EVOLUTION OF EXPERTS IN QUESTION ANSWERING COMMUNITIES

Aditya Pal, Shuo Chang, Joe Konstan
(University of Minnesota)
Motivation for Temporal Analysis

• Understanding user activity patterns
  • [Guo, 2009] analyzed hourly activity patterns to find that even though 80-20 contribution rule applies, yet top contributors’ participation is much flatter than power law

• Question routing schemes
  • [Liu and Agichtein, 2011] showed that temporal activity patterns can be used effectively to tune question routing algorithms to ensure that a question gets answered in a timely manner


Research Questions

• How do experts evolve and influence community members?

• What are the different evolutionary characteristics of experts?

• Can we identify different kinds of experts?

• Can we improve expert identification techniques by taking users’ evolution into account?
Dataset Description

- StackOverflow data (August 2008 – September 2010)
  - ~ 1M question asked by 165K users
  - ~ 2.4M answers by 156K users

- Expert labeling
  - Selected users with more than 9 answers (29K users)
  - Marked top 10% with highest reputation score

Data Preprocessing

- Divide data into bi-weekly buckets
  - First bucket = time of earliest question
  - 70 bi-weekly buckets

- Relative time series
  - Pick first 26 buckets of activity for a user
  - Normalize based on activity of other users during the same time period
Influence on Question Askers

![Graph showing best answer probability over time for experts and ordinary users. The graph displays a steady increase in best answer probability for experts as time progresses, while ordinary users show a more fluctuating pattern.]
Influence on Question Askers

Askers are wary in selecting newcomers’ answers as best
Influence on Question Askers

Askers are wary in selecting newcomers’ answers as best.

Experts get motivated as they get recognized for their work.
Influence of Experts on Ordinary Users

![Graph showing the number of answers by ordinary users over time buckets (bi-weekly). The graph compares the number of answers when experts answer versus when experts do not answer.]
Influence of Experts on Ordinary Users

Initially users participated vigorously on questions answered by experts.

As experts became distinguishable, participation propensity decreased.
Influence of Experts on Ordinary Users

• Prior work [Pal and Counts, ICWSM 2011]
  • Users get biased based on name value of experts

[ICWSM 2011] Pal and Counts: What's in a @name? How name value biases judgment of microblog authors.
Influence of Experts on Ordinary Users

• Prior work [Pal and Counts, ICWSM 2011]
  • Users get biased based on name value of experts

• Discussions in meta-StackOverflow
  • Enormous contributions by experts demoralized them a bit
  • Its intimidating initially but with time one can adapt amongst experts
  • Its intimidating to answer a question asked by an expert
  • Merits and demerits of allowing easy questions to be answered by beginners

[ICWSM 2011] Pal andCounts: What’s in a @name? How name value biases judgment of microblog authors.
Influence of Experts on Experts

• $p \sim$ probability of an expert answer ($\sim 0.4$)

• $n \sim$ number of answers to a question

• $ne \sim$ number of expert answers to a question

$$\sim \text{Binomial}(n, p) = \frac{n!}{ne! \cdot (n-ne)!} p^{ne} (1 - p)^{n-ne}$$
Influence of Experts on Experts

Expert Answer Probability on a Question

Number of Answers
Influence of Experts on Experts

Experts are less likely to collectively collaborate to answer a question
Influence of Experts on Experts

Experts are less likely to collectively collaborate to answer a question

Experts avoid each other, as they aim to have higher value/effort returns
Influence of Experts on Experts

![Graph showing the influence of experts on experts over time buckets.](image.png)
Influence of Experts on Experts

Initially, experts collaborated on more than 50% of the questions
Influence of Experts on Experts

Initially, experts collaborated on more than 50% of the questions.

As experts became distinguishable, collaboration declined drastically.
Expert Evolution

• Temporal Clustering Based on GMM

\[ P(X|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_{ik} \cdot P(x_i|\theta_k) \quad \text{[i.i.d time series]} \]

\[ P(x_i|\theta_k) \propto \frac{1}{\sqrt{|\Sigma_k|}} \exp \left\{ -\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right\} \]

• Bayesian Information Criteria

\[ BIC(K) = -2 \cdot \ln(\hat{P}(X|\theta)) + K \cdot \ln(N) \]
Number of Clusters

K=6 minimizes the BIC criteria
Expert Evolution Pattern

For question routing, experts in C are valuable

For finding churners, experts in E are valuable

For nurturing and motivation, experts in L are valuable
Identifying Different Types of Experts

Different types of experts can be found with 0.5 f-measure within 20-weeks of being in the community.

F-measure = \(\frac{2 \cdot p \cdot r}{p + r}\)

SVM with 10-fold cross validation
Identifying Experts

<table>
<thead>
<tr>
<th></th>
<th>StackOverflow</th>
<th></th>
<th>Intuit</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Static</td>
<td>Temporal</td>
<td>Static</td>
<td>Temporal</td>
</tr>
<tr>
<td>precision</td>
<td>90</td>
<td>94</td>
<td>70</td>
<td>73</td>
</tr>
<tr>
<td>recall</td>
<td>52</td>
<td>67</td>
<td>66</td>
<td>71</td>
</tr>
<tr>
<td>F-measure</td>
<td>66</td>
<td>78</td>
<td>68</td>
<td>72</td>
</tr>
</tbody>
</table>

Model based on temporal data outperform model based on static data
Summary

- Experts influence best answer selection of askers
- Ordinary users get intimated by experts
- Experts avoid other experts.
- Experts evolve with different patterns: C, E, L
- These experts can be found with satisfactory performance within 20 weeks
- Expert identification techniques can be improved by 5-15% by using their temporal data instead of static data
Take Aways

• Interface that anonymize user profiles initially on a question can be beneficial. These profiles could be revealed after a lapse of time.

• Different kinds of experts are useful for different objectives:
  - C (consistently active) – question routing schemes
  - E (early active) – churn prediction
  - L (late active) – nurturing and fostering

• Expert identification methods can be improved by using temporal data.
Thanks