes indicate a CQA community's long term review result for a given user's expertise under a specific topic

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Roadmap

- Motivation
- **Related Work**
- Our Method
  - Method Overview
  - Topic Expertise Model
  - CQARank
- Experiments
- Summery

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Related Work

- Link Analysis
  - HITS (Jurczyk and Agichtein, CIKM07)
  - Expertise Rank and Z-score (Zhang et al., WWW07)
  - Find global experts without model of user interests
- Latent Topical Analysis
  - UQA Model (Guo et al. CIKM08)
  - Fail to capture to what extent these users' expertise match the questions with similar topical interest
- Topic Sensitive PageRank
  - TwitterRank (Weng et al. WSDM10)
  - Topic-sensitive probabilistic model for expert finding (Zhou et al. CIKM12)

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Roadmap

- Motivation
- Related Work
- Our Method
  - **Method Overview**
  - Topic Expertise Model
  - CQARank
- Experiments
- Summery

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

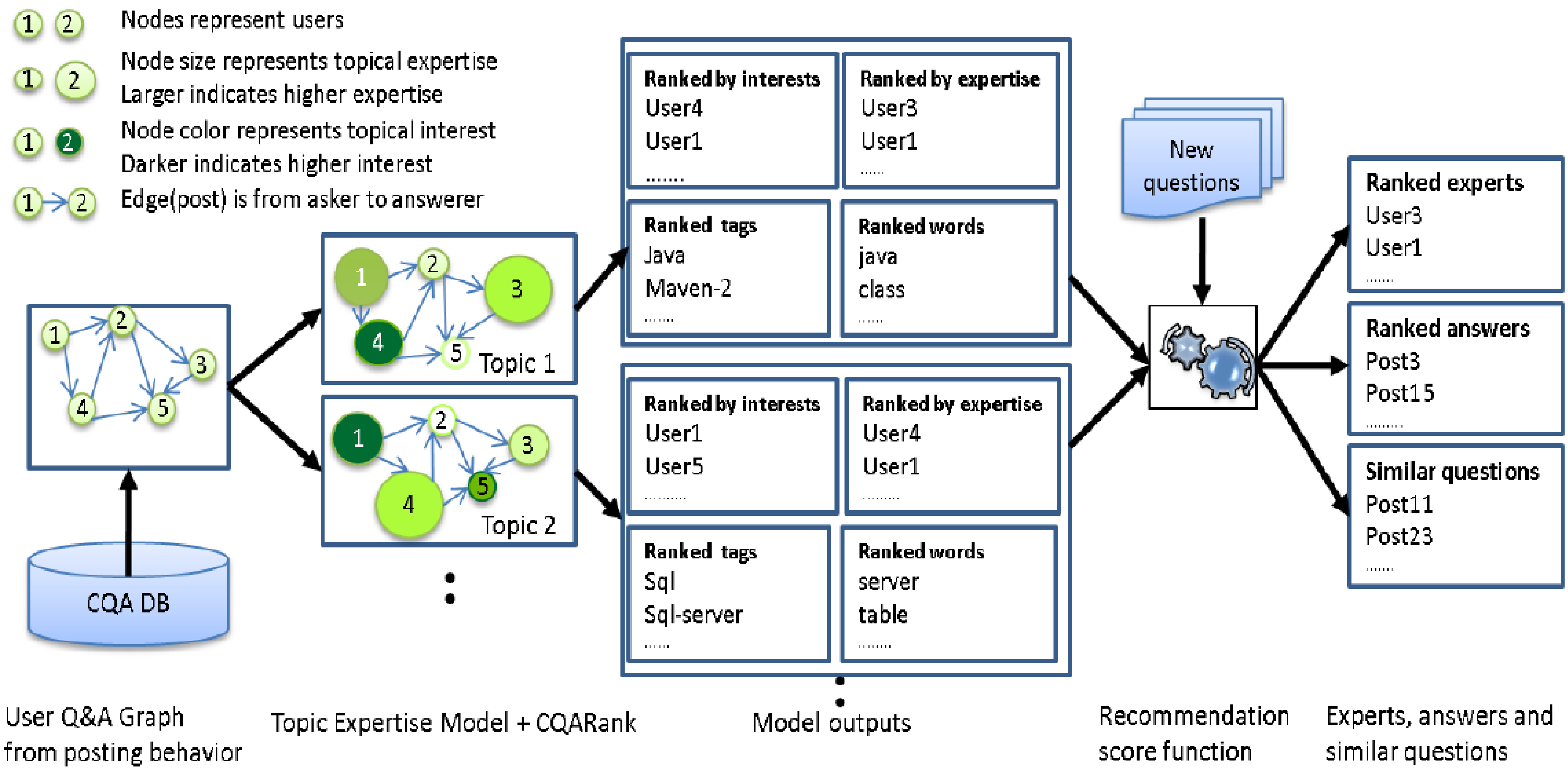
Method Overview

- Concepts
  - Topical Interest
  - Topical Expertise
  - Q&A Graph

- Our Approach
  - Topic Expertise Model
  - CQARank to combine learning results from TEM with link analysis of Q&A graph

Method Overview

- CQARank Recommendation Framework

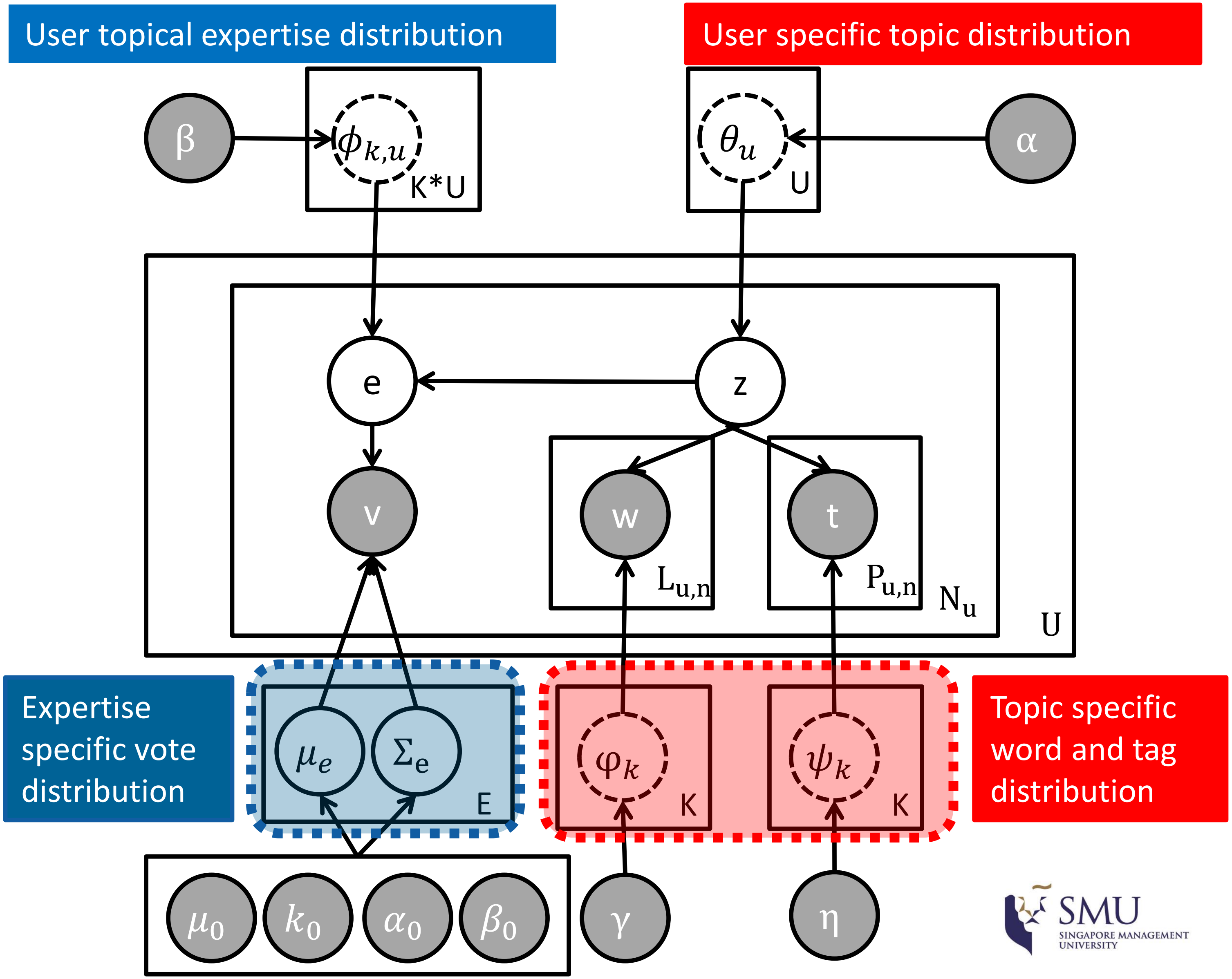


Roadmap

- Motivation
- Related Work
- Our Method
  - Method Overview
  - **Topic Expertise Model**
  - CQARank
- Experiments
- Summery

Topic Expertise Model

- U: # of users
- N_u: # of posts
- $L_{u,n}$: # of words
- $P_{u,n}$: # of tags
- z: topic label
- e: expertise label
- v: a vote
- w: a word
- t: a tag



Roadmap

- Motivation
- Related Work
- Our Method
  - Method Overview
  - Topic Expertise Model
  - **CQARank**
- Experiments
- Summery

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

CQARank

- CQARank combines textual content learning result of TEM with link analysis to enforce user topical expertise learning
- Construct Q&A Graph $G = (V, E)$
  - V is a set of nodes representing users
  - E is a set of directed edges from the asker to the answerer
    - $e = (u_i, u_j)$ $u_i \in V, u_j \in V$
    - Weight W_{ij} is the number of all answers answered by u_j for questions of u_i

CQARank

- For each topic z , the transition probability from asker u_i to answerer u_j is defined as:
  - $P_z(i \to j) = \frac{W_{ij} \cdot sim_z(i \to j)}{\sum_{k=1}^{|V|} W_{ik} \cdot sim_z(i \to k)} \;\; if \;\; \sum_m w_{i,m} W \neq 0$
  - $P_z(i \to j) = 0 \;\; otherwise$
- $sim_z(i \to j)$ is the similarity between u_i and u_j under topic z, which is defined as
  - $sim_z(i \to j) = 1 - \left|\theta'_{iz} - \theta'_{jz}\right|$
- The row-normalized transition matrix M is defined as
  - $\mathbf{M}_{ij} = P_z(i \to j)$

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

CQARank

- Given topic z , the CQARank saliency score of u_i is computed based on the following formula:
  - $\mathbf{R}_z(u_i) = \lambda \sum_{j: u_j \rightarrow u_i} \mathbf{R}_z(u_j) \cdot \mathbf{M}_{ij} + (1-\lambda) \cdot \theta_{u_i z} \cdot \mathbf{E}(z, u_i)$
  - $\mathbf{E}(z, u_i)$ is the estimated expertise score of u_i under topic z, which is defined as the expectation of user topical expertise distribution learnt by TEM.

$$\mathbf{E}(z, u_i) = \sum_e \phi_{z, u_i, e} \cdot \mu_e$$

  - $\lambda \in [0,1]$ is a parameter to control the probability of teleportation operation.

PEKING UNIVERSITY 北京大学

SMU SINGAPORE MANAGEMENT UNIVERSITY

Roadmap

- Motivation
- Related Work
- Our Method
  - Method Overview
  - Topic Expertise Model
  - CQARank
- **Experiments**
- Summery

Experiments

- Stack Overflow Data Set
  - All Q&A posts in three months (May 1^{st} to August 1^{st}, 2009)
  - Training data: 8,904 questions and 96,629 answers posted by 663 users.(10,689 unique tags and 135 unique votes)
  - Testing data: 1,173 questions and 9,883 answers
- Data Preprocessing
  - Tokenize text and discard all code snippets
  - Remove stop words and HTML tags in text
- Parameters Setting
  - $K = 15, E = 10, \alpha = \frac{50}{K}, \beta = 0.01, \gamma = 0.01, \eta = 0.001, \lambda = 0.2$
  - Norma-Gamma parameters
  - 500 iterations of Gibbs Sampling

TEM Results

- Topic Analysis - topic tags
  - Top tags provide phrase level features to distill richer topic information

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
java	c++	sql	python	css	svn	subjective	iphone	security	c#
eclipse	c	sql-server	linux	html	version-control	career-development	objective-c	c#	.net
spring	windows	mysql	windows	javascript	git	best-practices	cocoa-touch	encryption	visual-studio
maven-2	visual-c++	tsql	bash	jquery	mercurial	language-agnostic	iphone-sdk	php	asp.net
ant	visual-studio	database	perl	internet-explorer	tortoisesvn	project-management	cocoa	.net	visual-studio-2008
tomcat	linux	sql-server-2005	beginner	web-development	best-practices	learning	uikit	asp.net	sharepoint
jar	c#	php	unix	asp.net	visual-studio	design	xcode	cryptography	windows
jsp	delphi	database-design	vim	xhtml	tfs	jobs	uitableview	email	vb.net
j2ee	winapi	query	php	div	visual-sourcesafe	java	memory-management	authentication	c++
hibernate	gcc	oracle	java	best-practices	beginner	software-engineering	core-animation	java	iis

PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

TEM Results

- Topic Analysis - topic words
  - Top words have strong correlation with top tags under the same topic

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
java	library	table	command	css	git	time	view	user	project
class	files	query	line	html	files	software	method	server	application
spring	dll	sql	files	element	repository	make	object	password	web
jar	compiler	data	script	div	svn	project	class	key	files
eclipse	function	index	run	page	branch	programming	controller	data	visual
project	windows	column	directory	text	version	design	set	address	net
application	header	key	windows	width	control	development	make	hash	studio
files	make	rows	python	browser	changes	find	methods	security	windows
maven	source	database	output	image	source	job	objects	client	server
build	functions	tables	shell	elements	commit	problem	data	users	version

TEM Results

- Expertise Analysis
  - TEM learns different user expertise levels by clustering votes using GMM component.
  - 10 Gaussian distributions with various means for the generation of votes in data.
  - The higher the mean is, the lower the precision is.

	Expertise 1	Expertise 2	Expertise 3	Expertise 4	Expertise 5	Expertise 6	Expertise 7	Expertise 8	Expertise 9	Expertise 10
Mean	40.17	10.42	6.07	4.39	3.25	2.38	1.75	1.46	1.14	0.40
Precision	3.03E-04	1.97E-02	4.48E-02	1.07E-01	1.11E-01	2.43E-01	4.57E-01	5.92E-01	6.51E-01	3.14E+00

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Expert Users

- Task
  - Given a new question q and a set of users $\mathbf{U}$, Rank users by their interests and expertise to answer question q.
  - Recommendation score function

$$S(u,q) = Sim(u,q) \cdot Expert(u,q)$$
$$= (1 - JS(\theta_u, \theta_q)) \cdot \sum_z \theta_{q,z} \cdot Expert(u,z)$$

  - $\theta_{q,z}$ is the estimated posterior topic distribution of question q

$$\theta_{q,z} \propto p(z|\mathbf{w}_q, \mathbf{t}_q, u) = p(z|u)p(\mathbf{w}_q|z)p(\mathbf{t}_q|z)$$
$$= \theta_{u,z} \sum_{w:\mathbf{w}_q} \varphi(z,w) \sum_{t:\mathbf{t}_q} \psi(z,t)$$

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Expert Users

- Our method
  - **CQARank**
- Baselines

  Link analysis method
  - In Degree(**ID**)
  - PageRank(**PR**)

  Probabilistic generative model
  - **TEM**(Part of our method)
  - **UQA**(Guo et al. CIKM08)

  Combine link analysis and topic model
  - Topic Sensitive PageRank(**TSPR**)(Zhou et al. CIKM12)

北京大學 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Expert Users

- Evaluation Criteria
  - Ground truth: User rank list by average votes for answering q
  - Metrics: $nDCG$, Pearson/Kendall correlation coefficients
- Results

	nDCG@1	nDCG@5	nDCG	Pearson	Kendall
CQARank	**0.5858**‡	**0.7991**†	**0.8941**†	**0.1905**†	**0.1738**†
TEM	0.5757	0.7826	0.8920	0.1720	0.1429
UQA	0.4650	0.7548	0.8547	-0.0606	-0.0498
TSPR	0.4790	0.7551	0.8611	-0.0136	-0.0138
PR	0.5078	0.7875	0.8729	0.0575	0.0621
ID	0.5492	0.7710	0.8727	0.0920	0.0858

Recommend Answers

- Task
  - Give a new question q and a set of answers $\mathbf{A}$, Rank all answers in $\mathbf{A}$.
  - Recommendation score function

$$S(a,q) = Sim(a,q) \cdot Expert(u,q)$$
$$= (1 - JS(\theta_a, \theta_q)) \cdot \sum_z \theta_{q,z} \cdot Expert(u,z)$$

- Baselines and evaluation criteria are the same with expert recommendation task
- We use each answer's vote to generate ground truth rank list

PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Answers

- Result

	nDCG@1	nDCG@5	nDCG	Pearson	Kendall
CQARank	0.4748	**0.7857**†	**0.8194**†	**0.1644**‡	**0.1421**‡
TEM	0.4253	0.7830	0.8080	0.1289	0.1131
UQA	0.4010	0.7293	0.7661	-0.0840	-0.0709
TSPR	0.4007	0.7576	0.7924	0.0186	0.0091
PR	**0.5196**†	0.7791	0.8107	0.0718	0.0536
ID	0.4578	0.7756	0.8048	0.0572	0.0495

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Similar Questions

- When a user asks a new question(referred as *query question*), the user will often get replies of links to other *similar questions*
- Crawl 1000 questions as *query question* set whose similar questions exist in the training data set
- For each *query question* with n *similar questions*, we randomly select another m $(\mathrm{m} = 1000)$ questions from the training data set to form *candidate similar questions*

北京大學 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Similar Questions

- All comparing methods rank these $m + n$ *candidate similar questions* according to their similarity with the *query question*
- The higher the similar questions are ranked, the better the performance of the method is.
- Recommendation score is computed based on JS-divergence between topic distributions of the *query question* and *candidate similar questions*

PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Recommend Similar Questions

- Baseline
  - TSPR(LDA), UQA, SimTag
- Evaluation Criteria
  - Precision@K, Average rank of similar questions, Mean reciprocal rank (MRR), Cumulative distribution of ranks (CDR)

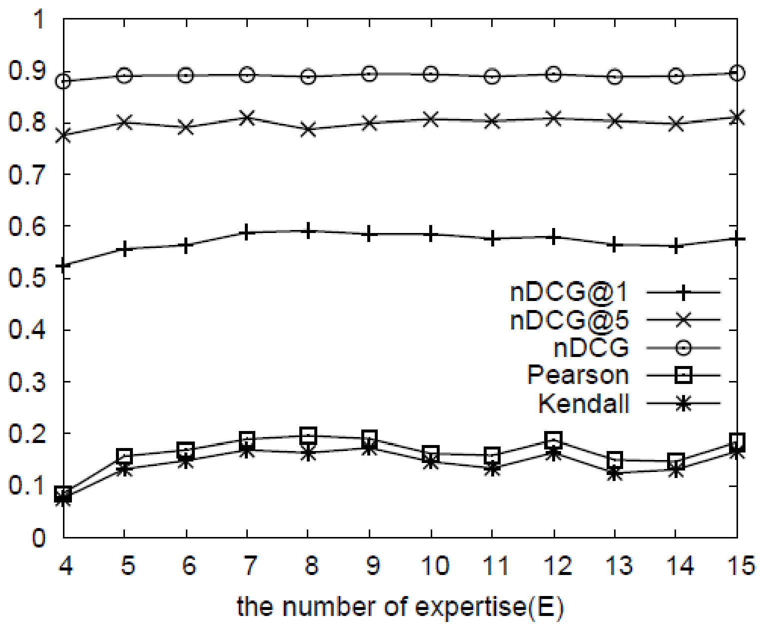
	$\bar{r}$	MRR	P@50	P@100	CDR@50	CDR@100
CQARank(TEM)	**161**$^\diamond$	**0.0713**$^\diamond$	**0.0089**$^\diamond$	**0.0061**$^\diamond$	**0.443**	**0.611**
TSPR(LDA)	547	0.0077	0.0009	0.0009	0.049	0.093
UQA	577	0.0069	0.0009	0.0009	0.044	0.091
SimTag	386	0.1143	0.0051	0.0028	0.257	0.285

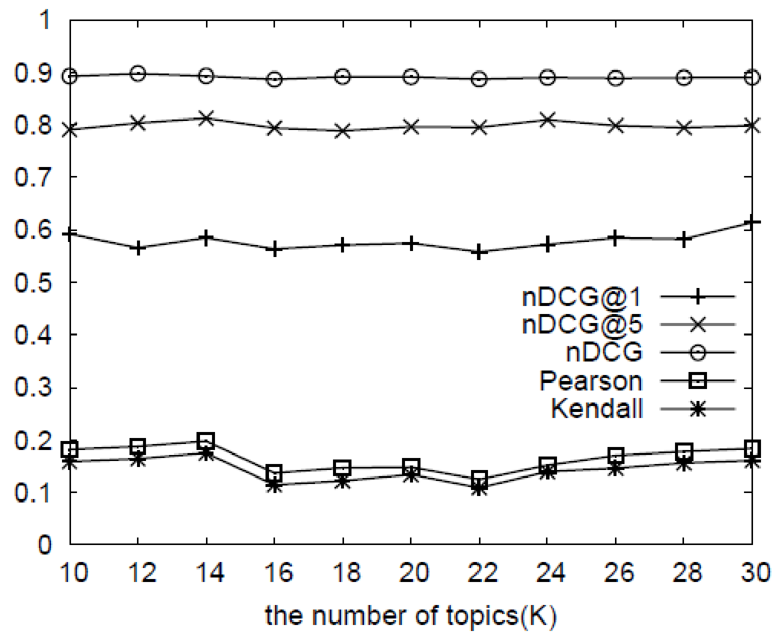
PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Parameter Sensitivity Analysis

- Performance in expert users recommendation of CQARank by varying the number of expertise (E) and topics (K)





北京大學 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Roadmap

- Motivation
- Related Work
- Our Method
  - Method Overview
  - Topic Expertise Model
  - CQARank
- Experiments
- **Summery**

北京大学 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Summery

- Conclusions
  - A probabilistic generative model to jointly model topics and expertise in CQA services
  - CQARank algorithm to combine textual content learning with link analysis
  - Our model is generalized and applicable for various CQA tasks
- Future Work
  - Temporal analysis of topic expertise and interests in CQA
  - Social influence of experts

北京大學 PEKING UNIVERSITY

SMU SINGAPORE MANAGEMENT UNIVERSITY

Thank you

Q&A