MiQs: Characterization and Prediction of Migrated Questions on StackExchange

Sangeeta Lal  
JIIT-Noida  
Uttar Pradesh, India  
sangeeta@jiit.ac.in

Denzil Correa  
IIITD  
New Delhi, India  
denzilc@iiitd.ac.in

Ashish Sureka  
IIITD  
New Delhi, India  
ashish@iiitd.ac.in

Abstract—StackExchange (SE) is a network of very popular Community based Question and Answer (CQA) websites. Each SE Q&A website is created to address questions on specific user interest or domain. However, several times it is noticed that user post questions on SE websites which do not match the domain of website. Such questions are considered as off-topic for the origin site. Off-topic questions need to be detected and migrated to more appropriate (on-topic) site in SE network. Off-topic questions are migrated manually to other sites by moderators or experts users (through voting process). The process of migrating questions from one site to other is known as Question Migration. Manual question migration is undesirable as it puts extra load on moderators or expert users. It can also potentially cause delay to the user seeking answer of his questions and can also lower down satisfaction level of the user.

This work presents first study of migrated questions on SE Q&A website. We present results of characterization study performed on five popular SE Q&A sites: Superuser, Serverfault, StackOverflow, dba.StackExchange, and programmers.StackExchange. We use publically available SE Q&A sites dataset (span of 4.7 years). We perform an in-depth characterization study on 38609 migrated questions. We present results of the temporal distribution, community structure of Q&A websites, owner reputation, amount of discussion, and popularity of migrated questions. We identify several distinguishing features of migrated questions and propose a machine learning based framework to predict migrating questions. Effectiveness of proposed model is tested on five SE Q&A sites. Result shows that proposed model is effective (shows maximum accuracy of 73%) in predicting migrating questions.

I. RESEARCH MOTIVATION AND AIM

StackExchange (SE) is a network of very popular Community based Question and Answer (CQA) websites1. SE network consists of 112 Q&A sites, 4.6 million users, 8 million questions and 14.2 million answers (on Jan-2014). All SE Q&A website are created to address questions of specific domain or user interest. For example StackOverflow2 is created for addressing programing questions, english.stackexchange3 is created for English language enthusiasts, travel.stackexchange4 is created for travel related questions. Each Q&A website explicitly mentions its domain & guidelines for posting questions. However, several times when user posts a question on a website it does not matches the domain of that website. Such questions are considered as ‘off-topic’ questions. ‘Off-topic’ questions needs to be detected and migrated to other more appropriate (‘on-topic’) website in SE network. Figure 1 shows snapshot of a migrated question5. Question present in Figure 1 is asked on StackOverflow but later on migrated to meta.StackOverflow. StackOverflow is a Q&A website for programing enthusiasts and hence expects questions related to programing. However this question is not related to programing, in this question user is asking some issue related to posting comments on StackOverflow website i.e issue related to a feature of the website. This questions does not match the interest domain of StackOverflow and hence migrated to meta.stackoverflow. meta.stackoverflow addresses meta questions related to SE Q&A websites and hence a perfect match for this question.

Fig. 1: Shows snapshot of a question migrated from StackOverflow to meta.StackOverflow

Identifying and migrating off-topic questions is an important activity for SE Q&A websites. SE websites does not only provide answer(s) to the question owner but also serves as knowledge base for the future users [1]. Many times it is noticed that an off-topic question on one website is an exact duplicate of an on-topic question on some other website in SE network [2]. Not-migrating or ignoring such off-topic questions will cause scattering of information. This may later on can create confusion and can potentially make it harder for future users to find similar information. Hence identifying and migrating appropriate questions is important to maintain content quality of SE websites.

Currently questions on SE websites are migrated manually by moderators and expert users. Manual migration is tedious.

---

1http://stackexchange.com/  
2http://stackoverflow.com/  
3http://english.stackexchange.com/  
4http://travel.stackexchange.com/  
5http://meta.stackoverflow.com/questions/143306/newline-in-stackoverflow-comments
process (see section III) and puts extra load on moderators and expert users. Also manual migration can only be done after posting the question i.e. after some users have spent their time and effort in reading and answering the question. In this work, our aim is to analyze migrated questions and to create an automated technique for questions migration (at the time of post creation). Automated techniques for predicting migrating questions at the time of post creation will not only reduce the load of moderators or expert users but will have multiple other benefits, some of them are listed below:

- Improving Response Time: Manual migration can be time consuming and user may not get appropriate answer/response to the question until it is migrated, and hence automated migration can help in improving response time to the user.
- Reputation Loss Issue: 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) [3]. This can create dissatisfaction among the user community.  
- 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. This process can help in improving the question quality, which over the period of time will improve quality of the Q&A website.

The broad research aim of the work presented in this paper is to analyze migrated questions on SE Q&A sites and to reduce load of moderator and expert users by providing framework for automated questions migration. The specific research aim of the work presented in this paper is following:

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 sites and to identify community structure of Q&A sites with respect to mutual question migration. To identify properties of migrated questions (such as reputation of owner, amount of discussion, popularity etc.).
2) 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).

II. RELATED WORK AND RESEARCH CONTRIBUTION

CQA sites are very helpful for developers. A recent study done by Gomez et al. [4] shows that StackOverflow helps in spreading innovations in programing as often developers share links related to API’s, tools and libraries [4]. Barua et al. [5] proposed an LDA based model to identify current trend of various programing tools such as languages, API’s, source code management systems etc. Lower quality questions on StackOverflow often receive lower quality answers [6] and hence to maintain quality of Q&A website it is necessary to remove lower quality questions. Lezina et al. [7] and Correa et al. [8] work on predicting ‘closed’ questions on StackOverflow. Lezina et al. and Correa et al. identified various features related to user-profile, post-feature, tags etc. and proposed machine learning based framework for predicting closed questions. Sometimes many low quality and high quality answers are available for a given question. Identifying best answer among them can be challenging for website community and future users. Tian et al. [9] work on identifying best answer from the given set of answer to a questions and reported accuracy of 72%. On StackOverflow it is mandatory to assign appropriate tags to the question before posting it on the website. However assigning suitable tags is challenging for new users. Xia et al. and Lipczak et al. work on predicting tags for Q&A sites and social networking sites [10], [11]. Xia et al. proposed ‘TagCombine’ model for tag recommendation on software information sites. Lipczak et al. proposed a feedback based tag recommendation method [10]. Lipczak et al. worked on multiple aspects (generality, efficiency and adaptability) of tag recommendation problem. Expert users are backbone of any CQA website. Identifying expert users at an early stage can be beneficial for the website community. Pal et al. and Attias et al. work on identifying experts on CQA sites [12], [13]. Gaussian model proposed by Pal et al. shows recall of 92% [12] as compared to activity pattern based model of Attias et al. which shows precision of 70%. In another work, Pal et al. [14] identified different classes of experts: C (consistently active), E (active initially), and L (active later) and proposed an SVM based model and reported precision of 94% on StackOverflow.

A. Research Contributions

In this paper, we work on analyzing and predicting migrated questions on SE Q&A sites. To the best of our knowledge this is the first study of migrated questions on SE Q&A sites, following are the specific and novel research contribution of this work:

- An in-depth characterizing 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 (refer Section V).
- 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 (refer Section VI).

III. QUESTION MIGRATION PROCESS ON SE SITES

This section gives brief detail about question migration process on SE Q&A websites. This sections describes various rules of question migration on SE sites. Figure 2 shows important aspects of question migration process.

What is question migration?

‘off-topic’ questions posted on SE Q&A sites are migrated to
other more appropriate (‘on-topic’) sites in SE network. This process of migrating questions from one site to another sites is known as question migration. After migration a stub of questions is left on origin site for 30 days, which is deleted thereafter. Question which are older than 60 days are not considered for migration [3].

Who can migrate a question?
- Moderators can migrate question to any site.
- 3000+ reputation users can cast vote to migrate a question to any of the five (actual number can be different for different sites) pre-listed sites. Question is migrated automatically after receiving four votes (to same site) from 3k reputation users [3].
- 250 reputation users can cast vote to migrate their own question to one of the five (actual number can be different for different sites) pre-listed sites [15].
- Others: In case desired site is not present in given list or user have less than 3k reputation, user can cast vote for moderators attention and can suggest desired target website [3]. Moderator can take decision to migrate such questions [3]. Questions cannot be migrated from meta sites to other sites through voting process as there is no predefined path, only moderators are allowed to migrate questions from meta site [16].

What are the side-effects of question migration?
All the answers and downvotes associated with the question on origin site are cleared after question migration. It may cause reversal any reputation earned or lost in the process of commenting or answering the migrated question on origin site [3].

What is migration-rejection?
Migrated questions can be rejected automatically after receiving desired number of votes from 3k reputation users. Community of the target websites may also decide to close the migrated question [17].

IV. Why ‘Off-Topic’ Questions Are Asked?
We manually analyze multiple migrated questions on SE website and identified reasons “why users ask ‘off-topic’ question?”. Following are our finding:

1) New Users: There are 112 Q&A websites (on Jan-2014) in the network of SE, belonging to various domains. New users joining the community many times are not aware of the domain (or guidelines) of Q&A website or about existence of other websites, and hence erroneously ask off-topic question. Table I shows an example (Type-a) of a user post, where a new-user on meta.SE is asking help and explanation about his migrated question.

2) Erroneously Posting on Wrong Website: Many times user erroneously asks off-topic question and after sometimes he want to migrate the question to more appropriate site. Table I show an example (Type-b) of a question on meta.StackOverflow where a user has asked question on Superuser and want to migrate his own question on webapps.SE.

3) Overlapping Site: Few sites in SE network have overlapping domain or one site is sub-domain of other site, for example {askubuntu,linux}, {StackOverflow, Programming.SE}, {CS.SE, CS.Theory.SE} etc. It is noticed that many times questions are migrated between these sites. Overlapping sites can create confusion for users and hence results in migration later on. Table I shows an example (Type-c) where a user is explicitly asking difference between unix.SE, Linux.SE and AskUbuntu.SE for identifying which site is more appropriate for asking the question.

4) Cross-Posting: SE community highly discourage unnecessary cross-posting of questions. However some users habitually cross-post all their questions on many SE websites, which later on results in migration of questions to other SE site. Table I shows an example (Type-d) where an expert-user is complaining about behavior of other user, who habitually post all his questions on three SE site.

5) Poor-Question Quality: ‘On-Topic’ questions having poor quality or unclear objective, are erroneously migrated sometimes. For example, Table I shows a question (Type-e) which is migrated from math.SE to physics.SE, later on owner cleared that he was actually interested in mathematical formulation, and hence asked question on math.SE.

V. Characterization Study
This section provides details of characterization study performed in this work.

A. Dataset Details
SE network consists of 112 Q&A websites. For this study we download SE data dump provided by SE community. We extract data of 41 sites, 21 technical sites and 20 corresponding meta sites. Table II shows basic characteristics of experimental dataset. Experimental dataset has a span of 4.7 years (Aug-2008 to Jun-2013), having 5.9 million questions out of which 50915 questions are migrated.

B. Temporal Distribution of Migrated Questions
We analyze temporal distribution of migrated questions on SE sites. Figure 3 shows monthly count and Figure 4 shows cumulative monthly count of questions migrated to SE sites.
TABLE I: Examples of Migrated Questions Showing Reasons of Question Migration

<table>
<thead>
<tr>
<th>Type</th>
<th>Category</th>
<th>ID</th>
<th>Website</th>
<th>Post Content (or Comment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) New User</td>
<td></td>
<td>3950</td>
<td>meta.tex.SE</td>
<td>I am a new user and my first question was migrated to this site of Meta - Tex.</td>
</tr>
<tr>
<td>(b) Erroneously Posting</td>
<td></td>
<td>85017</td>
<td>meta.StackOverflow</td>
<td>I've posted a question to superuser.com on Google Spreadsheets now I would like to move on webapps.stackexchange.com.</td>
</tr>
<tr>
<td>(c) Overlapping Sites</td>
<td></td>
<td>132659</td>
<td>meta.StackOverflow</td>
<td>What's the difference between Super User, Unix and Linux, and Ask Ubuntu? In general, what's a rule for deciding where to post a question when I have one (which might apply to any of those 3)?</td>
</tr>
<tr>
<td>(d) Cross-Posting</td>
<td></td>
<td>105847</td>
<td>meta.StackOverflow</td>
<td>I recently found a user who posted almost every one of his questions on at least 3 sites, and his latest question on 5 (!!!).</td>
</tr>
<tr>
<td>(e) Poor-Quality Question</td>
<td></td>
<td>11220</td>
<td>physics.SE</td>
<td>It would also help if you attempt to formulate the problem in a precise mathematical way rather than in terms of stars, etc.</td>
</tr>
</tbody>
</table>

TABLE II: Stack Exchange Dataset Details

<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>Jun-2013</td>
</tr>
<tr>
<td>Technical Site</td>
<td>41</td>
</tr>
<tr>
<td>Total Question</td>
<td>5940915</td>
</tr>
<tr>
<td>Number of Migrated Posts</td>
<td>102386</td>
</tr>
<tr>
<td>Number of Migrated Questions</td>
<td>50915</td>
</tr>
</tbody>
</table>

Fig. 3: Shows monthly question migration trend

Fig. 4: Shows cumulative monthly question migration trend

Fig. 5: Shows migrated question percentage with respect to total questions present for each site

C. Network Structure of Migrated Questions

For each Q&A website we compute percentage of the questions that are ‘migrated to’ and ‘migrated from’ these sites. Figure 5 shows that ~18% of the total questions present on dba.SE website are migrated from other sites. Figure 5 shows that meta sites also have a huge number of questions migrated from other websites. For example meta.superuser and meta.serverfault consists of 15% and 13% migrated questions, respectively.

For detailed characterization study we compute percentage of question migrated to each site out of total 50915 migrated questions. We select top five websites, having highest number of migrated questions-Superuser, Serverfault, StackOverflow, dba.SE, and Programmer.SE. Approximately 74% (38609) of all migrated question are migrated to these five websites.

Table III shows year-wise count of unique sites from which questions are migrated (in duration of our study). In the year 2008 and 2009 there were only 2 & 7 sites, respectively, from which questions are migrated. This count increases to 34 in 2010 and 84 in 2011. We also observe decrease in number of migrated questions after August 2011. We believe this happened because in July-2011 SE moderators increased number of votes to 4 (earlier was 3) for migrating a question on StackOverflow [18]. A huge drop in number of migrating questions is also observed during March-2012, we believe this happened as a side-effect of adding migration rejection feature in SE websites [19].

TABLE III: Year-wise Count of Question Migrating Sites

<table>
<thead>
<tr>
<th>Year</th>
<th>Migrated Questions</th>
<th>Count of Distinct Sites (from which questions are migrated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>469</td>
<td>2</td>
</tr>
<tr>
<td>2009</td>
<td>3812</td>
<td>7</td>
</tr>
<tr>
<td>2010</td>
<td>8114</td>
<td>84</td>
</tr>
<tr>
<td>2011</td>
<td>19103</td>
<td>84</td>
</tr>
<tr>
<td>2012</td>
<td>13322</td>
<td>100</td>
</tr>
<tr>
<td>2013</td>
<td>6095</td>
<td>73</td>
</tr>
</tbody>
</table>

We believe this happened because in July-2011 SE moderators increased number of votes to 4 (earlier was 3) for migrating a question on StackOverflow [18]. A huge drop in number of migrating questions is also observed during March-2012, we believe this happened as a side-effect of adding migration rejection feature in SE websites [19].
in the graph. Gephi\(^7\) tool detected three communities (shown by different colors) in the graph at resolution of 0.5. Two communities have three dominating (or power) nodes, namely StackOverflow, meta.StackOverflow and Superuser. In third community we did not observe any dominating node. In first community StackOverflow received questions from 54 distinct site belonging to various domains like webmasters, codereview, crypto, history etc. StackOverflow migrated questions to 21 different Q&A sites. meta.stackOverflow also received questions from 54 distinct Q&A websites, having 41 meta sites. Meta.StackOverflow\(^8\) is a Q&A website for meta-discussion about SE Q&A websites, and hence it is intuitive to see migrations from other meta websites. However, analysis shows that 86\% of all 2014 questions migrated to meta.StackOverflow, are migrated from StackOverflow. Second community has Superuser as the dominating node which recived question from 47 different Q&A websites and migrated questions to 22 Q&A websites. We observe that community of StackOverflow is denser as compared to meta.StackOverflow and Superuser. As 32 Q&A websites migrated atleast 10 questions to StackOverflow while for Superuser and meta.StackOverflow this count is equal to 15 and 10 respectively. We did not observe any one dominating node in third community as many Q&A websites like Security.SE, Apple.SE, Unix.SE etc have same size. Table V shows in-degree and out-degree distribution of Q&A websites. Table V shows that 19 websites have indegree less than 100, very few have degree between 100-500, and 15 sites have indegree greater than 500. This shows slight deviation from power-law as less than expected nodes have indegree between 100-500. Table V shows that questions migrated from 124 to 41 (dataset limit) Q&A websites, which shows complexity of question-migration problem. We perform further analysis on five sites, selected earlier. For each of the five websites we select top-5 websites from which questions migrated to these sites as well as top-5 sites to which questions are migrated from these five websites. Result shows that StackOverflow migrated highest number of questions to other four sites. This may be happening because of large user community of StackOverflow. Result also reveals that huge number of questions migrated between StackOverflow, Programmer.SE and codereview.SE. These three sites can be considered as overlapping (websites with similar user interest) site. This shows compliance with Section IV, where we list that overlapping interest domain can cause question migration. We also observe that for StackOverflow and programmers.SE their corresponding meta sites are among the top-5 site to which questions are migrated from these sites (not considering dba.SE due to insignificant number of questions).

**D. Questions of low reputation users are migrated more often?**

As discussed off-topic are questions which are migrated to other more appropriate site. We hypothesize that users with low reputation are more likely to post off-topic questions as compared to experienced or high reputation users. To test this hypothesis we plot reputation of owners of all the questions posted on selected Q&A websites and reputation of owners of all the questions migrated to that website (refer to Figure 7). For questions where specified userID is not available in the dataset, we assume reputation value 1 (starting reputation) for such questions. It is interesting to see that median value of reputation for owners of migrated question is consistently less as compared to general trend of reputation, for all the websites. For example, general median reputation of users on StackOverflow and programmers.SE is 296 and 229, which is decreased to 100 and 123 when considering reputation of owners of migrated questions. We compute Mann-Whitney-Wilcoxon (MWW) RankSum test \(^9\) for comparing two samples. Reputation of owners of migrated questions found to be significantly (0.05) different from reputations of owners other questions for all the five datasets. Hence we can infer that

---

\(^7\)https://gephi.org/

\(^8\)http://stackexchange.com/sites

\(^9\)http://www.randalolson.com/2012/08/06/statistical-analysis-made-easy-in-python/
questions posted by low reputation users are more likely to migrate as compared to questions posted by experienced users.

![Figure 7: Shows reputation of owners of migrated question is low as compared to general trend of reputation for all the websites](image)

**E. Migrated questions are discussed more?**

We analyze behavior of comments and answers posted on migrated questions to identify whether migrated questions are discussed more as compared to other questions. Figure 8 shows box plots for number of comments posted with respect to the question for various sites. Figure 8 shows that migrated questions have relatively higher number of comments. For example median value of number of comments for Superuser and Serverfault is 1 for migrated questions and 0 in general. We did not notice much difference in median value of number of comments posted on migrated questions for StackOverflow, dba.SE and programmers.SE. However, quartile-3 values of comment count for migrated questions is higher for both the sites. We observe that migrated questions have more number of comments because many time community ask owner of migrated question to give more explanation about the questions.

![Figure 8: Shows comment count of migrated questions is fractionally higher for all the websites](image)

Figure 9 shows box-plot for number of answer posted on a question (answer count) for all the sites. We observe slight increase in answer count of migrated questions (median count) for all cases except StackOverflow. For example median value of answer count for migrated questions on dba.SE and programmers.SE is 2 and 4 respectively. Answer count decrease to 1 and 3 for dba.SE and programmers.SE when considering all the questions posted on both sites. Result of comment count and answer count are found to be significantly (0.05) different using MWW Rank Sum test for all the five websites.

![Figure 9: Shows answer count of migrated questions is slightly higher for four the websites](image)

**F. Non-popular question are migrated more often?**

We conduct experiment to identify whether migrated questions are less popular as compared to other questions. We consider score count and view count as two parameters to measure popularity of questions. Figure 10 and Figure 11 shows box-plots for vote count and score count respectively. Figure 10 shows that median view count for all the websites is between 200-400 for both the categories, except for programmer.SE. For programmers.SE we observe increase in view count for migrated questions. Median view count of questions posted on programmers.SE is 400, median view count increases to 563 when considering only migrated questions. Figure 11 shows similar trend for score count plot. We observe that except programmers.SE all the Q&A websites have median score count between 1-2 with little or no difference between both the categories. For Programmers.SE median score count value of migrated questions is slightly high. Median score count value of migrated questions is 5 which decreases to 4 when considering all the question on Programmers.SE. We believe
that view count of migrated questions on Programmers.SE is high and it can be one of the factors in increasing score count of migrated questions. Difference between score count of migrated questions and other questions is significantly different for programmers.SE, StackOverflow and Superuser, while for dba.SE and Serverfault score count does not differ significantly. Similarly for view count results differ significantly for all the websites but superuser (using MWW Rank Sums test). We do not observe significant difference in the popularity of migrated questions for all the five websites.

VI. MIGRATED QUESTION PREDICTION

In second part of our study we work on predicting migrated questions. We consider migrated question prediction as a binary classification task i.e. predicting whether a question will migrate from a given site or not. Following subsections gives details of experimental setup and results.

A. Feature Selection

We identified 9 features based on post content and meta information available with the posted questions. Table VI shows various features with their respective categories. Category (B) shows 7 features based on post content. ‘Title Length’ and ‘Body Length’ feature are found to be useful for ‘closed’ question prediction by Correa et al. [8]. As off-topic questions are one of the types of ‘closed’ questions, we hypothesize that these two features can be helpful in classifying migrating questions. We observe that many times question posted on software development Q&A sites have names (or links) of tools or functions and hence we select ‘URL Count’, ‘Capital Letter Words Count’, and ‘First Letter Capital Words Count’ as features. ‘Special characters’ denotes presence of snippets from configuration files or source code. Category (C) shows two features: ‘Number of Tags’ and ‘Presence of Popular Tag’- based on meta information available with the question. To compute popular tags, we compute question count (questions having accepted answer) of all the tags present on a website, and select top-100 tags. We check presence of these top 100 tags in top-100 tag feature. There are many tags synonyms present on Q&A sites, for example microsoft-access and ms-access, ms-access is one of the popular tags (top-100) in dba.SE but microsoft-access is not a popular tag. We have use tags as they are present with the question without doing any synonym conversion. In this work, our aim is to predict question migration at the time of question creation and hence we do not use features based on comments or answers. It should also be noted that question content and tags present with the question often gets modified by expert users of SE community and original content of the posted question are not available, and hence prediction with the content present in the database may introduce biases in the results.

B. Experimental Setup

We work on five websites (selected earlier) for question migration prediction- StackOverflow, Serverfault, Superusers, dba.SE and programmers.SE - having total 44711 migrated (migrated from these websites) questions. Table VII shows count of migrated questions for individual sites. Table VII shows that our data set is imbalanced in nature, as count of migrated (+ve class) questions is much less than total number (1%-12%) of questions posted for each site. To make our data balanced we perform random sampling of questions present in -ve class (non-migrated questions). Each random sample is equal to the size of migrated questions i.e size of +ve class with respect to each site. To remove any sampling bias we perform random sampling 10 times and apply classifier for each (+ve,-ve) sample pair and report average accuracy, precision and recall over 10 samples. We work with four classifiers: KNN, Decision Trees (DT), AdaBoost, Gaussian Naive Baysian (GNB). We use sklearn11 implementation of the classifiers. Table VIII shows value of various parameters used for classification. We run classifier for multiple parameter values and select the value giving best results. For example for DT we obtain maximum accuracy for maximum depth value between 4-9, and hence we select maximum depth value as 5. We similarly select parameters for other classifiers. We use 70%-30% training-testing split with random permutation cross-validation with 10 iterations.

C. Classification Results

We apply classifier on individual features category as well as on combination of feature categories- {B}, {C}, {BC}. We notice that feature set {C} give better accuracy as compared to feature set {B} on four datasets. Accuracy results obtained by feature set {BC} and {C} are similar, except on StackOverflow and programmers.SE where feature set {BC} outperforms feature set {C}, increasing accuracy by 5% and 11%, respectively. Adaboost and DecisionTree classifier give better result as compared to other classifier, when comparing all the three parameters (accuracy, precision, recall). Maximum accuracy reported by both Adaboost and DecisionTree classifier is 73% on superuser using feature set {C}. Although accuracy value of feature set {C} and {BC} are same a huge drop is noticed in precision and recall while after dropping any of the feature set. This shows importance of considering both the features for classification. Table IX shows accuracy, precision and recall value obtained by applying Adaboost classifier on each dataset using feature set {BC}.

11http://scikit-learn.org/
TABLE VIII: Classifier parameters: DT: Decision Tree, KNN: K-Nearest Neighbor, GNB: Gaussian Naive Bayesian, ADA: AdaBoost

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>DT: Max-Depth=5</td>
</tr>
<tr>
<td>KNN: Leaf size=59, Neighbour = 6. Weighted=Uniform, Metric= Euclidean</td>
<td></td>
</tr>
<tr>
<td>GNB: No parameter</td>
<td></td>
</tr>
<tr>
<td>ADA: Estimators = 100</td>
<td></td>
</tr>
<tr>
<td>Cross validation</td>
<td>Random permutations cross-validation</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>

TABLE IX: 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>

D. Feature Importance Ranking

We use Adaboost classifier to rank various features to identify their importance in classification. Table X shows ranking of all the features for each Q&A website. Table X shows that ‘Body Length’ and ‘Special Character Count’ are among the most informative features for most of the datasets. ‘Title Length’ and ‘First Word Capital Count’ are also found to be very effective in classification. All these four features are related to quality of the questions posted. We observe that ‘Presence of Code’, ‘Number of Tags’ and ‘Presence of Popular Tags’ are among the least effective feature.

VII. Threats to Validity

This work presents results of characterization and prediction study performed on five popular Q&A sites. Characterization study shows that reputation of owners of migrated questions are considerably low and migrated questions engage more discussion. This work also presents a machine learning based framework to predict migrated questions. We present results of four different classifiers on dataset of five Q&A sites. AdaBoost gave maximum accuracy of 73% on Superuser. We rank various features on the basis of their importance in classification. Results shows that ‘Body Length’ and ‘Special Charcter Count’ are among the most important features for classifying migrated questions.

VIII. Conclusion

This work presents results of characterization and prediction study performed on five popular Q&A sites. Characterization study shows that reputation of owners of migrated questions are considerably low and migrated questions engage more discussion. This work also presents a machine learning based framework to predict migrated questions. We present results of four different classifiers on dataset of five Q&A sites. AdaBoost gave maximum accuracy of 73% on Superuser. We rank various features on the basis of their importance in classification. Results shows that ‘Body Length’ and ‘Special Charcter Count’ are among the most important features for classifying migrated questions.

REFERENCES

