Migrated Question Prediction on StackExchange

Sangeeta Lal\textsuperscript{1}, Denzil Correa\textsuperscript{2} and Ashish Sureka\textsuperscript{2}

Jaypee Institute of Information Technology (JIIT), India\textsuperscript{1}

sangeeta@jiit.ac.in

Indraprastha Institute of Information Technology, Delhi (IIITD), India\textsuperscript{2}

denizlc@iiitd.ac.in, ashish@iiitd.ac.in
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
Introduction to StackExchange (SE)

1. SE network consists of 112 Q&A sites
2. 4.6 million users, 8 million questions and 14.2 million answers (on Jan-2014)
3. All SE Q&A website are created to address questions of specific domain or user interest
4. StackOverflow\(^1\) is created for addressing programming questions
5. English.StackExchange\(^2\) is created for English language enthusiasts

\(^1\)http://StackOverflow.com/
\(^2\)http://English.StackExchange.com/
Research Motivation and Aim

StackExchange Community based Q&A website

Question *migrated* from StackOverflow to Meta.StackOverflow

![Figure 1](image_url)

**Figure 1:** Question is not related to programming - about posting comments on StackOverflow website (issue related to a feature of the website – Meta.StackOverflow)
Why Automate Question Migration (Currently done Manually)

- **Improving Response Time**: Manual migration can be time consuming and user may not get appropriate answer/response to the question until it is migrated.

- **Reputation Loss**: Question migration causes reversal of all the reputations earned (earned on origin site before migration) by contributing to a migrated question. This causes reputation loss to the people who have earned reputation by answering or commenting on migrated question (on origin site).

- **Improving Question Quality**: If a user gets an early feedback about his off-topic question, he may improve the question content to make it more clear.
Research Aim

1. To conduct an in-depth characterization study of migrated questions on SE Q&A sites. To investigate temporal distribution of migrated questions on SE websites and to identify community structure of Q&A website with respect to mutual question migration.

2. To identify properties of migrated questions (such as reputation of owner, amount of discussion and popularity).

3. To identify distinguishing features of migrated questions and to build a predictive model to identify migrating questions from Q&A sites (at the time of post creation).
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
Research Contributions

- An in-depth **characterization study of migrated questions** on five popular SE Q&A websites. We identify several research questions and conduct various experiment (on dataset of ≈38,000 migrated questions) to answer those questions.

- Machine learning based framework for **predicting migrating questions** at the time of post creation. We test the effectiveness of proposed framework on five popular SE Q&A websites. Proposed framework is found to be effective with maximum accuracy of 73%.

- We rank various features used in the proposed framework according to their importance in predicting migrating questions.
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
Figure 2: Who, How and What of Question Migration process on SE sites

- **Who can migrate a question?**
  - >250 Owner
  - >3000 Reputation
  - Moderator

- **How questions are migrated?**
  - 4 votes can migrate
  - Can vote to migrate their own question

- **What are the side effects of migration?**
  - Reversal of all the reputations
  - Answers & downvotes are deleted

- **What happens with migrated question?**
  - Close if duplicate/off-topic belongs on
  - Valid on-topic on target sites
  - Rejection
  - Closed as off-topic on source site
Experimental Dataset Details

Table 1: 41 sites, 21 technical sites and 20 corresponding meta sites. Span of 4.7 years (Aug-2008 to Jun-2013), having 5.9 million questions out of which 50915 are migrated.

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Date</td>
<td>Aug-2008</td>
</tr>
<tr>
<td>End Date</td>
<td>1st-Jun-2013</td>
</tr>
<tr>
<td>Technical Site</td>
<td>41</td>
</tr>
<tr>
<td>Total Question</td>
<td>5946915</td>
</tr>
<tr>
<td>Number of Migrated Posts (Only Technical Sites)</td>
<td>102386</td>
</tr>
<tr>
<td>Number of Migrated Question (Only Technical Sites)</td>
<td>50915</td>
</tr>
</tbody>
</table>
Temporal Distribution of Migrated Questions

Figure 3: Monthly Question Migration Trend
Temporal Distribution of Migrated Questions

Figure 4: Cumulative Monthly Question Migration
Network Structure of Migrated Questions
Network Diagram of All the Migrated Questions

1. Node represents Q&A website
2. Edge represents migrated question between the sites
3. Node size is proportional to in-degree of the node
4. Edge thickness is proportional to the number of questions migrated between two sites
5. We detect three communities (shown by different colors) in the graph at a resolution of 0.5
6. Two communities have three dominating (or power) nodes, namely StackOverflow, Meta.StackOverflow, and Superuser
In-degree and Out-degree Characteristics

Table 2: Indegree and Outdegree Characteristics

<table>
<thead>
<tr>
<th>In-Degree</th>
<th>No. of Sites</th>
<th>Out-Degree</th>
<th>No. of Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 100</td>
<td>19</td>
<td>&lt; 10</td>
<td>71</td>
</tr>
<tr>
<td>[100, 200)</td>
<td>4</td>
<td>[10, 20)</td>
<td>17</td>
</tr>
<tr>
<td>[200, 300)</td>
<td>0</td>
<td>[20, 50)</td>
<td>12</td>
</tr>
<tr>
<td>[300, 400)</td>
<td>0</td>
<td>[50, 100)</td>
<td>7</td>
</tr>
<tr>
<td>[400, 500)</td>
<td>2</td>
<td>[100, 500)</td>
<td>10</td>
</tr>
<tr>
<td>&gt;= 500</td>
<td>15</td>
<td>&gt;= 500</td>
<td>7</td>
</tr>
<tr>
<td>Total =</td>
<td>41</td>
<td>Total =</td>
<td>124</td>
</tr>
</tbody>
</table>
In-degree and Out-degree Characteristics

1. 19 websites have in-degree less than 100
2. Very few have in-degree between 400-500
3. 15 sites have in-degree greater than 500
4. Slight deviation from power-law as less than expected nodes have in-degree between 100-500
Reputation of Owners of Migrated Question

Figure 5: Reputation of Owners of Migrated Question is Low as Compared to General Trend of Reputation for all the Websites.
Comment Count of Migrated Question

Figure 6: Comment Count of Migrated Questions is Fractionally Higher for all the Websites
Answer Count of Migrated Question

Figure 7: Answer Count of Migrated Questions is Alightly Higher for the four Websites
View Count of Migrated Question

Figure 8: View Count of Migrated Questions is Same as General Trend of View Count
Score Count of Migrated Question

Figure 9: Score Count of Migrated Questions is Same as General Trend of View Count
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
Features used in Predictive Model with Respective Categories

<table>
<thead>
<tr>
<th>Type</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post Content (B)</td>
<td>Title Length&lt;br&gt;Body Length&lt;br&gt;URL Count&lt;br&gt;Consist of Code&lt;br&gt;Number of Words having First Letter Capital&lt;br&gt;Number of Words having All Capital Letters&lt;br&gt;Special Character Count (*.+[]/-:-)</td>
</tr>
<tr>
<td>Meta Features (C)</td>
<td>Number of Tags&lt;br&gt;Consist of Popular Tag</td>
</tr>
</tbody>
</table>
## Dataset Details

**Table 3:** MQC: Migration Question Count, TQC: Total Question Count (with Accepted Answer), RS: Random Sample Size, EDS: Experimental Dataset Size

<table>
<thead>
<tr>
<th>Data Set</th>
<th>MQC</th>
<th>TQC</th>
<th>RS (10 Times)</th>
<th>EDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackOverflow</td>
<td>32190</td>
<td>308950</td>
<td>32190</td>
<td>64380</td>
</tr>
<tr>
<td>Serverfault</td>
<td>5437</td>
<td>79507</td>
<td>5437</td>
<td>10874</td>
</tr>
<tr>
<td>Superuser</td>
<td>4972</td>
<td>85961</td>
<td>4972</td>
<td>9944</td>
</tr>
<tr>
<td>dba.SE</td>
<td>112</td>
<td>7559</td>
<td>112</td>
<td>224</td>
</tr>
<tr>
<td>programmers.SE</td>
<td>2000</td>
<td>15412</td>
<td>2000</td>
<td>4000</td>
</tr>
</tbody>
</table>
## Classifier Parameters

### Table 4: Classifier Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4*Classifier</td>
<td>DT: Max-Depth=5</td>
</tr>
<tr>
<td></td>
<td>KNN: Leaf size=59, Neighbor = 6, Weight=Uniform, Metric= Euclidean</td>
</tr>
<tr>
<td></td>
<td>GNB: No parameter</td>
</tr>
<tr>
<td></td>
<td>ADA: Estimators = 100</td>
</tr>
<tr>
<td>Cross validation</td>
<td>Random permutations cross-validation&lt;sup&gt;3&lt;/sup&gt;</td>
</tr>
<tr>
<td>Classification Iterations</td>
<td>10</td>
</tr>
<tr>
<td>Feature Set</td>
<td>{B}, {C}, {BC}</td>
</tr>
<tr>
<td>Train-Test Split</td>
<td>70%-30%</td>
</tr>
</tbody>
</table>

### Classification Results

**Table 5:** Classification Results of Proposed Prediction Model Using AdaBoost Classifier and Feature Set \{BC\}

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Accuracy(%)</th>
<th>Precision(%)</th>
<th>Recall(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>StackOverflow</td>
<td>71.61</td>
<td>71.44</td>
<td>72.06</td>
</tr>
<tr>
<td>Serverfault</td>
<td>58.33</td>
<td>59.12</td>
<td>53.56</td>
</tr>
<tr>
<td>Superuser</td>
<td>73.66</td>
<td>78.8</td>
<td>64.49</td>
</tr>
<tr>
<td>dba.SE</td>
<td>54.35</td>
<td>54.92</td>
<td>53.92</td>
</tr>
<tr>
<td>Programmer.SE</td>
<td>71.27</td>
<td>72.16</td>
<td>69.53</td>
</tr>
</tbody>
</table>
### Feature Importance Ranking

<table>
<thead>
<tr>
<th>Feature</th>
<th>SF</th>
<th>SU</th>
<th>PS</th>
<th>DS</th>
<th>SO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Length</td>
<td>0.2</td>
<td>0.19</td>
<td>0.11</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td>Body Length</td>
<td>0.23</td>
<td>0.31</td>
<td>0.27</td>
<td>0.2</td>
<td>0.33</td>
</tr>
<tr>
<td>URL Count</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td>0.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Number of Tags</td>
<td>0.05</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>Popular Tag</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td>All capital count</td>
<td>0.12</td>
<td>0.05</td>
<td>0.05</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>First Capital count</td>
<td>0.13</td>
<td>0.13</td>
<td>0.17</td>
<td>0.13</td>
<td>0.06</td>
</tr>
<tr>
<td>Special Count</td>
<td>0.18</td>
<td>0.22</td>
<td>0.27</td>
<td>0.18</td>
<td>0.14</td>
</tr>
<tr>
<td>have code</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
Conclusion

1. Reputation of owners of migrated questions are low
2. Migrated questions engage more discussion
3. *AdaBoost* gave maximum accuracy of 73% on SuperUser (Migrated Question Prediction)
4. Body Length and Special Character Count are among the most important features for classifying *migrated* questions
Presentation Outline

1. Research Motivation and Aim
2. Research Contributions
3. Characterization Study
4. Migrated Question Prediction
5. Conclusion
6. References
References

StackOverflow.
What is migration and how does it work?
[Online; accessed 24-Jan-2014].

Meta.StackExchange.
How to migrate question to programmers.stackexchange.com.
[Online; accessed 20-Jun-2014].

Meta.StackExchange.
How to migrate question to programmers.stackexchange.co.
[Online; accessed 20-Jun-2014].
References II

**Meta.StackOverflow.**
What is a rejected migration?  
[Online; accessed 24-Jan-2014].

**Meta.StackOverflow.**
Are moderators allowed to migrate question from MSO to SO?  
[Online; accessed 25-Jan-2014].

**Jeff Atwood.**
Require more than simple majority of close votes to migrate.  
[Online; accessed 31-Jan-2014].
Kevin Montrose.  
Allow diamond moderators to reverse question migrations?  
[Online; accessed 31-Jan-2014].

Denzil Correa and Ashish Sureka.  
Fit or unfit: analysis and prediction of 'closed questions' on stack overflow.  
In COSN, pages 201–212, 2013.

Galina Lezina and Artem Kuznetsov.  
Predict closed questions on stackoverflow.  


References

Dana Movshovitz-Attias, Yair Movshovitz-Attias, Peter Steenkiste, and Christos Faloutsos.  
Analysis of the reputation system and user contributions on a question answering website: Stackoverflow.  
In *ASONAM*, Niagara Falls, Canada, August 2013.

Aditya Pal and Joseph A. Konstan.  

Qiongjie Tian, Peng Zhang, and Baoxin Li.  
Towards predicting the best answers in community-based question-answering services.  
In *ICWSM*, 2013.
References VI

