A Computational Framework for Question Processing in Community Question Answering Services

Baichuan Li

January 9, 2014

Thesis Committee:

Prof. FU Wai Chee Ada (Chair)
Prof. LEE Ho Man Jimmy (Internal Examiner)
Prof. ZHU Xiaoyan (External Examiner)
Prof. KING Kuo Chin Irwin (Supervisor)
Prof. LYU Rung Tsong Michael (Supervisor)

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction
- Question Routing
  - Quality and Availability
  - Category
- Question Structuralization
- Conclusion and Future Work

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction
- Question Routing
  - Quality and Availability
  - Category
- Question Structuralization
- Conclusion and Future Work

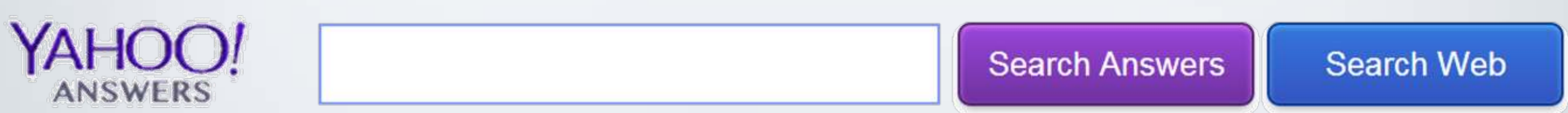
1 Community Question Answering

- What is CQA?
- Why CQA?

1

Example: Yahoo! Anwers

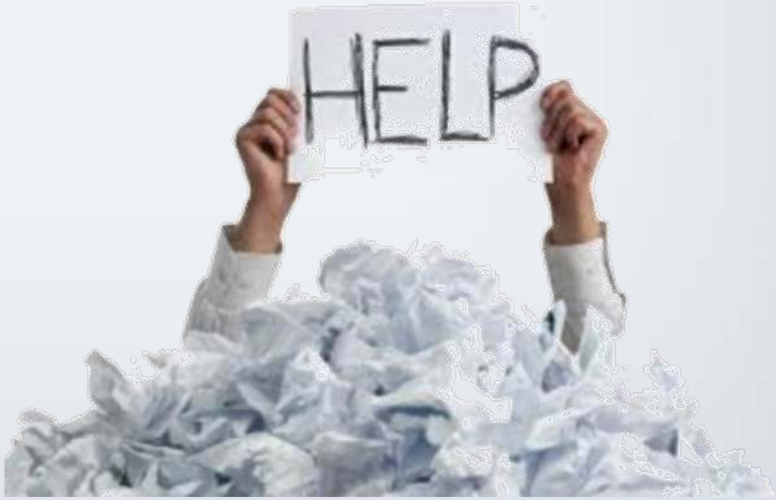
- The most popular CQA portal among the world
- Two questions are asked and six are answered every second
- 300 million questions have been asked by July, 2012



1

Challenges in CQA

- Inefficient Question Answering
  - Sharp increase of questions
  - Time lag between Q&A
- Straightforward Content Organization



1

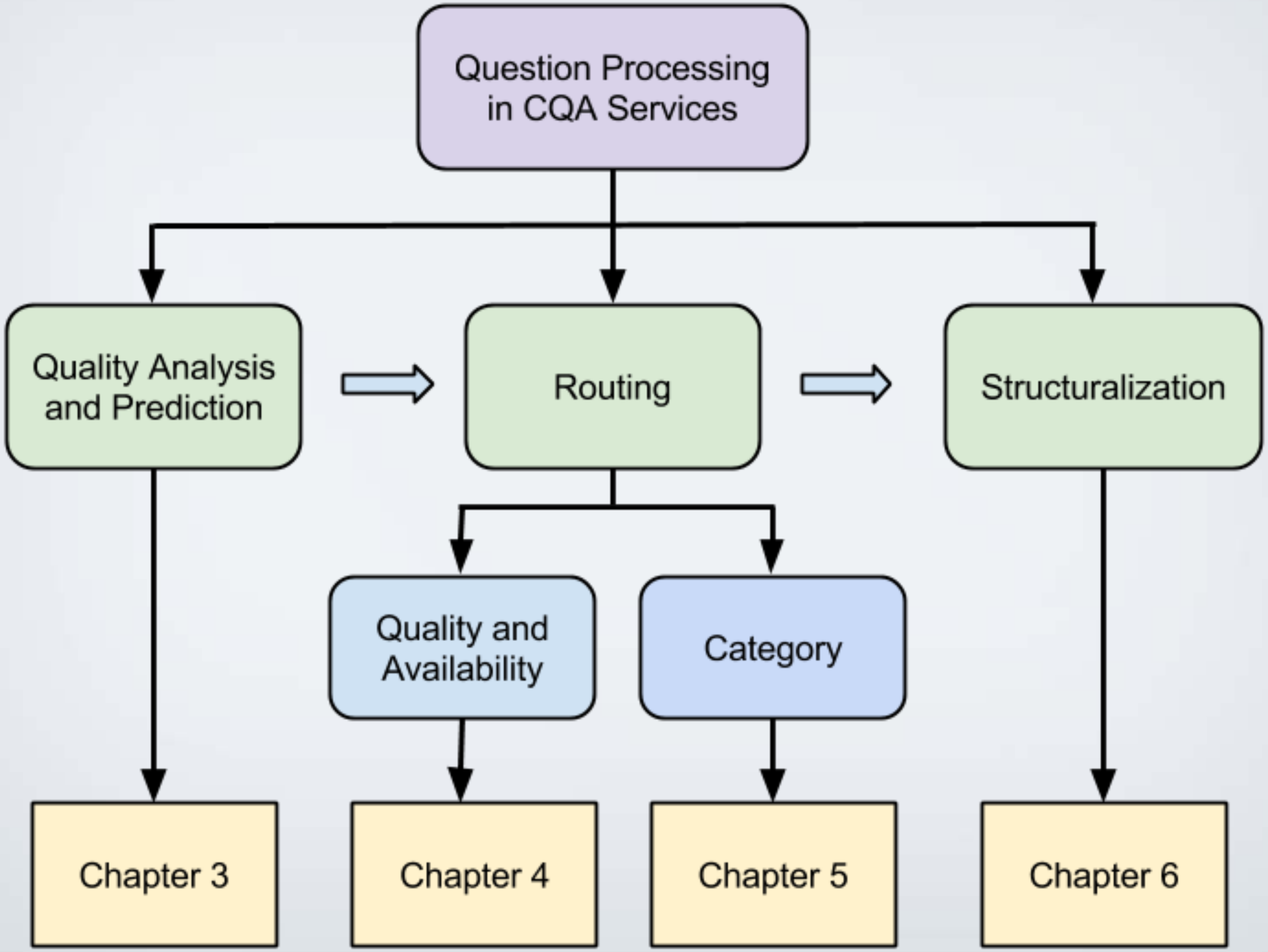
Objective of Thesis

- User
  - Facilitate answerers access to proper questions
  - Help askers obtain information more effectively
- System
  - Improve content organization
  - Enhance QA efficiency

Solution

A computational framework for question processing

Structure of Thesis

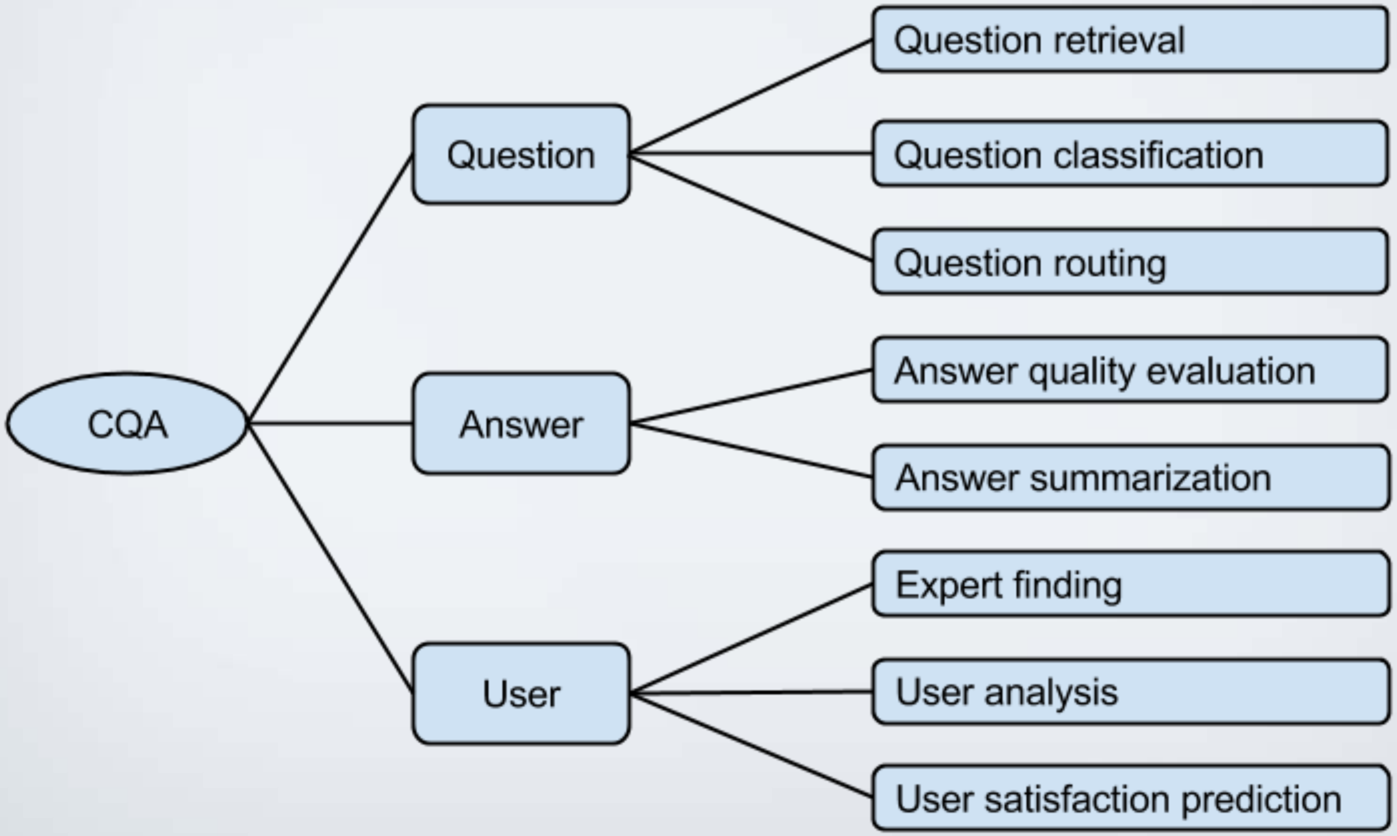


Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction
- Question Routing
  - Quality and Availability
  - Category
- Question Structuralization
- Conclusion and Future Work

2

Research topics in CQA



2

Question Processing

- Question Retrieval
  - Basic models (Jeon et al., 2005; Duan et al., 2008)
  - Extra information: category (Cao et al., 2010), syntactic knowledge (Wang et al., 2009), answer (Bian et al., 2008), etc.
- Question Classification
  - Properties: urgency, subjectivity
  - Models: SVM (Li et al., 2008), Co-training (Li et al., 2008), sequential minimal optimization (Harper et al., 2009)
- Question Routing
  - User Profiling
  - Question Profiling
  - Matching

2

Answer Processing

- Answer Quality Evaluation
  - Classification-based (Jeon et al., 2006; Eugene et al., 2008; Shah et al., 2010)
  - Ranking-based (Suryanto et al., 2009; Wang et al., 2009)
- Answer Summarization
  - Question type-based (Liu et al., 2008)
  - Constraint-based (Tomasoni et al., 2010; Liu et al., 2011)
  - Graph-based (Chan et al., 2012; Pande et al., 2013)

User Processing

- Expert Finding
  - Link analysis (Jurczyk et al., 2007; Zhang et al., 2007)
  - Content analysis (Liu et al., 2005; Budalakoti, 2013)
- User Analysis
  - User behavior (Gazan, 2006; Rodrigues et al., 2008)
  - Community (Li et al., 2012)
- User Satisfaction Prediction
  - Classification (Liu et al., 2008; Liu et al., 2010)

3

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction **(Chapter 3)**
- Question Routing
  - Quality and Availability **(Chapter 4)**
  - Category **(Chapter 5)**
- Question Structuralization **(Chapter 6)**
- Conclusion and Future Work

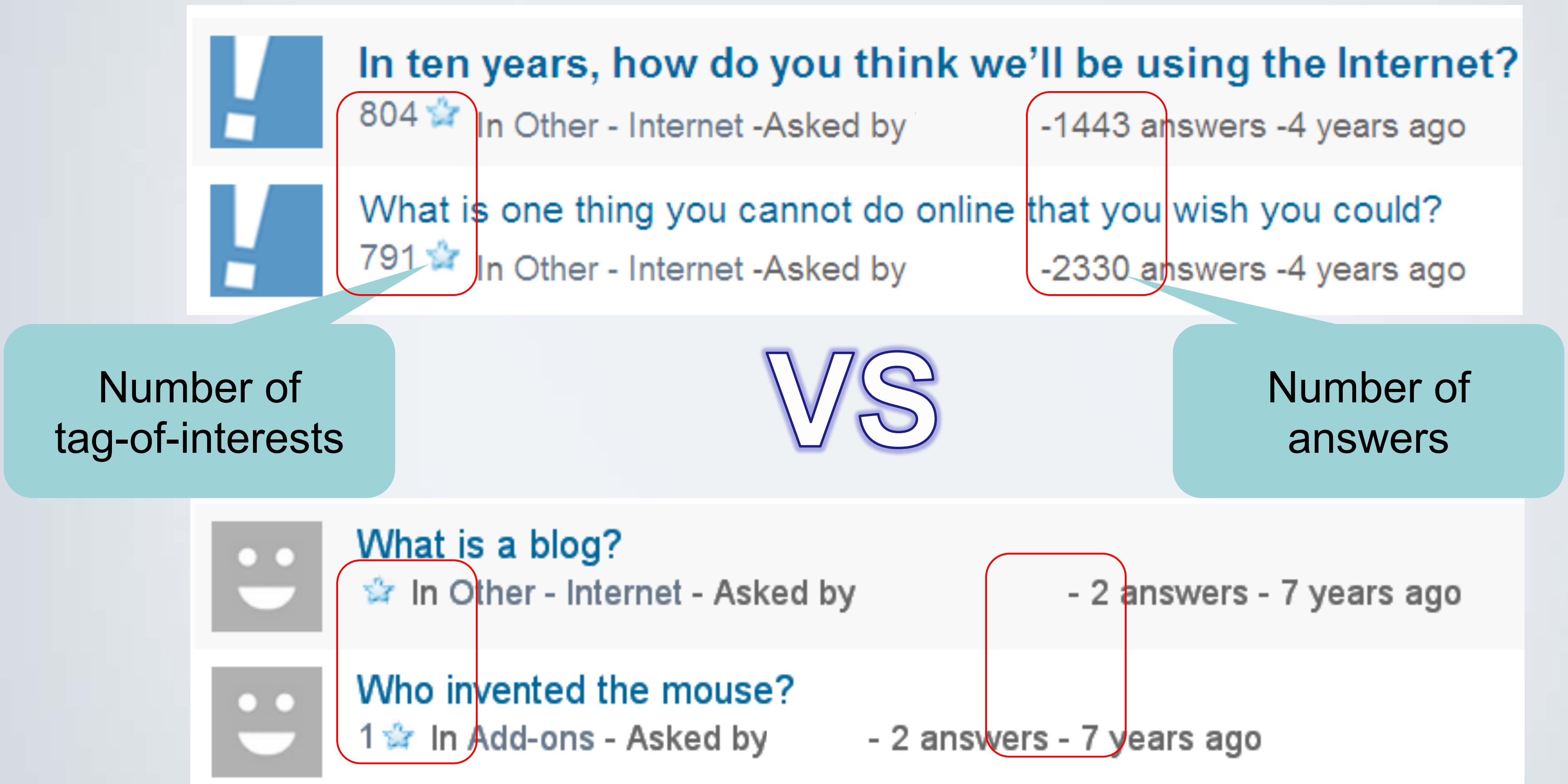
3

Question Quality Analysis and Prediction

- Motivation and Definition
- Study One: Factors Affecting Question Quality
- Study Two: Question Quality Prediction
- Summary

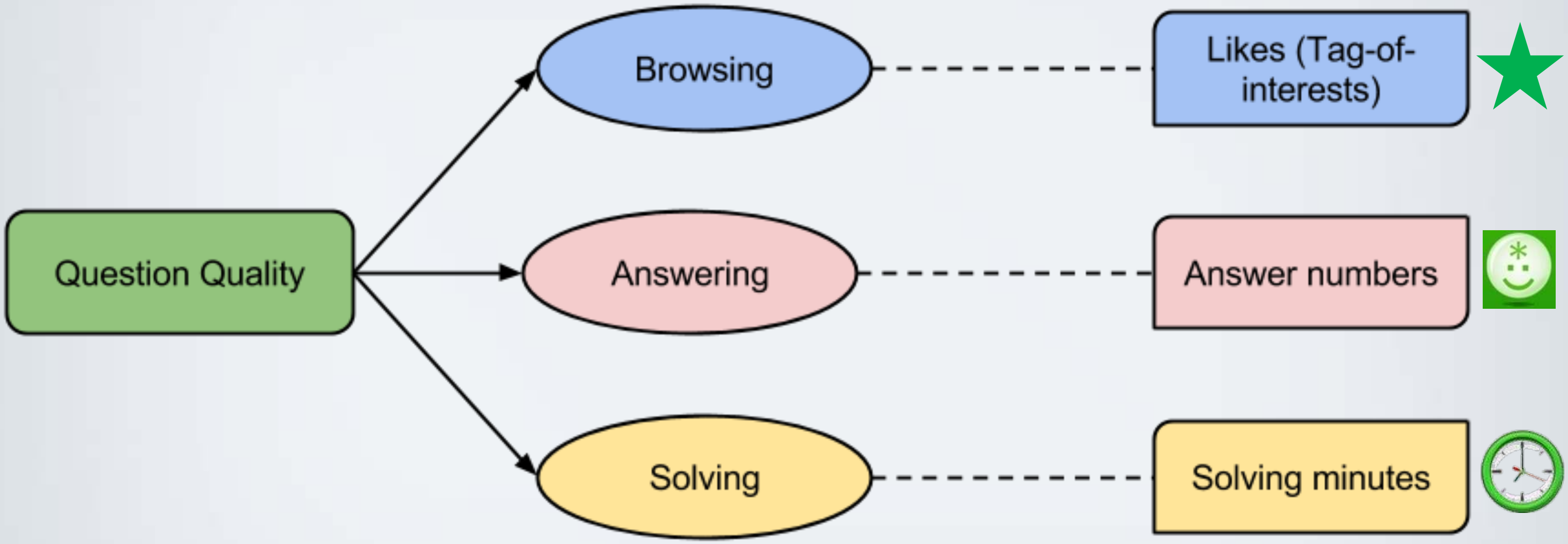
3

Question Quality



3

Definition of Question Quality



Construct of question quality in CQA

Motivation

- Question quality affects answer quality
  – Low quality questions hinder QA efficiency
  – High quality questions promote the development of the community
- Question routing
- Identifying question quality facilitates question search and recommendation

3

Study One: Factors Affecting Question Quality

- Factors

Topics

Askers

- Process
  - Select the two most popular subcategories (say, *Music* and *Movies*) and check their distributions of question quality
  - Track askers with at least five questions in both these two subcategories (22 in total)

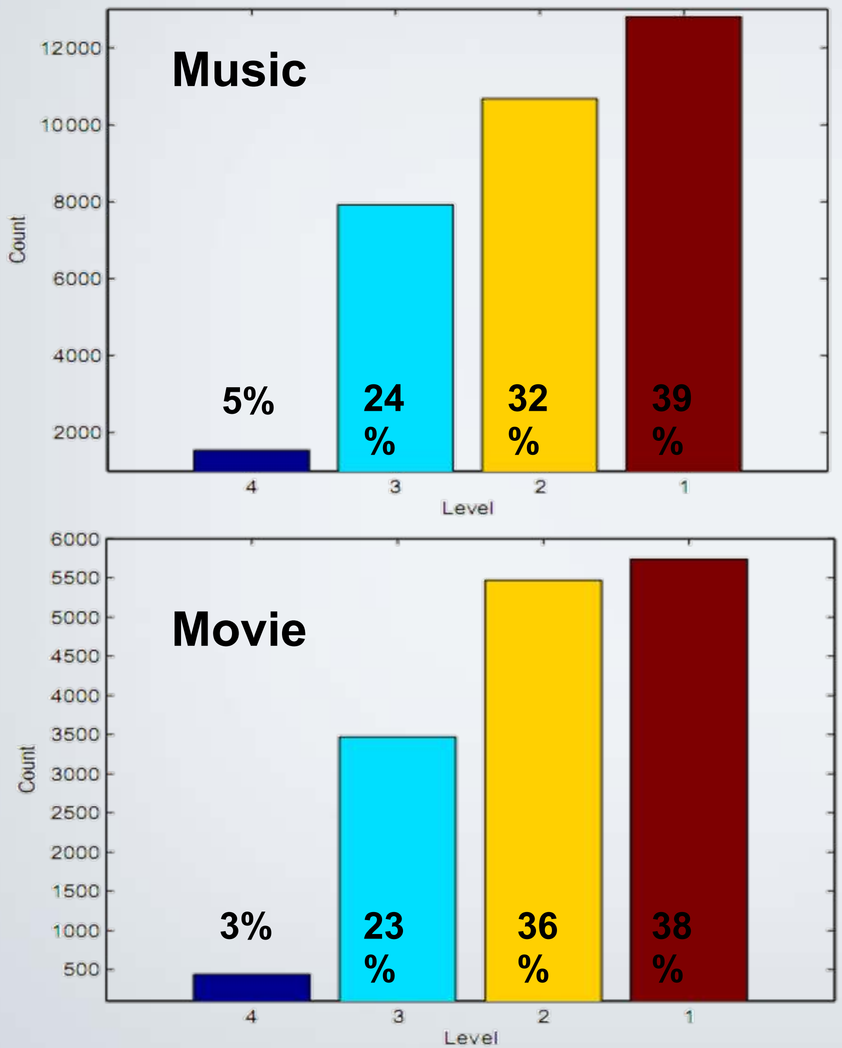
Data Description

Summary of data (crawled from Jul 7, 2010 to Sep 6, 2010)

Subcategory	# of questions	# of askers
Celebrities	11,817	7,087
Comics & Animation	11,327	6,801
Horoscopes	7,235	2,203
Jokes & Riddles	3,685	2,569
Magazines	548	462
Movies	15,121	10,996
Music	32,948	18,589
Other - Entertainment	2,244	2,003
Polls & Surveys	138,507	18,685
Radio	640	272
Television	14,477	10,146
All	238,549	62,853

Questions are assigned to four classes according to manually crafted rules

Observations



- The distributions of question quality in these subcategories are similar
- Topics only cannot distinguish good questions from bad ones

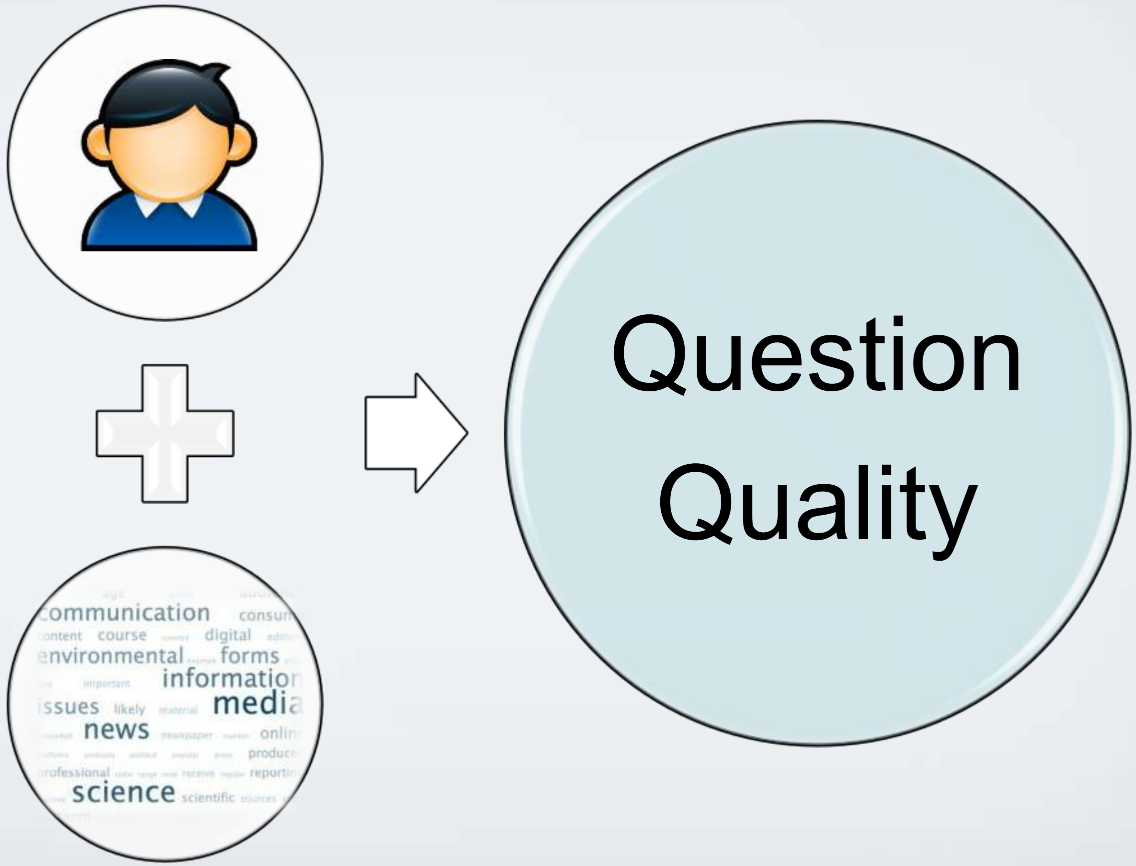
Observations

Summary of question quality for different askers

User	Music		Movies	
	Mean	Std	Mean	Std
1	2.50	0.93	2.17	0.41
2	2.45	0.52	2.57	0.98
3	1.86	0.90	1.45	0.82
4	2.65	0.72	2.60	0.55
5	1.90	0.74	2.00	0.71
6	2.62	0.87	1.83	0.86
7	2.48	0.68	2.20	0.84
8	2.86	0.92	2.14	0.90
9	2.38	0.92	2.30	1.06
10	2.50	0.53	2.40	0.55
11	2.00	0.71	1.50	0.55
12	2.48	0.95	2.47	0.84
13	2.84	0.68	2.83	0.41
14	1.33	0.52	2.40	0.89
15	1.90	0.74	1.83	0.75
16	1.80	0.84	1.83	0.75
17	2.15	0.55	2.50	1.05
18	2.36	0.92	1.67	0.87
19	2.00	1.00	2.00	1.00
20	2.00	0.67	2.00	1.00
21	2.69	0.68	2.80	0.45
22	2.13	0.99	2.57	1.27

- For the same topic
  - Different askers obtain various question quality
    - User 8 VS User 16 in *Music*
    - User 2 VS User 3 in *Movies*
- For the same asker
  - Question quality varies on different topics
    - User 14

Observations



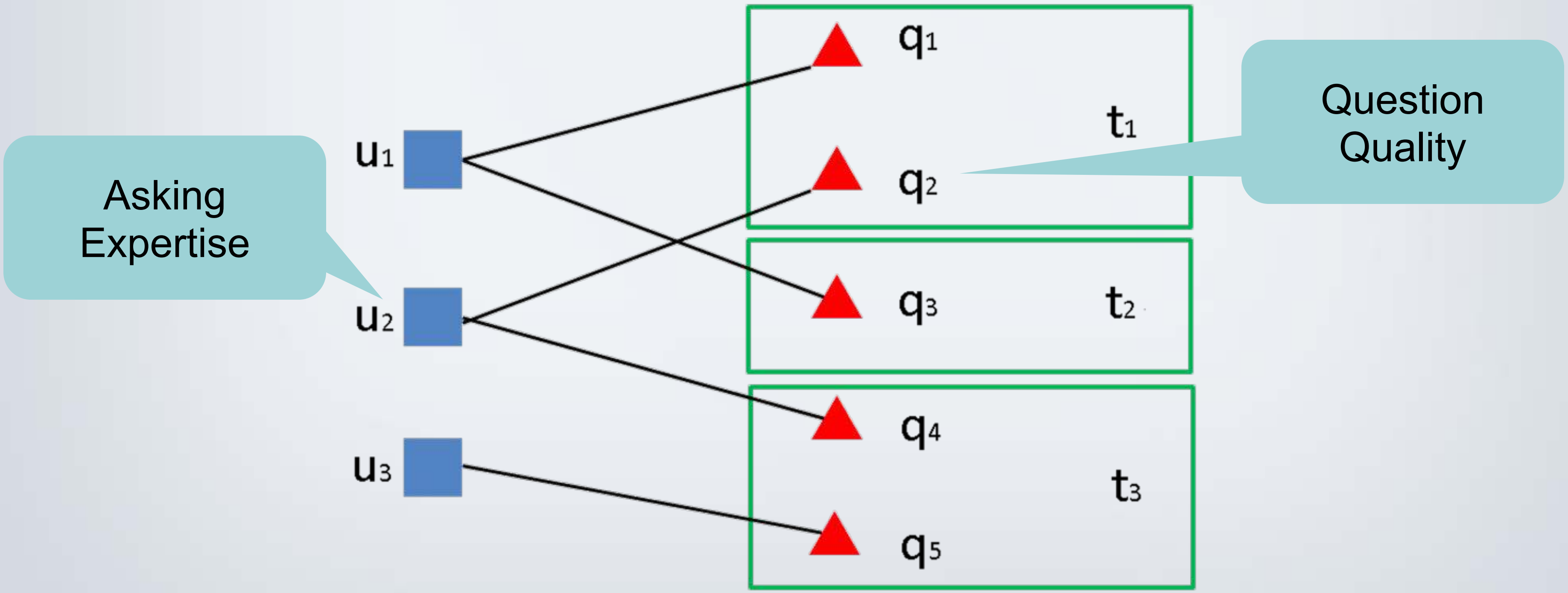
3

Challenges

- A new question comes…
- No answers, no tags
- Can we predict a new question’s quality?

Study Two: Question Quality Prediction

- Modeling the relationships among questions, topics and askers as a bipartite graph



3

Mutual Reinforcement Label Propagation for Predicting Question Quality

Algorithm 1 MRLP

Input: user asking expertise vector U_0^k, question quality vector Q_0^k, E, transition matrixes M and N, weighting coefficients α and β, some manual labels of U_0^k and/or Q_0^k.

similar users' asking expertise

Set $c = 0$.

while not convergence **do**

Propagate user expertise. $U_{c+1}^k = \alpha \cdot M \cdot U_c^k + (1 - \boxed{\alpha) \cdot E' \cdot Q_c^k.}$

question quality

Propagate question quality. $Q_{c+1}^k = \boxed{\beta \cdot N \cdot Q_c^k} + (1 - \beta) \cdot E^T \cdot U_{c+1}^k$, where E^T is the transpose of E.

asking expertise

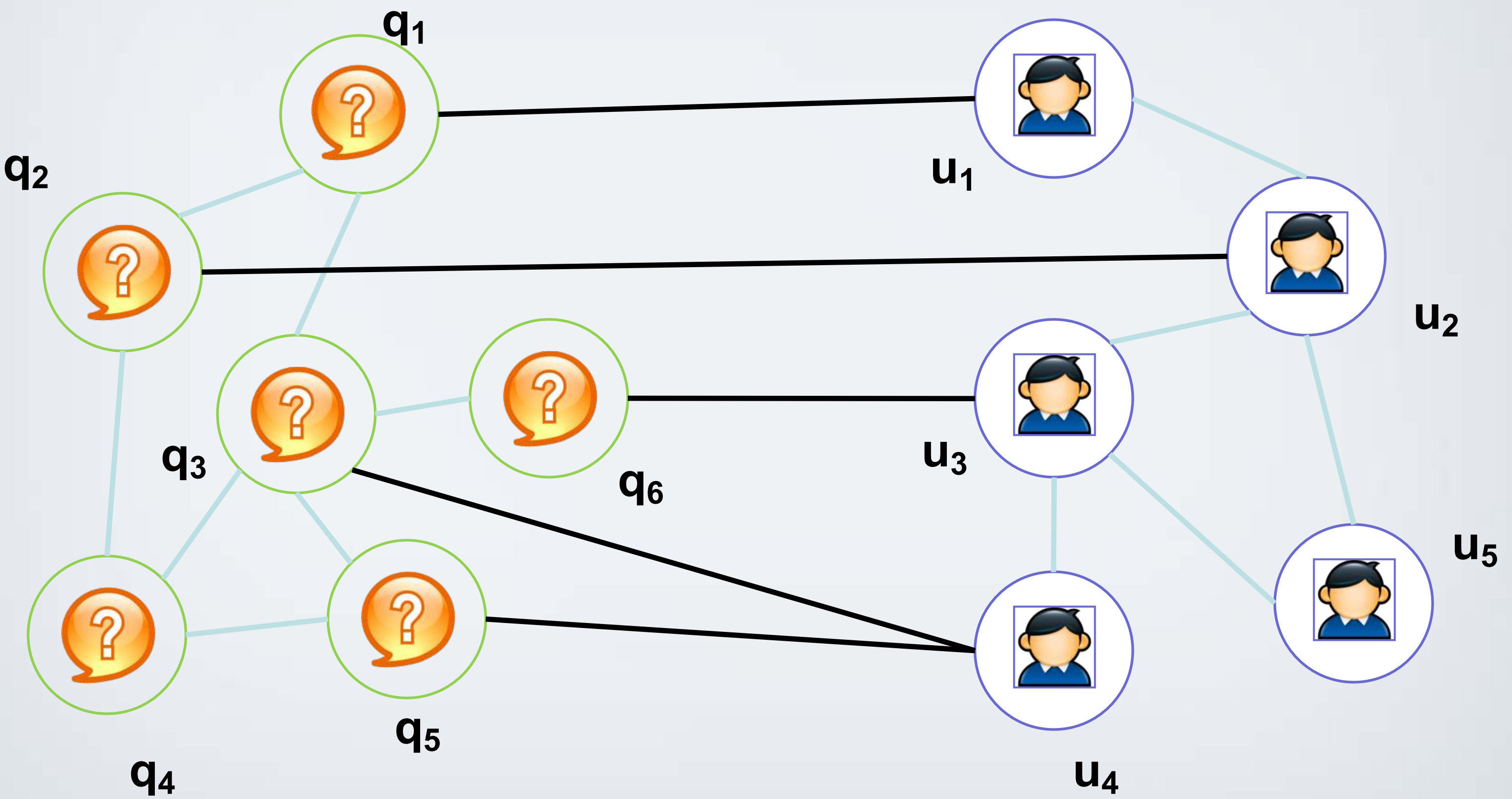
Clamp the labeled data of U_{c+1}^k and Q_{c+1}^k.

similar questions' quality

Set $c = c + 1$.

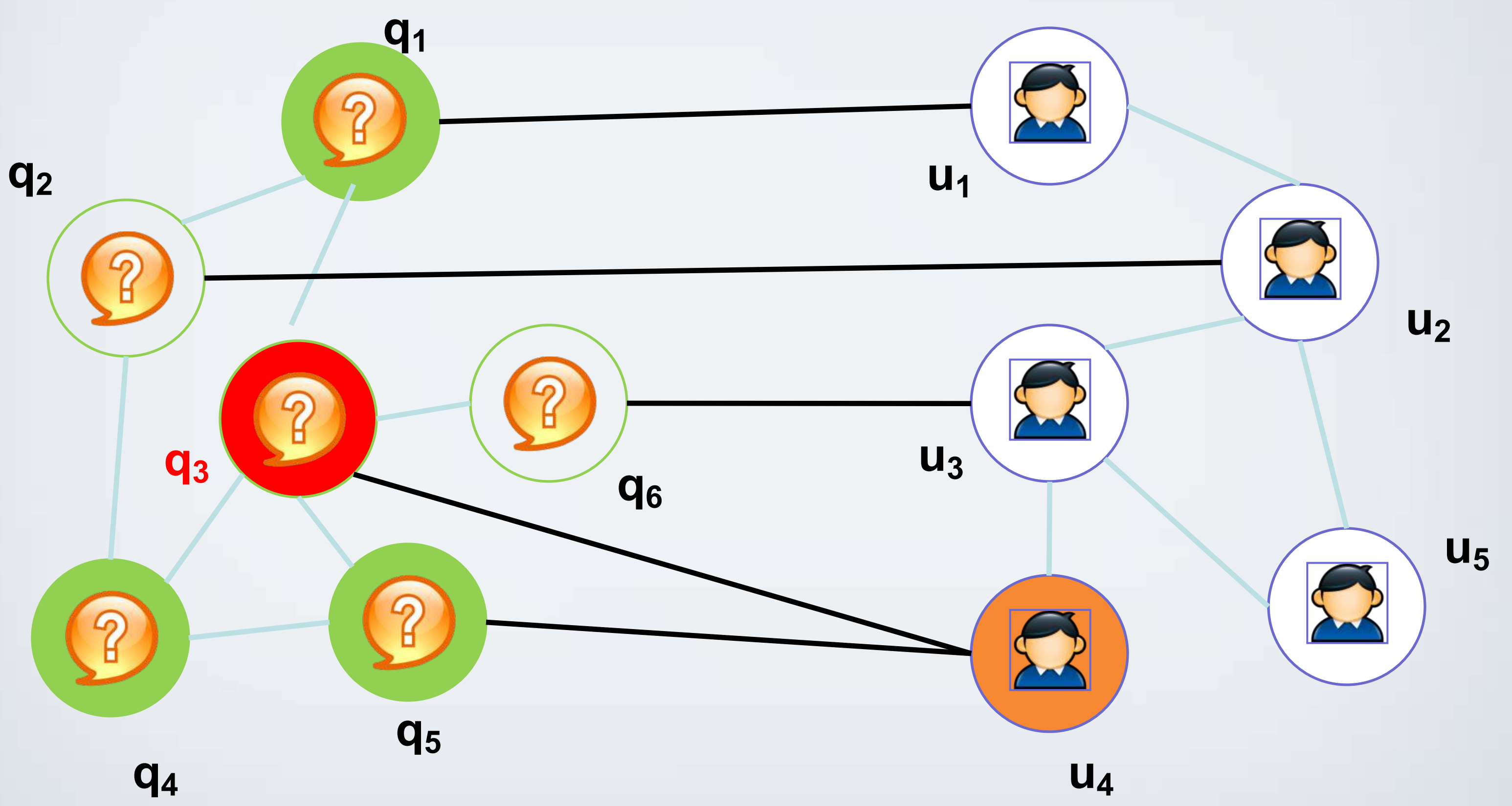
end while

Example



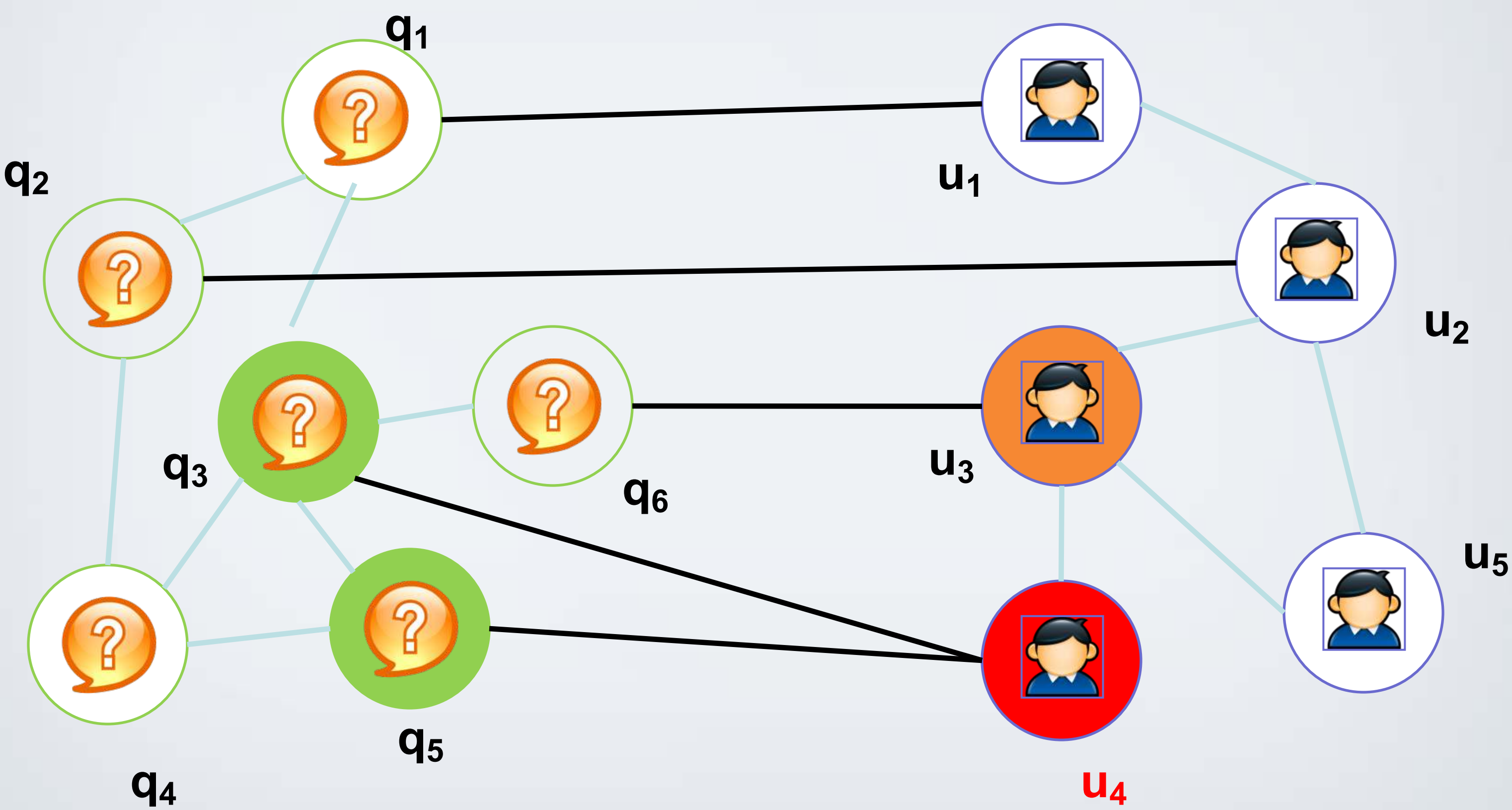
3

Question Quality Estimation



3

Asking Expertise Estimation



Features

Summary of features

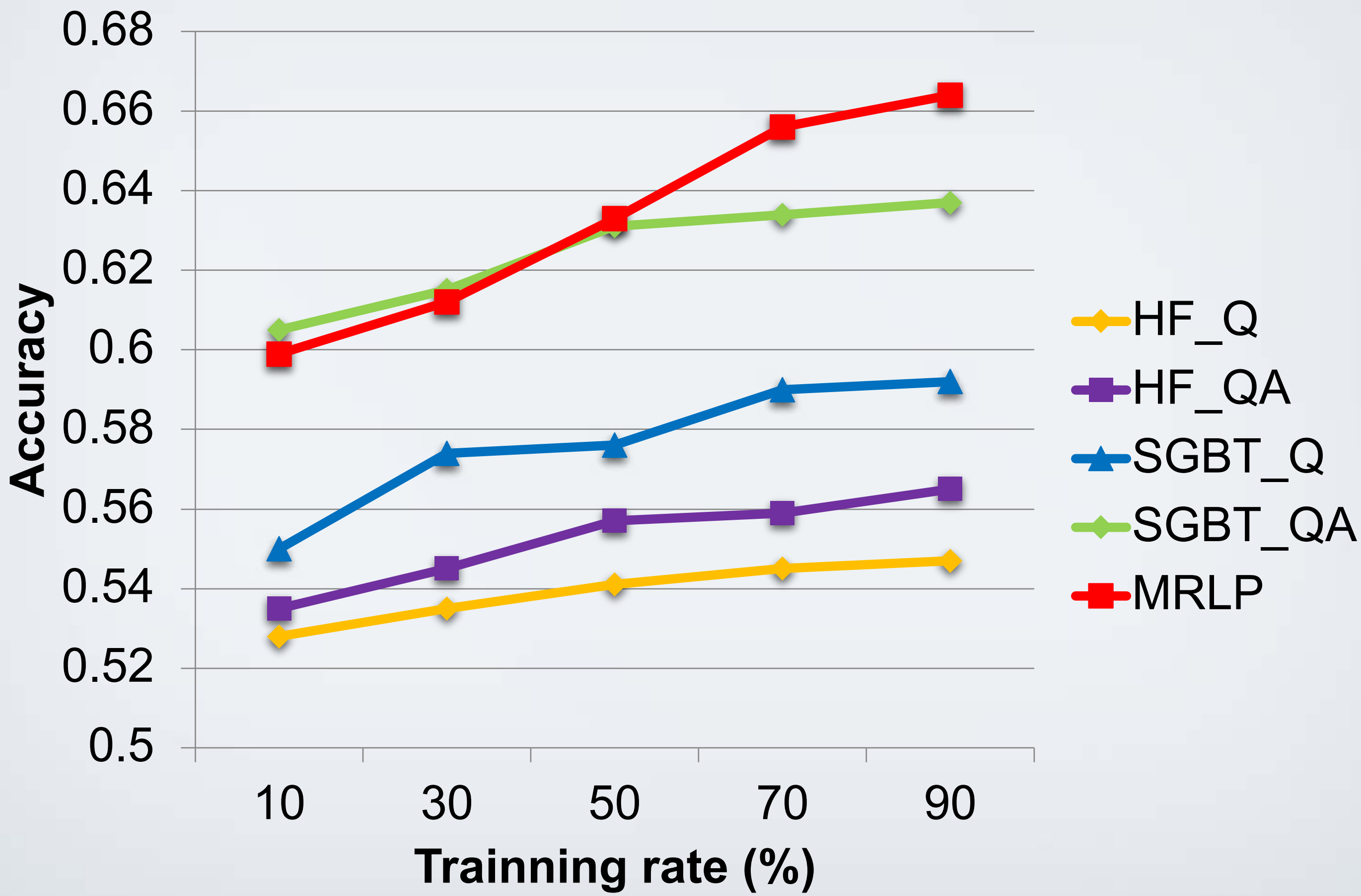
Name	Description	IG
	Question-related features	
Sub_len	Number of words in question subject (title)	0.0115
Con_len	Number of words in question content	0.0029
Wh-type	Whether the question subject starts with Wh-word (e.g., "what", "where", etc.)	0.0001
Sub_punc_den	Number of question subject's punctuation over length	0.0072
Sub_typo_den	Number of question subject's typos over length	0.0021
Sub_space_den	Number of question subject's spaces over length	0.0138
Con_punc_den	Number of question content's punctuation over length	0.0096
Con_typo_den	Number of question content's typos over length	0.0006
Con_space_den	Number of question content's spaces over length	0.0113
Avg_word	Number of words per sentence in question's subject and content	0.0048
Cap_error	The fraction of sentences which are started with a small letter	0.0064
POS_entropy	The entropy of the part-of-speech tags of the question	0.0004
NF_ratio	The fraction of words that are not the top 10 frequent words in the collection	0.0009
	Asker-related features	
Total_points	Total points the asker earns	0.0339
Total_answers	Number of answers the asker provided	0.0436
Best_answers	Number of best answers the asker provided	0.0331
Total_questions	Number of questions the asker provided	0.0339
Resolved_questions	Number of resolved questions asked by the asker	0.0357
Star_received	Number of stars received for all questions	0.0367

Methods for Comparison

- Logistic Regression
  - LG_Q and LG_QA
- Stochastic Gradient Boosted Tree (Friedman, J. H., 1999)
  - SGBT_Q and SGBT_QA
- Harmonic Function (Zhou et al., 2007)
  - HF_Q and HF_QA

3

Results: Accuracy



Different algorithms' accuracy (*Music*)

3

Sensitivity & Specificity

- Sensitivity measures the algorithm's ability to identify (recall) high-quality questions

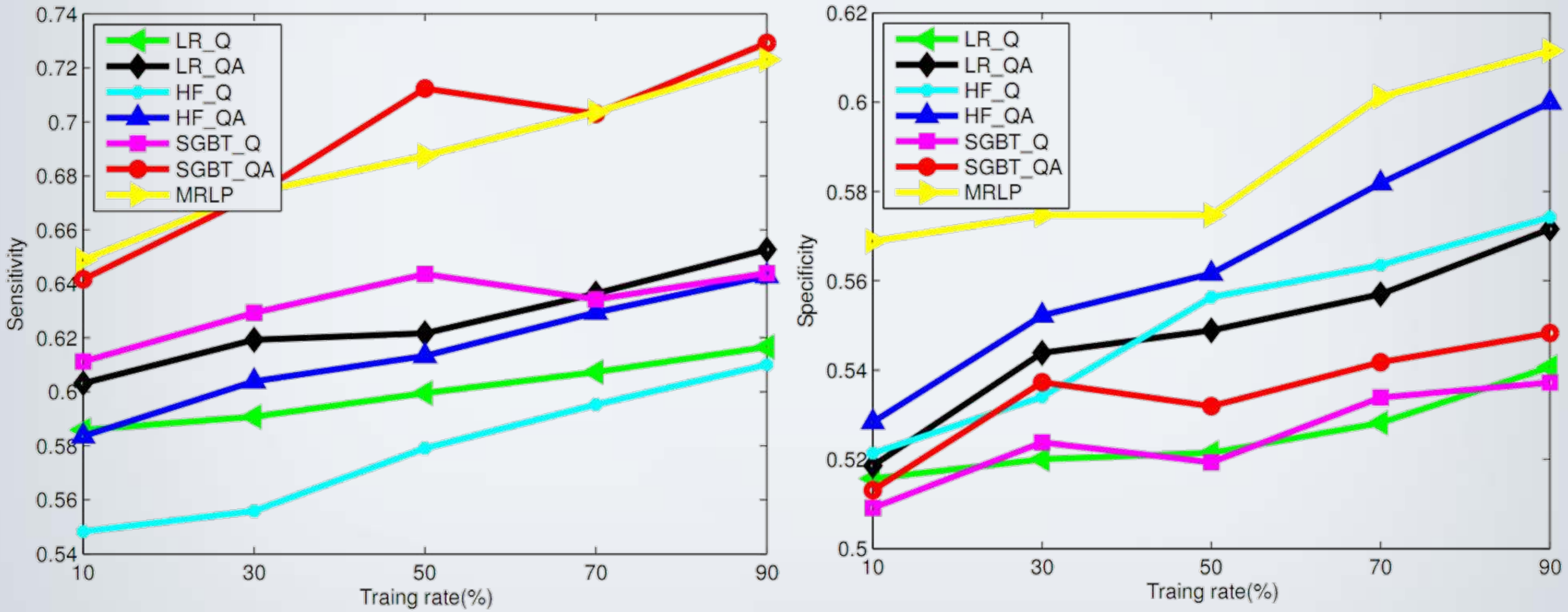
Sensitivity = TP/(TP+FN)

- Specificity measures the algorithm's ability to identify (recall) low-quality questions

Specificity = TN/(TN+FP)

3

Results: Sensitivity & Specificity



Different algorithms' Sensitivity and Specificity (*Music*)

3

Contribution of Chapter 3

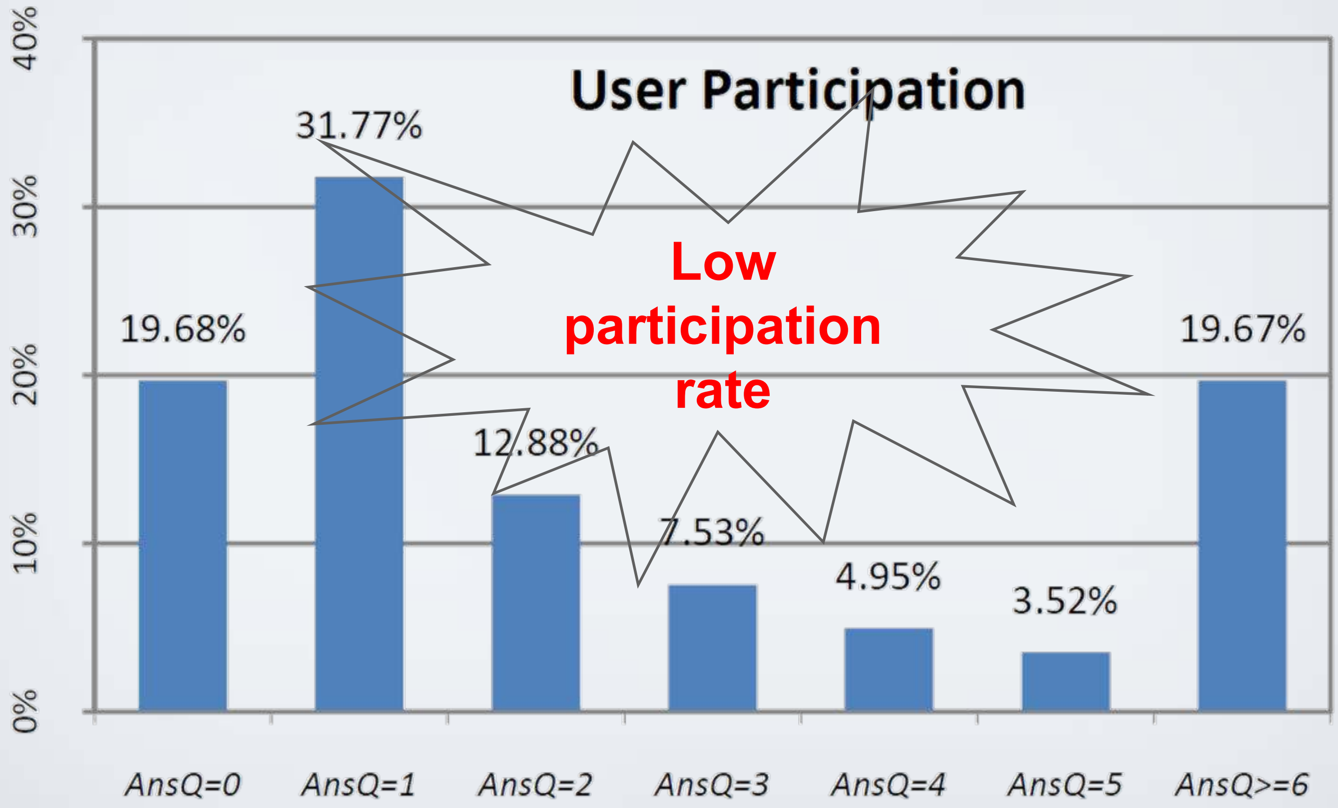
- First to investigate question quality in CQA
- Define question quality in CQA
- Conduct two studies
  - Analyze the factors influencing question quality
  - Propose a mutual reinforcement-based label propagation algorithm to predict question quality

4

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction **(Chapter 3)**
- Question Routing
  - Quality and Availability **(Chapter 4)**
  - Category **(Chapter 5)**
- Question Structuralization **(Chapter 6)**
- Conclusion and Future Work

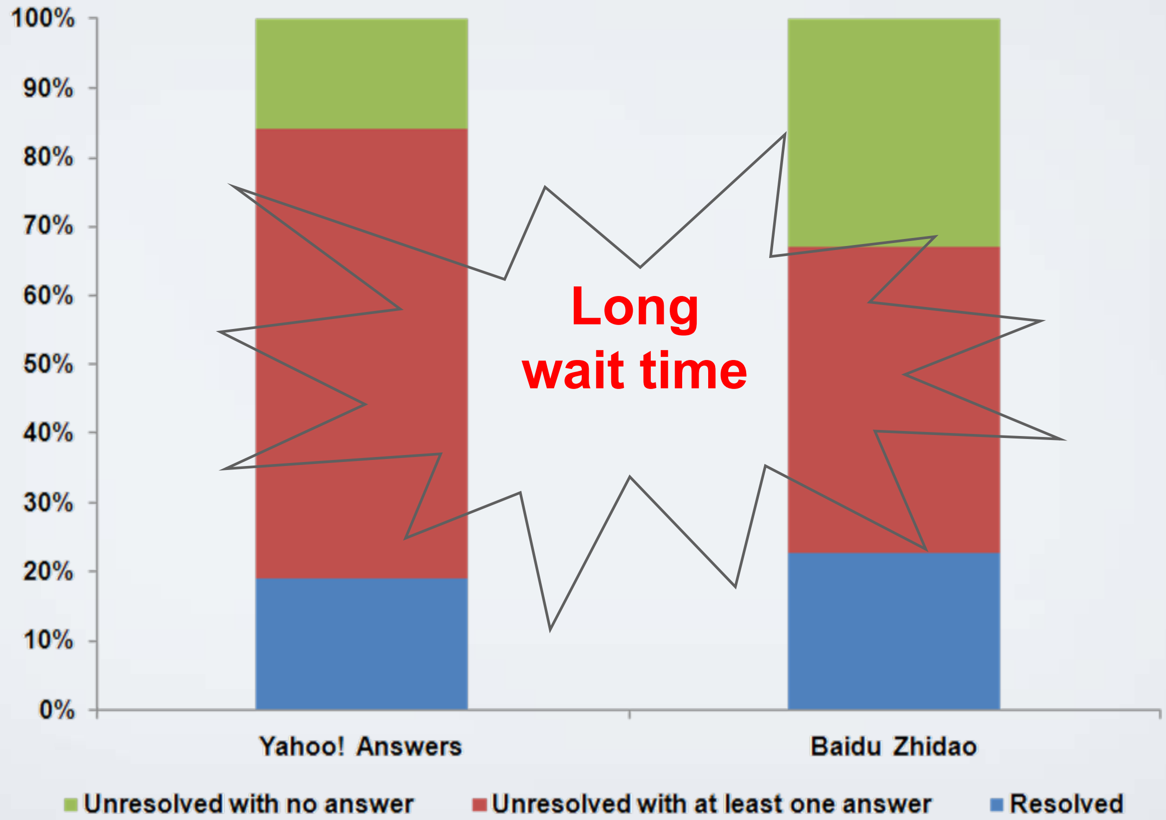
Motivation



User participation in Yahoo! Answers (Guo *et al*., 2008)

4

Motivation



Status of tracked questions in Yahoo! Answers and Baidu Zhidao within 48 hours

4

Question Routing

- Definition
- Framework
  - Expertise Estimation
  - Availability Estimation
- Experiments
- Summary

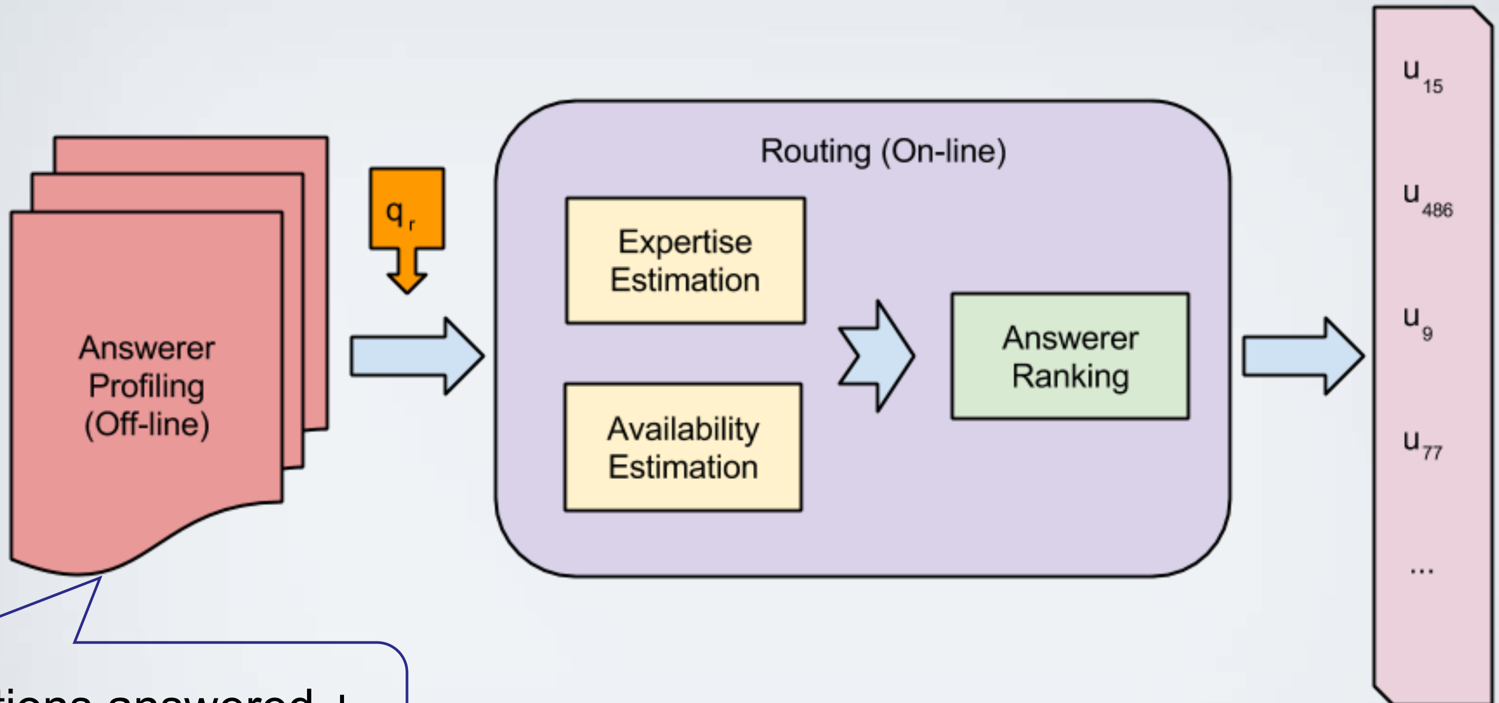
4

Question Routing (QR)

- What is QR?
  - The process of routing a new posted question to the users who are most likely to give **good** answers in a **short** period
- Two requirements
  - Expertise
  - Availability

4

Framework



questions answered + corresponding answers

The framework of Question Routing

$$QR(u_i, q_r, T) = \gamma \cdot E(u_i, q_r) + (1 - \gamma) \cdot A(u_i, T).$$

u_i's expertise on *q_r*

u_i's availability during *T*

4

Expertise Estimation

- Without answer quality
  - Query-likelihood language model

$$E(u_i, q_r) = P(q_r|q_{u_i}) = \prod_{\omega \in q_r} P(\omega|q_{u_i})$$

$$P(\omega|q_{u_i}) = (1-\lambda)P_{ml}(\omega|q_{u_i}) + \lambda P_{ml}(\omega|C)$$

all collection

$$P(\omega|q_{u_i}) = \frac{tf(\omega, q_{u_i})}{\sum_{\omega' \in q_{u_i}} tf(\omega', q_{u_i})}$$

$$P(\omega|C) = \frac{tf(\omega, C)}{\sum_{\omega' \in C} tf(\omega', C)}$$

term frequency of the term ω in q_{u_i}

Expertise Estimation

- With answer quality

quality score

$$E(u_i, q_r) = \alpha \cdot P(q_r|q_{u_i}) + (1-\alpha) \cdot Q(u_i, q_r)$$

- Quality score
  - Basic model
    - Weighted average answer quality of similar questions
  - Smoothed model
    - Leverage other similar users' answer quality of similar questions
  - Quality estimation
    - Logistic regression

	q_1	q_2	q_3	q_4	q_{new}
u_1		0.7			?
u_2		0.5			
u_3	0.9			0.8	
u_4			0.6		

Availability Estimation

- Model it as a trend analysis problem
- Employ an auto-regressive model

$$A(u_i, t) = \lambda_1 A(u_i, t-1) + ... + \lambda_p A(u_i, t-p) + \varepsilon$$

- The answerer u_i's availability for a period of time T

$$A(u_i, T) = 1 - \prod_{j=1}^{s} (1 - A(u_i, t_j))$$

4

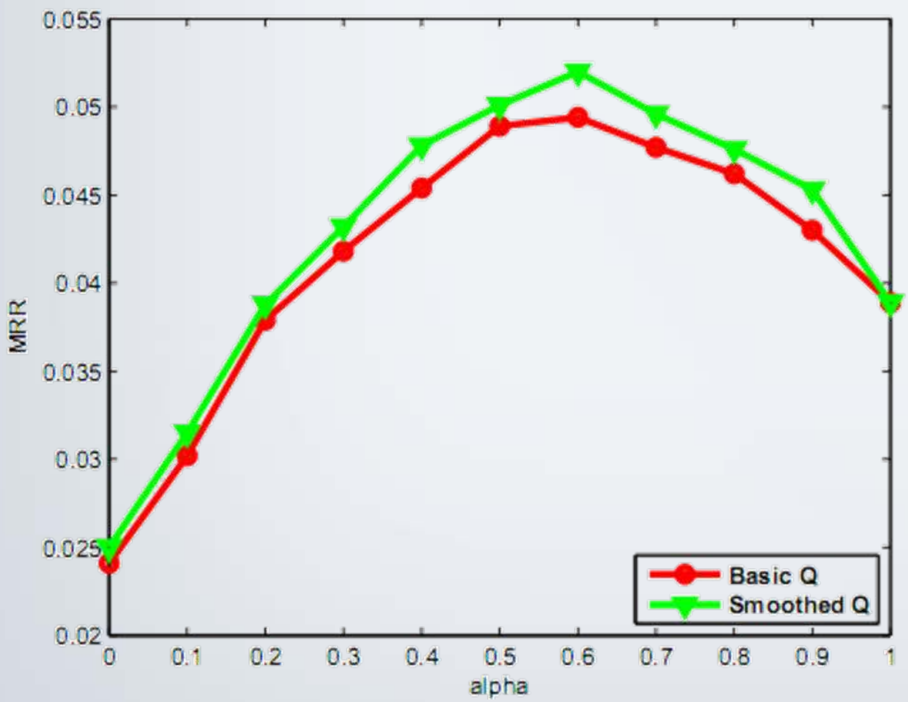
Methods

Method	QR score
QLL	$QR(u_i, q_r, T) = P(q_r \vert q_{u_i})$
Basic Q	$QR(u_i, q_r, T) = \alpha \cdot P(q_r \vert q_{u_i}) + (1-\alpha) \cdot Q_{BM}(u_i, q_r)$
Smoothed Q	$QR(u_i, q_r, T) = \alpha \cdot P(q_r \vert q_{u_i}) + (1-\alpha) \cdot Q_{SM}(u_i, q_r)$
QLL + AE	$QR(u_i, q_r, T) = \gamma \cdot P(q_r \vert q_{u_i}) + (1-\gamma) \cdot A(u_i, T)$
Basic Q + AE	$QR(u_i, q_r, T) = \gamma \cdot [\alpha \cdot P(q_r \vert q_{u_i}) + (1-\alpha) \cdot Q_{BM}(u_i, q_r)] + (1-\gamma) \cdot A(u_i, T)$
Smoothed Q +AE	$QR(u_i, q_r, T) = \gamma \cdot [\alpha \cdot P(q_r \vert q_{u_i}) + (1-\alpha) \cdot Q_{SM}(u_i, q_r)] + (1-\gamma) \cdot A(u_i, T)$

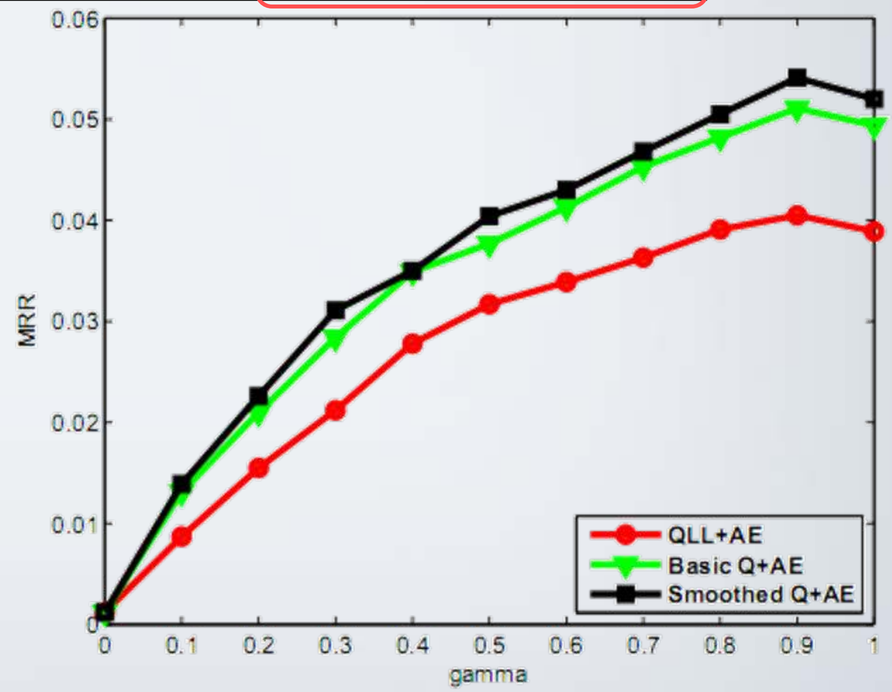
Results

Different methods' MRR for QR

QLL	Basic Q	Smoothed Q	QLL + AE	Basic Q + AE	Smoothed Q + AE
0.0389	0.0494	0.052	0.0405	0.0511	0.0541



MRR value of Basic Q and Smoothed Q versus various α



MRR versus γ across different methods

Contribution of Chapter 4

- Propose a *Question Routing* framework
  - User expertise
  - Answering availability
- Design user expertise estimation and availability estimation models
- Demonstrate the effectiveness of proposed framework

5

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction **(Chapter 3)**
- Question Routing
  - Quality and Availability **(Chapter 4)**
  - Category **(Chapter 5)**
- Question Structuralization **(Chapter 6)**
- Conclusion and Future Work

5

Motivation

- Previous Methods for Expertise Estimation
  - Language Models (Liu *et al.* 2005, Zhou *et al.* 2009)
  - PLSA (Qu *et al.* 2009)
  - LDA + LM (Liu *et al.* 2010)
- Limitations
  - Irrelevant answerers
    - All answerers' expertise is estimated
  - Irrelevant profiles
    - All previous answered questions are employed as user profile

5

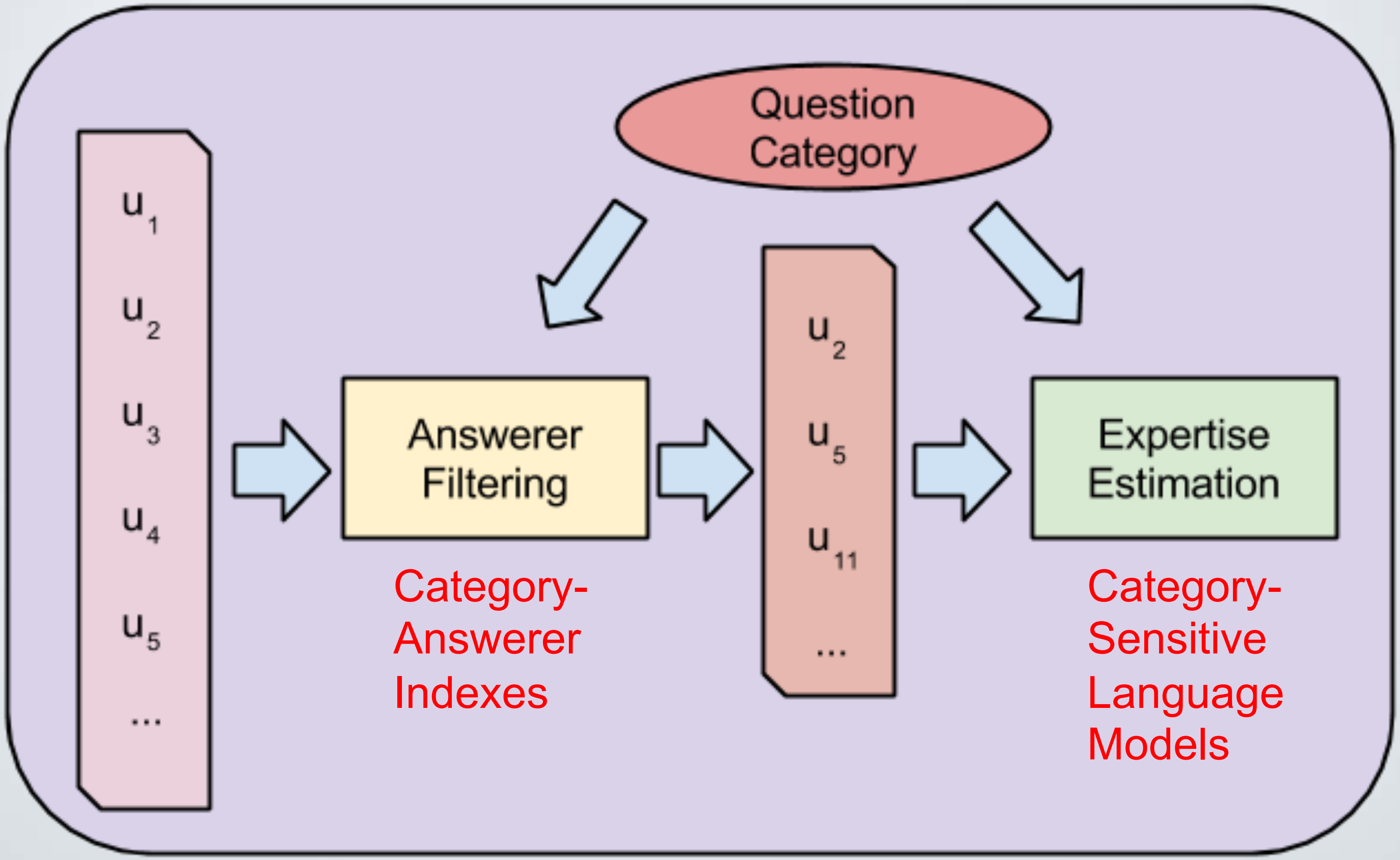
Category Information

- Two improvements in efficiency of QR
  - Higher accuracy
  - Lower cost

5 Category-Sensitive Question Routing

- Category for QR
  - Category-Answerer Indexes
  - Category-Sensitive Language Models
- Experiments
- Summary

Question Category for QR



Category-Answerer Indexes

- Severe index
  - Leaf category-based
- Lenient index
  - Top category-based

5

Category-Sensitive LMs

- Basic category-sensitive QLLM (BCS-LM)
  - Only consider profiles in the new question’s leaf category
- Transferred category-sensitive QLLM (TCS-LM)
  - Incorporate profiles in similar leaf categories

5

BCS-LM

$$E(u_i, q_r, \boxed{c_j}) \equiv P_{bcs}(u_i | q_r, \boxed{c_j}),$$

$$P_{bcs}(u_i | q_r, c_j) \propto P_{bcs}(q_r, c_j | u_i) P(u_i),$$

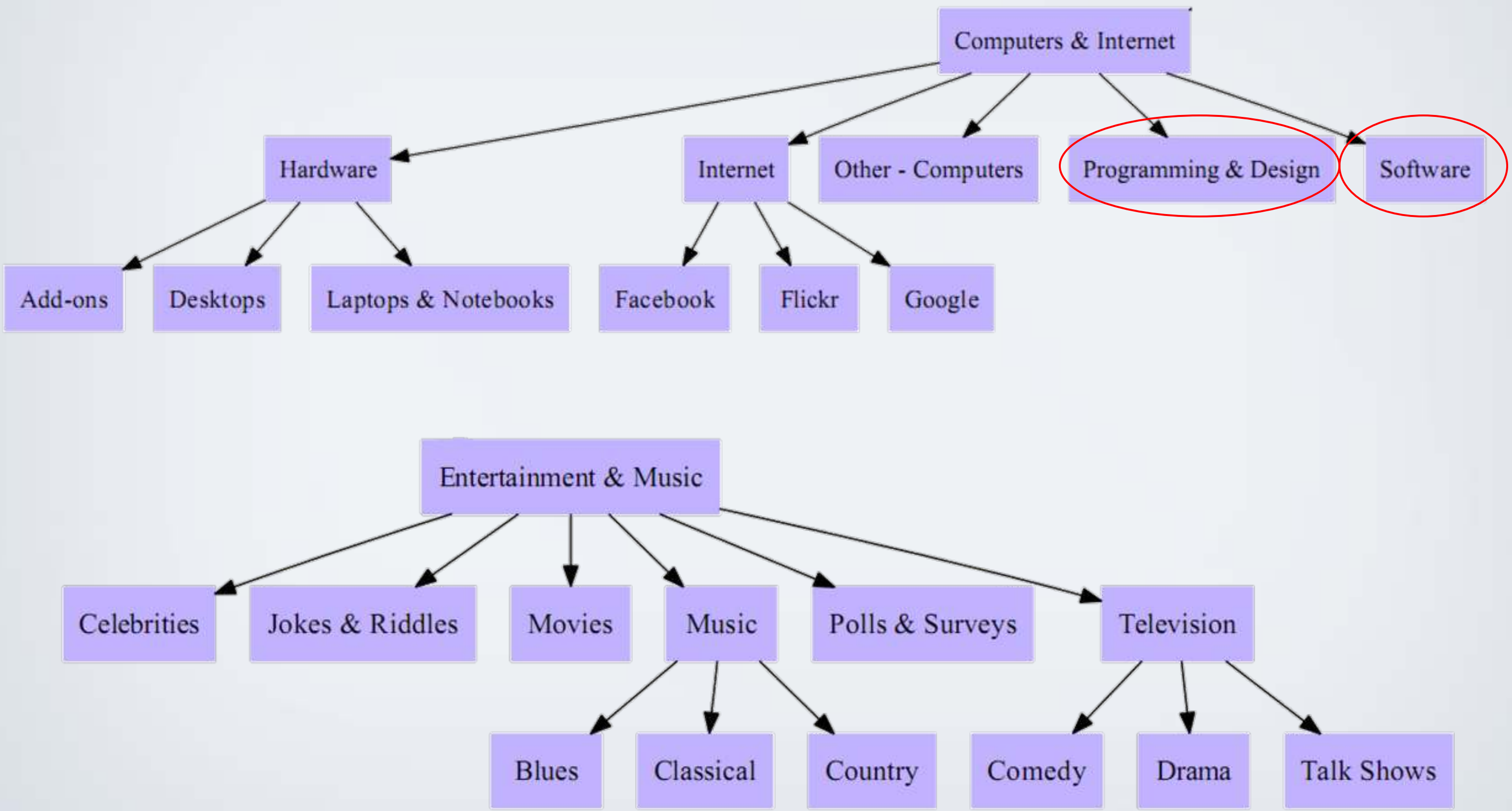
$$P_{bcs}(q_r, c_j | u_i) = P_{bcs}(q_r | c_j, u_i) P(c_j | u_i),$$

$$P_{bcs}(q_r | c_j, u_i) = P_{bcs}(q_r | c_j, q_{u_i}) = \prod_{\omega \in q_r} P(\omega | q_{u_{ij}}),$$

$$P(\omega | q_{u_{ij}}) = (1 - \lambda) P_{ml}(\omega | q_{u_{ij}}) + \lambda P_{ml}(\omega | Coll),$$

where c_j is q_r's category, $P(c_j|u_i)$ denotes the probability of answering questions in c_j for u_i, and $q_{u_{ij}}$ represents the question texts of all previously answered questions in c_j for u_i.

TCS-LM


Computers & Internet
Hardware
Internet
Other - Computers
Programming & Design
Software
Add-ons
Desktops
Laptops & Notebooks
Facebook
Flickr
Google
Entertainment & Music
Celebrities
Jokes & Riddles
Movies
Music
Polls & Surveys
Television
Blues
Classical
Country
Comedy
Drama
Talk Shows

TCS-LM

$$P_{tcs}(q_r, c_j|u_i) = \frac{\beta P_{bcs}(q_r, c_j|u_i) + \sum_{c_k \in Tran(c_j)} \boxed{T(c_k \to c_j)} P_{bcs}(q_r, c_k|u_i)}{\beta + \sum_{c_k \in Tran(c_j)} T(c_k \to c_j)},$$

$$c_k \in Tran(c_j) \ if \ T(c_k \to c_j) \geq \delta$$

$$E = \begin{array}{cc} \textit{Category} & \\ \begin{bmatrix} 2 & 0 & 3 & 4 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 4 & 8 & 4 & 5 & 1 \\ 7 & 0 & 6 & 0 & 2 \\ 0 & 4 & 5 & 6 & 1 \end{bmatrix} & \textit{Answerer} \\ \begin{array}{ccccc} & \mathbf{e}_j & & \mathbf{e}_k & \end{array} & \end{array}$$

$$T_{ans}(c_j \to c_k) = T_{ans}(c_k \to c_j) = \frac{\mathbf{e}_j \cdot \mathbf{e}_k}{|\mathbf{e}_j||\mathbf{e}_k|}.$$

5

Methods for Comparison

- Cluster-Based Language Model (CBLM)

$$P(q_r|u_i) = \sum_{Cluster} \prod_{\omega \in q_r} P(\omega|\theta_{Cluster})^{n(\omega,q_r)} con(Cluster, u)$$

and

$$\begin{aligned} P(\omega|\theta_{Cluster}) &= (1-\lambda)P(\omega|Cluster) + \lambda P(\omega|Coll) \\ con(Cluster, u_i) &= \sum_{\mathbf{qa}} con(\mathbf{qa}, u_i) \\ con(\mathbf{qa}, u_i) &= \frac{\prod_{\omega \in \mathbf{q}} P(\omega|\theta_{\mathbf{a}_{u_i}})}{\sum_{\mathbf{qa}'} \prod_{\omega \in \mathbf{q}'} P(\omega|\theta_{\mathbf{a}'_{u_i}})} \\ P(\omega|\theta_{\mathbf{a}_u}) &= (1-\lambda)P(\omega|\mathbf{a}_u) + \lambda P(\omega|Coll) \end{aligned}$$

- Mixture of LDA and QLLM (LDALM）

$$\begin{aligned} P(q_r|u_i) &= \prod_{\omega \in q_r} P(\omega|\theta_{u_i})^{n(\omega,q_r)} \\ P(\omega|\theta_{u_i}) &= \delta P_{LDA}(\omega|\theta_{u_i}) + (1-\delta)P_{LM}(\omega|\theta_{u_i}) \\ P_{LDA}(\omega|\hat{\theta}, \hat{\phi}, \theta_{u_i}) &= \sum_{z=1}^{Z} P(\omega|z, \hat{\phi})P(z|\hat{\theta}, \theta_{u_i}) \end{aligned}$$

5

Experimental Setting

- Data
  - Crawled from Yahoo! Answers
  - 433,072 questions and 270,043 answerers
- Ground Truth
  - GT-A: Answerers who answered the routed question
  - GT-BA: The answerer who gave the best answer of the routed question
- Evaluation Metrics
  - Precision at K (Prec@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

5

Experimental Results

Table 1 Different methods' $Prec@K$ in QR versus various Ks using GT-A (best results are shown in bold)

K	QLLM	BCS-LM	TCS-LM	LDALM	CBLM
1	0.0795	0.1114 (↑40.13%)	**0.1227** (↑54.34%)	0.0989 (↑24.40%)	0.0000
3	0.1659	**0.2364** (↑42.50%)	0.2340 (↑41.05%)	0.1950 (↑17.54%)	0.0000
5	0.2091	**0.2727** (↑30.42%)	0.2705 (↑29.36%)	0.2455 (↑17.41%)	0.0000
10	0.2705	0.3386 (↑25.18%)	**0.3455** (↑27.73%)	0.3102 (↑14.68%)	0.0000
20	0.3386	0.3909 (↑15.45%)	**0.3932** (↑16.13%)	0.3710 (↑9.57%)	0.0091
40	0.4136	0.4523 (↑9.36%)	**0.4591** (↑11.00%)	0.4392 (↑6.19%)	0.0273
60	0.4477	**0.4818** (↑7.62%)	0.4795 (↑7.10%)	0.4649 (↑3.84%)	0.0545
80	0.4727	**0.4955** (↑4.82%)	0.4909 (↑3.85%)	0.4867 (↑2.96%)	0.0727
100	0.4909	**0.5159** (↑5.09%)	0.5114 (↑4.18%)	0.4979 (↑1.43%)	0.0795

Table 2 Different methods' $Prec@K$ in QR versus various Ks using GT-BA (best results are shown in bold)

K	QLLM	BCS-LM	TCS-LM	LDALM	CBLM
1	0.0568	0.0682 (↑20.07%)	**0.0773** (↑36.09%)	0.0668 (↑17.61%)	0.0000
3	0.1091	**0.1477** (↑35.38%)	0.1409 (↑29.15%)	0.1258 (↑15.31%)	0.0000
5	0.1363	**0.1705** (↑25.09%)	0.1659 (↑21.72%)	0.1655 (↑21.42%)	0.0000
10	0.1705	0.2068 (↑21.29%)	**0.2091** (↑22.58%)	0.1950 (↑14.40%)	0.0000
20	0.2205	**0.2591** (↑17.51%)	0.2523 (↑14.42%)	0.2472 (↑12.11%)	0.0023
40	0.2750	0.3114 (↑13.24%)	**0.3136** (↑14.04%)	0.2891 (↑5.13%)	0.0091
60	0.3023	**0.3386** (↑12.01%)	**0.3386** (↑12.01%)	0.3109 (↑2.84%)	0.0295
80	0.3182	0.3432 (↑7.86%)	**0.3455** (↑8.58%)	0.3225 (↑1.35%)	0.0386
100	0.3364	**0.3614** (↑7.43%)	0.3591 (↑6.75%)	0.3365	0.0386

5

Experimental Results

Table 3 MRR and MAP of various models under GT-A (best results are shown in bold)

Method	MRR	MAP
QLLM	0.1460	0.1070
BCS-QLLM	0.1893 (↑29.66%)	0.1424 (↑33.08%)
TCS-QLLM	**0.1965** (↑34.59%)	**0.1469** (↑37.29%)
LDALM	0.1695 (↑16.10%)	0.1281 (↑19.72%)
CBLM	0.0031	0.0024

Table 4 Different methods' MQRT in QR (in seconds)

QLLM	BCS-QLLM	TCS-QLLM	LDALM	CBLM
10.4271	5.5098	8.9884	16.7689	4.2488

Table 5 Effects of using category-answerer indexes on answerer filtering

Type	Avg. num of potential answerers		Loss of recall
	Before filtering	After filtering	
Severe	243,167	19,235 (↓92.09%)	0.24
Lenient	243,167	137,171 (↓43.59%)	0.14

5

Contribution of Chapter 5

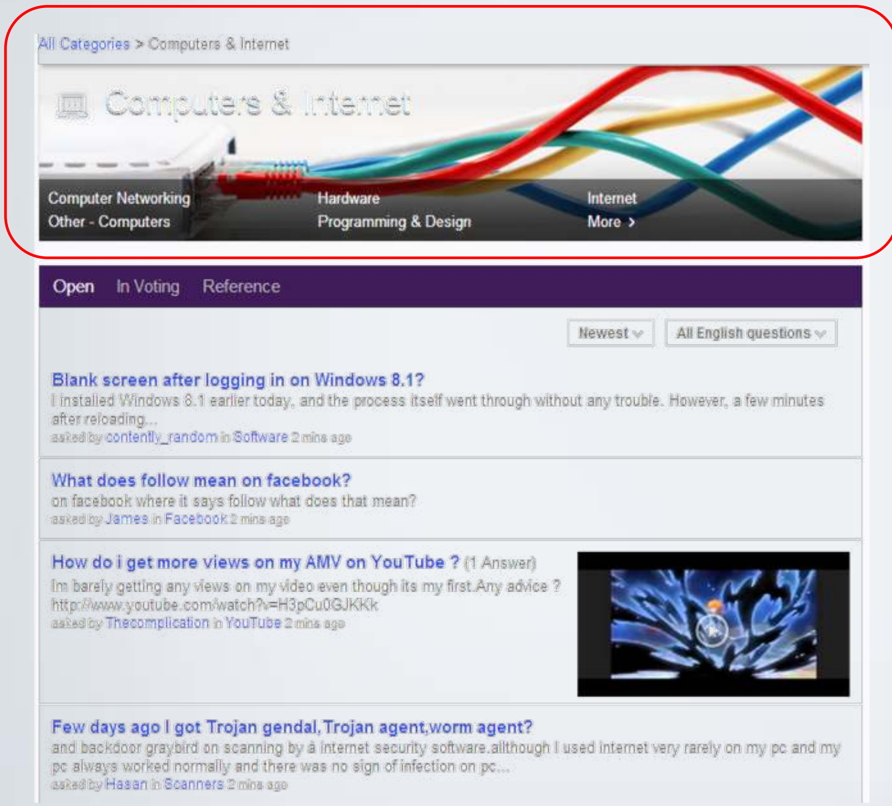
- Propose a novel QR approach which utilizes category information
  - Category-answerer indexes
  - Basic and transferred category-sensitive language models
- Empirical results
  - Much shorter list of candidate answerers
  - More accurate expertise estimation

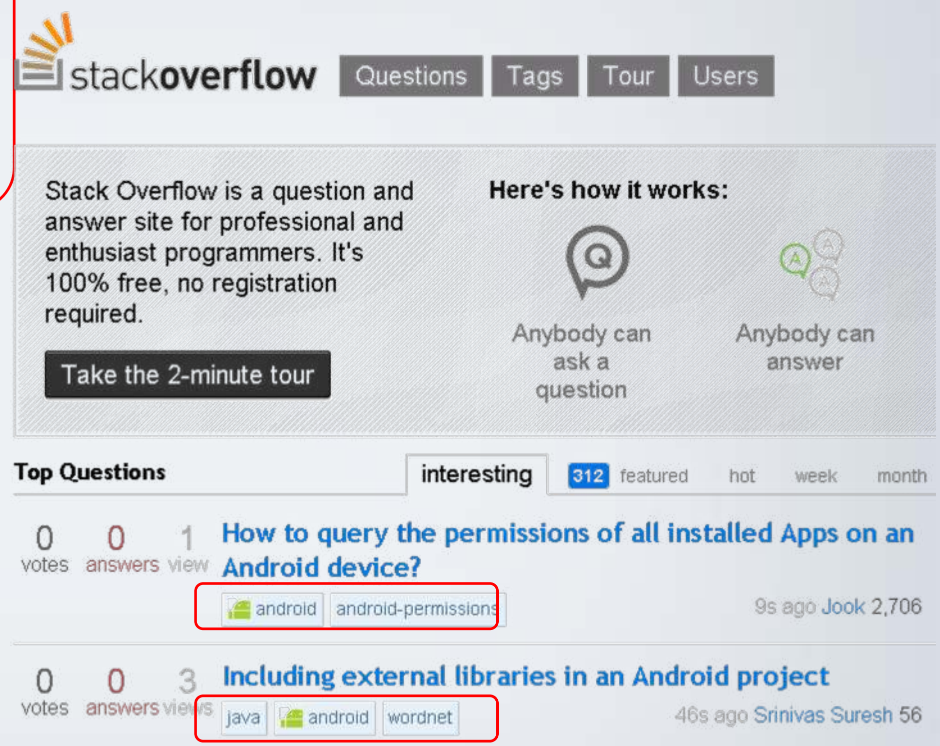
6

Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction **(Chapter 3)**
- Question Routing
  - Quality and Availability **(Chapter 4)**
  - Category **(Chapter 5)**
- Question Structuralization **(Chapter 6)**
- Conclusion and Future Work

Motivation





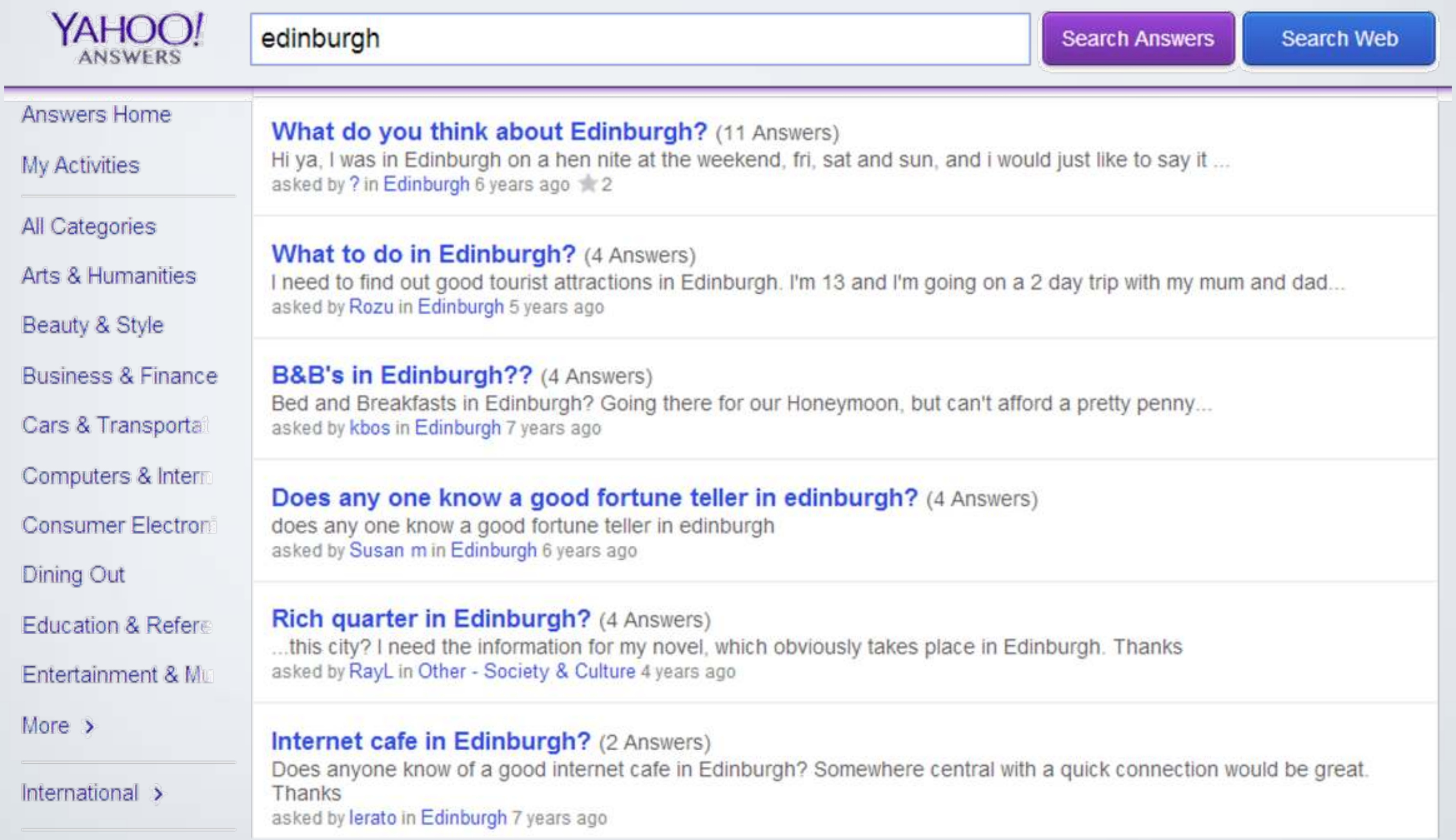
YAHOO! ANSWERS

List structure (with category hierarchy)

List structure (with social tags)

6

Example: Questions about *Edinburgh*



6

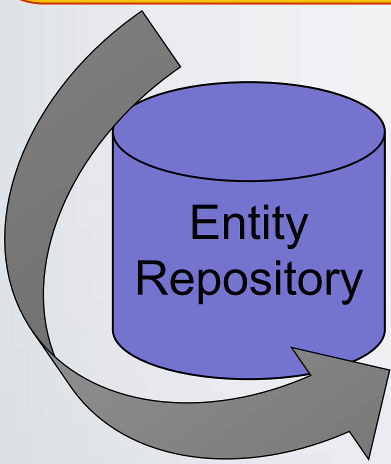
Question Structuralization

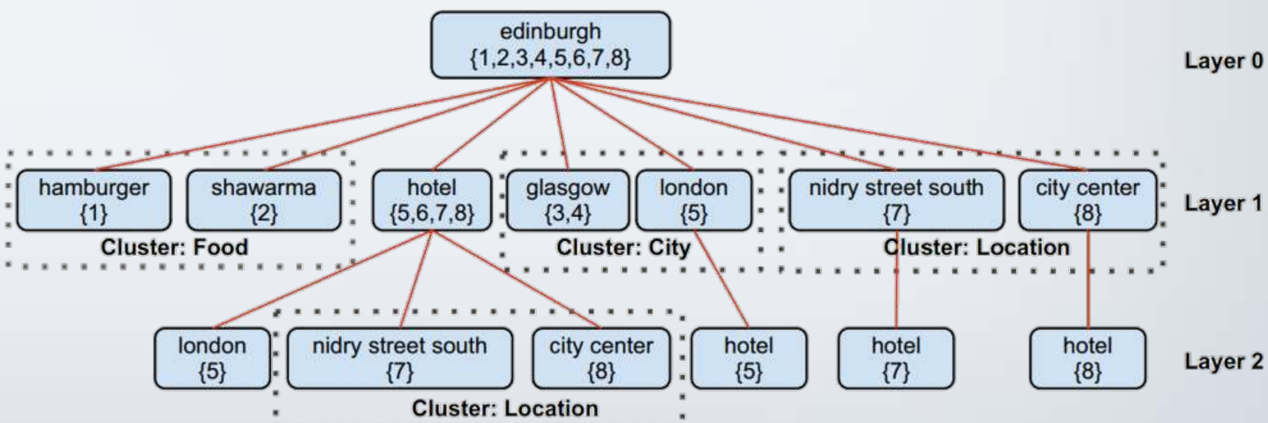
- Introduction to Cluster Entity Tree (CET)
- CET Construction
  - Entity extraction
  - Tree construction
  - Hierarchical entity clustering
- Evaluation
  - User study
  - CET-based question re-ranking
- Summary

6

Structuralize Questions: Cluster Entity Tree (CET)

1. Where can i buy a **hamburger** in **Edinburgh**?
2. Where can I get a **shawarma** in **Edinburgh**?
3. How long does it take to drive between **Glasgow** and **Edinburgh**?
4. Whats the difference between **Glasgow** and **Edinburgh**?
5. Good **hotels** in **London** and **Edinburgh**?
6. Looking for nice , clean cheap **hotel** in **Edinburgh**?
7. Does anyone know of a reasonably cheap **hotel** in **Edinburgh** that is near to **Niddry Street South** ?
8. Who can recommend a affordable **hotel** in **Edinburgh City Center**?





6

Challenges

- Question texts are usually ill-formed
- How to extract named entities with high precision and recall?
- How to efficiently cluster entities?

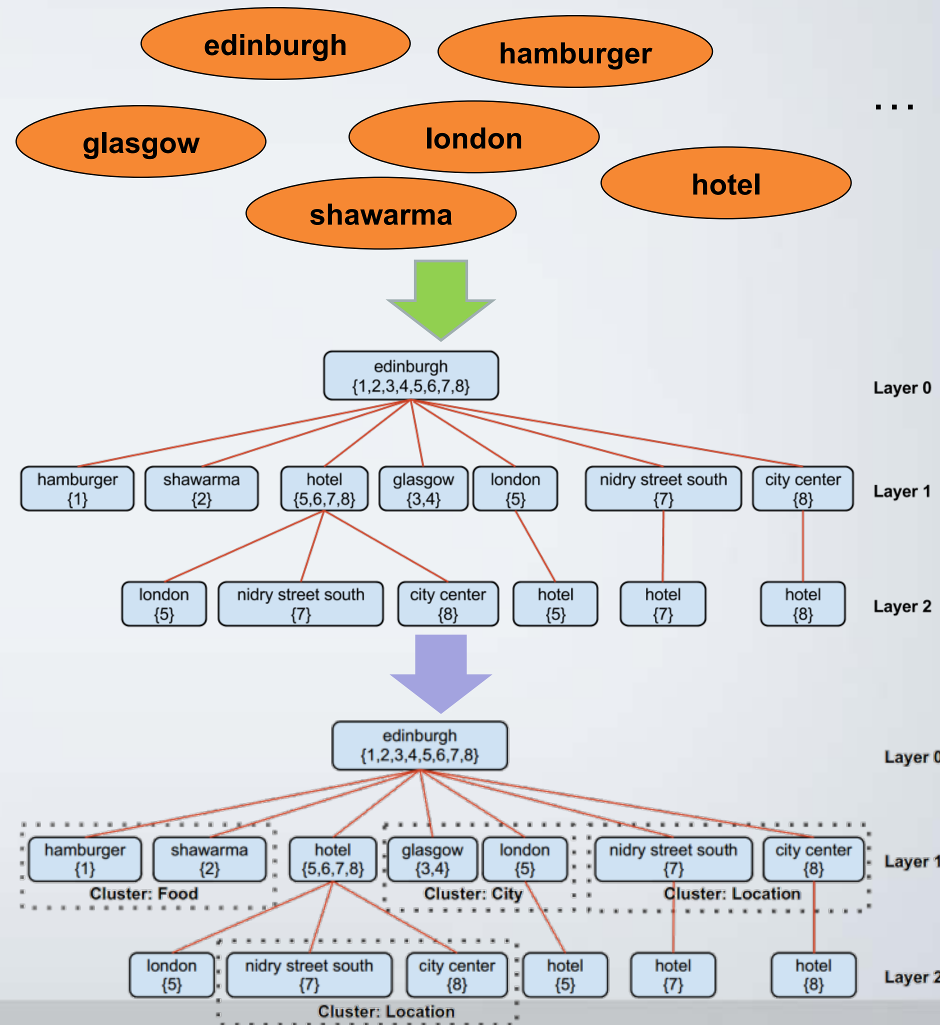
6

CET Construction

Entity Extraction

Tree Construction

Hierarchical Entity Clustering



6

Entity Extraction

- Candidate entity extraction
  - Parse each document to a parse tree
  - Extract all noun phrases, stem
  - Find the noun phrases included in our entity repository (NeedleSeek)
- Entropy-based filtering

total number of categories

$$Entropy(e_i) = -\sum_{c=1}^{|C|} \boxed{P_c(e_i)} log P_c(e_i)$$

$$\frac{number\ of\ e_i\ in\ category\ c}{all\ number\ of\ candidate\ entities\ in\ category\ c}$$

Evaluation

- 520 randomly sampled questions, 20 from each top category of Yahoo! Answers

Method	**Precision**	**Recall**	**F1**
Stanford NER	0.750	0.155	0.257
FIGER (Ling and Weld, 2012)	0.763	0.154	0.256
Freebase	0.644	0.595	0.619
Ours	0.647	0.809	0.719

6

Tree Construction

- Input: an entity and a set of documents
- Output: a hierarchical entity tree with the given entity as the root
- Method
  - Root node: the given entity + ids of documents containing the entity
  - Layer (1): entities that co-occur with the root entity + corresponding doc ids
  - …
  - Layer (n): for each entity on layer (n-1) nodes, all entities that co-occur with it and all its superiors + corresponding doc ids

6

Hierarchical Entity Clustering

- An agglomerative clustering algorithm modified from (Hu et al., 2012)
  - Efficient
  - No need to set the number of clusters
  - Good performance in practice

6

User Study

- 24 CETs from 70,195 questions
- 12 knowledge-learning tasks and 12 question-search tasks
  - A knowledge-learning task asks for some knowledge about an entity from question texts
    - “find the games running on **macbook pro**”
  - A question-search task asks users to find similar questions
    - “questions about who will win the MVP in **NBA** this year”

6

User Study

- 16 participants
- List-based program and CET-based program
- A questionnaire after each task
  - Familiarity
  - Easiness
  - Satisfaction
  - Adequate time
  - Helpfulness
  - Comments

6

User Study Results

	***Knowledge-learning* Tasks**		***Question-search* Tasks**	
	CET-based	List-based	CET-based	List-based
# Queries	2.99	4.47	2.56	3.38
# Answers	8.32	6.06	10.60	10.92
Precision	0.38	0.19	0.40	0.44
Time (secs)	136.44	121.87	103.71	87.75

6

Questionnaire Results

	***Knowledge-learning* Tasks**		***Question-search* Tasks**	
	CET-based	List-based	CET-based	List-based
Familiarity	3.18	3.22	3.07	3.28
Easiness	3.64	3.66	4.10	4.06
Satisfaction	3.70	2.94	3.86	3.44
Enough Time	3.87	3.83	4.44	4.54
Helpfulness	4.16	3.03	4.31	3.71

6 CET-based Question Re-Ranking

- Idea
  - Questions sharing similar topics should be ranked similarly
  - Traditional question retrieval models (Cao et al., 2010) cannot capture key semantics
  - By utilizing CET
    - Entities are given more weight while trivial words are not
    - Questions which are ranked lower will be brought higher by their top-ranked neighbors in the same cluster

6

Problem

Query q: Any **hamburger** to recommend in **Edinburgh** ?

Relevant Questions (Q_q):

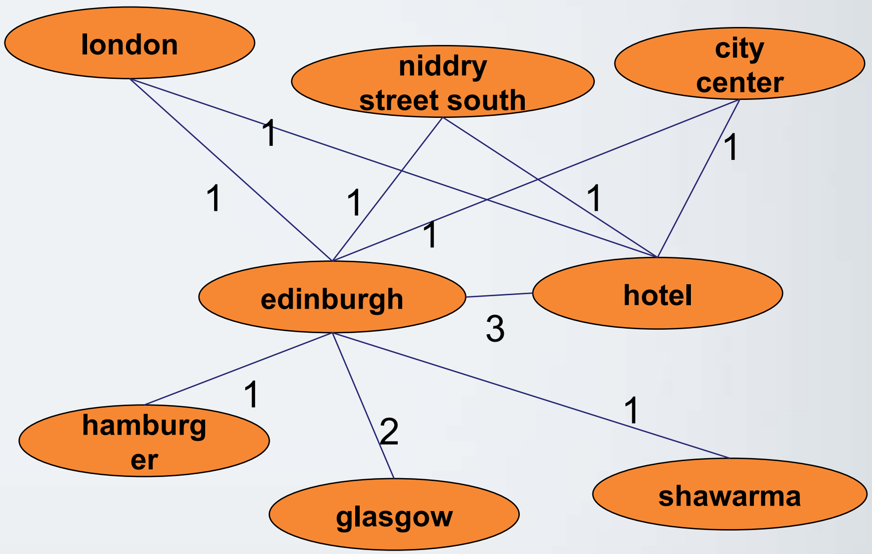
q_1: Any to recommend in **Edinburgh**?
q_2: Can anyone tell me where to buy a **hamburger** in **Edinburgh**?
q_3. Where to get something to eat like **shawarma** in **Edinburgh**? Thank you very much!

6

Step 1: PageRank

Question Collection (Q):

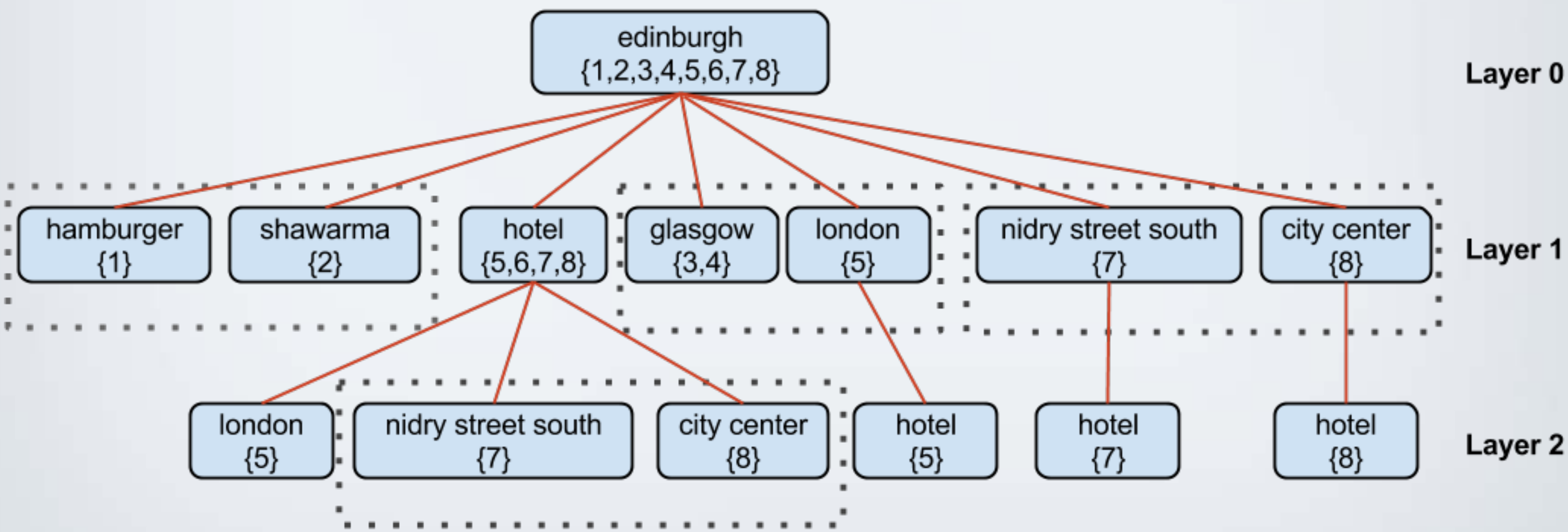
1. Where can i buy a **hamburger** in **Edinburgh**?
2. Where can I get a **shawarma** in **Edinburgh**?
3. How long does it take to drive between **Glasgow** and **Edinburgh**?
4. Whats the difference between **Glasgow** and **Edinburgh**?
5. Good **hotels** in **London** and **Edinburgh**?
6. Looking for nice , clean cheap **hotel** in **Edinburgh**?
7. Does anyone know of a reasonably cheap **hotel** in **Edinburgh** that is near to **Niddry Street South** ?
8. Who can recommend a affordable **hotel** in **Edinburgh City Center**?



6

Step 2: CET Construction

Query q: Any **hamburger** to recommend in **Edinburgh** ?



Step 3: CET-based Question Clustering

Relevant Questions (Q_q):

q_1: Any to recommend in **Edinburgh**?
q_2: Can anyone tell me where to buy a **hamburger** in **Edinburgh**?
q_3. Where to get something to eat like **shawarma** in **Edinburgh**? Thank you very much!

Entity Chains:

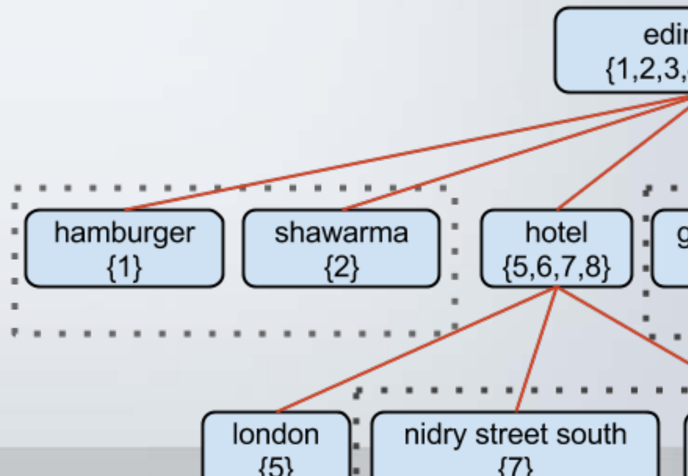
edinburgh X

edinburgh X — **hamburger** √

edinburgh X — **shawarma** √

Cluster 1
q_2 , q_3

Θ
q_1



6

Step 4: Question Re-ranking

Query q: Any **hamburger** to recommend in **Edinburgh** ?

Relevant Questions (Q_q):

q_1: Anything to recommend in **Edinburgh**?
q_2: Can anyone tell me where to buy a **hamburger** in **Edinburgh**?
q_3. Where to get something to eat like **shawarma** in **Edinburgh**? Thank you very much!

Cluster 1

q_2, q_3

$\ominus$

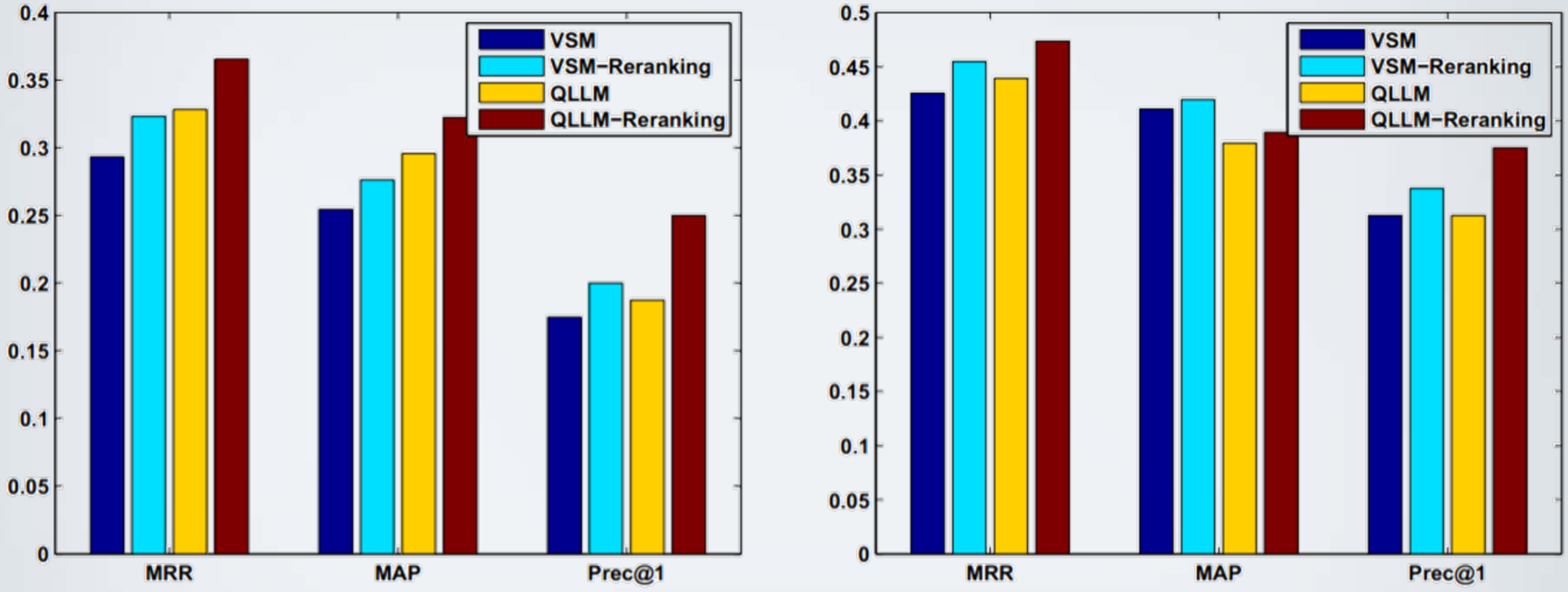
q_1

Re-ranking Results (Q'_q):

q_2: Can anyone tell me where to buy a **hamburger** in **Edinburgh**? (↑)
q_3. Where to get something to eat like **shawarma** in **Edinburgh**? Thank you very much! (↑)
q_1: Anything to recommend in **Edinburgh**?

6

Re-ranking Results



(a) Computer & Internet (b) Travel

6

Contribution of Chapter 6

- Propose a novel hierarchical entity-based approach to structuralize questions in CQA services
- Design a three-step framework to construct CETs and show its effectiveness from empirical results
- Demonstrate the great advantages of our approach in knowledge finding
  - User study (User aspect)
  - Question re-ranking (System aspect)

7

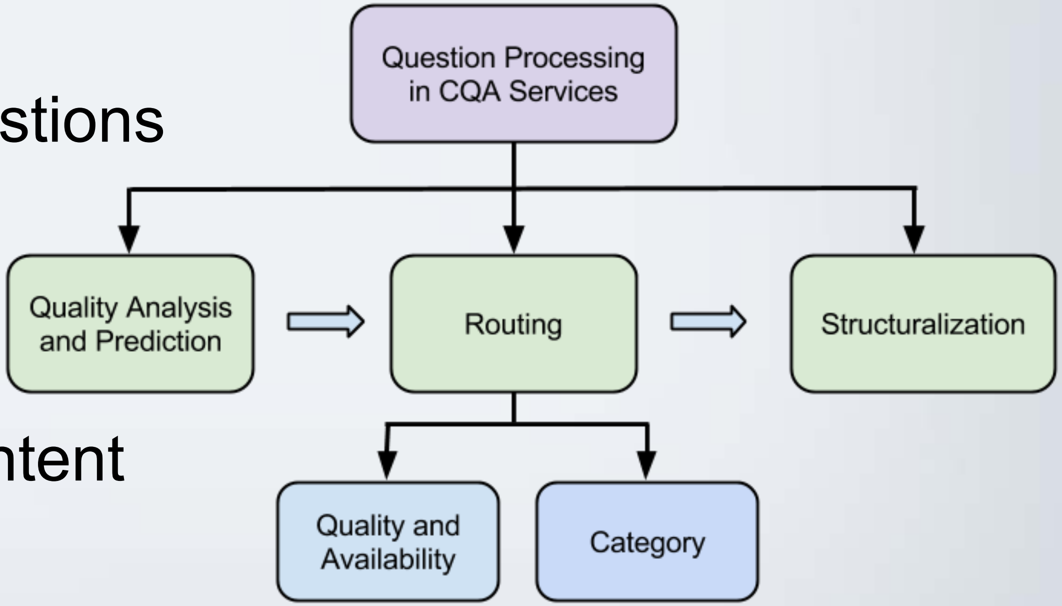
Agenda

- Introduction
- Background
- Question Quality Analysis and Prediction
- Question Routing
  - Quality and Availability
  - Category
- Question Structuralization
- Conclusion and Future Work

7

Conclusion

- A computational framework for question processing in CQA services
  - Facilitate answerers access to proper questions
  - Help askers obtain information more effectively
  - Improve system's content organization & QA efficiency



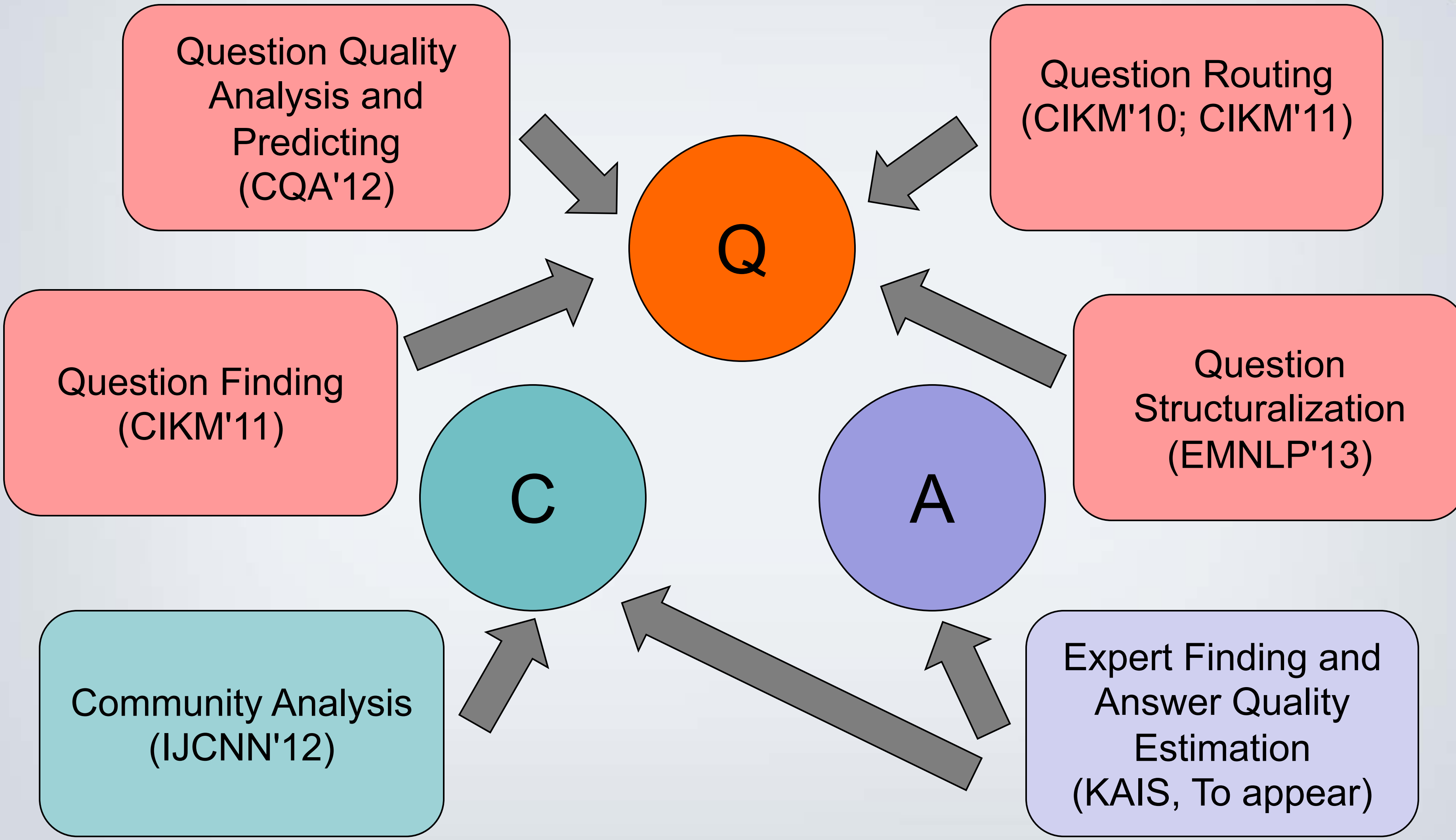
Future Work

- Quality Analysis and Prediction
  - More salient features
  - Question search and recommendation
- Routing
  - Category hierarchy
  - Diversity
- Structuralization
  - Entity normalization
  - Document summarization

THANK YOU!

Questions and comments
are welcome and appreciated.

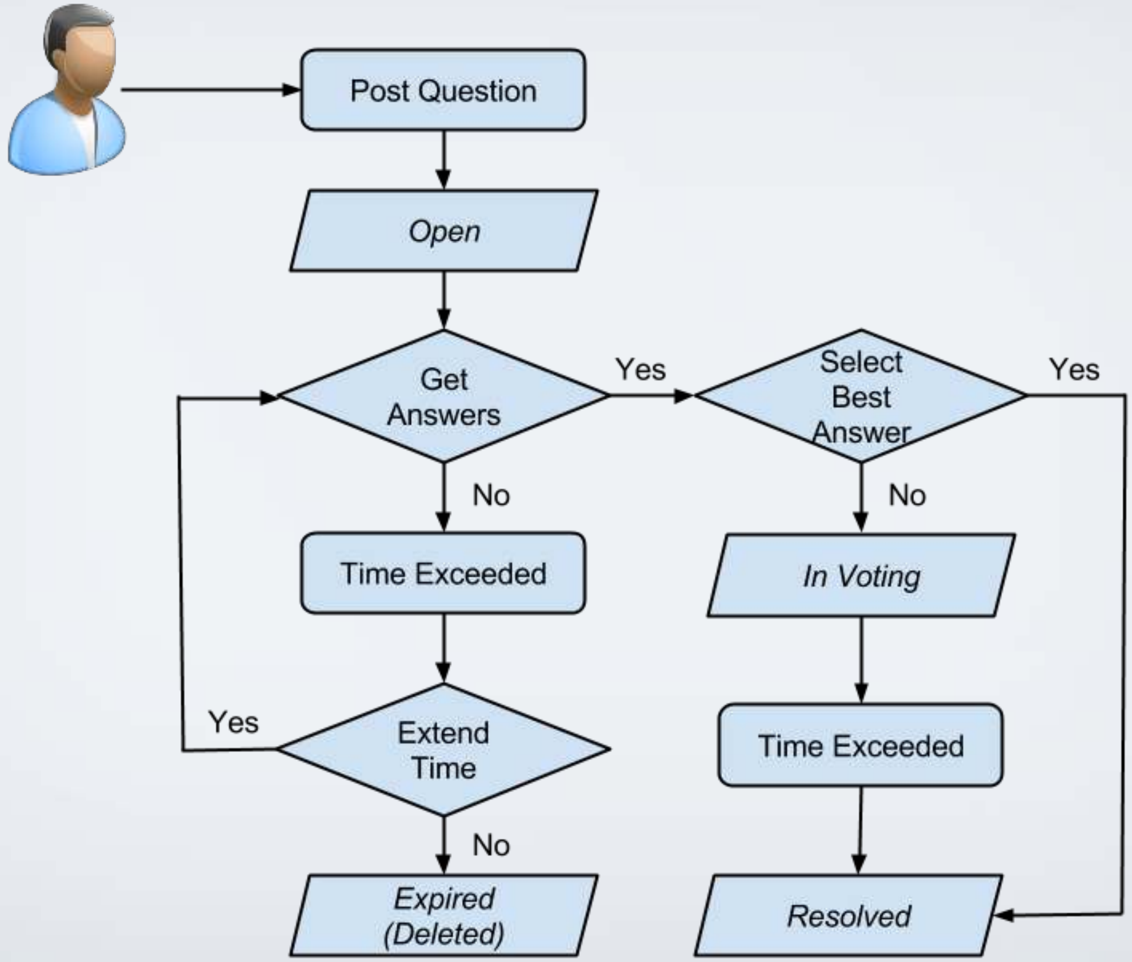
Publications



BACKUP SLIDES (FAQ)

- Chapter 3
- Chapter 4
- Chapter 5
- Chapter 6

A Question's Life in Yahoo! Answers



Back to FAQ

3

Question Analysis and Prediction

- How to set the ground truth of question quality?
- Features
- How to generate user similarity matrix M and question similarity matrix N?
- Why MRLP performs better?
- Why using sensitivity/specificity instead of precison/recall?
- Why the performance of MRLP is still not satisfying? How to improve it in the future?

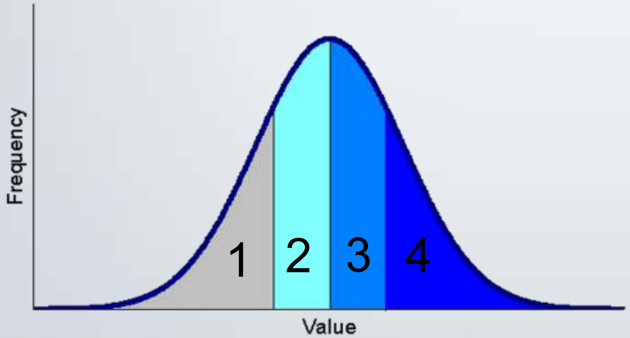
Back to FAQ

Ground Truth Setting

Rules for the ground truth setting

RM \ NTA	4	3	2	1
4	4	4	3	2
3	4	3	3	2
2	3	3	2	1
1	2	2	1	1

NTA: number of tag-of-interests + number of answers
RM: reciprocal of the minutes for getting the best answer



Summary of questions in four levels

Level	1	2	3	4
Count	53,806	62,192	69,836	52,715

Back to FAQ

Features

- Post-Solving features
  - Used for constructing the ground truth
    - Number of tag-of-interests
    - Number of answers
    - The minutes for getting the best answer
- Pre-Solving features
  - Used for predicting question quality
    - User related features: total points, number of questions asked, etc.
    - Question related features: text length, Wh-words, etc.

Back to FAQ

MRLP

Suppose there are m askers who ask n questions in t topics, let $U^1, U^2, ..., U^t$ denote the vectors $(m \times 1)$ of askers' asking expertise in these topics, and $Q(n \times 1)$ denote the vector of question quality, we define a $m \times n$ matrix E, where $e_{ij} = 1 (i \in [1, m], j \in [1, n])$ means u_i asks q_j, otherwise $e_{ij} = 0$. From E we get E':

$$E'_{ij} = \frac{e_{ij}}{\sum_{k=1}^{n} e_{ik}}.$$

n × n probabilistic transition matrix

For the question part of the bipartite graph, we create edges between any two questions within same topics:

$$w(q_i, q_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\lambda_q^2}\right) \quad N_{ij} = P(q_i \to q_j) = \frac{w(q_i, q_j)}{\sum_{k=1}^{n} w(q_i, q_k)}$$

For the asker part of the bipartite graph, we generate the probabilistic transition matrix M similarly.

Back to FAQ

MRLP VS Others

- It models the interaction between askers and topics explicitly
- It captures the mutual reinforcement relationship between asking expertise and question quality

Back to FAQ

Sensitivity & Specificity

- Sensitivity measures the algorithm's ability to identify high-quality questions (=recall)
- Specificity measures the algorithm's ability to identify low-quality questions
- Precision and recall focus on positive instances

		Condition (as determined by "Gold standard")		
		Condition positive	Condition negative	
Test outcome	Test outcome positive	True positive	False positive (Type I error)	Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	False negative (Type II error)	True negative	Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
		Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Accuracy

Back to FAQ

Discussion

- MRLP is more effective in distinguishing high quality questions from low quality ones than state-of-the-art methods
- At present, neither MRLP nor other methods achieves satisfactory performance due to the influence of features

Back to FAQ

Discussion

Table 5: Summary of features extracted from questions and askers

Name	Description	IG
	Question-related features	
Sub_len	Number of words in question subject (title)	0.0115
Con_len	Number of words in question content	0.0029
Wh-type	Whether the question subject starts with Wh-word (e.g., "what", "where", etc.)	0.0001
Sub_punc_den	Number of question subject's punctuation over length	0.0072
Sub_typo_den	Number of question subject's typos over length	0.0021
Sub_space_den	Number of question subject's spaces over length	0.0138
Con_punc_den	Number of question content's punctuation over length	0.0096
Con_typo_den	Number of question content's typos over length	0.0006
Con_space_den	Number of question content's spaces over length	0.0113
Avg_word	Number of words per sentence in question's subject and content	0.0048
Cap_error	The fraction of sentences which are started with a small letter	0.0064
POS_entropy	The entropy of the part-of-speech tags of the question	0.0004
NF_ratio	The fraction of words that are not the top 10 frequent words in the collection	0.0009
	Asker-related features	
Total_points	Total points the asker earns	0.0339
Total_answers	Number of answers the asker provided	0.0436
Best_answers	Number of best answers the asker provided	0.0331
Total_questions	Number of questions the asker provided	0.0339
Resolved_questions	Number of resolved questions asked by the asker	0.0357
Star_received	Number of stars received for all questions	0.0367

- Salient features?
  - User study via crowdsourcing sytems

[Back to FAQ]

Question Routing

- Statistic of tracked data
- Details of the Basic Model and the Smoothed Model for expertise estimation
- Why integrate expertise score and availability score directly?
- Experimental setup
- Impact of **β**

Back to FAQ

Tracked Data

- Many askers cannot get satisfied answers in time

	# resolved questions	# unresolved questions with at least one answer	# unresolved questions without answer
Yahoo! Answers	527	1,820	442
Baidu Zhidao	682	1,325	993

- Answerers have to find questions manually

Back to FAQ

Expertise Estimation

$$sim(a,b) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{T} w_{ai}\cdot w_{bi}}{\sqrt{\sum_{i=1}^{T} w_{ai}^2}\sqrt{\sum_{i=1}^{T} w_{bi}^2}}$$

- Basic Model

$$Q_{BM}(u_i,q_r) = \frac{\sum_{q_j\sim u_i} Q(u_i,q_j)\cdot sim(q_j,q_r)}{\sum_{q_j\sim u_i} sim(q_j,q_r)} \qquad w_{qt} = tf_{t,q}\times\log\frac{N}{df_t}$$

- Smoothed Model

$$Q_{SM}(u_i,q_r) = \beta Q_{BM}(u_i,q_r) + (1-\beta)\frac{\sum_{u_j\in U/u_i}\sum_{q_k\sim u_j} Q(u_j,q_k)\cdot sim(Q_{u_jq_k},Q_{u_iq_r})}{\sum_{u_j\in U/u_i}\sum_{q_k\sim u_j} sim(Q_{u_jq_k},Q_{u_iq_r})}$$

$$sim(Q(u_j,q_k),Q(u_i,q_r)) = \frac{1}{\sqrt{\frac{1}{sim(u_i,u_j)^2}+\frac{1}{sim(q_k,q_r)^2}}}$$

Back to FAQ

Example

	q_1	q_2	q_3	q_4	q_{new}
u_1		0.7			?
u_2		0.5			
u_3	0.9			0.8	
u_4			0.6		

Back to FAQ

Experimental Setup

- Data
  - Yahoo! Answers data (April 6, 2010 - May 14, 2010)
    - Objective: Predict the answerers of the questions posted after May 6, 2010
    - Training set: 17,182 questions, 48,663 answers and 16,298 answerers
    - Testing set: 1,713 questions, 5,403 answers and 2,891 answerers
    - Features: 7 answer-related and 5 user-related features
- Evaluation Metric
  - Mean Reciprocal Rank (MRR)

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{rank_i}$$

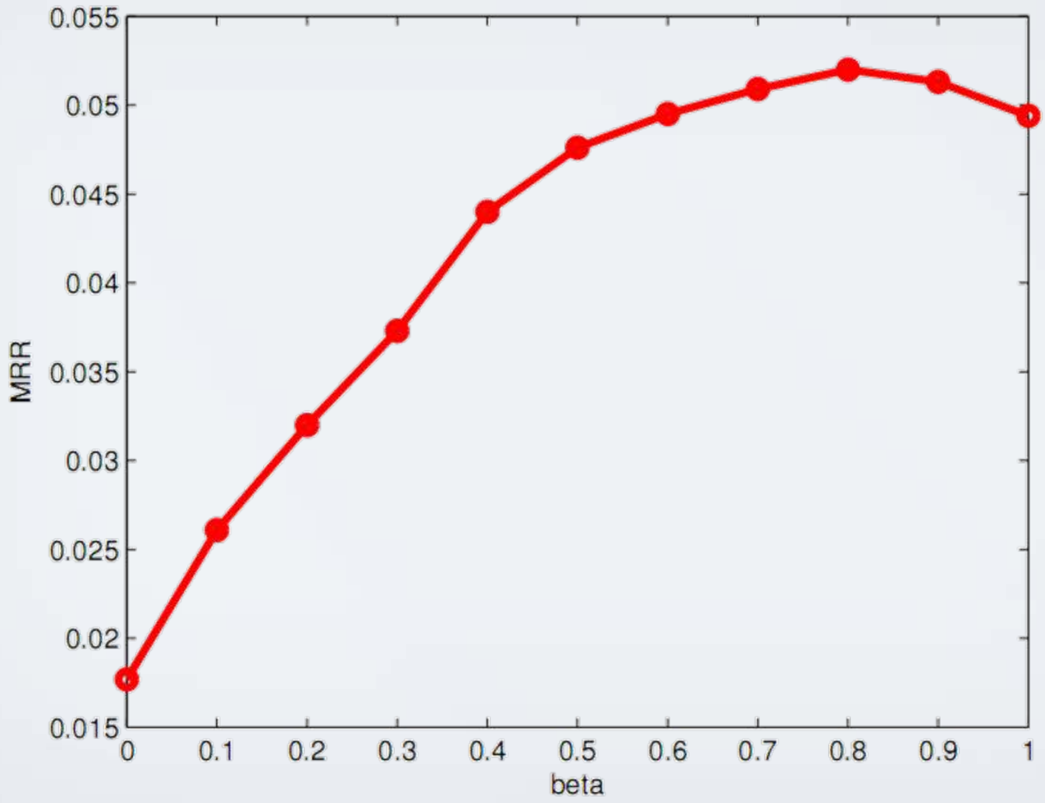
Back to FAQ

Integration of Expertise Score and Availability Score

- High expertise score doesn’t mean high availability score
- An active answerers doesn’t necessary obtain high expertise score (when considering answer quality)
- Expertise and availability are not totally independent

Back to FAQ

Impact of **β**



The MRR value of Smoothed Q versus various β

Back to FAQ

Category-sensitive QR

- Importance of category: an example
- Difference between question routing and question retrieval
- An example of category-answerer indexes
- Impact of user prior (*P(u)*) in language models
- Transferred probabilities between leaf categories
- Impact of δ on TCS-LM (Content VS User)
- LDA
- Data set statistics
- Definitions of evaluation metrics
  - Prec@K
  - MRR
  - MAP

Back to FAQ

One Example

- Alex, a senior Java programmer, is an active answer in Yahoo! Answers. He has answered more than 1,000 questions in terms of **Java programming** as well as 100 questions about **Java coffee**.
- Bob, a cafe manager, is also a frequent user of Yahoo! Answers. He answered around **300 questions about Java coffee**, but he knows little about Java programming.
- Carl, a college student, now asks a question "I met a problem in making Java, any ideas" in "**Food & Drink**" category.

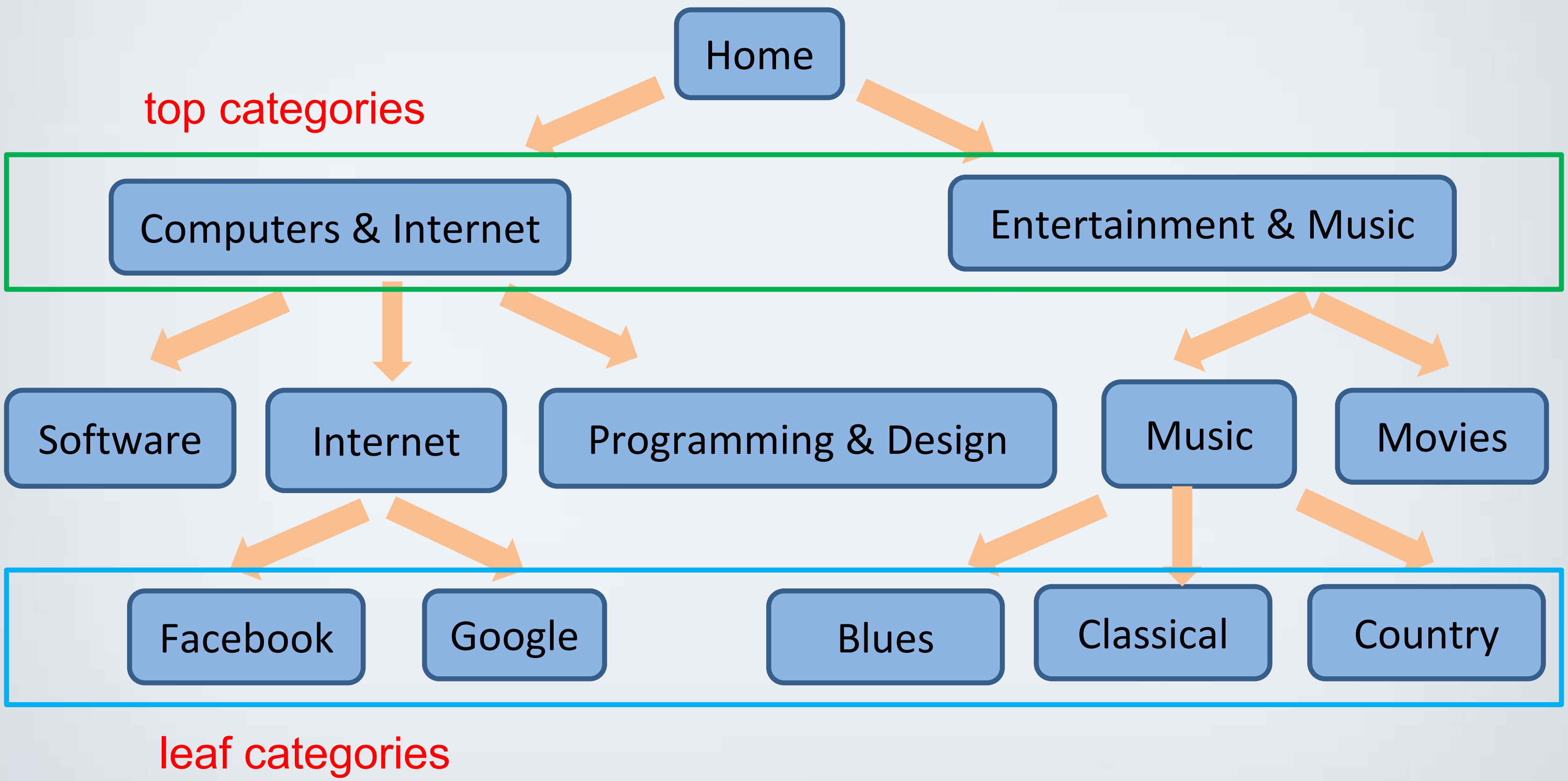
Back to FAQ

Question Routing and Question Retrieval

- Question routing
  - Steps
    - User Profiling
    - Question Profiling
    - Matching
  - Models for user and question profiling
    - Topic Model based, Language Model based, Classification-based, Diversity and Freshness aided, etc.
- Question retrieval
  - Models
    - language model, Translation-based Language model, VSM, BN25, etc.

Back to FAQ

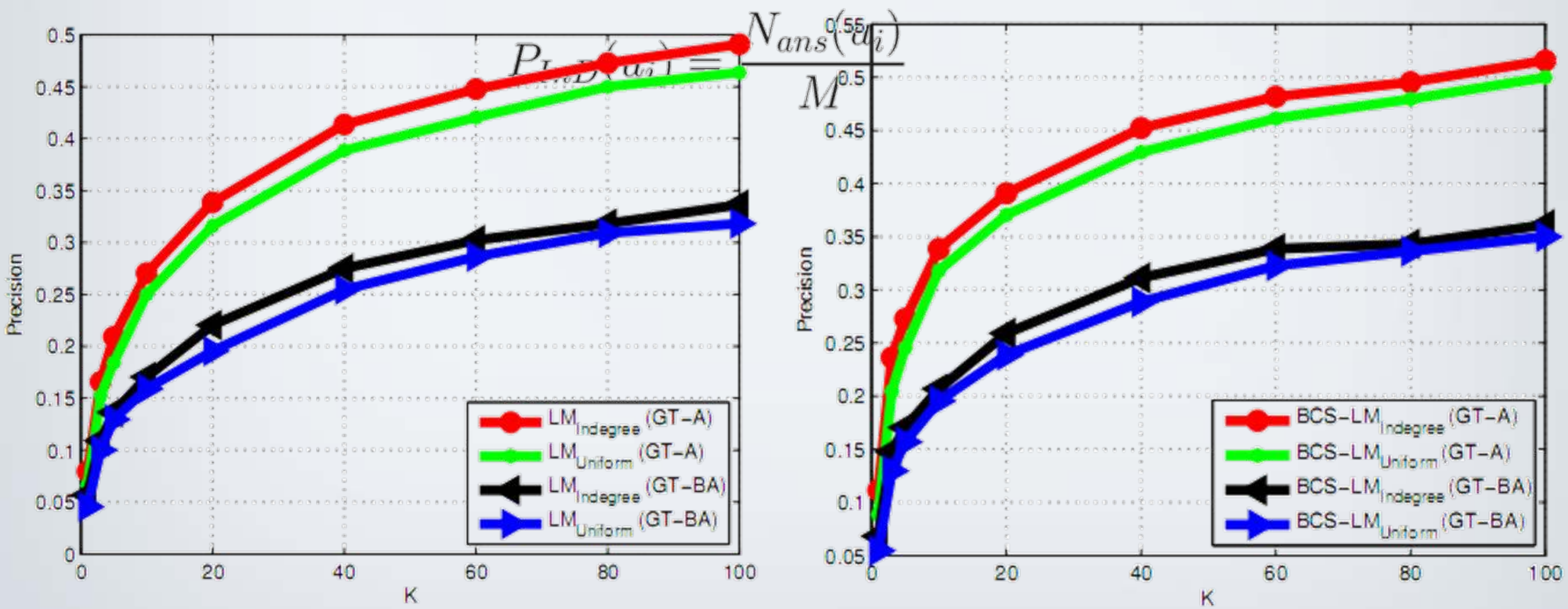
Category-Answerer Indexes



Back to FAQ

Impact of User Prior

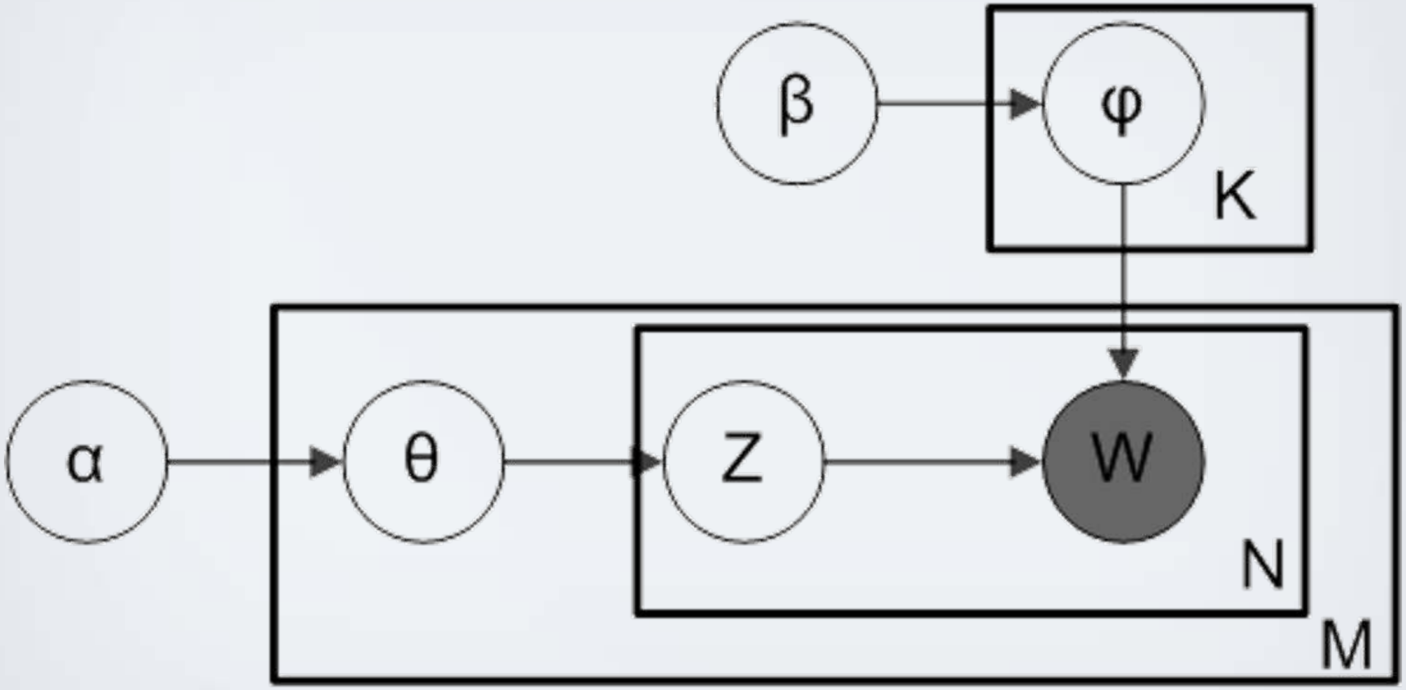
- Uniform distribution (Liu et al., 2004)
- In-degree (Bouguessa et at., 2008)



Prec@K of LM (left) and BCS-LM (right) with different answerer priors

Back to FAQ

Latent Dirichlet Allocation



Back to FAQ

Dataset

Number of questions	433,072
Number of answers	1,510,531
Average number of answers for one question	3.49
Maximum number of answers for one question	50
Mean first reply duration (in minutes)	197.32
Average question length in words (both subject and content)	43.87
Average answer length in words	30.08
Number of askers	240,277
Number of answerers	270,043
Number of both askers and answerers	68,551
Number of askers only	171,726
Number of answerers only	201,492

Back to FAQ

Prec@K

Precision at K (Prec@K): For a set of new questions Q_r, $Prec@K$ reports the fraction of successful QR when top K answerers of the ranking list are returned. The criteria of a successful QR, in the present study, is defined as at least one answerer in the top K of the ranking list actually answered the routed question. In this metric, the position of these users is not considered. The only key factor is whether there is at least one user in these K candidates who answered the routed question. $Prec@K$ is calculated as:

$$Prec@K = \frac{\sum_{q_r \in Q_r} S(q_r, K)}{|Q_r|},$$

$$S(q_r, K) = \begin{cases} 1, & if\, QR\ for\ q_r\ is\ successful; \\ 0, & otherwise. \end{cases}$$

Back to FAQ

MRR

Mean Reciprocal Rank (MRR): The reciprocal rank for an individual question q_r is the reciprocal of the rank at which the first user in the ranking list who actually answered q_r, or 0 if none of the users in the list answered q_r. The MRR value for a set of questions is the mean of each question's reciprocal rank. It is defined as:

$$MRR = \frac{1}{|Q_r|} \sum_{q_r \in Q_r} \frac{1}{Rank(q_r)},$$

where Q_r is a set of questions to be routed, $Rank(q_r)$ is the rank of the first user who actually answered q_r in the ranking list.

Back to FAQ

MAP

Mean Average Precision (MAP): For a set of new questions Q_r, MAP measures the mean of the average precision for each question q_r in QR:

$$MAP = \frac{\sum\limits_{q_r \in Q_r} AvgP(q_r)}{|Q_r|},$$

$$AvgP(q_r) = \frac{\sum_{k=1}^{N_r} (P_r(k) \cdot IsAns(k))}{NRA_r},$$

$$P_r(k) = \frac{NRA_r(k)}{k},$$

where Q_r is a set of questions to be routed, N_r is the number of potential answerers for q_r generated from answerer filtering, NRA_r is the number of real answerers for q_r, $IsAns(k)$ is a binary function to denote whether the k_{th} answerer actually answered q_r, and $NRA_r(k)$ denote the number of real answerers in top k ranked answerers for q_r.

Back to FAQ

Transferred Probabilities (Example)

Table 2 Transferred probabilities between partial leaf categories (answerer-based method)

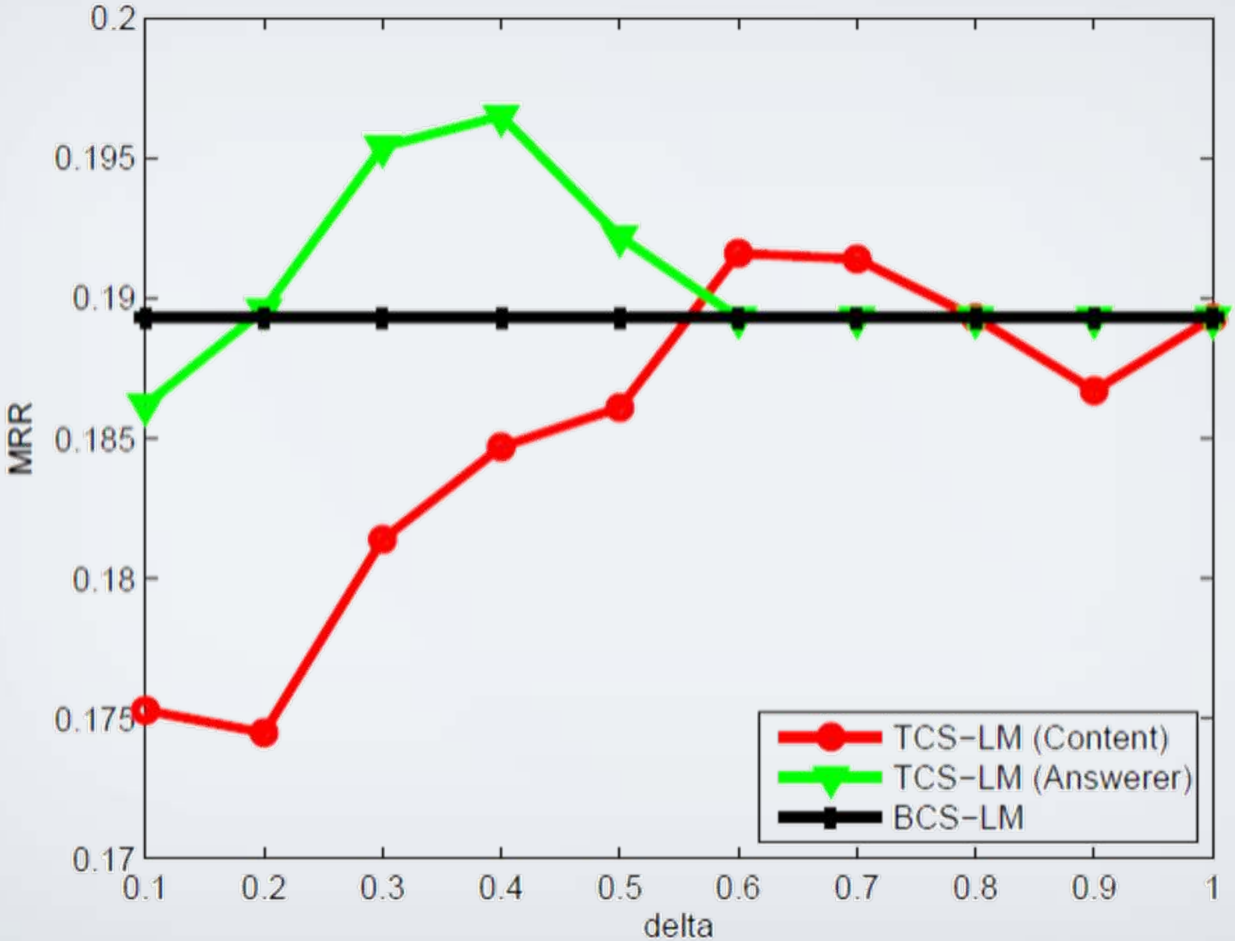
From \ To	Software	Printers	Comedy	Lyrics
Programming & Design	0.2975	0.0251	0.0026	0.0026
Scanners	0.1158	0.5604	0.0014	0.0008
Drama	0.0053	0.0006	0.2593	0.0137
Other - Music	0.0102	0.0019	0.0273	0.1683

Table 3 Transferred probabilities between partial leaf categories (content-based method)

From \ To	Software	Printers	Comedy	Lyrics
Programming & Design	0.2250	0.0236	0.0116	0.0116
Scanners	0.1676	0.2671	0.0049	0.0034
Drama	0.0136	0.0020	0.5481	0.0376
Other - Music	0.0443	0.0070	0.0748	0.2922

Back to FAQ

Impact of δ



MRR for TCS-LM using answerer-based and content-based approaches to estimate transferring probability under GT-A

Back to FAQ

Question Structuralization

- Why adopt entity-based approach for question structuralization?
- Definitions of ER and CET
- Tree construction example
- Detail of clustering algorithm
- What is the similarity function for clustering?
- How to evaluate the clustering results?
- Detail of category mapping
- Definition of B-Cubed Metrics
- What is the usage of Set EC?
- Program interface
- User study tasks

Back to FAQ

Structuralize Questions: Review

- Predefined category hierarchy
  - Coarse grained
  - Hard to maintain
- Topic models
  - Not trivial to control the granularity of topics (Chen et al., 2011).
  - Interpretation problem
- Social tagging
  - Not widely applicable
  - Sparsity (Shepitsen et al., 2008)

Back to FAQ

Advantages of CET

- CET avoids the granularity, interpretation, and sparsity problems by utilizing a large-scale entity repository
  - Entity repository contains millions of named entities on various topics
  - Usually give descriptions of entities
- Automatically build semantic hierarchy
  - Flexible & easy to maintain

Back to FAQ

Definitions

- Entity repository
  - ER = {R, g}
  - R is a set of named entities
  - g is a mapping function that defines the similarity of any two entities

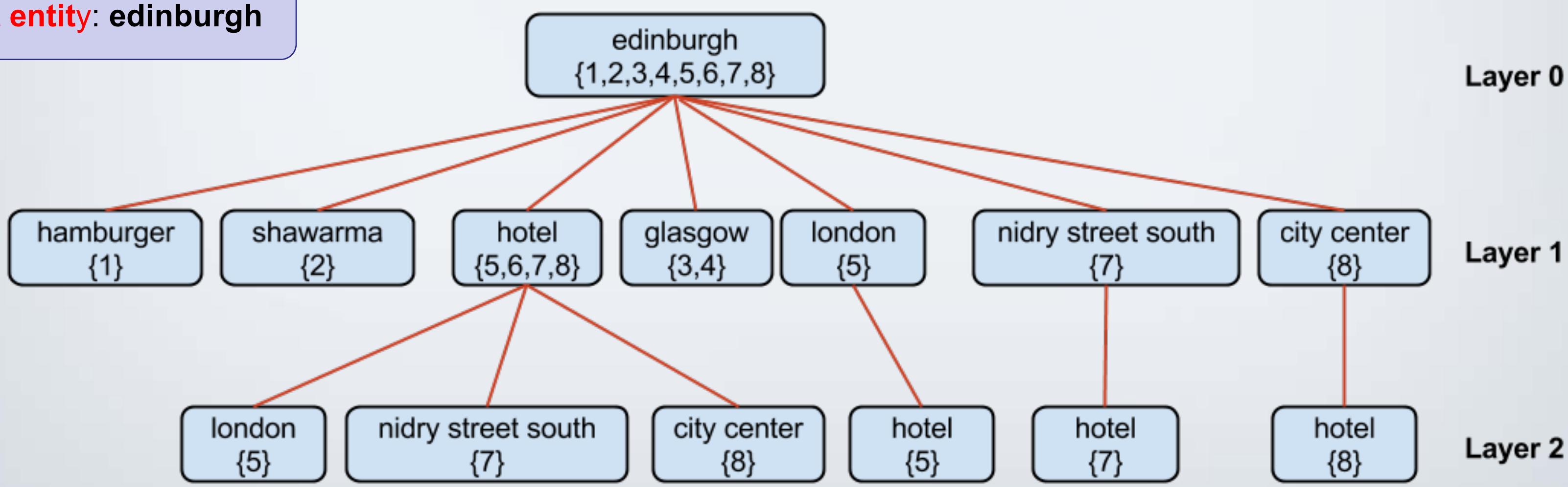
- Cluster Entity Tree (CET)
  - $CET_e = (v_e, V, E, C)$ is a tree structure
  - Each node $v_s \in V$ on CET_e includes
    - An entity extracted from the set of documents $D_e \in D$ containing e
    - A list L(s) which stores the indexes of documents containing entity *s* and its superior entities
  - If v_s is v_t's parent node, entity *t* must co-occur with *s* and *s*'s all superior entities at least once
  - Each $c \in C$ includes a set of similar nodes which share the same parent node

Back to FAQ

Tree Constuction Example

1. Where can i buy a **hamburger** in **Edinburgh**?
2. Where can I get a **shawarma** in **Edinburgh**?
3. How long does it take to drive between **Glasgow** and **Edinburgh**?
4. Whats the difference between **Glasgow** and **Edinburgh**?
5. Good **hotels** in **London** and **Edinburgh**?
6. Looking for nice , clean cheap **hotel** in **Edinburgh**?
7. Does anyone know of a reasonably cheap **hotel** in **Edinburgh** that is near to **Niddry Street South** ?
8. Who can recommend a affordable **hotel** in **Edinburgh City Center**?

Root entity: **edinburgh**



Back to FAQ

6

Modified Agglomerative Clustering

Input: a set of entities with the same parent
Output: clusters of entities

- Select one entity and create a new cluster which contains the entity
- Select the next entity e_i, calculate the similarity between the entity and all existing clusters
- Find $\arg\max_{c} sim(e_i, c),\ s.t.\ sim(e_i, c) > \theta$; otherwise, create a new cluster with e_i as the element
- Stop when all entities are clustered

Back to FAQ

Hierarchical Entity Clustering: Similarity Function

- Follow the approach in (Shi et al., 2010)
  - First-order co-occurrence: Pattern-based (PB)
  - Second-order co-occurrence: Distributional similarity (DS)
- PB
  - The set of terms extracted by applying a pattern one time is called a raw semantic class (RASC)
  - Given two entities a and b, calculate their similarity based on the number of RASCs containing both of them
- DS
  - Terms appearing in similar contexts tend to be similar
  - Given two entities a and b, calculate the similarity between their corresponding context feature vectors

If at least one entity is proper noun, PB is employed; otherwise DS is used.

Back to FAQ

PB

- Some well-designed patterns are leveraged to extract similar entities from a huge repository of webpages. The set of term s extracted by applying a pattern one time is called a raw semantic class (RASC)
- Given two entities t_a and t_b, PB calculates their similarity based on the number of RASCs containing both of them (Zhang et al. , 2009)

$$Sim(t_a, t_b) = \log(1 + \sum_{i=1}^{r_{ab}} P_{ab_i})) \cdot \sqrt{idf(t_a) \cdot idf(t_b)},$$

where $idf(t_a) = \log(1 + \frac{N}{C(t_a)})$, P_{ab_i} is a pattern which can generate RASC(s) containing both term t_a and term t_b, r_{ab} is the total number of such patterns, N is the total number of RASCs, and $C(t_a)$ is the number of RASCs containing t_a.

$$Sim_{PB}(t_a, t_b) = \frac{\log Sim(t_a, t_b)}{2 \log Sim(t_a, t_a)} + \frac{\log Sim(t_a, t_b)}{2 \log Sim(t_b, t_b)}$$

Back to FAQ

DS

- A term is represented by a feature vector, with each feature corresponding to a context in which the term appears
- The similarity between two terms is computed as the similarity between their corresponding feature vectors. Jaccard similarity is employed to estimate the similarity between two terms
- Suppose the feature vectors of t_a and t_b are **x** and **y** respectively:

$$Sim_{DS}(t_a, t_b) = \frac{\sum_i \min(x_i, y_i)}{\sum_i (x_i) + \sum_i (y_i) - \sum_i \min(x_i, y_i)}$$

Back to FAQ

Clustering Evaluation

- 8M questions from 4 top categories of Yahoo! Answers
- Ground truth setting
  - Map categories among YA and Freebase
  - Extract entities which appear exactly once in the corresponding Freebase categories
  - Attach each entity with a unique Freebase category label
- Three approaches
  - AC-MAX, AC-MIN, and AC-AVG
  - AC-MAX performs the best (F1 > 0.75)

Back to FAQ

Clustering Evaluation

Clustering results using AC-MAX (θ_{max}=0.1)

Level	*Travel*				*Cars & Transportation*				*Computer & Internet*				*Sports*			
	Count	P	R	F1	Count	P	R	F1	Count	P	R	F1	Count	P	R	F1
1	748	0.972	0.653	0.743	1281	0.948	0.868	0.897	3064	0.913	0.664	0.743	890	0.941	0.883	0.901
2	200	0.974	0.730	0.798	1202	0.989	0.956	0.965	11344	0.961	0.842	0.879	636	0.978	0.964	0.963
3	120	1.000	0.833	0.890	858	1.000	0.981	0.988	8184	0.978	0.899	0.920	492	0.965	0.882	0.899
4	NA	NA	NA	NA	1776	1.000	0.980	0.986	3648	0.990	0.908	0.934	1080	0.978	0.844	0.881
5	NA	NA	NA	NA	NA	NA	NA	NA	2520	1.000	0.952	0.968	NA	NA	NA	NA
Total	1068	0.976	0.688	0.770	5117	0.984	0.946	0.959	28760	0.968	0.857	0.891	3098	0.965	0.886	0.907

Back to FAQ

Category Mapping

- Goal: automatically evaluate clustering
  - Each entity is attached with a unique Freebase category label
- Two experts are asked to conduct category mapping from Yahoo! Answers to Freebase

Yahoo! Answers	FreeBase
Cars & Transportation	Aviation, Transportation, Boats Spaceflight, Automotive, Bicycles, Rail
Computers & Internet	Computer, Internet
Sports	Soccer, Olympics,Sports, American football, Baseball,Basketball,Ice Hockey,Martial Arts, Cricket,Tennis,Boxing,Skiing
Travel	Travel, Location, Transportation

Back to FAQ

Set EC

Category	Number of Questions	Number of Entities
Cars & Transportation	1,220,427	3,267,596
Computers & Internet	2,912,280	7,324,655
Sports	2,363,758	6,230,868
Travel	1,347,801	3,728,286

Back to FAQ

B-Cubed Metrics

- B-Cubed precision of an item is the proportion of items in its cluster which have the item's category (including itself)
- The overall B-Cubed precision is the averaged precision of all items

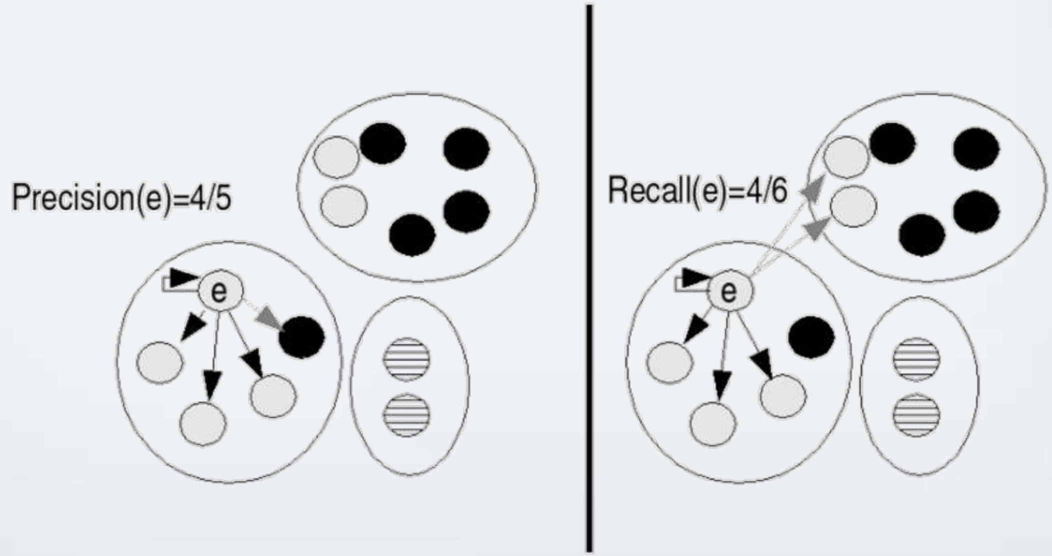


Figure : Example of computing the BCubed precision and recall for one item

Back to FAQ

Interface

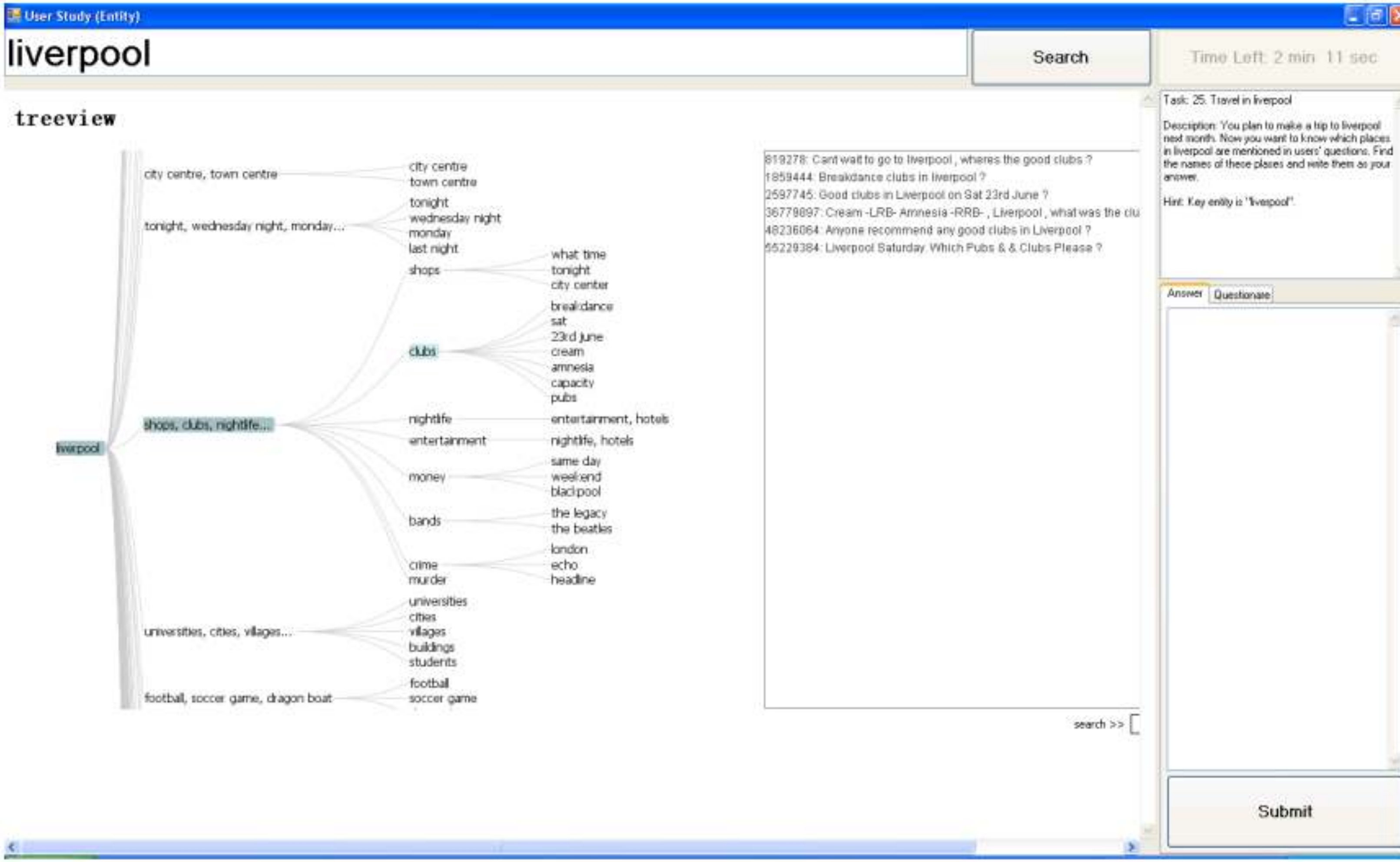


Figure : The interface of CET-based program

Back to FAQ

User Study Tasks

ID	Task Title	Category	Main Entity	Type
1	Find the names of games running on macbook pro	Computer & Internet	Macbook Pro	E
2	Find which components of thinkpad notebooks are usually asked	Computer & Internet	Thinkpad	E
3	How to ps body using photoshop cs2	Computer & Internet	Photoshop CS2	E
4	Questions about the best canon laser printer for a mac	Computer & Internet	Laser printer	S
5	Questions about how to connect Xbox 360 to Laptop or PC using a router	Computer & Internet	Xbox 360	S
6	Questions about green screen problem of Windows Movie Maker	Computer & Internet	Windows Movie Maker	S
7	Find the cities compared with Edinburgh	Travel	Edinburgh	E
8	Find the names of animals on myrtle beach	Travel	Myrtle Beach	E
9	Find the names of cities in Portugal	Travel	Portugal	E
10	Questions about looking for good hostels in Madrid	Travel	Madrid	S
11	Questions about how to get a low price ticket to Hong Kong Disneyland	Travel	Disneyland	S
12	Questions about how to go to Chinatown in Chicago	Travel	Chicago	S
13	Find the brand of running shoes that users have asked	Sports	Running shoes	E
14	Find football players that compared with messi	Sports	Messi	E
15	Find the names of skiing places that users have asked	Sports	Skiing	E
16	Questions asking horse racing website	Sports	Horse racing	S
17	Questions about who will win the MVP in NBA this year	Sports	NBA	S
18	Questions about when is the next match between Barcelona and Real Madrid	Sports	Real Madrid	S
19	Find the brand of cars that have been compared with Toyota	Cars & Transportation	Toyota	E
20	Which aspects of Jeep Wrangler have been asked	Cars & Transportation	Jeep Wrangler	E
21	Finding the names of sports cars being asked	Cars & Transportation	Sports cars	E
22	Questions which compare Mercedes Benz and BMW	Cars & Transportation	Mercedes Benz	S
23	Questions about the price to tow a suv from Newark to Florida	Cars & Transportation	SUV	S
24	Questions about How to reset the oil light for a 95 Honda civic	Cars & Transportation	Honda Civic	S

Back to FAQ