Expertise Prediction for Social Network Platforms to Encourage Knowledge Sharing

Nidhi Raj, Lipika Dey, Bhakti Gaonkar
TCS Innovation Labs Delhi
{nidhi1.r, lipika.dey, bhakti.g}@tcs.com

Abstract— **Knowledge sharing social platforms where users mutually benefit through question-answering are gaining popularity. The success of these platforms on the web has led to their adoption within the firewalls of enterprises also. In this paper we have presented some in-depth study about two such platforms – one open on the web and one which is within an enterprise to identify the similarities and dissimilarities of user behavior in the two platforms. We have proposed an algorithm to predict experts to improve the effectiveness of such platforms.**

Keywords- Expertise prediction; Attrition rate; Propensity to answer; Hubs

I. INTRODUCTION

With the growing popularity of social networks, an increasing number of organizations are considering the use of enterprise social networks to encourage knowledge sharing. Though a lot of research has been directed towards studying the nature and impact of public social networks, very little attention has been given towards documenting characteristics about enterprise social networks. The key challenge lies in the fact that enterprise data is usually not available outside the enterprise. Most of the internal reporting also concentrates on usage rather than assess effectiveness.

In this paper, we first present our study on detailed behavioral analysis of two question-answering platforms. One of these is a question-answering platform deployed within an enterprise, referred to as Enterprise Social Network Platform (ESNP) in the rest of the paper. The second one is an open social network platform named Stackoverflow. Functionally, the two platforms have a lot of similarities, though there are a few dissimilarities also. These features are presented in section III. Behavioral analysis of users of the two platforms reveals some interesting similarities and dissimilarities. These are presented in sections IV to VI. In section VII, we present a mechanism that can predict a possible expert for a given question. In section VIII we have presented results to show the effectiveness of the proposed expert prediction method.

II. REVIEW OF RELATED WORK

Social network analysis has received a lot of interest in recent times. Most of these are however targeted at understanding the network structures. [1] presents a good overview of most of the area. [2] presents a method to discover authoritative users in a question-answering platform based on the HITS algorithm [3]. [4] presents a framework to predict the quality of answers with non textual features like answerer's acceptance ratio, answer length, based on data from Yahoo Answers. [5] had proposed a model for calculating category wise experts based on the quality of movie reviews. [6] presented analysis of user behavior on an email social network for an open source software building community. It found that a few members account for the bulk of the messages sent, and the bulk of the replies.

Our expert prediction model builds on the mechanism proposed in [5]. However we take into account additional user behavior issues that are typical of knowledge-sharing platforms like temporal factors and also introduce additional factors to take care of randomness and bias in reviewing.

III. BRIEF OVERVIEW OF SOCIAL NETWORKS FOR QUESTION-ANSWERING

The functional features of open social network platforms like Yahoo Ask, Stackoverflow, and also enterprise social networks that are accessible only to enterprise employees are more or less similar. Users can ask questions, give answers and also provide their inputs about quality of questions or answers through votes and choosing favorites. All activities are usually incentivized in the form of reward or penalty points acquired by the users either for their actions or based on reviews given by other users. In the ESNP under consideration, a user could post a question in one of several pre-defined categories. These platforms are designed to ensure fruitful user participation in all kinds of roles through various mechanisms. However, detailed user behavior analysis on data collected from these sites reveal some interesting issues.

IV. BEHAVIORAL ANALYSIS OF USERS ON SOCIAL NETWORKS FOR QUESTION-ANSWERING

In this paper, we present analysis of results for study conducted over the two networks for a period of 13 months from August 2008 to August 2009. The ESNP had a total of 18,121 users from an organization, with 64% users posting questions, 77% users providing answers and 41% of the total doing both. In the same period, StackOverflow had 85,621 users with 65% users posting questions, 72% users providing answers and only 25% users doing both.

It was observed that for both the platforms the volumes of answers provided were far more than the volumes of questions asked. Table I shows that though the average number of answers was much higher for the ESNP, the percentage of questions not answered was also far higher in this than Stackoverflow. This is a kind of contradiction

978-0-7695-4513-4/11 $26.00 © 2011 IEEE
DOI 10.1109/WI-IAT.2011.93

which mainly arises due to the fact that Stackoverflow is strictly technical, while the ESNP had multiple categories and sub-categories, and not all of them get equal response from users. Fig. 1 shows the distribution of questions in multiple categories for ESNP. It can be concluded that knowledge-sharing is not the most favored activity on the ESNP. Table I further shows that though both platforms had provisions for the questioner to select the best answer, only 60% answers in stack overflow and a pathetic 16.7% answers in the ESNP had been marked as accepted or best answer. This raises some doubt about the usefulness of the answers in the ESNP. One of the reasons for this could be that the right questions were not reaching the right users. In order to get more insights about this, we conducted study on user behavior for the two platforms.

TABLE I. INSIGHTS INTO ESNP AND STACKOVERFLOW

	ESNP	Stack overflow
No of answers given per question	6.64	3.47
% of questions unanswered	11.12	3.85
% of questions with one answer	16.00	22.14
% of questions for which answers were accepted	16.70	60.05

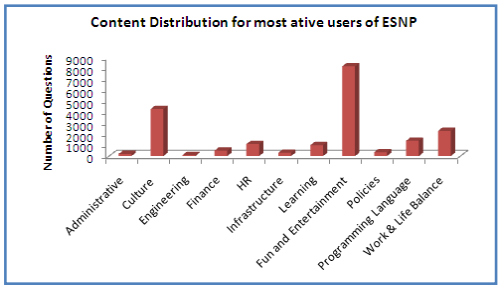

Figure 1. Activity distribution in multiple categories

We therefore looked into the possibility of finding experts within the community to whom the questions can be directed for answering. This would help the ESNP platform to cater to the knowledge-sharing experience in a meaningful way.

More detailed understanding of user behavior was required to design a suitable expert prediction model for such platforms. The next section presents some results on temporal behavioral patterns unearthed from the data of two platforms.

V. ANALYZING TEMPORAL BEHAVIOR OF USERS

It was observed that for both the platforms, majority of the users recorded low levels of activity, and very few users post large number of questions or provides large number of answers. The temporal behavior of users can be defined using a property called attrition rate, given by the percentage of users who were present in one month but absent in the consecutive month. Fig. 2 shows attrition rate plotted for the two networks for the period of thirteen months. Attrition rate for both turned out to be close to 50%, which implies that around half of the users of a social network platform on any given month, do not turn up in the next month.

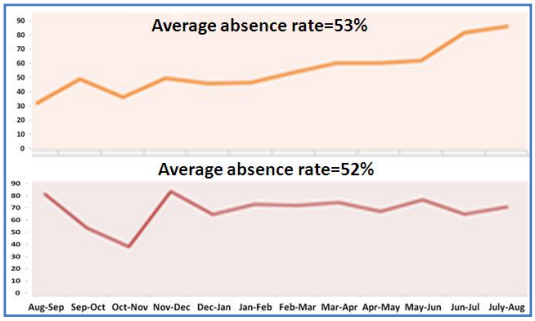

Figure 2. Both platforms show similar attrition rates - (top) Stackoverflow (bottom) ESNP

VI. ANALYZING VOTING BEHAVIOR ON SOCIAL NETWORKS

As discussed earlier, vote-ups and vote-downs by other users are assumed to be indicators of answer quality on question-answering platforms. Incentive schemes are attached for these activities which can affect both the voter and the vote-receiver. Our study of voting behavior on the ESNP revealed some startling facts.

It was observed that voting in the ESNP was quite random and uncontrolled. Fig. 3 shows the distribution of activities of the unique users present at the ESNP during the period. Interestingly, for a large number of users indicated by the blue circle in the figure, the only activity recorded was voting. It was also observed that the maximum number of votes cast by a single user was phenomenally high in some months. These two factors together reduce the credibility of voting, since there is no expertise required for voting, and some users may be doing this simply to gain easy points.

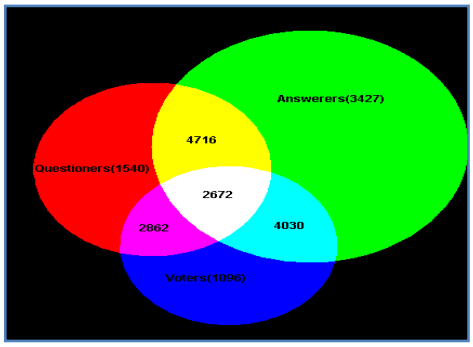

Figure 3. Activity distribution of users in the ESNP

Fig. 4 shows that some answers which receive a large number of vote-ups also received quite a few vote-downs. It is observed that vote-up and vote-down have a positive correlation of 0.1. This typically indicates polarization of users and happens very often within a closed environment where a small community is involved. Further probe revealed the existence of cliques which indicate motivated voting.

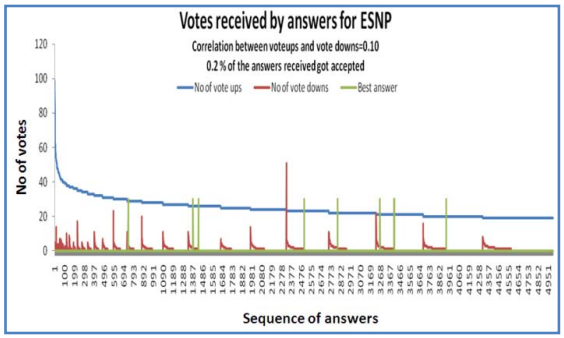

Figure 4. Distribution of the types of votes received by answers for ESNP

Fig. 4 also shows that in spite of a large number of vote ups by users, the questioner rarely accepts an answer as the best answer. The correlation between vote up and best answer was negative, -0.01.

Stackoverflow effectively puts a check on randomness of voting by introducing certain restrictions. Voting comes as a privilege only after the user gains credibility as an answerer. There is also an upper bound on the maximum number of votes by any user within a certain time-period. Also, the communities being large and geographically separated, the web-based platforms may not have the problems as observed in Enterprise Social Networks. We now state how experts can be identified on the ESNP after taking into account the user behavior.

VII. FINDING EXPERTS IN A KNOWLEDGE SHARING SOCIAL PLATFORM

The proposed methodology for predicting experts takes into account two different aspects of a user – subject matter expertise and propensity to answer. A user who has given quality answers in relevant areas in the recent past is more likely to provide a better answer to a question, than a user who had been active earlier but not in the recent past. It also associates takes into account the quality of voters to tackle randomness.

A. Computing Reputation of users as answerer

Let $\overline{U_k}$ be a vector that denotes the reputation of user k in all pre-defined categories. Let U_{kj} be an element of $\overline{U_k}$ which denotes the reputation of user k in j[th] category. U_{kj} is computed as follows:

1) U_{kj} is dependent on past answers given by user k in j^{th} category within a time-frame and the votes received by these. This is done in two steps.

a) Computing quality of each answer by user k

Let r_{kj} = Quality of an answer r in category j given by user k

Each vote received by the answer fetches a positive or negative score for it depending on the type of vote.

Let v denote the type of a vote and ρ_v is the weight associated to the type. Typically ρ_v can be either positive or negative depending on whether it is voting up or down.

Let N_r denote the total score obtained by answer r due to the votes received.

Then $N_r = \sum_{v=type\ of\ votes}(N_{vr} * \rho_v)$, where N_{vr} denotes the total number of votes of type v received by r

Let N_{jmax} = Maximum score received by any answer in j^{th} category

Let $S^E(r_j)$ = Set of reviewers who have cast a vote for r_j

U_{ij} = Reputation Score of a reviewer i in category j where reviewer i is a member of $S^E(r_j)$.

$$r_{kj} = \left(\frac{\sum_{i\,\epsilon S^E(r_j)} \bar{U}^r_{ij}\rho_V}{\sum_{i\,\epsilon S^E(r_j)} \bar{U}^r_{ij}}\right) X \left(\frac{N_r}{N_{jmax}}\right)$$

b) Computing expertise of user k in category j

Let n_{kj} = Total Number of answers given by user k in category j

Then U_{kj} is computed as

$$U_{kj} = \left(\frac{\sum_{l=1}^{nkj} r^l{}_{kj}}{n_{kj}}\right) \quad (1)$$

B. Propensity to answer

Considering users as nodes in the social network graph, one can think of good questioners as nodes with high "in-degree" of answers and have also received high number of vote-ups. Good answerers on the other hand are nodes with high out-degree to good quality questioner nodes.. By considering questioners as equivalent to "authorities" and answerers as equivalent to "hubs", we apply HITS-like algorithm [3] to compute the hub and authority scores for each user.

Let H_{ij} = Hub value of user i in category j.

Let A_{ij} = Authority value of user i in category j.

We compute H_{ij} as follows:

$H_{ij} = \sum_{k=1}^{n} A_{ij}$, where n denotes the total number of users.

Let,

$$\gamma_i{}^j = \frac{Number\ of\ questions\ by\ i\ chosen\ as\ favorite\ in\ cat\ j}{Total\ number\ of\ questions\ asked\ by\ i\ in\ cat\ j}$$

Then A_i^j is given as

$A_{ij} = \sum_{k=1}^{m} H_{ij} + \gamma_i{}^j$ where n denotes the total number of users.

The vectors H_{ij} and A_{ij} are initialized to 1 and 0 respectively, and are updated iteratively. After each iteration, the values in the H_{ij} and A_{ij} vectors are normalized, so that the values of hub and authority lies between 0 and 1. For user i, the propensity to answer in category j is given by H_{ij}.

C. Expertise calculation

Expertise of a user i in category j is computed as an average of reputation and propensity to answer. It can be computed as a weighted sum also.

$$Eij = \frac{(U_{ij}+H_{ij})}{2} \quad (2)$$

As was discussed earlier, a user who has given quality answers in the recent past is more likely to answer and

provide a good answer than a user who had been inactive in the recent past. Equation (3) presents the final equation used to compute expertise for each user in a given month as an exponential average of expertise computed over a fixed time-period. Users are ranked by expertise values.

$$E_{ij(t+1)} = \alpha E_{ijt} + (1-\alpha)E_{ij(t-1)} + (1-\alpha)^2 E_{ij(t-2)} + \cdots \quad (3)$$

VIII. Experiments and Results

The proposed methodology was tested using data from the platforms described earlier. We calculated the expertise of users for three consecutive months based on his or her performance in the preceding 8 months. The value of α was taken as 0.6. The expertise is then compared against the performance of users in a given month by correlating the predicted expertise value to the appreciations received. Fig. 5 shows that our method was able to predict experts who are highly likely to turn up. While the average attrition rate of these platforms is greater than 50%, it can be seen that for both the platforms an average of 80% of the top 100 experts predicted had actually turned up. Consequently, the possibility of requesting these users and receive an answer to an unanswered question also goes up. Table II shows the correlations between the predicted expertise and reputation of users declared by the Stackoverflow platform and also with the score received per answer in that given month. It can be seen for both the cases the correlation value is positive. The proposed method also ranks users who were not present in the system on the given month to identify people who were not present but could potentially provide good answers.

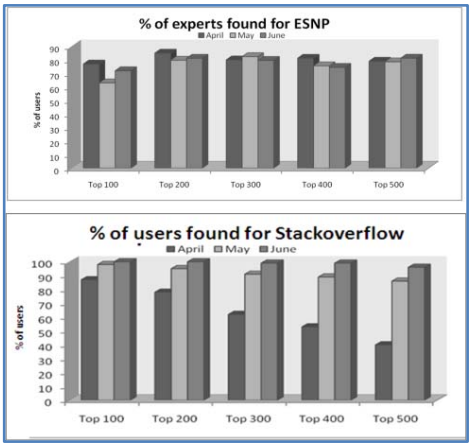

Figure 5. Analysis of turn ups of experts

For the ESNP platform, category-wise predicted expertise was compared with the users' actual performance on the system in case the user turned up. Fig. 6 shows the vote ups, vote downs and best answer status received by top 50 predicted experts who turned up and answered for 3 popular categories for month of June. It can be seen that the vote-ups received per answer is high while the number of vote downs received by these experts is mostly low.

Table II. Correlation Analysis for ESNP

	Correlation between system-given reputation and computed	Correlation between average score received per answer (system defined) and expertise
April	0.472048	0.501939
May	0.428441	0.473458
June	0.493456	0.326774

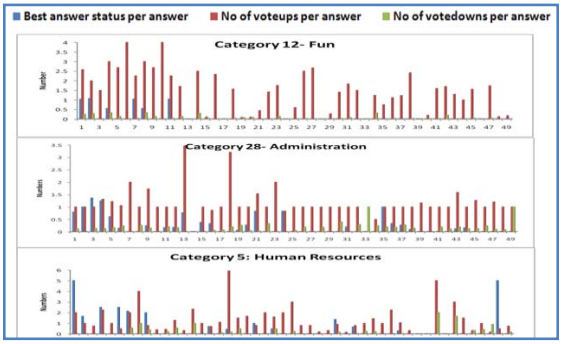

Figure 6. Actual votes received by top 50 predicted experts for top 3 categories in June

IX. Conclusions

In this paper, we have presented an in-depth analysis of user behavior on two different kinds of question-answering platforms, open and within enterprise. While the temporal behavioral patterns show a lot of similarity on the two, the key differences are in terms of core activities like questioning, answering and reviewing. Our study revealed that reviewing on enterprise social network platforms can be quite random. We have also proposed an expertise prediction algorithm, which can be used to identify experts within an enterprise. This can improve the effectiveness of such platforms by reducing the volumes of unanswered questions.

References

[1] John Scott., " Social Network Analysis: a handbook ", Sage publication inc., 2006.

[2] Pawel Jurczyk, Eugene Agichtein, "Discovering Authorities in Question Answer Communities by using Link Analysis", CIKM '07

[3] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46(5):604{632, 1999.

[4] Jiwoon Jeon, W. Bruce Croft, Joon Ho Lee, Soyeon Park, "A framework to Predict the Quality of Answers with non-textual Features", SIGIR '06 ISBN:1-59593-369-7

[5] Young Ae Kim; Minh-Tam Le; Lauw, H.W.; Ee-Peng Lim; Haifeng Liu; Srivastava, J.; "Building a Web of Trust without Explicit Trust Ratings", Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference , 978-1-4244-2161-9

[6] Bird, C., Gourley, A., Devanbu, P., Swaminathan, A., and Gertz, M. Mining Email Social Networks, ICSE 2006 Workshop on Mining Software Repositories (MSR 2006).