Quan Wang<sup>†</sup> Jing Liu<sup>‡</sup> Bin Wang<sup>†</sup> Li Guo<sup>†</sup>
<sup>†</sup> Institute of Information Engineering, CAS <sup>‡</sup> Harbin Institute of Technology

# Introduction

### Research Problem

- Question difficulty estimation in community question answering
- Applications
  - Question routing, incentive mechanism design, linguistic analysis

#### **Previous Solutions**

### Competition-based methods

- > Extract pairwise competitions from question answering threads
- > Estimate question difficulty based on extracted competitions
  - TrueSkill (Liu et al., 2013)
  - PageRank (Yang et al., 2008)

#### Drawbacks

- > Data sparsity issue: each question gets only two competitions
- Cold-start issue: cannot handle questions with no answers received

### **Our Solution**

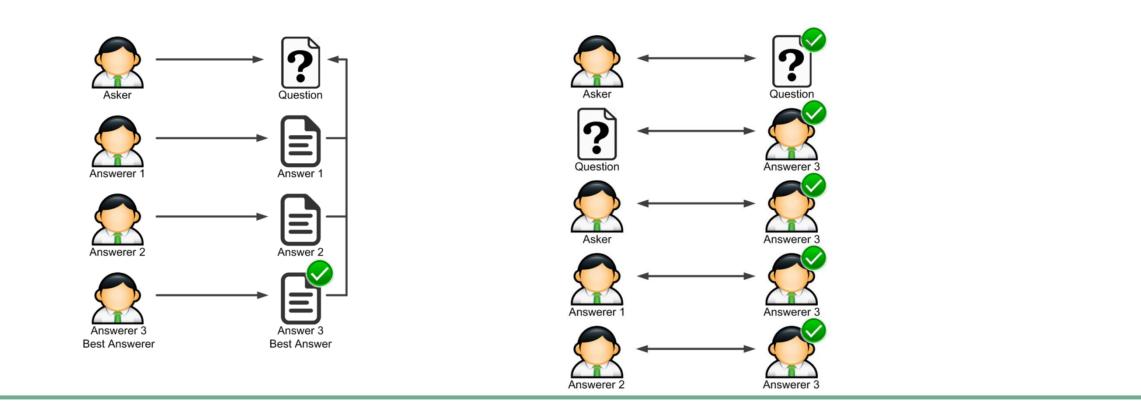
### Competitions + textual descriptions

- For data sparsity issue: textual descriptions provide additional information
- For cold-start issue: textual descriptions link cold-start questions to well-resolved ones

# Regularized Competition Model

### Assumption I: pairwise comparison assumption

- Question's difficulty > asker's skill
- Question's difficulty < best answerer's skill</p>
- Best answerer's skill > all other answerers' skill



# For assumption I: a margin-based loss

$$\ell\left(\bar{\theta}_{i}, \bar{\theta}_{j}\right) = \max\left(0, \delta - \left(\bar{\theta}_{j} - \bar{\theta}_{i}\right)\right)^{p}, \quad p = 1 \text{ or } 2$$

- Express question difficulty and user skill on the same scale
- If estimation is consistent with assumption, the loss is zero
- Otherwise, the loss is proportional to the violation

# Assumption II: smoothness assumption

Questions close to each other in textual descriptions have similar difficulty

differential-equations metric-spaces
group-theory differential-geometry

calculus real-analysis

field-theory combinatorics analysis complex-analysis limit
graph-theory measure-theory probability inequality
graph-theory abstract-algebra terminology
fourier-series probability-theory abstract-algebra terminology
fourier-series probability-theory geometry multivariable-calculus

algebraic-geometry logic
optimization integral homework
algebra-precalculus reference-request
linear-algebra statistics
algorithms general-topology

(a) Low difficulty

real-analysis group-theory elementary-set-theory elementary-set-theory integral soft-question algebra-precalculus statistics soft-question statistics polynomials elementary-number-theory algorithms reference-request functional-analysis trigonometry measure-theory matrices geometry commutative-algebra functions differential-equations differential-geometry

linear-algebra complex-an probability-theory general-topology multivariable-calculus calculus sequences-and-series

abstract-algebra algebraic-geometry

general-topology representation-theory homework

matrices probability-theory prime-numbers cardinals elementary-set-theory calculus integral complex-analysis probability analysis soft-question field-theory reference-request banach-spaces algebraic-topology limit graph-theory ring-theory set-theory logic group-theory

real-analysis sequences-and-series geometry number-theory functional-analysis measure-theory

(c) High difficulty

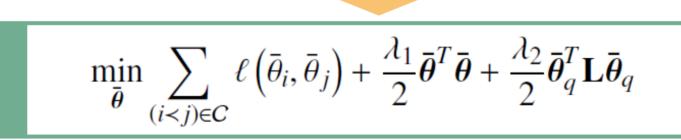
Figure: Tag clouds of SO/Math questions with different difficulty levels

(b) Medium difficulty



$$\mathcal{R} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \bar{\theta}_i^{(q)} - \bar{\theta}_j^{(q)} \right)^2 w_{ij} = \bar{\boldsymbol{\theta}}_q^T \mathbf{L} \bar{\boldsymbol{\theta}}_q$$

- > If textual descriptions are similar, difficulty gap will be small
- Can choose a variety of term weighting schemas
- Can choose a variety of similarity measures



# **Experimental Settings**

#### Datasets

- SO/Math: 10528 questions and 6564 users
- SO/CPP: 10164 questions and 14884 users
- For evaluation
  - 539 annotated SO/Math question pairs
  - 521 annotated SO/CPP question pairs
  - Development/test/cold-start split

#### Baselines

TrueSkill (TS), PageRank (PR), Competition Model (CM)

#### Evaluation metric

Accuracy: proportion of question pairs that are correctly judged

#### **Evaluation for Resolved Questions**

#### Results

- RCM preforms significant better on both datasets
- Improvements can be achieved by a variety of term weighting schemas and similarity measures
- Improvements on SO/Math are greater than those on SO/CPP

	PR	TS	CM		RCM	
			Н	Q	Н	Q
SO/CPP	0.5876	0.6134	0.6340	0.6753	0.7371	0.7268
SO/Math	0.6067	0.6109	0.6527	0.6820	0.7699	0.7699

#### **Evaluation for Cold-Start Questions**

#### Procedures

- Select k well-resolved questions closest in textual descriptions as nearest neighbors
- Calculate average difficulty of nearest neighbors

#### Results

RCM performs consistently better on both datasets with different k values

	PR	TS	C	M	RCM	
			Н	Q	Н	Q
SO/CPP	0.5870	0.5413	0.6120	0.6304	0.6380	0.6609
SO/Math	0.6411	0.6305	0.6653	0.7263	0.6958	0.7442

# **Difficulty Levels of Words**

#### Procedures

- Split questions into buckets according to their difficulty
- Calculate the frequency of a word in each bucket

#### Results

RCM might provide an automatic way to measure difficulty levels of words

