Early Detection of Potential Experts in Question Answering Communities

Aditya Pal[1], Rosta Farzan[2], Joseph A. Konstan[1], and Robert Kraut[2]

[1] Dept. of Computer Science and Engineering, University of Minnesota
[2] Human-Computer Interaction Institute, Carnegie Mellon University

Abstract. Question answering communities (QA) are sustained by a handful of experts who provide a large number of high quality answers. Identifying these experts during the first few weeks of their joining the community can be beneficial as it would allow community managers to take steps to develop and retain these potential experts. In this paper, we explore approaches to identify potential experts as early as within the first two weeks of their association with the QA. We look at users' behavior and estimate their *motivation* and *ability* to help others. These qualities enable us to build classification and ranking models to identify users who are likely to become experts in the future. Our results indicate that the current experts can be effectively identified from their early behavior. We asked community managers to evaluate the potential experts identified by our algorithm and their analysis revealed that quite a few of these users were already experts or on the path of becoming experts. Our retrospective analysis shows that some of these potential experts had already left the community, highlighting the value of early identification and engagement.

Keywords: Question Answering, Potential Experts, Expert Identification

1 Introduction

Question answering communities (QA) are excellent knowledge sources which enable their users to create value while participating in social interactions with one another. Prior studies [3] show that the quality of knowledge one can fetch from QA can exceed the quality of knowledge one can receive from information specialists. There is a core group of users also referred to as experts in these communities who are the key contributors of knowledge.

The experts constitute a small percentage of the community members and are responsible for a large percentage of the answers [10]. We also see the evidence of this in the TurboTax Live Community (TTLC) dataset. TurboTax Live Community (TTLC)[3] is a QA service that allows users to ask and answer tax-related and TurboTax product related questions. The TTLC dataset is a complete dump of questions, answers, and user IDs from the time period July 2006

[3] http://ttlc.intuit.com

Table 1. Participation characteristics of the two types of users in TTLC.

	no. users	no. questions	no. answers	no. best answers
superuser	83 (0.01%)	1,963 (0.31%)	177,427 (45%)	43,059 (78%)
user	604,900 (99.99%)	630,522 (99.69%)	218,366 (55%)	12,385 (22%)

- April 2009. It contains 83 superusers (interchangeably called experts), 604,900 ordinary users, 633,112 questions, and 688,390 answers. Superusers constitute 0.01% of the population yet they have provided 78% of the best answers and close to 45% of all answers. The superusers differ drastically from the ordinary users in terms of how they participate in the community, as depicted in table 1. Needless to say, the superusers are the drivers of community answer-production and are extremely important for this community to function.

Intuit[4] recognizes these superusers, making their status visible to other community members. This recognition adds a stamp of trust to their answers and keeps them motivated to carry on the good work. It is important to note that these users are not paid for their answers and do not have any association with Intuit. Intuit takes special care in identifying the superusers. They have employees that manually evaluate top answerers for qualities such as tax knowledge, quality of their answers, politeness and clarity of responses and writing ability. If a user has some professional experience in the tax domain, that is also a plus. Based on these assessments, a user can get promoted to superuser. Through April, 2009 Intuit has recognized 83 superusers; they acknowledge that there are likely many more qualified users, but due to the manual evaluation process, they have not yet identified them. The human evaluation process highlights two important limitations:

- Humans usually evaluate long-time contributors; as a result recently joined users with high potential are not considered.
- The evaluation process is slow, which leads to the risk of high-potential users leaving the community due to the lack of recognition of their efforts.

These limitations highlight the need for a screening tool to filter through tens of thousands of users to recommend potential superusers to human evaluators. Specifically, we use machine learning to identify high-potential users in the first few weeks of their participation. Early identification of potential experts can benefit the community in several ways. It enables measures to nurture experts and retain them. The proper training of potential experts could also improve their skills and improve the overall quality of the participation in the community.

The primary difficulty in finding potential experts early on is that the markers that reflect expertise of a person, e.g., number of answers, number of best answers, etc., are not that strong for a newly joined user. As a result not much prior work has been done in finding potential experts in early-stage in QA. Panciera et al. [7] show that initial contributions of experts are measurably different from contributions of ordinary users in communities like Wikipedia. The

[4] Intuit is the company that launched TurboTax live community (TTLC).

question arises whether early experts behavior is different in QA communities? Is there an untapped set of potential experts that we could develop - users who might otherwise leave the community due to lack of recognition? Our research seeks to address this challenge.

In this paper, we propose several different measures that could be used for identifying potential experts based on their early participation. We look at the behavioral characteristics of current experts when they joined the community and use predictive and ranking algorithms to estimate their potential. This helps answer several questions. Do the experts differ from ordinary users since the day they joined the community or did they improve over a period of time? What abstract qualities are required for users to become experts in general? What qualities are important to become an experts in a domain specific QA like TTLC? How effective can algorithms be in identifying potential experts early on?

2 Related Work

Several other researchers have addressed the question of expert identification in QA communities. Zhang et al. [10] modified PageRank [5] to propose an algorithm, ExpertiseRank. Their algorithm considers whom a person answered in addition to how many people a person answered. They combined the number of answers (a) and number of questions (q) of a user in one score, Z-score ($z = \frac{a-q}{\sqrt{a+q}}$). A person with high Z-score is considered to have higher expertise than a person with low Z-score value. They used a dataset from the Java developer forum to validate several ranking algorithms against a ground truth of human evaluation. Their analysis indicated that a simple measure such as Z-score outperforms complex graph based algorithms such as ExpertiseRank, PageRank, and HITS in the assessment of the expertise of the users.

Expertise measures such as Z-score and ExpertiseRank typically provide a ranking of users in terms of decreasing expertise levels. They do not instruct how many users should be selected as experts from the ranked list. Bouguessa et al. [1] addressed this issue. The authors considered number of best answers as an indicator of user expertise. Based on this indicator they modeled authority scores as a mixture of gamma distributions and used Bayesian Information Criteria (BIC) to estimate the appropriate number of mixture components and the parameters of each mixture component using the Expectation Maximization (EM) algorithm. Their results on datasets from Yahoo Answers resulted in two mixtures of users, suggesting that the Yahoo Answers community contains two types of users: {experts, non-experts}.

In other work, Jurczky et al. [4] performed link analysis over the question-answer interconnections among users of Yahoo Answers. Their analysis showed that the HITS algorithm outperforms classical graph measures such as in-degree, out-degree, and centrality for the identification of expertise. More recently, Pal et al. [6], proposed a model to estimate the question selection bias of the answerers. They showed that the experts differ from the normal users in that they avoid

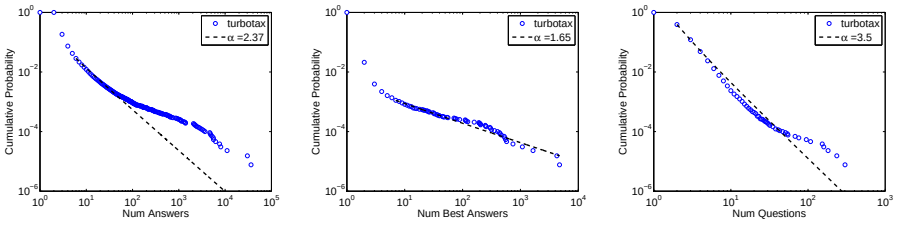

Fig. 1. Log-Log distribution of participation characteristics of users.

answering questions that already have a good answer; this bias was used to identify experts.

Our work differs from prior research as we focus on early identification of potential experts. As a result, we use user activity during the first few weeks of her participation in the community. We propose that motivation and ability to help others should be ideally present among potential experts, and model them using several abstract measures. As per our literature survey, we are the first to show that classification and ranking models can be built successfully to detect potential experts early on and provide a useful baseline for further work in this direction.

3 Dataset Description

TurboTax Live Community (TTLC) is a question and answering online service that allows users to ask and answer tax-related and TurboTax's product related questions. It has the same basic structure as popular Q&A websites like Yahoo Answers and Stackoverflow.com. We have TTLC question and answer data from July 2006 - April 2009. The dataset contains 633,112 questions asked by 525,143 users and 688,390 answers provided by 130,770 users. The dataset has 83 users explicitly marked as superusers who have been manually selected by Intuit employees.

Superusers play a vital role by answering tax questions of thousands of users (43,059 best answers - Table 1). Figure 1 shows the distribution of participation characteristics of users in TTLC. These plots follow power-law distribution as is the case with most online Q&A systems [10]. The power-law distribution is an indicator of an uneven participation where a large section of users contribute in a small proportion and a small section contribute in large proportion.

We selected users who gave 10 or more answers and discarded the remaining, to ensure that we have sizable data to evaluate each user. This led to a selection of 1,367 users out of which 83 were superusers and other 1,284 ordinary users. These 1,367 users represented less than 1% of all the community members yet they have provided 76% of all the answers.

Table 2. Indicators of user qualities.

Quality	Indicators
Motivation	M1: *Quantity of contributions.* M2: *Frequency of contributions.* M3: *Commitment towards the community.*
Ability	A1: *Domain knowledge of the user.* A2: *Trustworthiness of user's answers.* A3: *Politeness and clarity in response.*

4 Qualities of Potential Experts

Taking a cue from the participation characteristics of experts, a potential expert should be highly motivated to help others (*motivation*) and she should have the required capability to answer questions correctly (*ability*). Motivation in this context means the willingness of the person to help others. Ability aims to assess the quality of the help a person can provides. Table 2 mentions several indicators of these qualities in the context of a QA. We use several features to estimate these indicators of quality:

- **M1**: Quantity of the contributions made by the user is reflected from the number of answers, number of questions.
- **M2**: Frequency of the contributions is reflected from the *average time elapsed between two answers*. This parameter is estimated by taking the ratio of total number of answers given in a session by session time averaged for the sessions.
- **M3**: Commitment towards the community is indicated from *how many times a user logs into the system* (#login) and *how much time she spends in the community* (login span).

- **A1**: Domain knowledge of the user is hard to estimate as there are no direct measures to tell how much a user knows in the given domain. We use an indirect measure: *number of best answers* to approximate it.
- **A2**: Trustworthiness of user's answers can be determined from the *number of votes*, *number of positively voted answers*, *ratio of answers with negative votes to positively voted answers*. Votes are the ratings provided by the community members in QA and they can be positive as well as negative.
- **A3**: In order to estimate the politeness and clarity in response we perform a language analysis of the answers provided by the user and choose 56 language dimensions. The most prominent of those are 1) presence of typos, spelling mistakes, bad words, sms language, 2) usage of singular pronouns (I,You,They), 3) usage of negative terms like not, discard, reject, hate, etc, 4) usage of greetings like hi, hello, proper-noun (usernames). 5) usage of special characters (?,!,#,etc).

Next, we present our models for identifing potential experts using the above indicators of user qualities.

5 Early Identification of Potential Experts

To perform an early identification of potential experts, we select the first n weeks of data per user. For a given user, the start of her association is measured from the timestamp of her first answer. From the features defined previously, we compute the six abstract quality indicators: $M1$, $M2$, $M3$, $A1$, $A2$, and $A3$ to compute a feature vector per user. These feature vectors along with the user labels {superuser, ordinary user} are used by the learning models to predict superusers. The learning models are described below, followed by their performance.

5.1 Learning Model

We use Support Vector Machines (SVM) [2] and C4.5 Decision Tree (DTree) [9] over the features mentioned in the previous section to find potential experts. Both the algorithms are generally known to perform very well for the supervised learning problems. SVM is a maximum margin classifier that aims at maximizing the decision boundary margin. Maximization of decision boundary margin typically leads to better generalization performance. We use the sequential mining optimization approach to train SVM [8]. DTree splits the training set into subsets based on an feature using some splitting criteria. This process is repeated to create further subsets until all the subsets at a given level have same class or fall below a certain threshold. Then pruning is applied to the tree so that it doesn't overfit the training data. In order to construct the training data, we use 10-fold cross-validation. Cross-validation ensures that the models do not overfit the data and they report the true generalization accuracy.

5.2 Model Performance

Figure 2 shows the performances of the two models in predicting potential experts. We measure the performance of the models using three standard measures: Precision (p), Recall (r) and F-Measure ($\frac{2*p*r}{p+r}$). The precision of SVM is consistently better than of DTree but it has a lower recall and the F-measure of both the algorithms is nearly same.

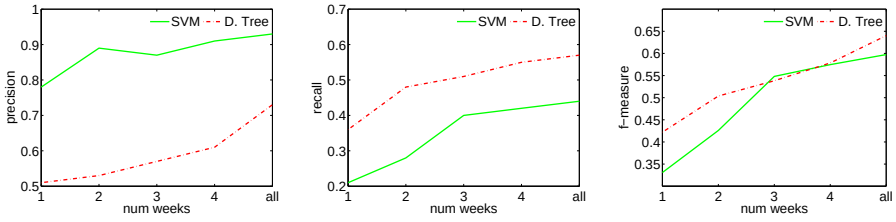

Fig. 2. Model performance over first N weeks of dataset.

Consider the DTree model over 2 weeks of data. It predicts 75 users as potential experts out of which 40 are actually labeled as superusers ($p = 40/75$, $r = 40/83$). Based on this, we make these two key observations:

- There are 35 ordinary users who showed the potential to be an expert but were not labeled as experts. They are either false positives or missed by the manual evaluators at Intuit. We asked Intuit to evaluate these users manually (described in next section), and found that 27 of these 35 ordinary users were almost ready to be a superuser. The lack of due recognition runs the risk of their disassociation from the community and indeed several of them had already either left or reduced their activity.
- There are 43 out of 83 (52%) superusers who did not show potential early on. Further, the recall performance over time suggests that even by considering activity over a time period as large as 1 or 2 years, we can successfully recall only 60% of the experts. This result suggests that our model only captures part of the behavior recognized in superuser status, but promisingly shows that this behavior ofter manifests early in a user's life-cycle.

The experiment described here shows promise that user modeling and machine learning can be useful in finding high potential users as early as within 2 weeks of their joining.

Furthermore, if we consider the entire time period for which the data is available to compute the features, the model performance expectedly improves significantly. However the performance still does not reach 100%. This suggests two possibilities: a) Our models do not precisely capture the human evaluation process to identify experts, and b) There are many more users worthy of superuser status than currently awarded that status. We believe both of these reasons to be true. The models only approximated the key indicators such as domain knowledge based on the participation characteristics without looking at the context of the question and other answers provided on it. Moreover the model does not take user's background and professional experience into account (not part of the dataset).

5.3 Balance between Quality Measures

In the previous subsection, we saw that the two qualities *motivation* and *ability* are important in identifying potential experts. In this subsection, we assess how these qualities perform individually. Table 3 shows the performance of the models

Table 3. Model performance over 2 week data over different user qualities.

	Motivation		Ability		Motivation + Ability	
	SVM	DTree	SVM	DTree	SVM	DTree
precision	0.70	0.56	0.70	0.50	0.89	0.53
recall	0.10	0.42	0.10	0.44	0.23	0.48
f-measure	0.18	0.48	0.18	0.47	0.37	0.50

over 2 week data. The two quality measures performed comparably when used individually - nearly identical for SVM. The combination of the two, however resulted in marginally better performance for DTree and substantially better for SVM. This result indicates that the two qualities represent two different aspect of user's potential and they are equally important in early identification of potential experts.

5.4 Ranking Users on Quality Measures

The purpose of early identification of potential experts is to provide mentoring opportunities and encouragement to the users. However this cannot be provided to all the users due to cost and time constraints. Hence it could be desirable to rank the users on their potential and provide these opportunities to the top ranked users only. We propose a ranking mechanism based on the observations drawn from the previous results.

As we saw earlier the two qualities are equally weighted by the classification algorithms, so it makes sense to have an unweighted ranking between them. We first pick 6 best features (using Information Gain) that represent each abstract quality (see Table 4). Other than the conventional features representing M1, A2, A1 the feature for M3 has high information gain and it indicates that the user's commitment towards the community is useful in predicting their potential. For all the features a higher score is preferred except however for the A3 feature a lower score is preferred. To handle this we multiply the value of the A3 feature by -1.

A Gaussian Cumulative Distribution Function based ranking is used which measures how high a user scores on a given feature in comparison to the overall population over that feature. The following formula shows how the scores for all the features are combined:

$$R_G(x_i) = \prod_{f=1}^{d} \int_{-\infty}^{x_i^f} N(x; \mu_f, \sigma_f) \tag{1}$$

where $N(x; \mu_f, \sigma_f)$ is the Gaussian distribution with parameters μ_f and σ_f. The integral is performed to compute the Gaussian CDF value for a given user feature. All the CDF values are then multiplied to get a one-dimensional score

Table 4. Information gain of different user qualities.

Abstract Quality	Feature	Info. Gain
M1	Number of answers	0.084
A2	Number of votes	0.081
A1	Number of best answers	0.076
M3	Frequency of login	0.074
M2	Avg. time elapsed between answers	0.024
A3	Usage of pronoun - I	0.017

Table 5. Average expertise scores of different types of users using 2 week data.

	mean	std	min	max
experts	0.28	0.12	0.14	0.44
35 potential experts (D. Tree)	0.22	0.12	0.1	0.41
35 top users (excluding potential experts)	0.09	0.10	0.01	0.34

for each user. Using this raking method, if we select the top 75 ranked users as potential experts then 33 of them are superusers ($p = 33/75$, $r = 33/83$) , which is slightly lower than that of DTree but still significant. Table 5 shows the expertise score of different types of users over the 2 week dataset. The average score indicates that potential experts have scores similar to experts and some of them are even better than the current experts in their expertise score. We see that ranking can also be used in conjunction with classification models to surface potential experts. This ranking further helps us in selecting users for human evaluation.

6 Human Evaluation

In order to estimate the effectiveness of the proposed prediction and the ranking models, we consider human evaluation of the identified potential experts. We created a stimuli sample by selecting the 35 potential users (DT) identified by DTree, top 35 ranked users (CDF) (excluding DT users) by the ranking algorithm, and 35 randomly selected users (RND) (excluding DT and CDF users) to create a stimuli sample for human evaluation.

Survey Users We asked those Intuit employees who actually evaluate users for promotion to superusers to take the survey. These evaluators exactly know what skills are required in a TTLC superuser. Each evaluator was presented with a list of 12-15 users selected equally from DT, CDF, and RND and were ordered randomly. The evaluators were not aware of the algorithms' predictions. The evaluators looked at all the answers (and not just first two weeks) provided by these users including the complete question thread (answers of other users provided on those threads) to estimate users' expertise. On an average they took 15 minutes per user and each evaluator took 3-4 hours to complete the survey. Every user was evaluated by 2 judges. The inter-rater reliability measured by Cronbach α was 0.86 which presents high agreement between the evaluators.

Survey Design The evaluators rated users on the two main criteria:

- **Q1**: "This user has the potential to become a successful superuser" - The evaluators marked their responses on a 5-point Likert scale where 1 indicates strong disagreement, 3 indicates that they neither agree or disagree (neutral), and 5 indicates strong agreement.

Table 6. Evaluators' rating of users over the questions: Q1 and Q2. We picked the maximum rating that a user received.

Q1: This user has the potential to become a successful superuser					
	strongly disagree	disagree	neither agree or disagree	agree	strongly agree
DT	3	3	11	10	8
CDF	8	8	10	7	2
RND	11	9	10	3	2
Q2: What is your assessment of this user's potential to become a TTLC SuperUser?					
	no potential	some potential	shows potential - not ready	almost ready	ready to be superuser
DT	1	2	5	18	9
CDF	2	7	8	14	4
RND	5	10	11	7	2

– **Q2**: "What is your assessment of this user's potential to become a TTLC SuperUser?" - The evaluators responded on a 5-point Likert scale where 1 indicates no potential demonstrated, 3 indicates user shows potential but is not ready yet and 5 indicates that the user is ready to be a superuser.

The questions Q1 and Q2 are similar in nature but are added for the robustness of the responses. Pearson correlation of evaluator ratings on Q1-Q2 is 0.867 which is significant at $p < 0.01$(2-tailed). We also asked participants to rate a user on a 5-point Likert scale, where 5 being the highest, for the following 6 criteria: 1) Tax knowledge, 2) Product knowledge, 3) Solving problems, 4) Writing, 5) Social skill, 6) Quality of responses.

6.1 User Ratings

Table 6 presents the evaluation details of users over Q1 and Q2. 18 (51%) *DT* users have demonstrated potential to become a superuser (Q1) and 27 (77%) of them are almost ready to be superuser (Q2). Evaluators suggested that they would like to wait a little and analyze a few more answers of the users coded as "almost ready" (4 on Q2). The result of the human evaluation strengthens our confidence in the classification models and indeed shows their effectiveness in identifying potential experts early on.

The ranking algorithm also found several worthy users which were missed by the DTree algorithm. Put together they discovered 45 out of 54 potential experts who were almost ready to be superusers (Q2) and 27 out of 32 users who showed potential to become successful superusers (Q1). Thus a conjunction of the two algorithms is an effective way to find potential experts.

The 32 users (30% of 105) were rated 4 or more on Q1 and 54 (51% of 105) were rated 4 or more on Q2. Our retrospective analysis shows that some of these potential experts had already left the community. The evaluators suggested that these users could be nurtured and retained by providing them some feedback and

Table 7. Wald Chi Square assessment of the likelihood of a user's potential over different rating aspects.

Rating Aspects	Sig.	Wald Chi-Square
Tax knowledge	.000	**26.11**
Product knowledge	.035	**4.43**
Solving problems	.012	**6.37**
Writing	.307	1.04
Social	.512	0.425
Quality of responses	.001	**10.50**

encouragement - highlighting the need of automated tools to find the potential experts early on.

6.2 Assessment of Rating Aspects

Survey takers evaluated users on 6 core aspects they consider necessary to become a TTLC superuser. We use them and the rating of users on Q2 to run a Wald Chi-Square test. The test assesses the likelihood of a user's potential to become a superuser with the user's assessment on the given rating aspect. Table 7 shows significant effect of tax knowledge, product knowledge, quality of responses, and solving problems on the judgment of the evaluators. The features proposed by us (see Table 4) also assess these qualities and indeed the high agreement between the human evaluators and our algorithms over the identified potential experts indicate that automated filters can be built to align as per the expectations of the humans.

7 Conclusion and Future Work

In this paper, we model users' behavior based on their early participation in the community and show that we could use classification as well as ranking algorithms to identify potential experts. The evaluation by community managers revealed that quite a few of the identified potential experts were already experts or on the path of becoming experts. A word of encouragement could put them to speed to reach the desired goal and stop them from leaving the community.

We showed that a person with potential should demonstrate *motivation* and *ability* to help others. These qualities are equally important and helps us devise a ranking algorithm to measure the expertise of the users. We advocate using a mix of classification and ranking algorithms to find potential experts. Our approach to select the predicted potential experts by the classification model and then using the ranking algorithm to sweep the top ranked from the remaining, would identify a majority of the potential experts if not all. The benefit of the early identification of experts could be long lasting. For a community like TTLC which has only 83 superusers, an addition of 32 worthy superusers is a significant addition. Even for communities with larger number of experts, it would only improve the quantity and the quality of the interactions.

Though our results are encouraging the study is exploratory and has certain limitations that may restrict the generality of its findings. We only consider TTLC - a single Q&A sites with a narrow purpose and an active team of professional behind the scene. TTLC has a very small number of hand labeled experts - it is not clear if we can generalize our findings to a communities with large number of experts. Our work depends on the human labeling and evaluation of user contributions, which used only two coders per user. Finally we do not attempt to exhaustively evaluate different models and machine learning strategies. This work focuses on demonstrating the potential for early detection of experts, and we leave optimization as future work.

Our results suggested that early identification of experts in QA communities is possible. We would like to see the application of this in different types of QA. We are conducting a second round of studies on TTLC, testing alternate volunteer-development strategies on users, such as, promotion to intermediate user status; providing training materials; providing mentoring; developing feedback and task driven mechanism.

Acknowledgments. To be inserted prior to publication.

References

1. Bouguessa, M., Dumoulin, B., and Wang, S. Identifying authoritative actors in question-answering forums: the case of Yahoo! answers. *ACM International conference on Knowledge Discovery and Data mining*, KDD, pg. 866–874, 2008.
2. Cortes, C., and Vapnik, V. (1995) Support-Vector Networks. *Journal of Machine Learning*, pp 273–297, Kluwer Academic Publishers.
3. Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A.: Predictors of answer quality in online Q&A sites. *ACM International conference on Human factors in computing systems*, CHI, pg. 865–874, 2008.
4. Jurczyk, P., and Agichtein, E. Discovering authorities in question answer communities by using link analysis. *ACM International conference on Information and Knowledge Management*, CIKM, pg. 919–922, 2007.
5. Lawrence, P., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report*, Stanford InfoLab, 1999.
6. Pal, A., and Konstan, J. A. Expert identification in community question answering: exploring question selection bias. *ACM International conference on Information and Knowledge Management*, CIKM, pp. 1505–1508, 2010.
7. Panciera, K., Halfaker, A., and Terveen, L. Wikipedians are born, not made: a study of power editors on Wikipedia, *ACM International conference on Supporting group work*, GROUP, pg. 51–60, 2009.
8. Platt, John C. Fast training of support vector machines using sequential minimal optimization, *Advances in kernel methods*, pg. 185–208, MIT Press, 1999.
9. Quinlan, J. R. C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc, 1993.
10. Zhang, J., Ackerman, M. S., and Adamic, L. Expertise networks in online communities: structure and algorithms. *ACM International conference on World Wide Web*, WWW, pp. 221–230, 2007.