QDEE: Question Difficulty and Expertise Estimation in Community Question Answering Services

Jiankai Sun

Sobhan Moosavi   Rajiv Ramnath   Srinivasan Parthasarathy
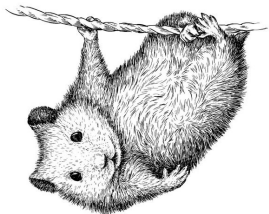
The Ohio State University

ICWSM, June 25 - 28, 2018

THE OHIO STATE
UNIVERSITY

Object-Oriented-Programming (OOP) vs StackOverflow-Oriented-Programming (SOP)

The Internet will do the remembering for you

Googling for the Regex

Every. Damn. Time.

O RLY?   @ThePracticalDev

Cutting corners to meet arbitrary management deadlines

Essential

Copying and Pasting from Stack Overflow

O'REILLY®   The Practical Developer
@ThePracticalDev

Left pic: https://goo.gl/5vKuaR

Right pic: https://goo.gl/XAG4DP

Motivation

- What if you could not find answers for your questions in Stack Overflow? 4.8m unanswered

Stack Overflow
Q&A for professional and enthusiast programmers

questions	16m
answers	24m
answered	71%
users	8.7m

"Declarations/definitions as statements in C and C++"
– asked 10 hours ago

Visit Site

0
Pic: https://goo.gl/3VEsV9
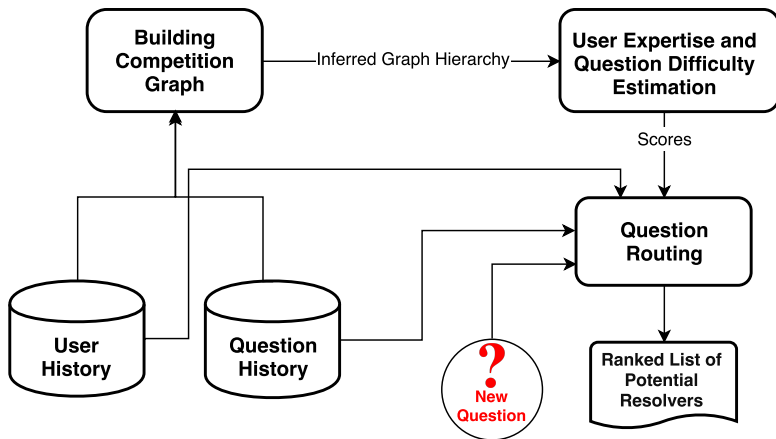
QDEE Framework

Goal: Routing questions to users with matching expertise based on question difficulty level
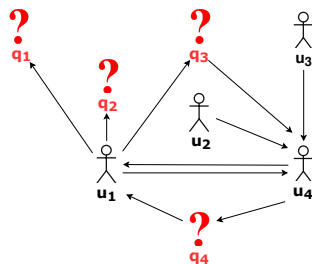
How to Build the Competition Graph [Liu et al. SIGIR'11]

Heterogeneous Nodes Type

- question
- users have two roles: asker and answerer

Construction of Edges
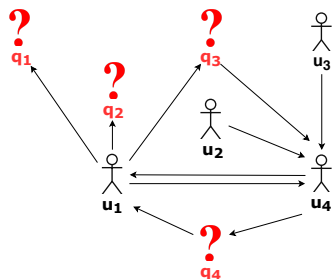
- asker $\Rightarrow$ question : (u_1, q_3)
- question $\Rightarrow$ best answerer : (q_3, u_4)
- asker $\Rightarrow$ best answerer : (u_1, u_4)
- *non-best answerers* $\Rightarrow$ best answerer : (u_2, u_4), and (u_3, u_4).

Data Sparseness Problem and EGA

Data Sparseness Problem

- Each question has only one in-edge (from asker) and one out-edge (to the best answerer)
- May not provide enough information to achieve an accurate estimation

Expertise Gain Assumption (EGA)

Users typically gain expertise across multiple interactions with the CQA and tend to ask more difficult questions within the same domain over time

Illustration of EGA

**The difficulty levels of questions are increasing over time**

Table: Questions asked by a Stack Overflow user in Python.

Questions in Python	Question-Date
q_a: use basic build-in function *sum* on a *list*	July 2013
q_b: changing a list element in multiple lists	Sept. 2013
q_c: *list comprehension* and *generator*	Oct. 2013
q_d: copying 2-D Python list of arbitrary length	Feb. 2014
q_e: using *regular expressions* in *Python*	Nov. 2014

Competition Graph with EGA
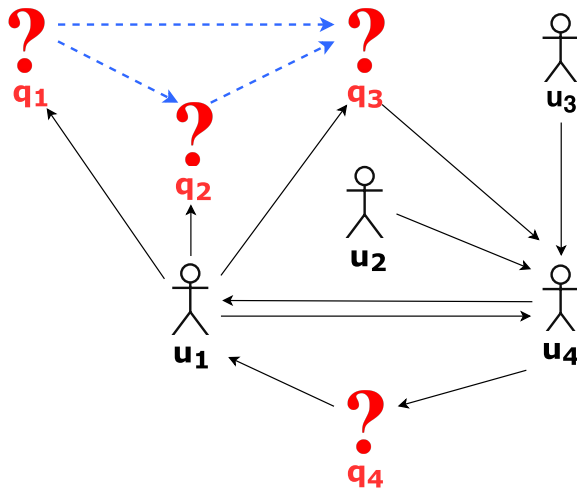
How to estimate question difficulty and user expertise
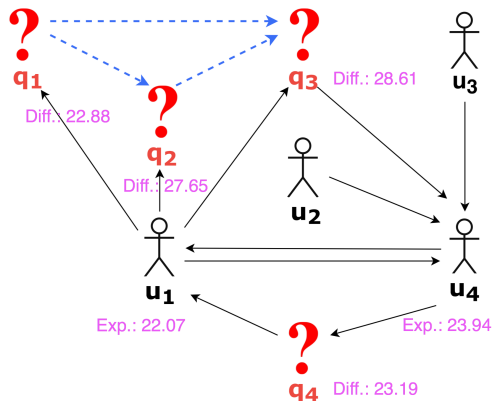
- Two ways: TrueSkill [Herbrich et al. NIPS'07, Liu et al. SIGIR'11] and Social Agony [Gupte et al. WWW'11, Tatti ICDM'15]
- The Biggest Advantage: Language Agnostic

Cold-Start Estimation and Routing

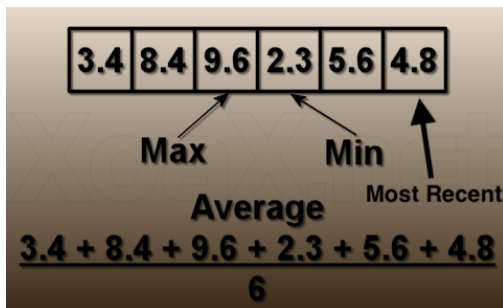- Cold-Start: newly posted questions with no answers (unseen nodes during the training)?
- A common problem in recommender systems [Cheng et al. WWW'17, Wang et al. TIST'14, CIKM'12, Sun et al. CIKM'12]

0
Pic: https://goo.gl/aVh8v5

Cold-Start Difficulty Estimation

- Given a cold-start (q^*) and well-resolved questions $q_1, q_2, ..., q_k$ asked by the same user within the same domain
- q^*'s difficulty level : maximum difficulty level of $q_1, ..., q_k$ $---Max$
- $Max(q_1, ..., q_k)$ $!= q_k$ $---Most Recent$

Cold-Start Routing: Identifying a Set of Potential Answerers to Route Newly Posted Questions

- Language agnostic
  - **Q:** Users who have answered questions with similar difficulty as q^*
- Language conscious
  - **T:** Users who have answered questions that are textually similar to q^*
  - Potential Answerers to Route **QT**: Combination of **Q** and **T**

---

[0] Pic: https://goo.gl/QiDGwM

Experimental Setup

- Datasets: Yahoo! Answers (Japanese) and Stack Overflow (English)
- Ground Truth: Question's *coin* (or *bounty*), given to the user who provided the best answer, as indicator of question difficulty level
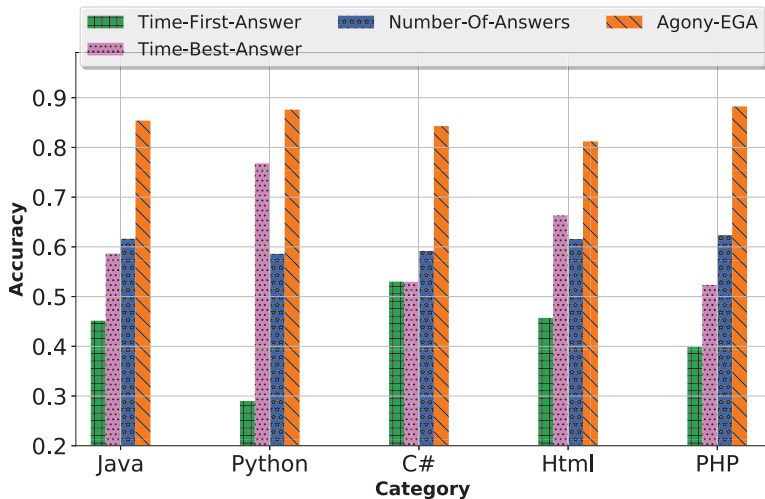- Accuracy Metric:

$$Accuracy = \frac{\# \text{ correctly predicted question pairs}}{\# \text{ valid question pairs}}$$

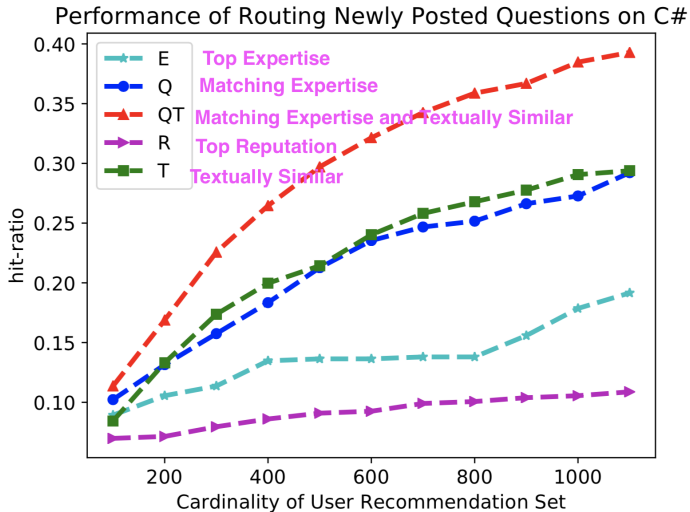# Yahoo! Answers: Advantages of Leveraging EGA
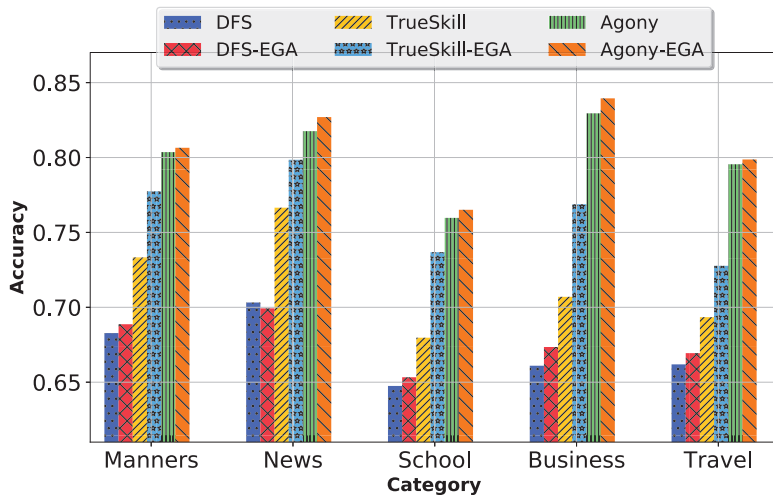
Stack Overflow: Social Agony vs State-of-the-art

Different Approaches for Cold-Start Routing

Performance of Routing Newly Posted Questions on C#

Legend:
- E — **Top Expertise**
- Q — **Matching Expertise**
- QT — **Matching Expertise and Textually Similar**
- R — **Top Reputation**
- T — **Textually Similar**

Patterns of User Answering Questions
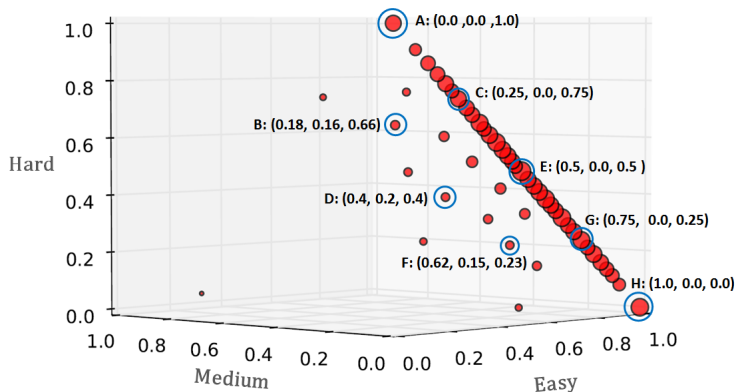
Conclusion

- Proposed Expertise Gain Assumption (EGA) to solve the data sparseness problem in CQAs
- Leveraged graph hierarchy (social agony) to estimate question difficulty and user expertise
- Proposed approaches to route cold questions to users with matching expertise
- Supported by NSF grants CCF-1645599 and IIS-1550302 and a grant from the Ohio Supercomputer Center (PAS0166)
- Code is available on GitHub:
 `https://github.com/zhenv5/QDEE`

Q & A

THE OHIO STATE
UNIVERSITY

Backup Slides

Inferring Graph Hierarchy by TrueSkill [Herbrich et al. NIPS'07, Liu et al. SIGIR'11]

- A skill based ranking system to rank Xbox players, developed by Microsoft Research
- Each player has two numbers
  - μ: average skill of the player
  - σ: degree of uncertainty in the player's skill
- a directed graph $G = (V, E) \Rightarrow$ a multi-player tournament with $|V|$ players and $|E|$ competitions

- an edge $(u, v) \in E \Rightarrow u$ loses the game between u and v
- a node v's ranking score in the graph hierarchy:
$$f_{ts}(v) = \mu_v - 3\sigma_v$$
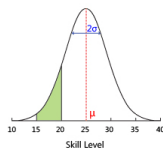
Inferring Graph Hierarchy by Social Agony [Gupte et al. WWW'11, Tatti ICDM'15]

- In social networks such as Twitter, people are not likely to follow people who are lower in the hierarchy
- Agony can be caused when people follow other people who are lower in the hierarchy
- Defined by the severity of their violation, agony to u caused by edge (u, v) is equal to $max(r(u) - r(v) + 1, 0)$
- The agony in the network given a ranking r
  - sum of agony on each edge
- Goal: find a ranking r that minimize the total agony in the graph

Cold-Start Difficulty Estimation

- what if the cold-start question q^* was asked by a new registered user who has no asking history?
- KNN: The difficulty score of q^* is predicted as the averaged difficulty scores of its nearest neighbors ($d_{knn}(q^*)$). [Wang et al. EMNLP'14]

Combine Them Together

$$d(q^*) = \alpha \cdot d_{knn}(q^*) + (1 - \alpha) \cdot d_{ega}(q^*)$$

Strategy Selection for Cold-Start Estimation

Acc of strategies for cold-start questions on C#